Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 6400 – 6409 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

BIG DATA ANALYSIS OF RAINFALL USING APACHE SPARK: A CASE STUDY OF INDONESIA

Arnold Paulinus Partogi Sinaga[1], Jhon Nobel Helsingki Turnip[2], Mohammad Irfan Fahmi[3]
Universitas Prima Indonesia, Medan
e-mail: arnoldpaulinus23@gmail.com

Abstract: *Rainfall is one of the most important climatological parameters that plays a significant role in supporting agriculture, water resource management, and hydrometeorological disaster mitigation in Indonesia. As the volume of meteorological data continues to increase, technologies capable of processing large-scale data efficiently are required. This study aims to analyze rainfall patterns in Indonesia using a Big Data approach by utilizing Apache Spark as the primary data processing framework. The data used in this research consist of daily rainfall data from 34 provinces in Indonesia obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) in CSV format. This research employs a descriptive quantitative method consisting of data ingestion, data cleaning, data transformation, and descriptive statistical analysis using Apache Spark with PySpark. The analysis was conducted to obtain information regarding the number of valid records, average rainfall, maximum rainfall, and minimum rainfall values for each province. The results indicate that Apache Spark is capable of processing rainfall data in a structured, efficient, and automated manner. Based on the analysis, West Sumatra Province recorded the highest average rainfall at 16.01 mm/day, while Central Sulawesi Province recorded the lowest average rainfall at 0.30 mm/day. In addition, the highest maximum rainfall was observed in Bengkulu Province at 186.4 mm. The use of Spark DataFrame and aggregation functions such as count(), avg(), max(), and min() proved effective in supporting data processing and analysis activities. The findings demonstrate that Apache Spark can serve as an effective solution for large-scale climatological data processing and provide valuable information to support decision-making in agriculture, water resource management, and disaster mitigation in Indonesia.*

Keywords: *Rainfall Analysis, Apache Spark, Big Data Analytics, PySpark,Climatological Data,Indonesia.*

Abstrak: Curah hujan merupakan salah satu parameter klimatologi terpenting yang berperan signifikan dalam mendukung pertanian, pengelolaan sumber daya air, dan mitigasi bencana hidrometeorologi di Indonesia. Seiring dengan terus meningkatnya volume data meteorologi, dibutuhkan teknologi yang mampu memproses data skala besar secara efisien. Studi ini bertujuan untuk menganalisis pola curah hujan di Indonesia menggunakan pendekatan Big Data dengan memanfaatkan Apache Spark sebagai kerangka kerja pemrosesan data utama. Data yang digunakan dalam penelitian ini terdiri dari data curah hujan harian dari 34 provinsi di Indonesia yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) dalam format CSV. Penelitian ini menggunakan metode kuantitatif deskriptif yang terdiri dari pengumpulan data, pembersihan data, transformasi data, dan analisis statistik deskriptif menggunakan Apache Spark dengan PySpark. Analisis dilakukan untuk memperoleh informasi mengenai jumlah catatan valid, curah hujan rata-rata, curah hujan maksimum, dan nilai curah hujan minimum untuk setiap provinsi. Hasil menunjukkan bahwa Apache Spark mampu memproses data curah hujan secara terstruktur, efisien, dan otomatis. Berdasarkan analisis, Provinsi Sumatera Barat mencatat curah hujan rata-rata tertinggi sebesar 16,01 mm/hari, sedangkan Provinsi Sulawesi Tengah mencatat curah hujan rata-rata terendah sebesar 0,30 mm/hari. Selain itu, curah hujan maksimum tertinggi diamati

di Provinsi Bengkulu sebesar 186,4 mm. Penggunaan Spark DataFrame dan fungsi agregasi seperti count(), avg(), max(), dan min() terbukti efektif dalam mendukung aktivitas pengolahan dan analisis data. Temuan ini menunjukkan bahwa Apache Spark dapat berfungsi sebagai solusi efektif untuk pengolahan data klimatologi skala besar dan memberikan informasi berharga untuk mendukung pengambilan keputusan di bidang pertanian, pengelolaan sumber daya air, dan mitigasi bencana di Indonesia.

