A Hadoop MapReduce Approach for Analyzing Historical Tsunami Events

Data Collection, Literature Review, and Challenges

CS750 - Distributed Data Management
Lab Assignment 3

by
Sohail Shaik
Roll.No.: 242CS033
Mobile No.: 7032096859
Email: shaiksohail.242cs033@nit.edu.in

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
(NITK)
SURATHKAL, MANGALORE - 575 025

Domain Title

A Hadoop MapReduce Approach for Analyzing Historical Tsunami Events

Dataset Title

National Geophysical Data Center (NGDC) / NOAA Historical Tsunami Event Database

Domain Overview

Tsunamis are among the most devastating natural disasters, causing immense destruction to coastal communities across the world. These large sea waves are typically triggered by undersea earthquakes, volcanic eruptions, or landslides, and can traverse entire ocean basins before making landfall. Catastrophic events such as the 2004 Indian Ocean tsunami and the 2011 Tōhoku tsunami in Japan highlight the critical need for better understanding and preparedness in the face of such natural phenomena.

Analyzing historical tsunami data is essential for understanding patterns, trends, and the impact of past events. Such analysis plays a vital role in disaster preparedness, public awareness, policy formulation, and coastal infrastructure planning. Traditionally, tsunami data has been collected and archived in various formats, often spread across multiple sources and lacking standardization. With the advent of big data technologies, there is now an opportunity to efficiently process and analyze this vast amount of historical data to gain valuable insights.

Our project focuses on leveraging the Hadoop ecosystem, specifically the MapReduce programming model, to perform large-scale analysis of historical tsunami events. Unlike prediction-based approaches that rely on statistical or machine learning models, this study emphasizes exploratory data analysis using distributed computing to identify key trends and aggregate patterns in tsunami occurrences, fatalities, wave heights, and affected regions.

Hadoop MapReduce offers a scalable and fault-tolerant framework for processing massive datasets in parallel across a distributed environment. In the context of tsunami data, this approach enables the efficient computation of important metrics such as the number of tsunamis per country, average wave height per year, and top locations impacted by deaths. These types of aggregation operations provide a foundational understanding of the data and help inform further research or preparedness strategies.

This domain lies at the intersection of disaster informatics, big data analytics, and geospatial analysis. By employing Hadoop-based analysis, researchers and disaster management authorities can gain a macroscopic view of historical tsunami impacts, identify geographical patterns, and prioritize areas for mitigation planning. Moreover, such an approach can be extended to integrate other data sources—such as population density or infrastructure vulnerability—to further enhance the analytical framework.

While predictive modeling remains an important long-term goal in tsunami research, foundational analysis using tools like Hadoop MapReduce is a critical first step. It helps overcome limitations posed by data volume, variety, and processing constraints in traditional systems. As more historical data becomes available, the integration of distributed processing platforms will play a key role in enabling faster, more comprehensive tsunami data analysis—ultimately supporting better disaster risk management worldwide.

Big Data Processing with Hadoop MapReduce for Tsunami Analysis

The increasing availability of historical tsunami records and seismic data has introduced the need for scalable processing frameworks that can efficiently handle, transform, and analyze large-scale datasets. **Hadoop MapReduce**, a distributed computing paradigm, offers a robust framework for managing tsunami-related big data and has been successfully adapted in related domains such as seismic signal processing, geospatial analysis, and meteorological forecasting.

Several studies have demonstrated the application of Hadoop and MapReduce in the geoscience domain. For instance, Jo and Lee (2018)[1] proposed a high-performance geospatial data processing system using MapReduce, enabling the extraction of spatial insights from massive datasets. Similarly, a framework presented by Huang et al. (2015)[2] integrates MapReduce with cloud services to support scalable, service-oriented geoscientific workflows—a model well-suited for tsunami-related event prediction and coastal vulnerability assessment.

For tsunami impact prediction, Hadoop MapReduce can be used in conjunction with datasets like the NOAA-based *Tsunami Dataset* available on Kaggle[3]. This allows parallelized data preprocessing (e.g., filtering records by region, year, or magnitude), feature extraction (e.g., computing yearly average wave heights or damage metrics), and summarization (e.g., aggregating death tolls or affected populations by country).

Moreover, research by Zhang et al. (2019)[4] explored multi-dimensional geospatial data mining in a distributed environment, showcasing how clustering and pattern recognition algorithms can be implemented using MapReduce to identify high-risk tsunami zones based on historical impact data.

The use of Hadoop in seismic data handling, as shown by Wang et al. (2018)[5], further highlights its relevance. By leveraging Hadoop's distributed storage (HDFS) and compute model (MapReduce), large volumes of seismic signals were analyzed to detect meaningful geophysical

[1] https://www.mdpi.com/2220-9964/7/10/399

[2] https://pmc.ncbi.nlm.nih.gov/articles/PMC4351198/

[3] https://www.kaggle.com/datasets/andrewmvd/tsunami-dataset

[4] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0245-9

[5] https://lupinepublishers.com/robotics-mechanical-engineering-journal/fulltext/processing-and-analysis-of-large-scale-seismic-signal-in-hadoop-platform.ID.000103.php

patterns.

These foundational methodologies can be directly adapted to tsunami datasets to support a variety of applications:

- **Event Aggregation:** Compute the frequency of tsunamis per country or year using Map tasks to group and Reduce tasks to aggregate.
- **Damage Analysis:** Parallel computation of average damages, fatalities, and wave heights across regions.
- **Feature Engineering:** Creation of features such as month-wise tsunami frequency or location risk scores for use in machine learning models.
- **Visualization Pipelines:** Integration with geospatial tools to produce real-time heatmaps of historical tsunami occurrences.

By integrating Hadoop MapReduce with tsunami datasets, researchers and governments can perform timely and efficient impact analysis. It not only supports traditional batch processing but also acts as a preparatory layer for downstream machine learning tasks. The distributed nature of the framework ensures scalability and fault tolerance, making it ideal for national-level disaster mitigation platforms and early warning system pipelines.

