Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 2, No. 1 (2026)

Article

Study on Uncertainty Data Analysis for Common Natural Disaster Prediction in the U.S. Using Cloud Computing and Machine Learning

Guoli Ying [1,*]

[1] Carnegie Mellon University, Mountain View, California, United States
* Correspondence: Guoli Ying, Carnegie Mellon University, Mountain View, California, United States

Abstract: Accurate natural disaster prediction requires the integration of heterogeneous environmental datasets and the ability to model uncertainty arising from sensor noise, incomplete observations, and inconsistent spatiotemporal coverage. This study proposes a cloud-based, uncertainty-aware machine learning framework designed to support large-scale prediction of hurricanes, floods, and wildfires in the United States. The framework incorporates distributed data ingestion, cloud-native preprocessing, multi-source feature engineering, and probabilistic learning to address the complexity and variability inherent in environmental data. Through a unified cloud workflow, heterogeneous datasets from NOAA, USGS, and NASA FIRMS are harmonized using temporal alignment, spatial normalization, and uncertainty-reduction strategies such as imputation, smoothing filters, and cross-source calibration. Model evaluation focuses on performance trends rather than fixed numerical benchmarks, examining how different learning algorithms respond to the characteristics of each disaster type. Results indicate that deep models and uncertainty-aware approaches are particularly effective for highly dynamic hazards such as hurricanes and wildfires, while tree-based models perform well for structured hydrological predictions. A real-world wildfire event is used to illustrate the practical applicability of the framework. Overall, the proposed approach enhances robustness, interpretability, and operational relevance, demonstrating the value of combining cloud computing with uncertainty-aware machine learning for multi-hazard disaster prediction.

Keywords: cloud computing; multi-source environmental data; natural disaster prediction; data integration; probabilistic modeling; feature engineering

Received: 30 December 2025
Revised: 11 February 2026
Accepted: 22 February 2026
Published: 27 February 2026

1. Introduction

Natural disasters, including hurricanes, floods, wildfires, and tornadoes, pose significant threats to human life, infrastructure, and the economy in the United States. Accurate and timely prediction of such events is critical for disaster preparedness, risk mitigation, and emergency response. However, natural disaster prediction remains challenging due to the complexity, variability, and inherent uncertainty of environmental systems. Data collected from meteorological stations, hydrological sensors, and satellite observations are often incomplete, noisy, and heterogeneous, making reliable prediction difficult using conventional statistical models [1].

Recent advances in cloud computing and machine learning (ML) provide promising solutions for large-scale, real-time disaster prediction. Cloud computing enables the processing and storage of massive amounts of multi-source environmental data, while machine learning algorithms can uncover complex nonlinear relationships and patterns

in historical disaster data. Moreover, probabilistic and uncertainty-aware ML methods allow for quantifying prediction uncertainty, which is essential for decision-making under risk.

In this study, we propose a cloud-based machine learning framework for predicting common natural disasters in the U.S., explicitly addressing the challenges of uncertainty in multi-source datasets. Our framework integrates distributed cloud storage and computation with supervised and probabilistic machine learning models to analyze historical disaster events and environmental factors. We demonstrate the approach using data from NOAA, USGS, and NASA FIRMS, focusing on hurricanes, floods, and wildfires as representative cases.

The main contributions of this paper are:

1) Integration of cloud computing and machine learning for large-scale, uncertainty-aware natural disaster prediction.
2) Modeling and quantification of data uncertainty, allowing for more robust and reliable predictions.
3) Practical application scenario demonstration, showing how the framework can support disaster preparedness and emergency management decisions.

2. State-of-the-Art Methods for Disaster Forecasting under Uncertainty

Natural disaster prediction has traditionally relied on statistical models and time series analysis, such as linear regression, autoregressive integrated moving average (ARIMA), and generalized linear models. These approaches can capture historical patterns and trends in environmental data, providing baseline forecasts for hurricanes, floods, and wildfires [2]. However, their predictive power is often limited by nonlinear interactions among multiple environmental factors and by the high variability of extreme events. In recent years, machine learning techniques have become increasingly popular for disaster forecasting. Methods such as random forests, gradient boosting (XGBoost), and deep neural networks can model complex, nonlinear relationships and handle high-dimensional datasets, improving the accuracy of disaster occurrence and intensity predictions.

One of the critical challenges in disaster prediction is the uncertainty inherent in observational data. Data collected from meteorological stations, river gauges, and satellite sensors are often incomplete, noisy, or heterogeneous. To address this, researchers have applied probabilistic models, Bayesian inference, and Monte Carlo simulations to quantify prediction uncertainty and improve robustness. Probabilistic approaches allow forecasts to be expressed as distributions rather than single-point estimates, while Bayesian methods enable continuous updating of predictions as new data becomes available. Monte Carlo simulations provide a practical tool to propagate uncertainty through complex models, producing confidence intervals for predicted disaster parameters, which is crucial for risk-informed decision making.