Kata kunci: Analisis Curah Hujan, Apache Spark, Analisis Big Data, PySpark, Data Klimatologi, Indonesia

INTRODUCTION

Rainfall is one of the most important climatological parameters that plays a crucial role in supporting the agricultural sector, water resource management, and hydrometeorological disaster mitigation. In tropical countries such as Indonesia, rainfall distribution exhibits highly dynamic characteristics and is influenced by various atmospheric and oceanographic factors, including the monsoon system, El Niño–Southern Oscillation (ENSO), and

the Indian Ocean Dipole (IOD) . High rainfall variability leads to differences in climatic conditions across regions and time periods, thereby affecting agricultural productivity, water availability, and national food security.

Changes in rainfall patterns can also generate various adverse impacts on both the environment and society. High-intensity rainfall may trigger floods and landslides, while reduced rainfall during certain periods can lead to droughts that negatively affect agriculture and water resources. Therefore, a comprehensive understanding of rainfall patterns is essential as a foundation for accurate and data-driven decision-making processes .

Along with the advancement of weather and climate observation technologies, the volume of meteorological data generated has increased significantly over time . Data collected routinely by the Meteorology, Climatology, and Geophysics Agency (BMKG) produce large-scale datasets characterized by high volume, velocity, and variety, which classify them as Big Data. Processing climatological data using conventional approaches often encounters limitations in terms of computational efficiency, processing speed, and the capability to handle massive datasets .

Big Data technology offers a promising solution to address these challenges through distributed computing systems capable of processing data in parallel. One of the most widely adopted frameworks for Big Data processing is Apache Spark. Apache Spark provides advantages such as in-memory computing and distributed processing, enabling faster data analysis compared to traditional approaches. Previous studies have demonstrated that Apache Spark can significantly improve data processing efficiency across various domains, including weather, climate, and environmental data analysis. Furthermore, the integration of Apache Spark with Python through PySpark offers flexibility in data analysis while supporting a wide range of statistical and data transformation functions efficiently.

Several previous studies have investigated rainfall analysis using statistical methods, machine learning techniques, and geographic information systems . Other studies have also shown that the application of Big Data technologies enhances the capability to analyze large and complex meteorological datasets. However, research specifically

focusing on the utilization of Apache Spark for descriptive analysis of Indonesian rainfall data using official BMKG datasets remains relatively limited . This condition indicates a research opportunity to further explore the

implementation of Big Data technologies for processing and analyzing national climatological data.

Based on this background, this study aims to analyze rainfall data in Indonesia using a Big Data approach based on Apache Spark. The dataset employed in this research consists of daily rainfall records from 34 provinces in Indonesia obtained from BMKG. The analytical process includes data ingestion, data cleaning, data transformation, and descriptive statistical analysis using PySpark. The results are expected to provide insights into the characteristics of rainfall patterns across Indonesia while demonstrating the effectiveness of Apache Spark in supporting large-scale climatological data processing for decision-making purposes in agriculture, water resource management, and disaster mitigation.

METHOD

The data processing and analysis method in this study employed a Big Data processing approach through Apache Spark with the PySpark API. The research process began with data acquisition, which involved downloading the rainfall dataset from the BMKG in CSV format and loading it into a Spark DataFrame.The next stage was descriptive statistical analysis, which included calculating the mean, maximum, minimum, standard deviation, and frequency distribution to identify rainfall patterns and trends. The analysis results were then presented in graphical visualizations to facilitate interpretation and conclusion drawing.

