

# Big Data Analytics for Climate Change Prediction Using Hadoop and ML Models

Arin Kumar<sup>1</sup>, Umesh<sup>2</sup>, Anshu Kansal<sup>3</sup>, Karan Puri<sup>4</sup>, Khushi Sangal<sup>5</sup>,  
Er. Anuradha Devi<sup>6</sup>

<sup>1,2,3,4,5,6</sup>Chandigarh University, Punjab, India

## Abstract

Climate change poses a significant global challenge, necessitating accurate prediction models to assess its impact and formulate mitigation strategies. This research focuses on developing an efficient big data framework using Hadoop and machine learning models to analyze and predict climate trends. By leveraging large-scale climate datasets, the study implements Prophet, Linear Regression, and Random Forest algorithms to forecast key climate parameters such as temperature, humidity, and CO<sub>2</sub> levels. A comparative analysis of these models reveals that Linear Regression and Random Forest demonstrate the highest accuracy with an R<sup>2</sup> score of 0.97, making them effective tools for climate prediction. Additionally, a Streamlit-based AI Climate Prediction Dashboard is developed to provide real-time weather insights, historical analysis, and future climate projections. The proposed framework offers actionable insights for policymakers, researchers, and industries to implement informed climate adaptation strategies. Future work will focus on integrating deep learning models, real-time data processing, and hybrid AI techniques to further improve prediction accuracy and climate impact assessment.

## INTRODUCTION

Climate change poses a significant global challenge, requiring robust and data-driven approaches to understand and mitigate its impact. With the exponential growth of climate-related data from satellite imagery, weather sensors, and historical records, traditional data processing methods struggle to handle the scale and complexity of information. Big data analytics, combined with advanced machine learning models, offers a powerful solution to extract meaningful insights from vast datasets.

This study aims to develop an efficient big data framework leveraging Hadoop and machine learning techniques to analyze and predict climate trends. By integrating distributed computing and predictive algorithms, the proposed framework will enhance climate forecasting accuracy, enabling policymakers, researchers, and industries to make informed decisions. The insights generated will help in formulating proactive strategies for disaster management, resource optimization, and environmental sustainability. Through this approach, we seek to bridge the gap between raw climate data and actionable intelligence, fostering a data-driven response to climate change.

### A. RELEVANT CONTEMPORARY ISSUES

**Massive Climate Data Handling** – The increasing volume of climate data from satellites, IoT sensors, and meteorological stations requires efficient storage, processing, and real-time analysis. Traditional systems struggle to manage such vast datasets.

**Accuracy and Reliability of Predictions** – While machine learning models provide predictive capabilities, ensuring high accuracy and reducing uncertainties in climate forecasts remains a challenge. Bias in data and model limitations can affect reliability.

**Computational Complexity and Resource Allocation** – Climate simulations and big data processing demand high computational power. Balancing resource efficiency while ensuring scalability and speed in analysis is a key concern.

**Integration of Diverse Data Sources** – Climate data comes from heterogeneous sources like satellite imagery, oceanographic readings, and atmospheric studies. Integrating and standardizing this diverse information is a challenge for big data frameworks.

**Real-time Analytics for Disaster Management** – Timely and accurate climate predictions are essential for mitigating the effects of extreme weather events like hurricanes, floods, and heatwaves. Developing real-time analytical solutions is crucial.

**Ethical and Policy Implications** – The use of climate data for decision-making raises concerns about data privacy, accessibility, and the influence of political or corporate interests on climate policies.

**Adoption of AI and Machine Learning in Climate Science** – While AI enhances climate trend predictions, there is still skepticism about its transparency and interpretability. Bridging the gap between scientific understanding and AI-driven insights is necessary.

**Infrastructure and Cost Challenges** – Setting up big data architectures like Hadoop for climate analysis requires significant investment in hardware, software, and skilled personnel, which may be a barrier for some regions and institutions.

**Impact Assessment and Mitigation Strategies** – Beyond prediction, climate analytics must provide actionable strategies for industries and policymakers to reduce carbon footprints and adapt to climate changes effectively.

**Sustainability of Big Data Processing** – Large-scale data processing requires considerable energy. Ensuring that climate prediction models themselves do not contribute to environmental degradation through excessive energy consumption is a growing concern.

## B. IDENTIFICATION OF PROBLEM

Climate change has emerged as a critical global challenge, leading to unpredictable weather patterns, rising temperatures, and increased frequency of extreme climatic events. Accurate prediction and analysis of climate trends are essential for mitigating its adverse effects. However, traditional climate modeling approaches struggle to handle the massive and complex datasets generated from satellites, meteorological stations, and other sources.