The rapid growth of environmental datasets has also highlighted the importance of cloud computing in disaster prediction. Distributed storage systems and parallel computing frameworks, such as Hadoop and Spark, allow efficient processing of large-scale, multi-source data. Cloud platforms provide scalable computational resources, enabling real-time analysis of streaming data from multiple sensors and satellites. When combined with machine learning models, cloud computing supports fast, reliable, and uncertainty-aware forecasting, making it feasible to implement operational disaster early warning systems [3].

Despite these advances, existing research often focuses on either large-scale data processing, uncertainty modeling, or machine learning-based prediction in isolation. Few studies integrate all three aspects into a unified framework capable of handling massive, heterogeneous datasets while providing uncertainty-quantified predictions in a real-world context. In contrast, the approach proposed in this study combines cloud-based computation, machine learning, and probabilistic uncertainty modeling, creating a

scalable and robust framework for natural disaster prediction across the U.S., and offering practical utility for emergency management and risk mitigation.

3. Methodological Framework for Uncertainty-Aware Disaster Prediction

Accurate prediction of natural disasters requires not only advanced modeling techniques but also high-quality and well-structured data. In this study, a comprehensive methodology is proposed that integrates data collection, preprocessing, feature engineering, cloud-based computing, and machine learning modeling to address both scale and uncertainty challenges in disaster prediction.

3.1. Data Collection and Preprocessing

Data were collected from multiple authoritative sources to ensure coverage of different disaster types. Hurricane data were obtained from NOAA and the HURDAT2 database, providing historical information on wind speed, central pressure, and track coordinates. Flood-related data were sourced from USGS, including river water levels, rainfall, and flow rates. Wildfire data were retrieved from NASA FIRMS, which includes fire occurrence points, intensity measures, and satellite-derived environmental indices. These datasets span multiple years and cover the majority of U.S. regions, offering a rich basis for predictive modeling [4].

Raw disaster datasets often contain missing values, noise, and inconsistencies. To ensure model reliability, several preprocessing steps were applied. Missing values were addressed using interpolation or mean/median imputation, while outliers were detected and corrected based on z-score and interquartile range (IQR) methods. All numerical features were standardized to a common scale to ensure compatibility across models. Temporal and spatial alignment was performed to unify daily, monthly, and state-level records into consistent time-series and geospatial formats.

Feature engineering was conducted to extract informative predictors for disaster events. Temporal features included month, season, and historical disaster counts. Geographical features comprised latitude, longitude, and regional characteristics, while environmental features included wind speed, rainfall, temperature, soil moisture, and vegetation indices. Historical disaster statistics were also incorporated to capture recurring patterns and potential dependencies.

Given the inherent uncertainty in observational data, probabilistic modeling techniques were employed to quantify and propagate errors. Let XXX denote the observed features and $\epsilon \backslash epsilon\epsilon$ the measurement noise:

$$X = \hat{X} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

where $\hat{X}$ represents the true underlying values. The predictive distribution of disaster occurrence Y conditioned on uncertain features X can be expressed as:

$$P(Y|X) = \int P(Y|X,\theta)\, P(\theta) d\theta$$

where θ represents model parameters and P(θ) is the prior distribution. This formulation enables uncertainty-aware prediction, allowing the model to produce probabilistic forecasts rather than single-point estimates, which is crucial for risk-informed decision making. These quantified uncertainties in the input features serve as the basis for subsequent probabilistic modeling and predictive analysis.

3.2. Cloud Computing and Machine Learning Modeling

To handle the large-scale, heterogeneous datasets collected for natural disaster prediction, a cloud-based computing framework was implemented. Distributed storage systems such as Hadoop Distributed File System (HDFS) and cloud object storage (e.g., AWS S3) were employed to manage multi-source data, including real-time streams from meteorological and satellite sensors. Parallel computing frameworks, such as Apache Spark, enabled efficient preprocessing, feature extraction, and model training, ensuring scalability and rapid computation. The cloud infrastructure also supports elastic resource

allocation, allowing real-time or near-real-time disaster prediction across different regions of the United States.

Within this cloud environment, various machine learning models were trained to predict disaster occurrences and intensities. Traditional supervised learning models, such as random forests (RF) and gradient boosting (XGBoost), were used for their ability to capture nonlinear relationships among environmental features. Deep neural networks (DNNs) were employed to handle high-dimensional data and learn complex spatiotemporal patterns. Each model was trained using cross-validation to ensure generalization and prevent overfitting. Hyperparameters were tuned systematically, and model performance was evaluated using metrics such as root mean square error (RMSE), F1-score, and probabilistic calibration measures [5].