This research method utilizes several software and technologies, as follows:

1. Apache Spark is used as the primary framework for Big Data processing. Spark's distributed computing and in-memory processing capabilities enable efficient processing of large datasets. In this study, Spark was used to: Read and load CSV datasets in Spark Data Frame format; Perform filtering, grouping, and aggregation operations; Calculate descriptive statistics; Spark was run in local mode (single node) because the research was conducted on a single computer.
2. PySpark is a Python-based Apache Spark API. PySpark was chosen because: Python syntax is easier to understand; It supports flexible data exploration.
3. Python was used as the primary programming language in this study. Python supports various data analysis libraries such as: Pandas (for additional data manipulation); Matplotlib (for visualization); and NumPy (for numerical calculations).
4. Jupyter Notebook was used as the development environment because: It supports interactive code execution; It facilitates documentation of code and analysis results; It facilitates integration between code and visualization.
5. Operating System and Device Specifications: The study was conducted using the Windows operating system with device specifications that support local-scale Big Data processing. The device specifications used allow Apache Spark to run optimally in local mode without the need for an additional cluster.

RESULTS AND DISCUSSION

DATA PROCESSING AND ANALYSIS FRAMEWORK

Research Design

This study employed a quantitative descriptive approach to analyze rainfall data in Indonesia using Big Data technology. The quantitative approach was selected because the analyzed data consisted of numerical climatological observations obtained from the Meteorology, Climatology, and

Geophysics Agency (BMKG). Descriptive analysis was conducted to identify and describe rainfall characteristics, distribution patterns, and statistical trends without developing predictive models. Exploratory data analysis was utilized to understand rainfall patterns and data characteristics before statistical summarizations.

Dataset Description

The dataset used in this study was obtained from BMKG and stored in Comma-Separated Values (CSV) format. The data consisted of daily climatological observations from various provinces in Indonesia, including rainfall measurements, observation dates, and location information. The dataset exhibits Big Data characteristics, including large data volume, time-series structure, and the possibility of containing missing values, making efficient data processing techniques necessary. Due to the extensive spatial coverage and continuous temporal observations, the dataset provides valuable information for analyzing rainfall variability across different regions and periods. However, the large size and heterogeneous nature of the data also present challenges related to data quality, storage, processing efficiency, and analytical scalability. Therefore, appropriate data preprocessing procedures, including data cleaning and transformation, are required to ensure data consistency and reliability before analysis. The utilization of a large-scale climatological dataset in this study enables a more comprehensive assessment of rainfall patterns and supports the application of Big Data technologies for extracting meaningful insights from complex environmental data.

Apache Spark Processing Framework

Apache Spark was utilized as the primary Big Data processing framework, while PySpark was employed as the programming interface for implementing data processing tasks using Python. Apache Spark was selected because of its distributed computing architecture and in-memory processing capability, which enable efficient processing of large-scale datasets.

Statistical Analysis

Descriptive statistical analysis was performed using Spark DataFrame aggregation functions. The analysis included the calculation of the number of valid records, average rainfall, maximum rainfall, and minimum rainfall values using functions such as count(), avg(), max(), and min(). Descriptive statistics were employed to summarize rainfall characteristics and facilitate comparisons among provinces in Indonesia. The analysis provides a baseline understanding of rainfall variability across Indonesia. By providing an overview of central tendencies and extreme values, the results help reveal variations in precipitation patterns among provinces. This information serves as an important foundation for understanding regional climatic conditions and supports further investigation of rainfall variability across the country. The results highlight differences in precipitation characteristics among provinces. These findings provide useful insights into regional rainfall dynamics across Indonesia.

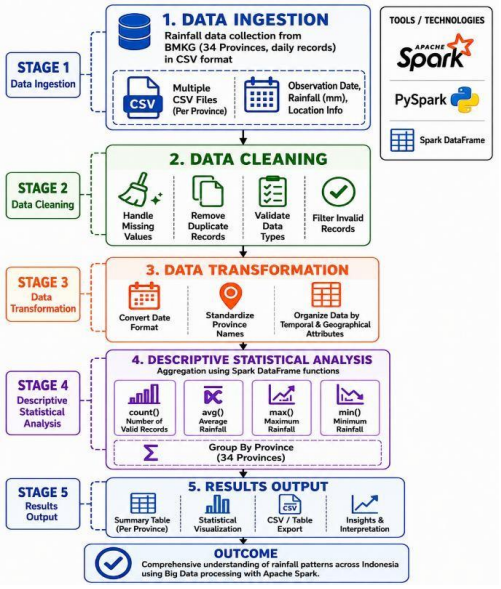

Figure1Research framework for rainfall data analysis using Apache SparkRESULTS AND DISCUSSION