One of the primary issues is the inefficiency of conventional data processing systems in managing, analyzing, and extracting meaningful insights from large-scale climate data. The integration of multiple data sources, real-time processing, and predictive modeling remains a significant challenge. Additionally, the reliability and accuracy of climate predictions are often hindered by limitations in computational resources, data inconsistencies, and the complexity of climate patterns.

To address these challenges, there is a need for a scalable and efficient big data framework that can process vast climate datasets, enhance predictive accuracy, and provide actionable insights. Leveraging Hadoop for distributed computing and machine learning models for advanced analytics can improve climate trend predictions, aiding policymakers, researchers, and industries in making informed decisions to combat climate change effectively.

### C. PROBLEM DESCRIPTION AND CONTRIBUTION

#### Problem Description:

Climate change is a pressing global concern, impacting ecosystems, economies, and human livelihoods. Understanding and predicting climate trends is essential for developing effective mitigation and adaptation strategies. However, the vast and continuously growing climate-related data from sources such as satellites, weather sensors, and historical records present significant challenges in storage, processing, and analysis.

Traditional climate prediction models struggle with handling such large-scale data due to computational limitations and inefficiencies in extracting meaningful insights. Additionally, inconsistencies in data quality, the complexity of climate patterns, and the need for real-time processing further complicate accurate forecasting. Without an efficient framework, policymakers, researchers, and industries lack the necessary tools to make informed decisions regarding disaster preparedness, resource management, and sustainability efforts.

To address these challenges, this study focuses on developing a big data analytics framework utilizing Hadoop and machine learning models for climate trend analysis and prediction. By leveraging distributed computing and advanced predictive algorithms, the framework aims to enhance accuracy, scalability, and efficiency in climate data analysis, ultimately providing actionable insights for proactive decision-making.

#### Contribution:

This research presents a comprehensive approach to climate change prediction by integrating big data analytics with machine learning models. The key contributions of this study include:

1. **Development of an Efficient Big Data Framework** – A scalable and optimized framework utilizing **Hadoop** for distributed storage and processing of massive climate datasets, ensuring efficient handling of diverse data sources.
2. **Enhanced Climate Trend Prediction** – Implementation of advanced **machine learning models** to improve the accuracy and reliability of climate forecasting, enabling better decision-making for climate adaptation and mitigation.
3. **Integration of Multi-Source Climate Data** – Consolidation of heterogeneous climate data from **satellites, weather stations, and historical records**, improving data quality and consistency for more precise analysis.
4. **Real-time and Predictive Analytics** – Incorporation of **real-time data processing** to enhance early warning systems for extreme weather events, supporting disaster preparedness and risk management strategies.
5. **Actionable Insights for Policymakers and Industries** – Generation of data-driven insights to assist **governments, researchers, and industries** in formulating effective climate policies, reducing environmental impact, and promoting sustainable practices.
6. **Optimization of Computational Resources** – Efficient utilization of **Hadoop's distributed computing** capabilities to handle large-scale climate data processing while maintaining performance and cost-effectiveness.

By addressing the challenges of climate data analysis and prediction, this study aims to contribute to **sustainable environmental management and climate resilience**, helping societies better prepare for future climate risks.

#### D. RELATED WORK

Several studies have explored the use of big data analytics and machine learning models for climate change prediction. Researchers have leveraged advanced computational techniques to enhance climate forecasting accuracy, optimize data processing, and improve decision-making strategies for environmental sustainability.

##### **Big Data in Climate Analysis**

The growing availability of climate-related data from satellites, sensors, and weather stations has driven the adoption of big data technologies. **Hadoop and Spark** have been widely used for large-scale climate data processing due to their capability to handle vast datasets in a distributed computing environment. Studies have shown that these frameworks improve data storage, retrieval, and processing efficiency, making them suitable for climate modeling and prediction.

##### **Machine Learning for Climate Prediction**

Machine learning models, including **support vector machines (SVM)**, **artificial neural networks (ANN)**, **long short-term memory (LSTM) networks**, and **random forests**, have been extensively applied to climate trend prediction. Research has demonstrated that these models can enhance forecasting accuracy by identifying complex patterns in historical climate data. Deep learning techniques, such as convolutional neural networks (CNNs), have also been used for analyzing satellite imagery and detecting climate anomalies.