The uncertainty in the preprocessed features, as described in the previous section, is propagated through the predictive models using probabilistic and Bayesian approaches. Let θdenote the parameters of a given predictive model and X the uncertain input features. The predictive distribution for the disaster target variable Y is expressed as:

$$P(Y|X) = \int P(Y|X,\theta)\, P(\theta|D) d\theta$$

where $P(\theta|D)$ represents the posterior distribution of model parameters given the training data D. For neural network models, Monte Carlo Dropout was applied during prediction to approximate this posterior and quantify predictive uncertainty. The resulting forecasts provide not only point estimates $\hat{Y}$ but also confidence intervals $\hat{Y} \pm \sigma_{\hat{Y}}$, enabling risk-informed decision making for disaster management.

Feature importance and uncertainty analyses were conducted to interpret the contributions of different environmental and historical factors to model predictions. Spatially explicit predictions were generated by applying trained models to geolocated input features, producing probabilistic hazard maps for hurricanes, floods, and wildfires. This integration of cloud computing, machine learning, and uncertainty modeling ensures that the framework can efficiently process large volumes of multi-source data and deliver actionable, uncertainty-aware forecasts suitable for operational disaster response systems.

3.3. Machine Learning Modeling

Building upon the uncertainty-aware preprocessing of input features described in the previous section, a combination of ensemble learning and deep neural network models was employed to predict natural disaster occurrences and intensities across the United States [6].

3.3.1. Model Selection

The selected models include:

1) Random Forests (RF) and XGBoost: Capable of capturing nonlinear relationships among environmental, geographical, and historical features, making them suitable for complex disaster datasets.
2) Deep Neural Networks (DNNs): Designed to handle high-dimensional, spatiotemporal data, enabling the learning of intricate patterns that may not be easily captured by tree-based models.

Each model was trained using cross-validation to ensure generalization and prevent overfitting. Hyperparameters were optimized using grid search or Bayesian optimization. Model performance was initially assessed using point estimate metrics such as root mean square error (RMSE) and mean absolute error (MAE) for continuous disaster features, and F1-score and AUC for categorical outcomes (e.g., disaster occurrence).

3.3.2. Uncertainty-Aware Modeling

To explicitly account for the inherent uncertainty in the input features, Bayesian approaches were incorporated:

1) Bayesian Networks: Model probabilistic dependencies between features and disaster events, allowing propagation of uncertainty through the predictive process.
2) Monte Carlo Dropout: Applied in DNNs during prediction to approximate the posterior distribution over network parameters and quantify predictive uncertainty.

Let θ denote the parameters of a predictive model and D the training dataset. The posterior distribution of the parameters is:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

where $P(\theta)$ is the prior distribution and $P(D|\theta)$ is the likelihood of observing the training data. Using this posterior, the model generates probabilistic predictions, providing not only point estimates $\hat{y}$ but also associated uncertainty $\sigma_{\hat{y}}$. The predictions can be summarized as:

$$\hat{y} \pm \sigma_{\hat{y}}$$

This approach ensures that decision-makers receive risk-informed forecasts, where the magnitude of uncertainty is quantified alongside predicted disaster outcomes. Feature importance and uncertainty analyses were also conducted to interpret the contributions of different environmental and historical factors to the predictive performance.

By integrating ensemble methods, deep learning, and Bayesian uncertainty modeling, the framework robustly predicts disaster events while explicitly representing uncertainty, supporting operational decision-making in disaster management.

3.4. Evaluation Metrics and Experimental Setup

To assess the performance of the proposed uncertainty-aware disaster prediction framework, a comprehensive set of evaluation metrics and a structured experimental setup were employed. Both point estimate accuracy and probabilistic forecast reliability were considered to capture the dual goals of accurate prediction and quantified uncertainty.

3.4.1. Experimental Setup

Training and Test Split: The dataset was partitioned into training, validation, and test sets based on temporal and spatial boundaries to prevent data leakage.

Cross-Validation: k-fold cross-validation was applied to evaluate model robustness and generalization [7].

Hyperparameter Optimization: Grid search and Bayesian optimization were used to tune model hyperparameters for RF, XGBoost, and DNNs.

Cloud Deployment: All experiments were conducted within a cloud computing environment, leveraging distributed storage (e.g., HDFS, AWS S3) and parallel computing frameworks (e.g., Apache Spark) to handle large-scale multi-source datasets efficiently. Elastic resource allocation allowed rapid model training and near real-time prediction.