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
August 2026, IX (4): 6400 – 6409 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Rainfall Data Analysis Results

The rainfall dataset obtained from the Meteorology, Climatology, and Geophysics Agency (BMKG) consisted of daily rainfall observations from 34 provinces in Indonesia during April 2026. After the data cleaning process, invalid records containing special codes representing unavailable or unmeasured observations were removed, resulting in a valid dataset suitable for statistical analysis.

Descriptive statistical analysis was conducted using Apache Spark to determine the number of valid records, average rainfall, maximum rainfall, and minimum rainfall values for each province. The results revealed substantial variation in rainfall characteristics among provinces. West Sumatra recorded the highest average rainfall with a value of 16.01 mm/day, followed by Bengkulu with 15.67 mm/day and Yogyakarta with 13.99 mm/day. In contrast, Central Sulawesi exhibited the lowest average rainfall at 0.30 mm/day, indicating considerably drier conditions during the observation period. This finding suggests a relatively limited contribution of precipitation to regional water availability, which may influence agricultural productivity and water resource management in the area

The analysis also showed that Bengkulu Province experienced the highest maximum rainfall value of 186.4 mm, suggesting the occurrence of an extreme rainfall event during the study period. Meanwhile, the minimum rainfall value in most provinces was 0 mm, indicating the presence of days without rainfall. These findings demonstrate the spatial variability of rainfall distribution across Indonesia and highlight differences in regional climatic conditions. Such variability reflects the influence of local geographic characteristics, atmospheric circulation patterns, and seasonal climate dynamics on rainfall occurrence. Understanding these regional differences is essential for improving climate-related planning, water resource management, and disaster risk reduction strategies across Indonesia.

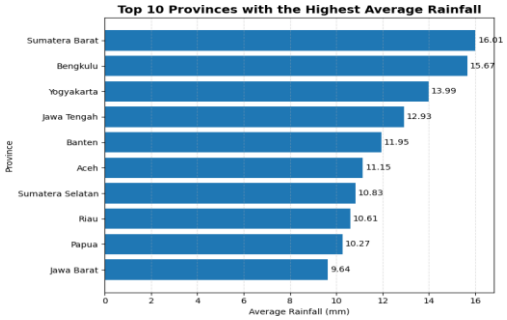

Figure 2 top 10 provinces with the highest average rainfall in Indonesia during the observation period.

The figure shows that Sumatera Barat, Bengkulu, and Yogyakarta recorded the highest average rainfall values among all provinces analyzed

Figures 2 and 3 demonstrate the spatial variability of average rainfall across Indonesian provinces during the observation period. The results indicate a marked contrast between provinces with high and low rainfall characteristics. West Sumatra and Bengkulu were identified as the provinces with the highest average rainfall values, whereas Central Sulawesi and South Sulawesi exhibited the lowest averages. These findings highlight the heterogeneous nature of rainfall distribution in Indonesia, which is influenced by regional climatic and geographical factors. The observed differences suggest that rainfall patterns vary considerably across regions, reflecting the complex interactions between topography, atmospheric circulation, and local environmental conditions. Such variability emphasizes the importance of regional-scale rainfall analysis for supporting climate-related planning and resource management.