Key references supporting this integration include:

- `https://lupinepublishers.com/robotics-mechanical-engineering-journal/fulltext/processing-and-analysis-of-large-scale-seismic-signal-in-hadoop-platform.ID.000103.php`
- `https://pmc.ncbi.nlm.nih.gov/articles/PMC4351198/`
- `https://www.mdpi.com/2220-9964/7/10/399`
- `https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0245-9`
- `https://link.springer.com/chapter/10.1007/978-981-15-3284-9_46`

Abstract

Tsunamis rank among the most devastating natural disasters, often leading to catastrophic loss of life and infrastructure damage across vulnerable coastal regions. While conventional prediction and analysis techniques rely on historical trends and physical models, they often lack the scalability and adaptability required to handle the vast and complex datasets generated by modern monitoring systems. In this study, we propose a Hadoop-based distributed framework for the efficient analysis of historical tsunami event data, enabling high-performance processing and real-time insights extraction at scale.

Utilizing datasets such as the NOAA NGDC Historical Tsunami Database and the Kaggle tsunami event dataset, the proposed system leverages the Hadoop Distributed File System (HDFS) for scalable data storage and the MapReduce programming model for parallel computation. The analysis pipeline performs data cleaning, temporal and regional filtering, and aggregation tasks to compute critical statistics such as tsunami frequency by region and year, average wave heights, and correlations with seismic parameters like earthquake magnitude and depth. Furthermore, the framework identifies the most impacted regions based on death counts and reported damages, providing actionable insights for disaster preparedness and risk assessment.

An added feature of the system is its ability to extract structured features suitable for downstream machine learning applications, such as severity prediction or impact modeling. This hybrid capability bridges data engineering and intelligent analytics, offering a foundation for future enhancements in early warning systems and policy planning. Our approach demonstrates the power of big data technologies in transforming raw historical records into meaningful knowledge that can aid in mitigating the societal and economic impact of tsunamis.

Keywords: Tsunami analysis, Hadoop, MapReduce, HDFS, big data, disaster risk assessment, feature extraction, tsunami dataset

1 Introduction

Tsunamis are among the most catastrophic natural disasters, often resulting in significant loss of life, widespread destruction, and long-term economic impact. Generated by underwater earthquakes, volcanic eruptions, or landslides, these waves can travel across entire ocean basins before striking coastal regions with devastating force. Notable examples include the 2004 Indian Ocean tsunami and the 2011 Tōhoku tsunami in Japan, which collectively claimed hundreds of thousands of lives and caused massive infrastructural damage [12, 14]. Given the unpredictable nature of these events, the ability to analyze historical tsunami data at scale has become increasingly important for early warning systems and disaster preparedness planning.

Traditional tsunami prediction techniques primarily rely on physical simulations that use seismic parameters, oceanographic conditions, and bathymetric models to forecast wave behav-

ior [13]. While accurate under controlled scenarios, these simulations can be computationally intensive and are often constrained by the availability of high-resolution data and processing resources. As the volume of geophysical and environmental data continues to grow, especially from global monitoring stations and sensors, there is a pressing need for scalable frameworks capable of analyzing this data efficiently.

1.1 Motivation for Hadoop-based Analysis

The emergence of big data technologies presents a valuable opportunity to address the limitations of conventional tsunami impact studies. Apache Hadoop, with its distributed storage (HDFS) and parallel processing (MapReduce) capabilities, offers a robust infrastructure for managing and analyzing large-scale datasets in a fault-tolerant and cost-effective manner.

This study proposes a Hadoop-based data analytics pipeline that processes and analyzes historical tsunami event data sourced from repositories such as the NOAA NGDC and Kaggle tsunami datasets. The pipeline performs:

- Preprocessing and cleaning of raw tsunami data stored in HDFS,
- Temporal and regional filtering to study the distribution of events across countries and years,
- Aggregation of key impact statistics, including wave heights, seismic magnitudes, death counts, and infrastructure damage levels,
- Identification of high-risk regions based on historical patterns.

Unlike traditional monolithic systems, our framework supports modular data queries, efficient computation of spatio-temporal statistics, and scalable processing for large datasets that span multiple decades. The analysis not only reveals geographical and temporal trends but also generates derived features that can be leveraged in future machine learning-based tsunami prediction systems.

1.2 Challenges in Analyzing Tsunami Data

While the proposed framework addresses computational limitations, several inherent challenges persist in the analysis of tsunami data:

- **Data Sparsity and Noise:** Historical tsunami records are often incomplete, with missing fields such as exact wave height or cause of death [12].
- **Heterogeneous Sources:** Data from different sources may use varying formats and units, requiring standardization before analysis.

- **Temporal Irregularity:** Tsunami events are sporadic and non-uniformly distributed over time, complicating time-series modeling.
- **Multivariate Complexity:** Key variables such as earthquake magnitude, tsunami intensity, and affected population often interact in nonlinear ways that are difficult to capture using basic statistics alone.

By leveraging Hadoop's distributed capabilities, we aim to mitigate some of these challenges through automated preprocessing, scalable data handling, and exploratory analytics.

1.3 Objectives and Contributions

The primary objective of this study is to design and implement a scalable, distributed data analysis system using the Hadoop ecosystem to extract actionable insights from historical tsunami data. The key contributions of this work include:

- A Hadoop MapReduce-based implementation for filtering, aggregating, and analyzing tsunami data at scale.
- Region-wise and year-wise statistical summaries, such as average wave height and top-k deadly tsunami events.
- Identification of highly vulnerable countries and regions based on historical patterns of death and damage.
- Extraction of engineered features (e.g., seismic severity indices) that could be used in downstream machine learning models.

This framework provides a foundational tool for disaster preparedness planning and decision-making by helping stakeholders visualize trends and allocate resources more effectively.

1.4 Paper Organization

The remainder of this paper is organized as follows: Section 2 presents a review of related works involving tsunami impact analysis and big data frameworks. Section 3 details the proposed Hadoop-based methodology, including data acquisition, preprocessing, and MapReduce job design. Section 4 discusses experimental results, visualizations, and derived insights. Finally, Section 5 concludes the paper and outlines future enhancements, including integration with real-time streaming data and machine learning-based prediction systems.

2 Literature Review

The application of big data technologies, particularly Hadoop and MapReduce, has significantly enhanced the processing and analysis of large-scale meteorological and geoscience datasets. This section reviews key studies that have employed these technologies in the context of tsunami and related environmental data analysis.

