

Article

Research on the Construction of Insurance Trigger Index for Lightning Risk Based on Satellite Monitoring Data

Guanhua Hao ¹, Shanshan Jiang ^{1,*}, Yuxi Chen ^{2,3} and Min Xia ^{3,4}

¹ School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202383390012@nuist.edu.cn

² Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK; tp803778@student.reading.ac.uk

³ Collaborative Innovation Center on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; xiamin@nuist.edu.cn

⁴ Jiangsu Provincial University Key laboratory of Big Data Analysis and Intelligent Systems, Nanjing University of Information Science and Technology, Nanjing 210044, China

* Correspondence: jss@nuist.edu.cn

Abstract

Thunderstorm disasters are one of the major meteorological disasters in China, causing significant human casualties and economic losses each year. Traditional loss compensation insurance is confronted with difficulties such as inspection and assessing, causing low claim processing efficiency, while index insurance can effectively overcome these deficiencies by triggering payment through objective indices. This paper is based on satellite remote sensing monitoring data, using a combination of principal component analysis, random forests, and fuzzy mathematical theory to construct a lightning risk index and design a complete index insurance product. Experimental validation based on historical satellite monitoring data has shown that the risk indices constructed in this paper can effectively capture the temporal and spatial variability of lightning activity. Random forest models have a relatively low fitting error of training labels, and the SHAP values reveal a characteristic weight of importance consistent with physical perception. The insurance product has a reasonable distribution of amount and compensation, and premium pricing balances actuarial fairness with market acceptability. The present methodology provides a transferable design path to monitor and transfer the lightning risk using multi-source remote sensing data, with some outreach value in the field of lightning and other natural disasters.

Keywords: lightning risk; index insurance; principal components analysis; random forests; fuzzy mathematics



Academic Editor: Atsushi Mase

Received: 23 May 2026

Revised: 26 June 2026

Accepted: 29 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Lightning, as a powerful convective weather phenomenon, is closely related to many meteorological disasters and poses a significant threat to modern society's production and life. According to incomplete statistics, there were 437 lightning-related disasters in China, including six cases of fire or explosions and 13 cases of personal accidents, resulting in 15 deaths and five injuries, with direct economic losses of approximately 12 million yuan and indirect economic losses of about 3.02 million yuan in 2023 [1]. In response to the increasingly urgent demand for prevention and control, insurance is an important means to transfer the risks of lightning disasters. Index insurance, as an innovative risk management tool, triggers compensation based on preset objective indices without the need for on-site investigation

and assessment, effectively overcoming the shortcomings of traditional insurance. This model has been successfully applied in areas such as agricultural drought [2], typhoons [3], and floods [4,5], but research in the field of lightning disasters is still relatively limited.

Currently, the widely adopted lightning risk assessment standard internationally is IEC 62305-2. This standard assesses the lightning risk levels of various facilities from multiple dimensions such as lightning strike frequency, lightning current amplitude, and building characteristics through quantitative calculations and has played an important role in engineering practice. However, in specific scenarios such as large residential areas and large industrial zones, IEC 62305-2 is unable to accurately reflect the real lightning risk [6]. Moreover, traditional assessment methods are mostly static. They perform a one-time risk classification based on long-term statistical data and cannot respond to the diurnal and seasonal variations of lightning activities.

In recent years, the rapid development of satellite remote sensing technology [7–9] has provided a new data source for real-time monitoring of lightning activities. Satellite data have advantages such as high temporal and spatial resolution, wide coverage, and the ability to be continuously observed [10–12], which can more timely and precisely depict the temporal and spatial distribution characteristics of lightning activities, laying a data foundation for the dynamic assessment of lightning risks. At the same time, machine learning methods have been widely applied in the field of natural disaster risk assessment. Principal component analysis (PCA), as a classic unsupervised dimensionality reduction method, can compress multiple related variables into a few principal components, retaining the main information of the original data while reducing the feature dimension, and can be applied to the extraction and integration of lightning activity information [13]. Random forests, as a type of ensemble learning method, can effectively handle high-dimensional features and nonlinear relationships and can output feature importance scores, demonstrating good performance in disaster vulnerability assessment [14]; SHAP (SHapley Additive exPlanations) values further address the lack of interpretability of “black box” models such as random forests, decomposing model predictions into the contribution values of each feature and revealing the direction and degree of influence of different factors on risks [15–17]. The combination of remote sensing data and machine learning methods provides technical possibilities for constructing dynamic and refined lightning risk indices.