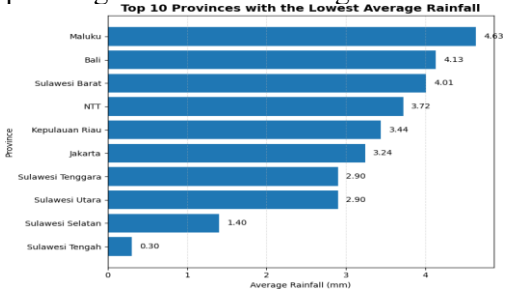

Figure 3 top 10 provinces with the lowest average rainfall in Indonesia

during the observation period. Sulawesi Tengah and Sulawesi Selatan exhibited the lowest average rainfall values compared with other provinces

Table I reveal considerable

Category	Province	Rainfall (mm/day)
Highest Average Rainfall	West Sumatra	16.01
Lowest Average Rainfall	Central Sulawesi	0.30
Highest Maximum Rainfall	Bengkulu	186.4

Discussion of Rainfall Characteristics

The statistical results presented in Table I reveal considerable differences in rainfall characteristics among Indonesian provinces during the observation period. The gap between the highest average rainfall in West Sumatra and the lowest average rainfall in Central Sulawesi indicates substantial spatial variability in precipitation patterns. Such differences reflect the diverse climatic and geographical conditions across the Indonesian archipelago. The findings also suggest that rainfall distribution is not uniform and may be influenced by regional atmospheric processes, topographical features, and proximity to major moisture sources. This variability highlights the importance of province-level rainfall analysis for understanding local climate behavior and supporting regional environmental planning.

The results indicate that rainfall distribution in Indonesia during April 2026 varied significantly among provinces. Provinces located on the western side of Indonesia, particularly West Sumatra, Bengkulu, and Aceh, generally recorded higher rainfall values than many other regions. This condition can be associated with their geographical position and proximity to the Bukit Barisan mountain range, which enhances orographic rainfall through the uplift of moist air masses. As moist air is forced to rise along mountainous terrain, it cools and condenses, leading to increased cloud formation and precipitation. In addition, the western regions of Indonesia are directly influenced by moisture transport from the Indian Ocean, which can further contribute to higher

rainfall intensity and frequency. These factors collectively explain the relatively high rainfall observed in these provinces and highlight the strong influence of geographical and atmospheric conditions on regional rainfall patterns across Indonesia.

Conversely, several provinces in Sulawesi recorded relatively low rainfall values during the observation period. Central Sulawesi exhibited the lowest average rainfall among all provinces, which may be influenced by regional atmospheric circulation patterns and uneven cloud distribution. These results demonstrate that geographical and atmospheric factors play an important role in determining rainfall variability across Indonesia. Differences in topography, proximity to moisture sources, prevailing wind patterns, and local climatic conditions can significantly affect the amount and distribution of rainfall received in each region. The relatively low rainfall observed in several provinces may also reflect seasonal climatic influences that reduce precipitation intensity during specific periods. Understanding these regional differences is important for improving climate assessments and supporting effective planning in sectors such as agriculture, water resource management, and environmental conservation.

The occurrence of a maximum rainfall value of 186.4 mm in Bengkulu further indicates that extreme rainfall events can occur even when average rainfall values remain within moderate ranges. Therefore, descriptive statistical analysis provides useful information for understanding both general rainfall conditions and extreme weather events that may affect environmental management, agriculture, and disaster mitigation efforts. These findings emphasize the

importance of examining not only average rainfall values but also maximum and minimum observations to obtain a more comprehensive understanding of regional rainfall characteristics. The presence of extreme rainfall events may increase the risk of flooding, landslides, and other hydrometeorological hazards, particularly in areas with vulnerable geographical and environmental conditions. Consequently, continuous monitoring and analysis of rainfall data are essential for improving preparedness and supporting the development of effective early warning systems. In addition, the identification of provinces experiencing high rainfall variability can assist government agencies and related stakeholders in implementing more targeted mitigation and adaptation strategies. Such information is also valuable for agricultural planning, water resource allocation, and infrastructure development, as rainfall patterns directly influence crop productivity, water availability, and the resilience of public facilities to extreme weather conditions.