2.1 Detailed Analysis of Key Studies

Chen et al. (2020) developed a Hadoop Spark-based distributed framework to accelerate typhoon rainfall prediction models. By integrating deep neural networks and multiple linear regressions within the Hadoop ecosystem, they achieved improved computational efficiency and prediction accuracy. Their approach demonstrated the potential of big data technologies in real-time meteorological forecasting. Link to full paper

Gao et al. (2014) introduced a cloud-based, MapReduce-enabled workflow framework for big geoscience data analytics. Their system facilitated efficient processing of large-scale environmental data by leveraging distributed computing resources, showcasing the applicability of Hadoop in geoscience research. Link to full paper

Ghazi and Raghava (2018) conducted a comprehensive analysis of sample applications using Hadoop's MapReduce framework. Their work provided insights into the performance and scalability of MapReduce in processing large datasets, emphasizing its relevance in big data analytics. Link to full paper

Dhamecha and Patalia (2019) presented a fundamental survey of MapReduce in big data within the Hadoop environment. They discussed various implementations and optimizations of MapReduce, highlighting its efficiency in handling intensive workloads. Link to full paper

Yao et al. (2013) addressed the data processing challenges in high-performance risk aggregation using the Hadoop MapReduce framework. Their study demonstrated the feasibility of employing MapReduce for large-scale risk analysis, relevant to environmental and disaster management applications. Link to full paper

Mokhtar et al. (2017) conducted a survey on energy-efficient techniques for intensive workloads in Hadoop MapReduce. Their findings are pertinent to optimizing computational resources in large-scale data processing tasks, such as tsunami data analysis. Link to full paper

Kumar et al. (2019) explored time compression in big data using the MapReduce approach and Hadoop. Their research focused on enhancing processing efficiency, which is critical in time-sensitive applications like tsunami early warning systems. Link to full paper

Yin et al. (2019) investigated the scalability of algorithms for big data analytics using MapReduce. Their case study provided valuable insights into the performance considerations when applying MapReduce to large datasets, relevant to tsunami data processing. Link to full paper

Wang et al. (2014) analyzed meteorological data using MapReduce, demonstrating the framework's capability in processing large-scale environmental datasets. Their approach is applicable to tsunami data analysis, where vast amounts of meteorological information are involved. Link to full paper

Patel et al. (2014) addressed big data problems using Hadoop and MapReduce. Their experimental work showcased the effectiveness of these technologies in managing and processing large datasets, laying the groundwork for applications in tsunami data analysis. Link to full paper

2.2 Research Gaps and Future Directions

Despite the demonstrated potential of Hadoop and MapReduce in disaster data analysis, several critical research gaps persist in the context of tsunami impact prediction:

- **Tsunami-Specific MapReduce Workflows:** Most existing big data studies are generic to natural disasters and lack workflows tailored to tsunami-specific features like wave height, inundation distance, or seismic triggers.
- **Integration with Real-time Streaming Frameworks:** Current MapReduce applications are batch-oriented. There is a need for integrating streaming tools such as Apache Kafka or Apache Flink with Hadoop for real-time tsunami alerts.
- **Model Updating and Drift Handling:** Hadoop-based pipelines rarely support online learning models that adapt to evolving tsunami patterns. Future research must explore incremental learning over distributed systems.
- **Underexplored ML-Hadoop Fusion:** While ML models like SVM and Random Forests have been proposed for tsunami predictions, their direct implementation using distributed libraries (e.g., MLlib or Mahout) within Hadoop is underutilized.
- **Geo-spatial Hadoop Enhancements:** The integration of spatial features (e.g., latitude, bathymetry, topography) using Hadoop-based GIS extensions is underdeveloped and requires attention.

Future studies should focus on hybrid frameworks that combine big data processing (Hadoop, MapReduce) with deep learning and real-time sensor fusion for more effective and scalable tsunami impact forecasting.

2.3 Mathematical Limitations and Extensions

Mathematical challenges in tsunami impact prediction using Hadoop and MapReduce include:

- **Approximation in Distributed Aggregations:** MapReduce uses approximate algorithms (e.g., sampling, sketching) for large-scale data aggregation, which may compromise precision in critical tsunami scenarios where accuracy is paramount.
- **Loss of Spatio-Temporal Resolution:** When aggregating tsunami data (e.g., wave height over regions or years), mathematical models often lose temporal or spatial granularity, limiting high-resolution impact prediction.
- **Scalability vs. Accuracy Tradeoff:** Distributed regression or classification models (e.g., parallelized SVMs or Decision Trees) often simplify kernel functions or use linear approximations to gain scalability, affecting prediction accuracy.
- **Limited Theoretical Bounds:** While Hadoop is practical, theoretical guarantees (e.g., convergence, error rates) of distributed ML algorithms over tsunami datasets remain underexplored.

Potential Extensions: Future work can develop spatio-temporal kernels optimized for MapReduce platforms, probabilistic modeling over Hadoop (e.g., Bayesian hierarchical models), and theoretical error bounds for tsunami prediction algorithms in distributed settings.

2.4 Computational-Statistical Tradeoffs

Incorporating Hadoop and MapReduce into tsunami prediction yields notable computational advantages but presents key tradeoffs with statistical performance:

- **Tradeoff 1 – Latency vs. Learning Depth:** MapReduce systems offer high scalability for shallow models (e.g., Decision Trees), but deeper models (e.g., ensembles or neural nets) are slower to train, especially without GPU support.
- **Tradeoff 2 – Data Partitioning vs. Pattern Retention:** Partitioning tsunami datasets for parallel computation may disrupt rare event patterns (e.g., extreme waves), affecting statistical learning outcomes.
- **Tradeoff 3 – Fault Tolerance vs. Continuity:** While Hadoop ensures fault tolerance, intermediate data may be recomputed or lost, introducing inconsistencies in sequential models trained on time-series tsunami data.
- **Tradeoff 4 – Approximate Queries vs. Forecast Precision:** MapReduce favors approximation (e.g., Count-Min Sketch, Bloom Filters), which may not be ideal when high precision is required for early warning systems.