Currently, relevant research mainly focuses on the monitoring and warning of lightning activities and has not yet systematically applied the above methods to the product design of lightning index insurance. To address the above issues, this paper comprehensively utilizes remote sensing monitoring data, machine learning methods, and fuzzy mathematics theory to propose a set of index insurance product design schemes based on the lightning risk index. The main work includes the following three aspects:

1. Construction of the lightning risk index. Based on the lightning monitoring data derived from satellite inversion, principal component analysis is used to extract the main variation information, and the multi-dimensional features are compressed into a comprehensive score as the training label. A random forest regression model is trained, and the SHAP values are used to explain the contribution direction and degree of each feature to the risk. On this basis, the fuzzy mathematics comprehensive evaluation method is introduced to combine the risk weights output by machine learning, achieving the construction of the lightning risk index and the classification of risk levels.
2. Design of the index insurance product. Firstly, based on the historical distribution of the index and the risk level thresholds, determine the trigger threshold for the insurance; secondly, in combination with the multi-layer arrangement for catastrophe risk transfer, set the upper limit of the insurance company's claim and the claim amount; finally, based on the trigger probability and expected payout, determine a reasonable

premium standard. The insurance payout is determined by daily assessment and aggregation, using the preset index threshold as the trigger basis. Based on the above steps, this paper designs a complete index insurance product scheme.

3. Experimental verification of product feasibility. Utilize historical satellite monitoring data to conduct backtesting and validation of the aforementioned insurance scheme. Through consistent parameter setting, we test the compensation levels at different times, evaluate the trigger frequency, different percentile parameters, various model settings and premium rationality of the insurance product. We also analyze the financial sustainability of the insurance product.

The structure of this article is as follows. Section 2 reviews the relevant research work and summarizes the main methods for constructing lightning risk index and index insurance; Section 3 elaborates on the system framework of lightning risk index insurance products, forming a complete insurance product plan; Section 4 conducts experimental verification to test the stability of the insurance design model; Section 5 summarizes the main work and conclusions of this article, points out the limitations of the research, and looks forward to future research directions.

2. Related Works

The study of insurance trigger indices based on monitoring data plays a crucial role in insurance design, as it can provide insurers, reinsurance institutions, and regulatory authorities with objective and quantifiable risk assessment bases. By accurately constructing trigger indices, this field helps enhance the transparency of insurance products, reduce moral hazard, and optimize the compensation efficiency of special insurance types such as catastrophe risks insurance. In recent years, scholars from various disciplines have adopted interdisciplinary methods and diverse data sources to conduct in-depth research on lightning disaster risk assessment, lightning risk index construction, and the application of parametric insurance and extreme value theory, achieving significant progress. This section will categorize and present representative theories and research based on the technical paths and research methods of the above three aspects.

2.1. Lightning Disaster Risk Assessment Research

IEC 62305-2:2010 is the initial and most authoritative method for assessing the risk of lightning disasters, and it is widely used in engineering design and practice. In IEC 62305-2:2010, the lightning risk is the sum of different risk components, expressed by the following general equation:

$$R_X = N_X \cdot P_X \cdot L_X, \quad (1)$$

where N_X is the number of dangerous events per year, P_X is the probability of damage to a structure, L_X is the consequent loss. Mohammad [18] developed a software application that complies with the IEC 62305-2 standard and is specifically designed for lightning protection design and risk assessment analysis in photovoltaic power stations. However, this standard does not fully consider the additional loss impact of building structures and their surrounding layouts on these areas; moreover, the IEC 62305-2 standard is directly formulated for individual buildings and thus, by its very scope and methodology, is not applicable for directly assessing the magnitude of lightning strike risk over a larger regional area.