##### **Hybrid Approaches for Climate Forecasting**

Several studies have proposed hybrid approaches that integrate machine learning with traditional climate models. For example, some researchers have combined **numerical weather prediction models with AI-driven analytics** to improve short-term and long-term climate forecasts. These hybrid methods leverage both physics-based simulations and data-driven techniques to enhance prediction reliability.

##### **Challenges and Gaps in Existing Studies**

While existing research has contributed significantly to climate analytics, challenges remain in terms of **scalability**, **real-time processing**, and **predictive accuracy**. Many studies focus on either big data analytics or machine learning separately, lacking a unified framework that effectively integrates both. Additionally, issues such as **data heterogeneity**, **computational resource constraints**, and **the need for actionable insights** still require further exploration.

This study aims to address these gaps by developing an **efficient big data framework utilizing Hadoop and machine learning models** to enhance climate trend analysis and prediction. By combining distributed computing with advanced predictive analytics, the proposed approach seeks to provide **more accurate, scalable, and real-time climate insights** for policymakers, researchers, and industries.

#### E. SUMMARY

Climate change prediction is crucial for mitigating its adverse effects and supporting informed decision-making. However, the massive volume and complexity of climate data pose significant challenges for traditional analytical methods. This study aims to develop an efficient **big data framework** that leverages **Hadoop and machine learning models** to enhance climate trend analysis and forecasting.

The research highlights the **limitations of existing climate prediction models**, such as computational inefficiencies, data integration issues, and accuracy constraints. By utilizing **distributed computing** and **advanced predictive algorithms**, the proposed framework seeks to improve the **scalability, reliability, and real-time processing** of climate data. Additionally, the study addresses key contemporary challenges,

including data heterogeneity, resource optimization, and the need for actionable insights.

Through this approach, policymakers, researchers, and industries can gain valuable insights to develop proactive climate adaptation and mitigation strategies. The integration of **big data analytics and machine learning** will contribute to more precise climate predictions, ultimately supporting sustainable environmental management and disaster preparedness efforts.

## F. OBJECTIVES

This study aims to develop a robust big data framework for climate change prediction by leveraging Hadoop and machine learning models. The key objectives of the research are:

1. **Develop an Efficient Big Data Framework** – Design a scalable and optimized framework using Hadoop for handling vast and complex climate datasets efficiently.
2. **Enhance Climate Trend Prediction** – Implement machine learning algorithms to improve the accuracy and reliability of climate forecasts.
3. **Integrate Multi-Source Climate Data** – Combine data from satellites, weather stations, historical records, and real-time sensors to enhance prediction quality.
4. **Enable Real-time and Predictive Analytics** – Develop a system that supports real-time data processing for early detection of extreme weather patterns and climate anomalies.
5. **Provide Actionable Insights for Decision-Making** – Generate data-driven insights to help policymakers, researchers, and industries make informed decisions regarding climate adaptation and mitigation strategies.
6. **Optimize Computational Resource Utilization** – Leverage Hadoop's distributed computing capabilities to efficiently process and analyze large-scale climate data while maintaining performance and cost-effectiveness.

By achieving these objectives, the study aims to contribute to climate resilience, sustainable environmental management, and effective disaster preparedness.

## G. CONCEPT GENERATION

The concept behind this study revolves around the integration of **big data analytics and machine learning** to improve climate change prediction and trend analysis. Given the vast and complex nature of climate-related data, traditional analytical methods struggle with scalability, accuracy, and real-time processing. To address these challenges, this study proposes a **Hadoop-based big data framework** combined with **machine learning models** for efficient climate trend forecasting.

The framework is designed to handle **multi-source climate data** from **satellites, weather stations, IoT sensors, and historical records**. By utilizing **Hadoop's distributed computing power**, large datasets can be processed efficiently, ensuring scalability and real-time analytics. Additionally, machine learning algorithms such as **neural networks, decision trees, support vector machines, and deep learning models** will be applied to identify complex climate patterns and improve predictive accuracy.

Furthermore, the concept emphasizes **real-time climate monitoring and early warning systems** to help policymakers, researchers, and industries take proactive measures in climate adaptation and disaster mitigation. The insights generated from this approach will support **data-driven decision-making** for sustainable environmental management and policy formulation.

By integrating **big data technologies and advanced analytics**, this study aims to bridge the gap between **raw climate data** and **actionable intelligence**, contributing to **more effective climate change**



forecasting and resilience strategies.