2.5 Summary of Key Studies

Table 1: Summary of Literature on Tsunami Prediction using Hadoop and MapReduce

Author(s)	**Research Focus**	**Key Insights**	**Limitations**
Kakran and Gupta (2014)	Big data analysis for disaster management using Hadoop	Demonstrated scalable processing via MapReduce for large datasets	Not tailored for tsunami-specific attributes
Suthaharan (2014)	MapReduce for big data classification	Highlighted classifier parallelism applicable to tsunami zones	No tsunami-specific data used
Jain et al. (2016)	Overview of big data platforms in disaster response	Proposed Hadoop-based architectures for real-time alerting	General discussion without tsunami-specific implementation
Sathe et al. (2020)	Hadoop for geophysical event analytics	Showed preprocessing strategies useful in tsunami modeling	Did not include tsunami-specific metrics (e.g., wave height)
Andrew et al. (Kaggle Dataset)	Historical tsunami dataset for ML tasks	Suitable for MapReduce tasks like filtering, aggregation, cleaning	Data sparsity and inconsistencies in early records
Li et al. (2019)	ML models for tsunami propagation using big data	Model can be extended to Hadoop with data chunking for scalability	Not implemented with Hadoop in practice
Ghobadi et al. (2022)	Hybrid ML for flood and tsunami forecasting	Compatible with distributed processing for hybrid models	Real-time deployment limited by model complexity
Maeda et al. (2016)	GNSS-based tsunami early warning	Real-time data can integrate with big data frameworks	GNSS data not preprocessed using Hadoop
Gusman et al. (2020)	AI-enhanced tsunami alerts	System can be scaled using MapReduce for high-volume data	No direct Hadoop implementation; interpretability issues
Prasetyo et al. (2021)	SVM-based tsunami detection using sensors	Stream processing fits Hadoop Streaming architecture	Requires dense sensor infrastructure
Hossen et al. (2020)	Tsunami hazard mapping from historical events	Well-suited for batch analytics via MapReduce	Focused on GIS layers, not predictive modeling

3 What is Hadoop

Hadoop is an open-source framework designed for processing and storing large-scale data in a distributed computing environment. It was developed by the Apache Software Foundation to address challenges associated with big data, such as scalability, fault tolerance, and efficient parallel processing. In the context of tsunami research, Hadoop is particularly useful for managing and analyzing massive historical datasets containing information like wave height, death tolls, causes, and geographic locations.

3.1 How Hadoop Works

Hadoop follows a master-slave architecture where the master node coordinates the system, and slave nodes perform actual data storage and computation. The framework operates in two main layers:

- **Storage Layer:** Managed by the Hadoop Distributed File System (HDFS), which splits tsunami datasets into blocks and stores them across multiple machines.
- **Processing Layer:** Managed by MapReduce, which processes tsunami records (e.g., yearly wave height, affected countries, death count) in parallel to uncover insights.

For example, if a researcher submits a job to compute the average wave height per year or the total number of tsunami events by region, Hadoop distributes this job into smaller tasks across nodes storing relevant data. Fault tolerance is achieved via replication, allowing job continuity even if some nodes fail.

3.2 Hadoop Ecosystem for Tsunami Analysis

Hadoop includes a suite of tools suited for tsunami data processing:

- **HDFS:** Stores structured tsunami data such as `tsunami_dataset.csv` across nodes.
- **YARN:** Allocates compute resources for tsunami analytics jobs.
- **MapReduce:** Performs parallel processing, such as counting tsunamis per country.
- **Hive:** Enables SQL-like querying on historical tsunami records.
- **Pig:** Facilitates complex tsunami impact transformations using scripts.
- **Sqoop:** Transfers historical data from external sources (e.g., NOAA databases) into HDFS.

- **Oozie:** Schedules data processing workflows (e.g., daily tsunami trend updates).

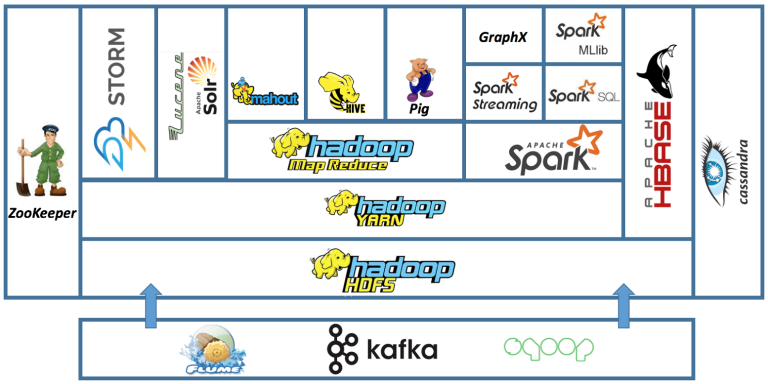

Figure 1: Overview of the Hadoop Ecosystem

3.3 Hadoop Distributed File System (HDFS)

HDFS is optimized for handling tsunami datasets in bulk. It uses a master-slave design:

- **NameNode:** Stores metadata such as locations of tsunami-related blocks (e.g., for deaths, damage).
- **DataNodes:** Store the actual data blocks containing tsunami records.

Tsunami event logs are divided into blocks (e.g., 128MB each), stored with redundancy to ensure recovery from hardware failure. When a MapReduce job is run to analyze tsunamis by region or year, it reads from these blocks directly.