Moreover, the identification of provinces experiencing both high average rainfall and extreme precipitation events can provide valuable information for regional development planning. Areas characterized by high rainfall intensity may require improved drainage infrastructure, flood control systems, and enhanced disaster preparedness measures. Conversely, regions with relatively low rainfall may face challenges related to water availability and agricultural sustainability. Therefore, rainfall variability analysis not only contributes to climatological research but also supports practical decision-making for sustainable regional development and resource management.

Apache Spark Performance in Rainfall Data Processing

Apache Spark was successfully utilized to process rainfall data from 34 provinces in Indonesia through an integrated workflow consisting of data ingestion, data cleaning, data transformation, and statistical analysis. The framework enabled automatic processing of multiple CSV files and simplified the management of large climatological datasets. Its distributed computing architecture allowed data processing tasks to be executed efficiently across multiple computational resources, reducing the time required for analysis. The use of Spark DataFrame also facilitated structured data manipulation and streamlined the implementation of statistical aggregation functions. These capabilities enhanced the overall efficiency of the analytical process and demonstrated the suitability of Apache Spark for handling large-scale climatological datasets in a reliable and scalable manner..

The implementation of Spark DataFrame facilitated efficient data manipulation and aggregation operations. Statistical calculations, including record counting, average rainfall, maximum rainfall, and minimum rainfall values, were performed using built-in Spark functions such as count(), avg(), max(), and min(). These functions allowed rainfall statistics to be generated efficiently without requiring complex manual processing. The use of built-in aggregation functions reduced computational complexity and simplified the analytical workflow. Furthermore, Spark DataFrame provides an optimized execution engine that improves processing performance when handling large datasets. This capability enables researchers to perform statistical analyses more efficiently while maintaining data accuracy and consistency throughout the analytical process..

In addition, Apache Spark provided a structured, scalable, and efficient environment for rainfall data analysis. Compared with conventional spreadsheet-based approaches and single-machine computation, Spark offers greater flexibility in handling multiple datasets simultaneously, enabling researchers to process, integrate, and analyze large volumes of climatological data in a

systematic manner. Its distributed computing architecture and in-memory processing capabilities contribute to improved computational performance, reduced processing time, and enhanced scalability when dealing with increasingly complex datasets. Furthermore, Spark supports parallel data processing and automated analytical workflows, allowing statistical computations to be performed more efficiently while maintaining data consistency and reliability.

Another important advantage of Apache Spark is its ability to support future integration with advanced analytical techniques, including machine learning, predictive modeling, and real-time data processing. As climatological datasets continue to expand due to the increasing number of observation stations and automated monitoring systems, the need for scalable analytical platforms becomes increasingly critical. Apache Spark provides a flexible ecosystem thatenables researchers to process large volumes of data efficiently while maintaining analytical accuracy. This capability is particularly beneficial for climate-related studies that require continuous monitoring and rapid analysis of environmental changes. Consequently, the adoption of Apache Spark can facilitate the development of more sophisticated climate information systems and support data-driven environmental decision-making.

CONCLUSIONS

This study successfully demonstrated the application of Apache Spark as a Big Data processing framework for analyzing rainfall data in Indonesia using BMKG datasets. The research was conducted using a quantitative descriptive approach with data processing stages including data ingestion, data cleaning, data transformation, and descriptive statistical analysis. The implementation of PySpark enabled efficient processing of rainfall data from 34 provinces and supported large-scale data handling in a structured and automated manner.

The results of the analysis show significant spatial variability of rainfall across Indonesia during the observation period. West Sumatra Province recorded the highest average rainfall at 16.01 mm/day, while Central Sulawesi recorded the lowest average rainfall at 0.30 mm/day. In addition, Bengkulu Province experienced the highest maximum rainfall value of186.4 mm, indicating the occurrence of extreme rainfall events. These findings highlight the diversity of rainfall distribution patterns across different regions in Indonesia and provide important information for understanding climatological conditions

Furthermore, the implementation of Apache Spark proved to be effective in processing and analyzing large-scale climatological data. The use of Spark DataFrame and built-in aggregation functions such as count(), avg(), max(), and min() enabled efficient statistical computation and reduced processing complexity compared to conventional methods. The distributed and in-memory computing capabilities of Apache Spark also contribute to faster data processing and improved scalability for large datasets.