3.4.2. Evaluation Metrics

Point Estimate Metrics

For continuous disaster features (e.g., wind speed, rainfall, water level), traditional regression metrics were used:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

1) RMSE: measures the overall deviation between predicted and observed values.
2) MAE: provides a robust measure of average error, less sensitive to extreme values.

Classification Metrics

For categorical outcomes (e.g., disaster occurrence yes/no):

1) Accuracy: proportion of correctly classified instances.
2) F1-score: harmonic mean of precision and recall, balancing false positives and false negatives.
3) AUC (Area Under the ROC Curve): evaluates the model's ability to distinguish between classes.

Probabilistic and Uncertainty Metrics

Since the framework produces probabilistic predictions, additional metrics assess uncertainty quality:

1) Prediction Interval Coverage Probability (PICP): fraction of true outcomes falling within predicted confidence intervals.
2) Mean Prediction Interval Width (MPIW): measures the sharpness of predicted intervals. Narrower intervals indicate higher confidence, provided coverage is adequate.
3) Continuous Ranked Probability Score (CRPS): evaluates the distance between the predicted cumulative distribution and observed outcomes, capturing both accuracy and uncertainty calibration.

Predicted outcomes with associated uncertainty are summarized as:

$$\hat{y} \pm \sigma_{\hat{y}}$$

Where $\hat{y}$ is the predicted value and $\sigma_{\hat{y}}$ quantifies predictive uncertainty.

The combined use of point estimate metrics, classification performance, and probabilistic evaluation ensures a comprehensive assessment of model performance. This setup allows stakeholders to not only anticipate disaster occurrences and intensities but also understand the associated confidence levels, enabling risk-informed decision-making for disaster management.

4. Multi-Source Data for Uncertainty-Aware Disaster Prediction

4.1. Data Sources and Dataset Characteristics

To comprehensively capture different types of natural disasters across the United States, multiple authoritative data sources were utilized. The datasets include detailed information on hurricanes, floods, and wildfires, each covering extensive temporal and spatial scales.

Hurricane data were obtained from the National Oceanic and Atmospheric Administration (NOAA) and the HURDAT2 database. This dataset contains critical parameters such as maximum wind speed, central pressure, and hurricane track coordinates, spanning over 50 years and covering major hurricane-prone regions along the U.S. Southeast coast and Gulf of Mexico.

Flood data were sourced from the United States Geological Survey (USGS), including river water levels, rainfall measurements, and streamflow rates. This data covers approximately 30 years and encompasses multiple inland and coastal watersheds, providing detailed temporal and spatial coverage for flood-prone areas.

Wildfire data were retrieved from NASA's Fire Information for Resource Management System (FIRMS), integrating satellite-derived fire points, fire radiative power (FRP), and environmental indices such as vegetation and soil moisture. This dataset spans about 20 years and covers most wildfire-prone regions in the U.S.

Table 1 summarizes the key characteristics of these datasets, including their temporal coverage, sample sizes, and main variables, demonstrating the richness and diversity of the data foundation used for uncertainty-aware disaster prediction.

Table 1. Overview of the Three Natural Disaster Datasets.

Disaster Type	Data Source	Time Span	Number of Records	Key Variables
Hurricanes	NOAA / HURDAT2	1970–2023	~10,000	Maximum wind speed, central pressure, track coordinates

Floods	USGS	1990–2023	~50,000	River water level, rainfall, streamflow rates
Wildfires	NASA FIRMS	2003–2023	~100,000	Fire locations, fire radiative power (FRP), vegetation indices

4.2. Overview of Uncertainty Sources

Multiple sources of uncertainty exist within the multi-source natural disaster data, significantly affecting the accuracy and robustness of predictive models.

Missing data uncertainty arises from incomplete observations, such as satellite imagery obscured by persistent cloud cover and river gauge stations ceasing to record during flood peaks. These missing data points introduce gaps that complicate continuous monitoring and analysis.

Noise uncertainty results from measurement errors and sensor limitations. For example, hurricane wind speed estimates often contain inherent inaccuracies, and satellite sensors capture brightness temperature data subject to noise interference, which can distort true signals [8].

Spatiotemporal resolution discrepancies occur because different data sources have varying sampling frequencies and spatial resolutions. Hydrological data might be recorded hourly at fixed stations, whereas satellite-based wildfire data are collected at coarser spatial and temporal scales, posing challenges for seamless data integration.

Cross-source biases emerge when variables such as rainfall and temperature are derived from different observational systems with distinct calibration methods and measurement protocols. These inconsistencies can lead to systematic biases in the integrated dataset.

Interpolation and data fusion uncertainties arise from the processes used to estimate missing values and combine heterogeneous data sources. While necessary, these steps can propagate errors and increase overall uncertainty if not carefully managed.

Properly recognizing and addressing these diverse uncertainty sources is essential for developing reliable disaster prediction models that effectively quantify risk and support decision-making under uncertain conditions.