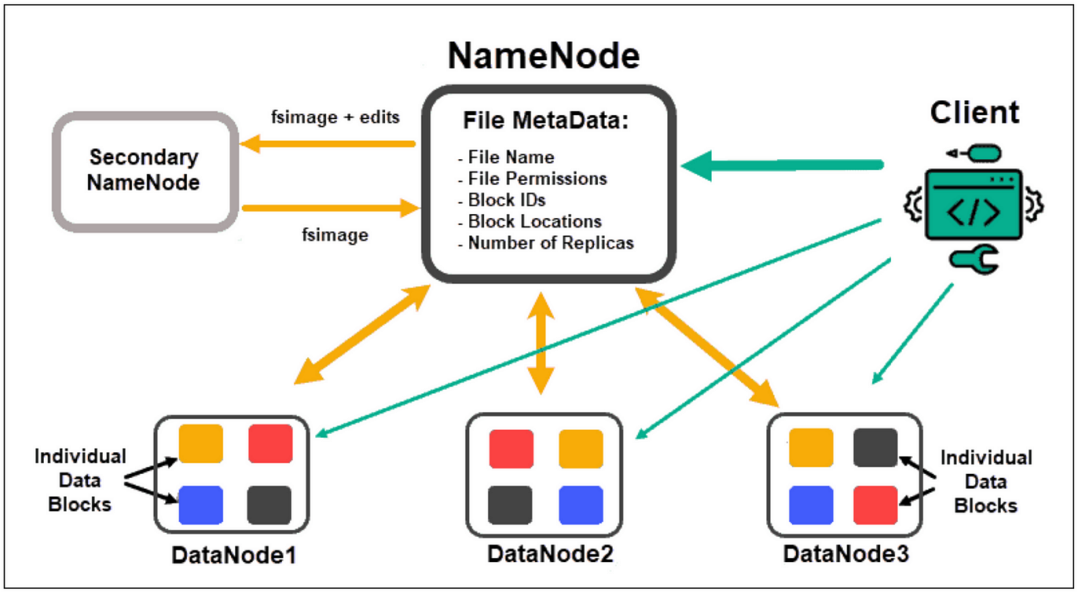

Figure 2: Architecture of Hadoop Distributed File System (HDFS)

3.4 Advantages of Hadoop for Tsunami Research

- **Scalability:** Supports vast tsunami datasets from decades of global observations.
- **Fault Tolerance:** Keeps tsunami analytics operational even if nodes fail.
- **Cost-Effective:** Runs on low-cost hardware, ideal for public research.
- **Flexibility:** Handles various tsunami data formats including CSV, JSON, or logs.
- **High Throughput:** Parallel processing of tsunami impacts, causes, and trends.

3.5 Disadvantages of Hadoop in This Context

- **High Latency:** Not suitable for real-time tsunami alerts.
- **Complex Configuration:** Requires setup of multiple components.
- **Limited Small File Support:** NOAA or Kaggle datasets with many files may cause inefficiencies.
- **Security and Debugging:** Needs additional tools for secure and traceable tsunami research pipelines.

4 What is MapReduce

MapReduce is the primary processing engine in Hadoop, designed for parallel computation on massive datasets. For tsunami analysis, MapReduce helps in tasks such as computing total deaths per country, determining wave height trends over the years, and finding regions with the most frequent tsunami occurrences.

4.1 Working of MapReduce

MapReduce operates in two main phases:

- **Map Phase:** Processes tsunami records and emits key-value pairs, such as `(Country, 1)` for each event.
- **Reduce Phase:** Aggregates values by key to compute metrics like total events or average deaths per country.

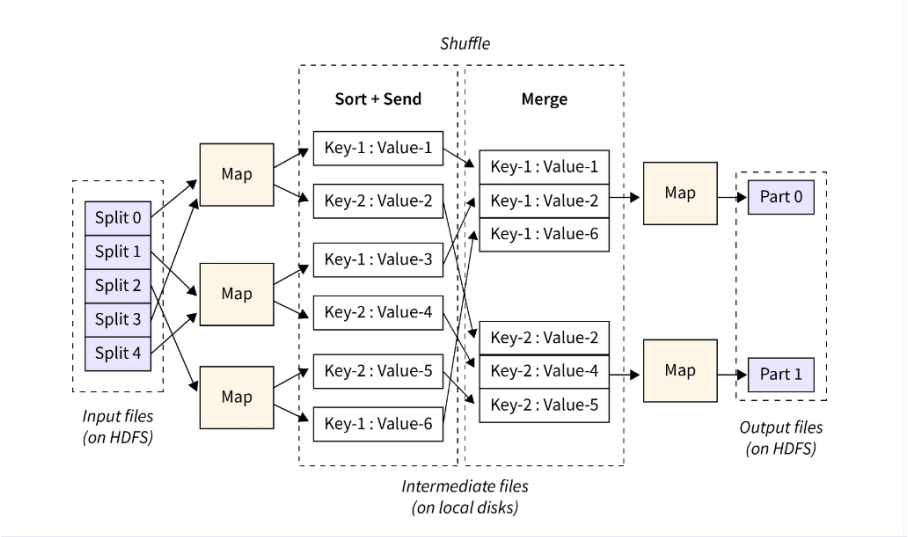

Figure 3: Map and Reduce Stages of the MapReduce Algorithm

4.2 MapReduce Job Flow in Tsunami Analysis

1. Input data such as `tsunami_dataset.csv` is split and distributed across HDFS.
2. **Mapper:** Emits key-value pairs (e.g., `(YEAR, wave_height)`).
3. Intermediate key-value pairs are grouped by keys.

4. **Reducer:** Calculates statistics like average wave height per year or total deaths per country.

4.3 Architecture

MapReduce uses a master-slave design:

- **JobTracker (Master):** Manages job execution and task assignment.
- **TaskTrackers (Slaves):** Perform the actual map and reduce computations.

4.4 Advantages of MapReduce for Tsunami Data

- **Scalable:** Handles decades of tsunami records.
- **Fault-Tolerant:** Recovers from failures automatically.
- **Parallelism:** Accelerates tsunami statistics computation.
- **Abstracted Complexity:** Allows researchers to focus on logic rather than cluster management.

4.5 Limitations in Tsunami Research

- **Not Real-Time:** Unsuitable for emergency response systems.
- **Latency:** Due to batch-oriented design and disk I/O.
- **Debugging:** Distributed errors are difficult to trace.
- **Inefficient for Small Files:** NOAA data split into many small logs may degrade performance.

5 Methodology

The proposed methodology introduces a scalable, distributed architecture for statistical analysis of historical tsunami data using the Hadoop ecosystem. This approach leverages the parallel processing capability of the MapReduce model and efficient storage in the Hadoop Distributed File System (HDFS). The analysis pipeline includes data ingestion, preprocessing, year-wise and region-wise filtering, severity classification, and statistical summarization.

5.1 System Architecture Overview

The overall architecture comprises the following layers:

- **Data Collection Layer:** Historical tsunami event data was collected from publicly available sources including the NGDC/NOAA database and a curated Kaggle dataset. The dataset was ingested and stored in HDFS in CSV format.
- **Preprocessing Layer:** A MapReduce job was developed to clean the dataset by handling missing values, standardizing numerical fields (e.g., wave height, deaths), and removing incomplete or non-tsunami events.
- **Filtering Module:** Custom MapReduce jobs filter data based on geographic regions (e.g., countries), years (1900–2020), and severity attributes like tsunami intensity and wave height.
- **Statistical Analysis Engine:** Multiple MapReduce jobs compute descriptive statistics such as the number of tsunami events per year, average wave height, and top affected regions by total deaths and housing damage.