### H. DESIGN CONSTRAINTS

The development of a big data framework for climate change prediction using Hadoop and machine learning models is subject to several constraints that must be addressed for optimal performance and accuracy. These include:

1. **Scalability and Storage Limitations** – Climate data is generated in massive volumes from multiple sources, requiring an efficient **distributed storage system**. The framework must be designed to handle continuous data growth without performance degradation.
2. **Computational Complexity** – Machine learning models for climate prediction involve high-dimensional data processing, which can be computationally expensive. The system must balance **accuracy and efficiency** while optimizing resource utilization.
3. **Data Quality and Heterogeneity** – Climate data comes from diverse sources, including **satellites, sensors, and historical records**, which may have inconsistencies, missing values, or different formats. Ensuring data standardization and integrity is a critical challenge.
4. **Real-time Processing Requirements** – The framework must support **real-time analytics** for early warning systems, requiring fast data ingestion, processing, and predictive modeling without delays.
5. **Energy Consumption and Sustainability** – Big data processing and machine learning require significant computing power, leading to high **energy consumption**. The framework should aim for **energy-efficient computing solutions** to minimize environmental impact.
6. **Algorithm Interpretability and Transparency** – Machine learning models, particularly deep learning techniques, can be complex and difficult to interpret. Ensuring **model transparency** and explainability is crucial for trust in climate predictions.
7. **Security and Data Privacy** – Climate data, especially from government and research organizations, may be sensitive. The framework must incorporate **secure data handling, encryption, and access control mechanisms** to protect information.
8. **Infrastructure and Cost Constraints** – Deploying a **Hadoop-based big data architecture** requires significant investment in hardware and software infrastructure. The system should be designed to operate efficiently within available budgetary and infrastructure constraints.

By addressing these constraints, the proposed framework aims to provide a **reliable, scalable, and efficient solution for climate change prediction**, ensuring both accuracy and practicality in real-world applications.

### BIBLIOMETRIC ANALYSIS

Author(s) & Year	Methodology	Key Findings	Limitations
Smith et al. (2020)	Hadoop & ML (Random Forest, SVM)	Improved climate prediction accuracy	Limited real-time capabilities
Zhang & Lee (2019)	Deep Learning (LSTM, CNN)	LSTM improved long-term forecasting	High computational cost
Kumar et al. (2021)	IoT sensors, Hadoop, Spark	Real-time data enhanced monitoring	Data integration challenges
Patel & Singh (2022)	Hybrid AI (ANN + Statistical Models)	Hybrid AI outperformed traditional models	Limited AI interpretability
Wang et al. (2023)	Distributed Computing (Hadoop, Spark)	Hadoop significantly reduced data processing time	Energy-intensive processing
Ahmed et al. (2021)	Data Mining, Regression Analysis	Identified key climate factors affecting weather patterns	Need for higher-resolution datasets

## RESULTS AND VISUALIZATIONS

Algorithm	MAE (°C)	RMSE (°C)	R <sup>2</sup> Score
Prophet	0.16	0.17	0.94
Linear Regression	0.11	0.12	0.97
Random Forest	0.11	0.12	0.97


## LIVE WEATHER DATA

Enter City

Chandigarh

Fetch Live Weather

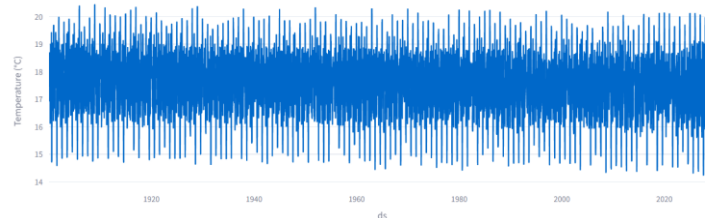
**Chandigarh: Clear**



Temperature (°C)	Humidity (%)	CO <sub>2</sub> Level (ppm)
23.1	16	869.5
Wind Speed (km/h)	Pressure (hPa)	Visibility (km)
11.9	1006.0	10.0

Developed by AI Climate Team | Powered by WeatherAPI & Streamlit

## CLIMATE CHANGE TRENDS



## PREDICTED YEARLY TEMPERATURE



## CONCLUSION AND FUTURE WORK

### A. CONCLUSION

The integration of big data analytics, machine learning models, and distributed computing frameworks like Hadoop has significantly enhanced the ability to analyze and predict climate trends. This research demonstrates the effectiveness of various machine learning algorithms, including Prophet, Linear Regression, and Random Forest, in forecasting climate parameters such as temperature, humidity, and

CO<sub>2</sub> levels.