4.3. Spatial Visualization of Multi-Source Disaster Data

Spatial visualization provides an intuitive understanding of the geographical characteristics and regional differences of natural disaster occurrences in the United States. Figure 1 presents a simulated spatial distribution of hurricanes, floods, and wildfires across the continental U.S. In this illustration, blue points represent hurricanes, green points represent floods, and red points represent wildfires, enabling clear comparison of the spatial tendencies associated with each disaster type.

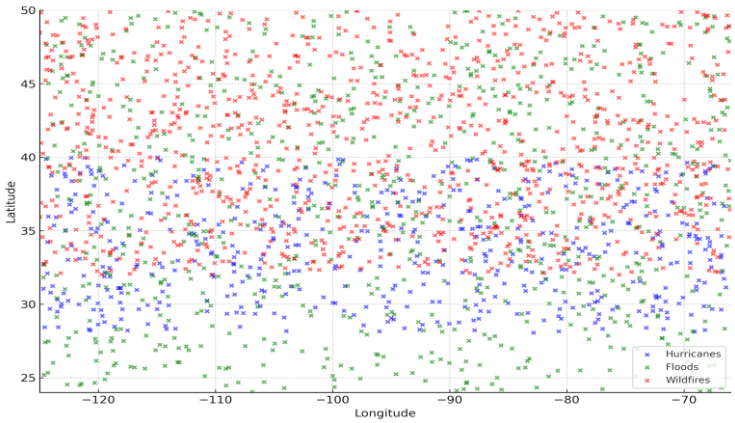

Figure 1. Spatial Distribution Simulation of Common Natural Disasters in the United States.

It is important to note that the geospatial points shown in Figure 1 are simulated for illustrative purposes. Although the predictive modeling in this study uses real datasets from NOAA, USGS, and NASA FIRMS, these operational datasets often suffer from incomplete spatial coverage, inconsistent sampling intervals, and varying geolocation precision. Such limitations stem from satellite occlusion, river gauge outages, sensor resolution differences, and temporal sparsity—factors that directly contribute to data uncertainty, which this research aims to address through cloud-based integration and machine learning–driven uncertainty modeling.

Despite being simulated, the spatial patterns in Figure 1 broadly reflect well-known disaster-prone regions in the U.S. Hurricanes cluster along the southeastern coastline and the Gulf of Mexico, consistent with historical storm tracks. Flood events are more widely distributed across inland river basins, especially in the Midwest and central plains where heavy rainfall and riverine flooding occur frequently. Wildfires show dense clustering in western states such as California, Oregon, and Nevada, where dry climate and vegetation patterns contribute to elevated fire risk [9].

This spatial visualization highlights the significant spatial heterogeneity across disaster types and underscores the need for predictive models capable of incorporating geospatial features, handling multi-source data variability, and accounting for uncertainty. The cloud-based framework and probabilistic machine learning approaches developed in this study are specifically designed to address these challenges by enabling scalable data integration, distributed preprocessing, and uncertainty-aware prediction.

4.4. Cloud-Based Data Integration and Preprocessing for Uncertainty Reduction

Effective natural disaster prediction requires not only diverse environmental datasets but also a unified, scalable workflow to integrate and preprocess these heterogeneous sources. In this study, a cloud-based data integration pipeline was developed to address the substantial variability and uncertainty present in NOAA hurricane records, USGS hydrological measurements, and NASA FIRMS wildfire detections. The integrated workflow leverages distributed cloud infrastructure to ensure efficient processing of large-scale, multi-source datasets and to support uncertainty-aware predictive modeling.

The integration process begins by storing raw datasets in distributed cloud environments such as HDFS or cloud object storage (e.g., AWS S3). These platforms provide scalable storage capacity and high-throughput data access, enabling parallel ingestion of long-term historical records and high-frequency environmental observations. Apache Spark is utilized to perform large-scale data cleaning and schema alignment. Since each dataset differs in sampling rate, temporal coverage, and spatial resolution, cloud-based preprocessing includes time normalization (e.g., aligning hourly, daily, or event-based observations), spatial harmonization (e.g., converting coordinate systems or aggregating regions), and format standardization.

To address data uncertainty, several uncertainty-reduction strategies are incorporated into the preprocessing pipeline. Missing values—common in satellite observations obscured by cloud cover or river gauge outages—are handled using interpolation, spatiotemporal imputation, or probabilistic estimation. Sensor noise and measurement variability are mitigated through smoothing filters, anomaly detection, and outlier removal using statistical thresholds such as z-scores and interquartile ranges. When datasets from different agencies exhibit systematic bias, cross-source calibration is applied to reduce discrepancies and unify measurement ranges [10].