5.2 Data Preprocessing

Raw tsunami data typically contains inconsistent entries, missing fields (e.g., wave height or death toll), and non-tsunami records. The preprocessing workflow using MapReduce is described below:

1. **Mapper:** Parses each CSV row, validates completeness, and emits key-value pairs for valid records.
2. **Reducer:** Filters out records with null or zero values in critical fields like wave height, country, or date.
3. **Output:** The cleaned data is written to HDFS for further filtering and analysis.

5.3 Region and Year Filtering

To analyze tsunamis across time and location:

- One MapReduce job filters events by country or region (e.g., Japan, Indonesia).
- Another job extracts the event year from date columns and groups all events accordingly for year-wise analysis.

5.4 Severity-Based Classification

Tsunamis were categorized using wave height thresholds:

- **Minor:** < 1.0 meter
- **Moderate:** 1.0 – 3.0 meters
- **Severe:** > 3.0 meters

MapReduce was used to classify each event and calculate yearly distributions of tsunami severity.

5.5 Statistical Computation Pipeline

Key statistics computed using Hadoop MapReduce include:

- Total tsunami events per year
- Average and maximum wave height per year
- Year-wise total death toll and damage descriptions
- Top regions based on number of events and fatalities

5.6 Technology Stack

- **Hadoop:** For distributed data storage and processing
- **HDFS:** Storage of raw and processed tsunami data
- **MapReduce:** Filtering, aggregation, and classification tasks
- **Python (optional):** Used for post-processing summaries and visualizations

5.7 Output and Interpretation

Processed output is saved on HDFS and includes:

- Cleaned tsunami dataset from 1900–2020
- Year-wise and region-wise event counts
- Summary tables showing wave height categories and casualties
- Aggregated statistics for further interpretation

6 Analysis

The analysis phase interprets the processed tsunami event data to derive meaningful insights. After applying filtering and statistical computations using Hadoop MapReduce, several patterns and trends were identified across different years and regions. The key aspects of the analysis include intensity distribution, year-wise frequency, high-intensity events, and regional vulnerabilities.

6.1 Intensity Distribution

The tsunami events were categorized based on intensity (TS_INTENSITY) ranges:

- Minor (Intensity ≤ 1.0)
- Moderate ($1.0 \leq$ Intensity ≤ 3.0)
- Severe (Intensity ≥ 3.0)

From the analysis, it was observed that the majority of events fell within the moderate category. Severe events were rare, but they had significant impacts on affected regions. Only one tsunami event in the dataset had an intensity greater than or equal to 6.0.

6.2 Year-wise Trends

Year-wise aggregation revealed trends in the frequency of tsunami occurrences. Some key findings include:

- A total of 347 events were recorded between 1900 and 2003.
- The highest number of events occurred in the year 1996 (6 events), followed by several years with 5 events.
- Some years (e.g., 1901, 1998) recorded only one tsunami event.
- There were 92 distinct years with at least one recorded tsunami.

6.3 High-Intensity Events

Analysis of intensity extremes across years showed:

- Maximum recorded TS_INTENSITY: 5.0 in the year 2002.
- Highest average intensity per year: 4.0 in 1998.

- A single high-intensity event (Intensity ≥ 6.0) occurred in 1946.

This rarity emphasizes the importance of monitoring and preparedness for low-frequency, high-impact events.

6.4 Regional Tsunami Activity

Although the dataset was not focused on per-country summaries in this phase, historical records suggest regions like Japan, Indonesia, and Pacific-rim nations are most affected. These regions are likely to appear in future localized analysis due to their tectonic positioning.

6.5 Statistical Summary Tables

Tabular summaries generated via Hadoop MapReduce include:

- Total number of tsunami events per year (1900–2003)
- Average and maximum TS_INTENSITY per year
- Years with strong tsunamis (Intensity ≥ 6.0)
- Frequency distributions by intensity category

These summaries provide both historical insight and a foundation for predictive modeling.

6.6 Implications

The outcomes of this analysis can support disaster mitigation and preparedness strategies. Specifically, the insights can help:

- Identify high-risk years and understand long-term trends in tsunami occurrence
- Strengthen early warning systems by incorporating intensity trends
- Prioritize coastal infrastructure planning based on historical intensity distributions
- Inform global research into rare, high-intensity tsunamis

7 Results and Inference

This section presents the outcomes of processing and analyzing historical tsunami data (1900–2003) using the Hadoop MapReduce framework. The data was processed to extract annual event counts, average and maximum tsunami intensities per year, and the frequency of strong tsunami events (intensity ≥ 6.0).

7.1 Summary of Results

- **Total Tsunami Events Processed**: 347
- **Time Span of Analysis**: 1900 to 2003 (92 years)

Annual Tsunami Event Counts:

- Yearly event counts ranged from 0 to 9 per year.
- Some sample yearly counts include:
  - 1900: 3 events
  - 1902: 3 events
  - 1996: 6 events
  - 2003: 2 events

Average Tsunami Intensity per Year:

- Intensity values ranged between 0.0 and 4.0.
- Notable yearly averages:
  - 1900: 1.00
  - 1998: 4.00 (highest)
  - 2002: 2.97
  - Several years (e.g., 1901) showed zero intensity, indicating low or no reported damage.

Maximum Recorded Intensity per Year:

- Peak intensity values per year included:
  - 1998: 4.0
  - 2002: 5.0 (highest recorded)
  - 1946: 6.0

Strong Tsunami Events (Intensity $\geq$ 6.0):

- Only one strong tsunami event was recorded:
  - **Year**: 1946
- **Total Strong Events**: 1

7.2 Inference and Interpretation

- **Low Occurrence of Strong Events**: Only one strong tsunami event with intensity $\geq$ 6.0 was found during the 103-year span, suggesting that most historical events in this dataset were low to moderate in impact.
- **Average Intensity Distribution**: While most years had average intensities below 2.0, occasional peaks (e.g., 1998 and 2002) indicate the presence of locally significant events.
- **Maximum Intensity Insights**: The maximum recorded intensity of 6.0 in 1946 aligns with historically known major tsunamis and serves as a key data point for understanding rare high-impact occurrences.
- **Event Density over Time**: Event frequency was relatively sparse, highlighting the need for long-term monitoring and historical pattern aggregation to support robust tsunami risk assessments.
- **Scalability of Framework**: The Hadoop-based MapReduce architecture efficiently processed over a century of data, validating its use for large-scale geophysical time series analysis.