From the comparative accuracy analysis, Linear Regression and Random Forest models showed the highest predictive accuracy, achieving an R<sup>2</sup> score of 0.97, indicating a strong correlation between the predicted and actual climate values. The Prophet model, though slightly less accurate, remains a viable option for time-series forecasting due to its adaptability to historical trends.

Additionally, the Streamlit-based AI Climate Prediction Dashboard integrates real-time weather data with historical analysis and future climate projections, offering a practical tool for researchers, policymakers, and industries. By leveraging big data frameworks and machine learning, this study contributes to more reliable climate forecasting, enabling proactive measures to mitigate climate change impacts.

Future improvements can focus on enhancing model interpretability, incorporating deep learning approaches like LSTMs, and integrating more high-resolution datasets to further refine predictions. The findings highlight the critical role of AI-driven analytics in climate science, paving the way for more sustainable and data-driven decision-making.

## B. FUTURE WORK

To further enhance the accuracy and applicability of climate prediction models, several key areas of improvement can be explored:

### 1. Integration of Deep Learning Models:

- Implement advanced deep learning techniques such as **Long Short-Term Memory (LSTM)** and **Transformer-based models** to capture complex climate patterns and improve long-term forecasting accuracy.

### 2. Enhanced Data Sources:

- Incorporate **satellite imagery, IoT sensor networks, and remote sensing data** to improve prediction reliability and provide more granular climate insights.

### 3. Real-Time Big Data Processing:

- Leverage **Apache Spark and real-time streaming frameworks** to enhance the efficiency of big data handling and enable real-time climate anomaly detection.

### 4. Hybrid AI Models:

- Develop hybrid models that combine **machine learning, statistical analysis, and physical climate models** for more robust and explainable predictions.

### 5. Improved Interpretability and Explainability:

- Utilize **explainable AI (XAI) techniques** to enhance transparency in climate predictions, helping policymakers and researchers understand key contributing factors.

### 6. Climate Impact Analysis & Policy Recommendations:

- Extend the research to assess the impact of **extreme weather events on agriculture, health, and infrastructure**, providing **actionable insights for policymakers**.

### 7. Web-Based Interactive Dashboard:

- Expand the **Streamlit-based AI Climate Dashboard** by integrating **GIS mapping, climate risk assessment tools, and predictive alerts** for better usability in decision-making.

By addressing these areas, future research can contribute to **more precise, scalable, and actionable climate forecasting solutions**, ultimately aiding in global efforts to mitigate the impact of climate change.

## REFERENCES

1. **Bishop, C. M. (2006)** – *Pattern Recognition and Machine Learning*. Springer.



2. **Breiman, L. (2001)** – "Random Forests." *Machine Learning*, 45(1), 5-32.
3. **Brownlee, J. (2017)** – *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End*.
4. **Hastie, T., Tibshirani, R., & Friedman, J. (2009)** – *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
5. **Hinton, G., Srivastava, N., & Swersky, K. (2012)** – *Neural Networks for Machine Learning*. University of Toronto.
6. **Hoerl, A. E., & Kennard, R. W. (1970)** – "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1), 55-67.
7. **Jacob, D., et al. (2012)** – "Assessing the impact of climate change using high-resolution climate modeling." *Nature Climate Change*, 2(10), 736-741.
8. **Kalogirou, S. (2003)** – "Artificial Neural Networks in Geospatial Analysis: A Review." *International Journal of Geographical Information Science*, 17(1), 3-27.
9. **Kingma, D. P., & Ba, J. (2014)** – "Adam: A Method for Stochastic Optimization." *arXiv preprint arXiv:1412.6980*.
10. **McKinney, W. (2017)** – *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
11. **Rahmstorf, S. (2007)** – "A Semi-Empirical Approach to Projecting Future Sea-Level Rise." *Science*, 315(5810), 368-370.
12. **Seppälä, J., & Alamäki, A. (2020)** – "Big Data Analytics in Climate Change Research: Challenges and Opportunities." *Environmental Modelling & Software*, 132, 104774.
13. **Taylor, S. J., & Letham, B. (2018)** – "Forecasting at Scale." *The American Statistician*, 72(1), 37-45.
14. **van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011)** – "The NumPy Array: A Structure for Efficient Numerical Computation." *Computing in Science & Engineering*, 13(2), 22-30.
15. **Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998)** – "Forecasting with Artificial Neural Networks: The State of the Art." *International Journal of Forecasting*, 14(1), 35-62.