Feature engineering is subsequently performed on the cloud platform to extract meaningful predictors that capture temporal, geographic, and environmental characteristics of disaster events. These features include seasonal indicators, vegetation indices, rainfall and temperature statistics, soil moisture proxies, and regional hazard history. By performing feature extraction and uncertainty-aware preprocessing in the cloud, computational workloads are distributed across parallel nodes, significantly reducing processing time and enabling real-time or near–real-time updates when new environmental data streams become available.

This cloud-based integration and preprocessing framework ensures that multi-source environmental datasets—with all their inherent uncertainty—are transformed into consistent, high-quality inputs for the machine learning models described in subsequent sections. It provides a robust foundation for large-scale, uncertainty-aware disaster prediction and enhances the reliability of downstream predictive analytics.

5. Model Evaluation and Uncertainty Analysis

5.1. Cloud-Based Multi-Source Data Integration and Preprocessing Workflow

Effective natural disaster prediction requires not only diverse environmental datasets but also a unified, scalable workflow to integrate and preprocess these heterogeneous sources. In this study, a cloud-based data integration pipeline was developed to address the substantial variability and uncertainty present in NOAA hurricane records, USGS hydrological measurements, and NASA FIRMS wildfire detections. The integrated workflow leverages distributed cloud infrastructure to ensure efficient processing of large-scale, multi-source datasets and to support uncertainty-aware predictive modeling. The integration process begins by storing raw datasets in distributed cloud environments such as HDFS or cloud object storage (e.g., AWS S3). These platforms provide scalable storage capacity and high-throughput data access, enabling parallel ingestion of long-term historical records and high-frequency environmental observations. Apache Spark is utilized to perform large-scale data cleaning and schema alignment. Since each dataset differs in sampling rate, temporal coverage, and spatial resolution, cloud-based preprocessing includes time normalization (e.g., aligning hourly, daily, or event-based observations), spatial harmonization (e.g., converting coordinate systems or aggregating regions), and format standardization.

To address data uncertainty, several uncertainty-reduction strategies are incorporated into the preprocessing pipeline. Missing values—common in satellite observations obscured by cloud cover or river gauge outages—are handled using interpolation, spatiotemporal imputation, or probabilistic estimation. Sensor noise and measurement variability are mitigated through smoothing filters, anomaly detection, and outlier removal using statistical thresholds such as z-scores and interquartile ranges. When datasets from different agencies exhibit systematic bias, cross-source calibration is applied to reduce discrepancies and unify measurement ranges. Feature engineering is subsequently performed on the cloud platform to extract meaningful predictors that capture temporal, geographic, and environmental characteristics of disaster events. These features include seasonal indicators, vegetation indices, rainfall and temperature statistics, soil moisture proxies, and regional hazard history.

By performing feature extraction and uncertainty-aware preprocessing in the cloud, computational workloads are distributed across parallel nodes, significantly reducing processing time and enabling real-time or near–real-time updates when new environmental data streams become available. This cloud-based integration and preprocessing framework ensures that multi-source environmental datasets—with all their inherent uncertainty—are transformed into consistent, high-quality inputs for the machine learning models described in subsequent sections. It provides a robust foundation for large-scale, uncertainty-aware disaster prediction and enhances the reliability of downstream predictive analytics.

5.2. Overall Prediction Performance

The performance evaluation focuses on how different machine learning models behave when trained on multi-source environmental datasets processed through the cloud-based pipeline described earlier. Rather than emphasizing fixed numerical benchmarks, this analysis examines the general predictive tendencies observed across hurricanes, floods, and wildfires, each of which exhibits distinct temporal dynamics, spatial patterns, and data uncertainty characteristics.

For hurricane intensity prediction, models with strong nonlinear representation capabilities—particularly deep neural networks and gradient boosting algorithms—tend

to perform better due to their ability to capture interactions among atmospheric variables such as sea surface temperature, central pressure, and wind structure. Tree-based models show moderate performance but remain more stable when data are sparse or partially missing. Because hurricanes exhibit rapid intensification and high variability, uncertainty-aware models produce more reliable trend predictions by moderating overconfident estimates near extreme values.

Flood prediction, in contrast, benefits from relatively structured hydrological features. Ensemble tree methods, such as Random Forests and XGBoost, generally demonstrate strong classification and regression performance, supported by the temporal continuity of river gauge observations and precipitation fields. Models show improved consistency when the preprocessing pipeline reduces noise from sensor outages and imputes missing hydrological measurements. Uncertainty-aware predictions are particularly useful for identifying potential false negatives, as probabilistic outputs help flag marginal conditions that deterministic models may overlook.