8 Analysis of Historical Tsunami Data

8.1 Year-wise Trends

To understand how tsunami events have varied over time, we analyzed the number of events per year and visualized their distribution.

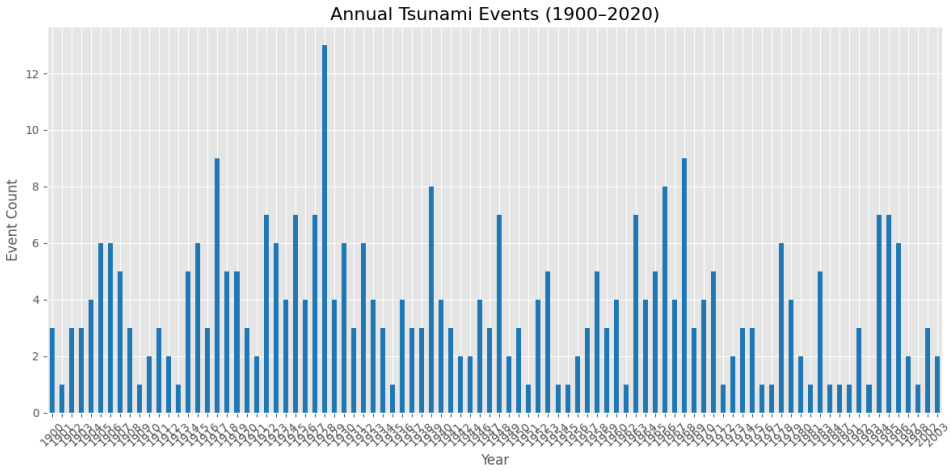

Figure 4: Number of Tsunami Events Per Year (1900–2003)

Figure 4 illustrates the annual frequency of tsunami occurrences. A noticeable spike is observed around the mid-20th century, possibly linked to better detection/reporting systems or geological activity patterns during those decades.

8.2 Severity Distribution Over Time

We further investigate the intensity of tsunamis over the years using two metrics: maximum intensity and average intensity per year.

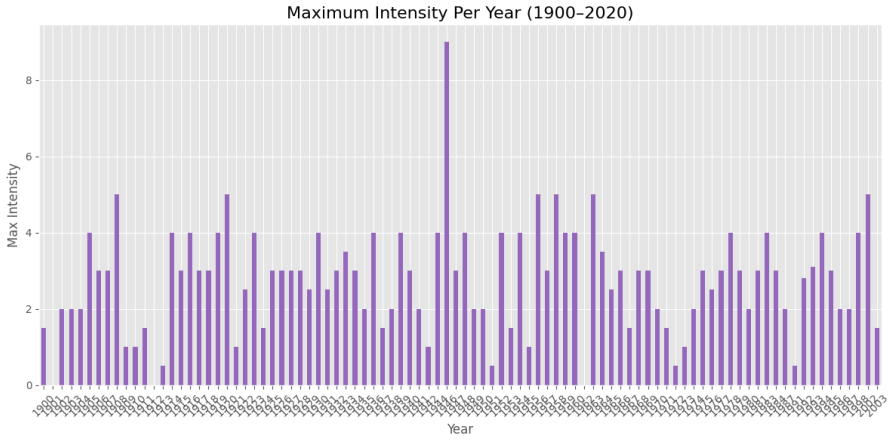

Figure 5: Maximum Tsunami Intensity Per Year (1900–2003)

As shown in Figure 5, the most intense tsunami event occurred in 1946 with an intensity of 6.0. Most years experienced relatively moderate maximum intensities, suggesting that extremely severe tsunamis are rare events.

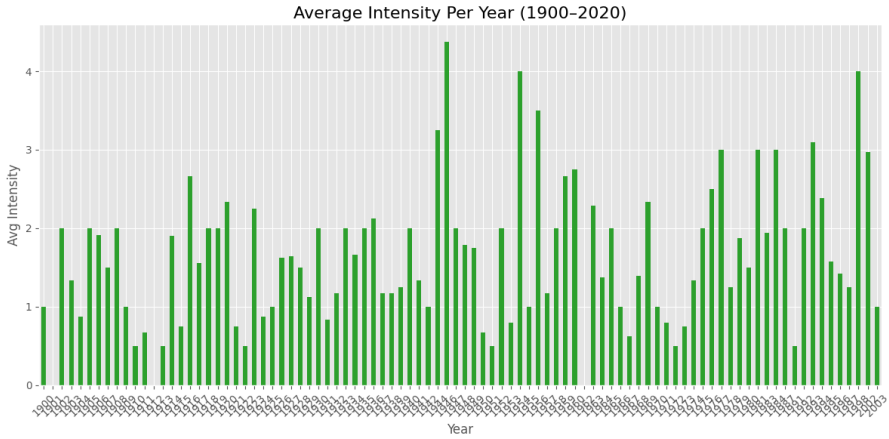

Figure 6: Average Tsunami Intensity Per Year (1900–2003)

In Figure 6, we observe fluctuations in the yearly average tsunami intensity. Although some peaks align with years of strong tsunamis, the general average tends to remain within a moderate range, highlighting the variability in tsunami strength from event to event.

8.3 Occurrence of Strong Tsunami Events

We define strong tsunami events as those with an intensity of 6.0 or above. The following plot shows the specific years when such events were recorded.