Overall, the findings of this study indicate that Apache Spark is a reliable and efficient framework for rainfall data analysis and has strong potential for supporting decision-making in agriculture, water resource management,and disaster mitigation Future research is recommended to extend this approach by integrating predictive analytics or machine learning models to further enhance rainfall forecasting and climate pattern analysis. In addition, the results obtained in this study demonstrate that Big Data technologies can play a significant role in improving the efficiency and effectiveness of climatological data analysis, particularly when dealing with large and continuously growing datasets. The proposed analytical framework can serve as a reference for

future environmental and climate-related studies that require scalable data processing solutions. Moreover, the insights generated from rainfall data analysis can contribute to the development of evidence-based policies and strategic planning aimed at enhancing climate resilience and sustainable resource management across Indonesia.

Furthermore, this study highlights the potential of integrating Big Data technologies into climatological research to address the increasing challenges posed by climate variability and environmental change. By utilizing Apache Spark for large-scale rainfall data processing, the study demonstrates a practical and scalable approach that can support more comprehensive climate assessments and data-driven decision-making.

REFERENCES

World Meteorological Organization (WMO), State of Climate Services 2021: Water. Geneva, Switzerland: WMO, 2021

Intergovernmental Panel on Climate Change (IPCC), Climate Change 2021: The Physical Science Basis. Cambridge, U.K.: Cambridge University Press, 2021

World Meteorological Organization (WMO), State of the Global Climate 2023. Geneva, Switzerland: WMO, 2024.

United Nations, Global Water and Climate Assessment Report. New York, NY, USA: United Nations, 2023.

Food and Agriculture Organization (FAO), The State of Food and Agriculture 2023. Rome, Italy: FAO, 2023.

World Meteorological Organization (WMO), Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2021). Geneva, Switzerland: WMO, 2023

United Nations Office for Disaster Risk Reduction (UNDRR), Global Assessment Report on Disaster Risk Reduction 2022. Geneva, Switzerland: UNDRR, 2022.

Intergovernmental Panel on Climate Change (IPCC), Climate Change 2022: Impacts, Adaptation and Vulnerability. Cambridge, U.K.: Cambridge University Press, 2022.

World Meteorological Organization (WMO), State of Climate Services 2022. Geneva, Switzerland: WMO, 2022

A. Pratama, D. S. Wibowo, and L. Kurniawan, “Big data characteristics in meteorological datasets for climate monitoring,” Journal of Environmental Informatics, vol. 39, no. 2, pp. 88–97, 2023.

X. Liu, Y. Sun, and M. Chen, “Distributed computing frameworks for large-scale climate data analytics,” Future Generation Computer Systems, vol. 143, pp. 275–288, 2023.

R. Singh and P. Gupta, “Statistical analysis of climate data using distributed computing frameworks,” Big Data Research, vol. 31, 2024.

M. Zaharia et al., “Apache Spark: A Unified Engine for Large-Scale Data Processing,” Communications of the ACM, vol. 64, no. 6, pp. 56–65, 2021.

T. Brown and R. Wilson, “Climate data processing using Apache Spark: A scalable framework for meteorological analysis,” Journal of Big Data, vol. 10, no. 1, 2023

M. Rodriguez and A. Torres, “Efficient meteorological data processing using Apache Spark and cloud computing,” IEEE Access, vol. 13, pp. 11245–11260, 2025

Apache Software Foundation, Apache Spark Documentation: PySpark API Reference, 2025

C. M. Liyew and H. A. Melese, “Machine learning techniques to predict daily rainfall amount,” Journal of Big Data, vol. 8, no. 1, 2021