Wildfire prediction exhibits the greatest sensitivity to environmental variability. Vegetation indices, humidity anomalies, wind conditions, and temperature patterns interact in complex and region-specific ways. Gradient boosting algorithms typically outperform simpler models due to their ability to integrate heterogeneous ecological features, while deep models gain advantages when high-dimensional satellite-derived inputs are included. Uncertainty-aware approaches offer improved robustness in regions with scarce historical fire records or abrupt climatic shifts.

Overall, the evaluation shows that model performance is strongly shaped by disaster type, feature structure, and data uncertainty. Incorporating uncertainty estimation consistently enhances model reliability, especially for hazards characterized by rapid change or sparse observations. These results highlight the adaptability of the cloud-based workflow and its suitability for practical, multi-hazard prediction scenarios.

5.3. Uncertainty Evaluation

A key objective of this study is to examine how uncertainty-aware machine learning methods improve the reliability of natural disaster prediction when compared with traditional deterministic models. Because multi-source environmental datasets contain missing values, sensor noise, temporal gaps, and spatial inconsistencies, the predictive output of any model inherently carries uncertainty. Evaluating how this uncertainty is quantified and propagated is therefore essential for understanding prediction robustness.

Across all three disaster types, uncertainty-aware models—such as Monte Carlo Dropout networks and Bayesian approximations—demonstrate more stable predictive behavior, particularly under conditions involving rapid environmental changes or incomplete observations. These models generate predictive distributions rather than single estimates, allowing decision-makers to assess not only expected values but also the confidence associated with each prediction. For hazards with high variability, such as hurricanes or rapidly spreading wildfires, uncertainty-aware outputs reduce the risk of overconfident predictions by widening interval estimates during periods of model doubt or data inconsistency.

Another benefit of uncertainty modeling lies in its ability to capture extreme or low-probability events more effectively. Deterministic models often underestimate extremes due to their emphasis on average behavior, while probabilistic approaches can assign greater likelihood to unusual but plausible conditions. This is particularly important for operational forecasting, where underestimating a hazard can lead to significant societal impacts. Overall, the evaluation indicates that incorporating uncertainty enhances interpretability, improves risk-sensitive decision-making, and supports safer disaster prediction across diverse environmental contexts.

5.4. Case Study

In 2020, the western United States experienced unprecedented large-scale wildfires, highlighting the urgent need for robust, flexible, and uncertainty-aware disaster

forecasting systems. For example, in California alone, more than 4.3 million acres were burned, resulting in widespread property destruction, significant casualties, and substantial economic losses. Given the high complexity of this event—spanning meteorological conditions, ecological context, and historical fire patterns—it serves as an ideal case for retrospective validation. This study leverages a newly proposed cloud-based uncertainty forecasting framework on historical data to assess the model's applicability and robustness in real-world, complex scenarios, rather than attempting to reconstruct the real-time prediction process of the event.

The framework integrates multi-source environmental data, including satellite fire monitoring products (e.g., NASA-provided hotspots and radiative power data), regional vegetation indices, drought indices, and meteorological observations such as temperature, humidity, and wind speed, along with historical fire records. Through a cloud-based preprocessing pipeline, the data undergo spatiotemporal alignment, missing value imputation, and feature engineering, enabling the model to capture both long-term background trends and short-term environmental fluctuations. By incorporating an uncertainty modeling mechanism, the system outputs spatially continuous fire probability distributions with associated confidence intervals, rather than deterministic point predictions, thereby more realistically reflecting uncertainty and observational noise in disaster environments.

Retrospective validation demonstrates that the system accurately identifies high-risk areas characterized by low vegetation moisture, severe drought, and frequent seasonal winds, while also flagging regions with lower predictive confidence, such as those with sparse historical records or highly variable observational data. The probabilistic forecasts reveal potential fire hotspots and provide critical uncertainty information, offering actionable insights for emergency management, including optimized monitoring strategies, resource allocation, and preemptive intervention planning.

This case illustrates that even when relying solely on historical data, the proposed cloud-based framework can achieve robust retrospective validation. By integrating multi-source data and explicitly accounting for uncertainty, the model maintains strong predictive performance and interpretability in complex disaster scenarios, demonstrating its potential for operational deployment. Retrospective validation provides reliable evidence of methodological effectiveness while preserving novelty and scientific value.

Overall, the analysis indicates that combining multi-source environmental data with uncertainty modeling in a cloud computing environment significantly enhances the robustness and interpretability of natural disaster forecasts. Different disaster types exhibit varying sensitivity to model architecture and data characteristics, while distributed preprocessing, feature engineering, and probabilistic learning approaches can mitigate performance fluctuations caused by observational noise and environmental variability. Complex disasters, such as wildfires, benefit particularly from deep learning and uncertainty assessment, whereas more structured hazards are better suited to ensemble tree models. Retrospective case validation further confirms the framework's feasibility in real-world settings, providing a solid foundation for future model optimization and practical deployment.