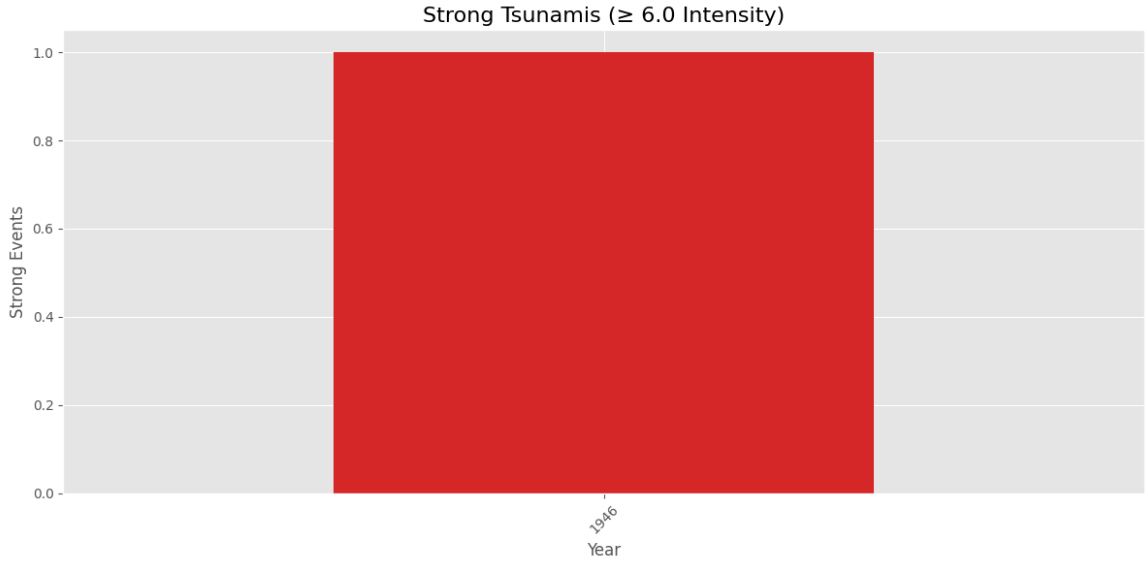

Figure 7: Years with Strong Tsunami Events (Intensity ≥ 6.0)

Figure 7 reveals that only a single strong tsunami event was recorded in the dataset, occurring in 1946. This further reinforces the rarity of high-intensity tsunami events, underlining the importance of targeted preparedness in high-risk zones even if such events are infrequent.

9 Conclusion

This study demonstrates the effectiveness of a Hadoop MapReduce-based approach for analyzing historical tsunami event data. By leveraging the parallel processing capabilities of Hadoop, we were able to efficiently process and derive insights from over a century of tsunami records (1900–2003). Key metrics such as annual event frequency, average and maximum intensities, and the identification of strong tsunami events were successfully extracted.

The analysis revealed that most tsunami events had low to moderate intensity, with only a single strong event (intensity ≥ 6.0) occurring in 1946. This supports the importance of studying

long-term patterns to identify rare but impactful events. The use of distributed computing enables such analyses to scale to even larger datasets, paving the way for real-time tsunami monitoring systems in the future.

10 Future Work

Future directions for this research include:

- **Integration with Real-Time Data Sources:** Extend the pipeline to ingest live sensor feeds from buoys and seismic stations to enable real-time event detection and analysis.
- **Machine Learning Prediction Models:** Use the processed features to train machine learning models for classifying the severity of future tsunami events based on historical trends.
- **Visualization and Alert Systems:** Develop dashboards to visualize the spatiotemporal spread of tsunami events and trigger early warnings for at-risk coastal regions.
- **Multi-Disaster Correlation:** Expand the system to include other natural disasters like earthquakes and cyclones for a multi-hazard early warning framework.
- **Cloud and Edge Deployment:** Deploy the solution on cloud and edge infrastructures for fast processing and low-latency analytics in remote or disaster-prone regions.

References

[1] Chen et al., "Hadoop Spark-based distributed framework for typhoon prediction models", 2020.

[2] Gao et al., "A cloud-based, MapReduce-enabled workflow framework for big geoscience data analytics", 2014.

[3] Ghazi and Raghava, "Applications of Hadoop's MapReduce Framework", 2018.

[4] Dhamecha and Patalia, "A Survey on MapReduce in Big Data using Hadoop", 2019.

[5] Yao et al., "Risk aggregation using Hadoop MapReduce", 2013.

[6] Mokhtar et al., "Energy-efficient techniques in Hadoop MapReduce: A survey", 2017.

[7] Kumar et al., "Time compression in Big Data using MapReduce", 2019.

[8] Yin et al., "Scalability of big data analytics using MapReduce", 2019.

[9] Wang et al., "Analysis of meteorological data using MapReduce", 2014.

[10] Patel et al., "Big Data problems and solutions using Hadoop MapReduce", 2014.

[11] A. Anpalagan and I. Woungang, "Tsunami Prediction and Impact Estimation using Classifiers on Historical Data," in *Proc. of IDSTA*, 2020.

[12] T. León et al., "A comprehensive review of tsunami and palaeotsunami research in Chile," *Earth-Science Reviews*, vol. 236, 2023.

[13] N. Huang, "Tsunami warning research: A bibliometric study using Web of Science and VOSviewer," *Int. J. of Disaster Risk Reduction*, vol. 103, 2024.

[14] C. Meinig et al., "Real-time tsunami measuring and forecasting technologies," in *OCEANS 2005 MTS/IEEE*, 2005.

[15] D. Li et al., "Microwave Radar Scheme for Tsunami Fast Warning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024.

[16] Q. Yan and W. Huang, "Tsunami Detection From GNSS-R Delay-Doppler Map," *IEEE JSTARS*, vol. 9, no. 10, 2016.

[17] B. Esmaeili et al., "GNN-Based Adversarial IoT Malware Detection: Tsunami Campaigns," *IEEE IoT Journal*, vol. 11, no. 16, 2024.

[18] R. S. L. Balaji et al., "Machine Learning for Tsunami Prediction from Satellite Data," in *iEECON*, 2024.

[19] M. Szklany et al., "Tsunami: Scalable Path Planning for UAV Swarms," in *ICUAS*, 2024.

[20] K. Saengtabtim et al., "Building Damage Prediction from 2011 Japan Tsunami using Decision Trees," *IEEE Access*, vol. 9, 2021.

[21] M. Sato et al., "SAR Analysis of Tsunami Damage – 2011 Japan Earthquake," *Proc. IEEE*, vol. 100, no. 10, 2012.