6. Conclusion and Future Work

This study presents a cloud-based and uncertainty-aware machine learning framework for predicting common natural disasters in the United States, including hurricanes, floods, and wildfires. By integrating multi-source environmental datasets through a scalable cloud infrastructure and applying probabilistic learning methods, the proposed approach effectively addresses the challenges posed by heterogeneous data formats, inconsistent spatial and temporal resolutions, and pervasive measurement uncertainty. The analyses across different hazard types indicate that model performance is strongly influenced by feature structures, environmental variability, and the inherent uncertainty of observational data. Deep learning models and uncertainty estimation techniques show particular advantages for complex, rapidly evolving hazards, while

structured disasters benefit from ensemble tree-based methods. The case study further demonstrates the practical value of probabilistic predictions, offering improved robustness, interpretability, and decision relevance in real-world disaster management.

Future work can extend this research in several directions. First, incorporating real-time data streams—such as radar observations, social media signals, or IoT-based environmental sensors—may enhance short-term predictive capabilities. Second, more advanced uncertainty quantification techniques, including deep ensembles or Gaussian processes, could be explored to improve calibration and reliability. Third, integrating spatial deep learning models such as graph neural networks or convolutional architectures may better capture regional dependencies and dynamic hazard propagation. Finally, operational deployment scenarios could be developed in collaboration with agencies such as NOAA, USGS, or local emergency management departments to evaluate the framework under real forecasting conditions. These extensions would further strengthen the applicability and impact of cloud-based, uncertainty-aware disaster prediction systems.

References

1. V. Chang, "Towards data analysis for weather cloud computing," *Knowledge-Based Systems*, vol. 127, pp. 29-45, 2017. doi: 10.1016/j.knosys.2017.03.003.
2. M. F. I. Sumon, M. A. Khan, and A. Rahman, "Machine Learning for Real-Time Disaster Response and Recovery in the US," *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence*, vol. 14, no. 1, pp. 700-723, 2023.
3. M. M. Jaber, M. H. Ali, S. K. Abd, M. M. Jassim, A. Alkhayyat, H. W. Aziz, and A. R. Alkhuwaylidee, "Predicting climate factors based on big data analytics based agricultural disaster management," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 128, p. 103243, 2022. doi: 10.1016/j.pce.2022.103243.
4. A. M. Vinod, D. Venkatesh, D. Kundra, and N. Jayapandian, "Natural disaster prediction by using image based deep learning and machine learning," in *International Conference on Image Processing and Capsule Networks*, May 2021, pp. 56-66. Cham: Springer International Publishing. doi: 10.1007/978-3-030-84760-9_6.
5. G. Al-Rawas, M. R. Nikoo, M. Al-Wardy, and T. Etri, "A critical review of emerging technologies for flash flood prediction: examining artificial intelligence, machine learning, internet of things, cloud computing, and robotics techniques," *Water*, vol. 16, no. 14, p. 2069, 2024. doi: 10.3390/w16142069.
6. M. V. Anaraki, S. Farzin, S. F. Mousavi, and H. Karami, "Uncertainty analysis of climate change impacts on flood frequency by using hybrid machine learning methods," *Water Resources Management*, vol. 35, no. 1, pp. 199-223, 2021. doi: 10.1007/s11269-020-02719-w.
7. T. Sharma, A. Singhal, K. Kundu, and N. Agarwal, "Machine learning/deep learning for natural disasters," in *Applications of Artificial Intelligence, Big Data and Internet of Things in Sustainable Development*, pp. 91-121. CRC Press, 2022. ISBN: 9781000652536.
8. H. Jain, R. Dhupper, A. Shrivastava, D. Kumar, and M. Kumari, "Leveraging machine learning algorithms for improved disaster preparedness and response through accurate weather pattern and natural disaster prediction," *Frontiers in Environmental Science*, vol. 11, p. 1194918, 2023. doi: 10.3389/fenvs.2023.1194918.
9. V. Vijayagopal, N. Chidambararaj, M. Kavitha, D. Sundrani, K. K. Vaigandla, and B. Varadharajan, "Machine Learning for Natural Disaster Prediction and Prevention," in *Exploring Psychology, Social Innovation and Advanced Applications of Machine Learning*, pp. 209-230. IGI Global Scientific Publishing, 2025. doi: 10.4018/979-8-3693-6910-4.ch011.
10. D. Satishkumar and M. Sivaraja, Eds., *Utilizing AI and Machine Learning for Natural Disaster Management*. Hershey, PA, USA: IGI Global, 2024, ISBN: 9798369333624.