Fuzzy mathematics is more accurate and feasible for regional lightning risk assessments. Fuzzy logic, as a powerful analytical tool, can combine linguistic variables with numerical variables to estimate the inherent subjectivity in risk analysis and thereby determine whether the risk level is acceptable [19]. Luis et al. [20] selected typical lightning disaster events and specific scenarios, focusing on the types of lightning and the electro-

magnetic coupling methods, and analyzed the relationship between harmful events and lightning parameters. Among them, the number of sources of damage is mainly reflected by the value of ground flash density. Regarding the impact and severity of fires caused by lightning strikes, Luis E. suggested using single-element fuzzy values assigned within the range of 0-1 to represent them. Yu et al. [6] took Chongqing as an example and analyzed the lightning risk from three dimensions: hazard, exposure degree, and vulnerability of the risk-bearing body. They derived the lightning risk assessment formula for this region. The lightning risk assessment based on fuzzy mathematics is more suitable for the language description of value and loss. It is more targeted, having a more reasonable analysis logic, and applicable to a wider range of applications; however, this method mainly relies on expert experience rather than actual measurement data when determining the weights of the formulas, and the selection criteria for parameters have not yet been unified.

In recent years, multi-method combined research has gradually emerged. For instance, Murphy et al. [21] combined spatial lightning map data, the probability risk calculation method derived from IEC 62305-2, and weighted average interpolation technology to obtain the regional risk quantity, and compared it with the tolerable threshold to issue lightning warnings.

2.2. Construction Methods of Lightning Risk Index

2.2.1. Analytic Hierarchy Process

When constructing the lightning risk index, a relatively simple and convenient method is to use the analytic hierarchy process. Mahdariza [22] proposed determining several main influencing variables by reviewing the literature and then using the analytic hierarchy process to determine the weights of these variables. However, the index formula designed according to this method is only applicable to the corresponding area, and the coverage area cannot be too large; for regions with significant climate differences, other completely different types of variables may need to be added. The design of this index formula relies on the description in the literature and subjective experience and still needs to be tested and reviewed by the authorities of the corresponding area. It does not have strong objectivity and practicality.

2.2.2. Principal Component Analysis

Principal component analysis reflects the contribution of each feature to the overall information structure of the dataset, thereby facilitating the identification of features that play a primary role in lightning destructiveness. Montanya et al. [23] incorporated electric field, temperature, humidity, wind speed, and air pressure into PCA and found that in the planar space formed by the first and second principal components, the data could be clearly distinguished into three clusters corresponding to the pre-discharge, during-discharge, and recovery phases of thunderstorms, based on which they designed a decision-making logic for lightning protection warnings. Thomas et al. [24], using PCA on their research dataset, found that the first principal component reflected the presence of convection, while the second described whether microphysical processes were sufficient to trigger lightning. However, PCA merely condenses the information from the original data; its derived scores lack physical or practical meaning and do not directly reflect the contribution of individual features to lightning-induced risk. Therefore, the PCA score alone should not be directly adopted as a risk index. If regression is performed using nonlinear models such as random forests, the utility of PCA-concentrated information may be more fully exploited.

2.2.3. Machine Learning

Although the analytic hierarchy process and fuzzy mathematics theory take into account various lightning parameters and meteorological factors based on lightning knowl-

edge, these methods have strict requirements for scenario assumptions and are difficult to generalize in diverse scenarios. Moreover, traditional methods may struggle to handle the massive amounts of data available today, and the quantitative relationship between lightning risk and indices is relatively monotonous. In contrast, machine learning methods loosen the assumptions of scenarios and no longer strictly rely on linear relationships or specific distribution forms. They can efficiently process large datasets and automatically extract effective information from high-dimensional features; moreover, the discovered quantitative relationships are more flexible and can capture complex patterns such as non-linearity and interactivity. Therefore, machine learning methods are expected to enhance the applicability and interpretability of risk indices, providing a new technical path for lightning disaster risk assessment.

In machine learning algorithms, the support vector machine (SVM) achieves the maximum separation of samples in the feature space by finding the optimal hyperplane. It has unique advantages in having small sample, with nonlinear and high-dimensional pattern recognition. Currently, in the risk assessment of lightning disasters, the application of SVM mainly focuses on binary classification problems, simplifying the assessment results to a binary output of whether a lightning disaster will occur or not. Sheng et al. [25] trained a classifier using historical monitoring data such as longitude and latitude, lightning type, and disaster level to achieve qualitative discrimination. However, a single binary output cannot depict the gradual characteristics and threshold intervals of the disaster-causing intensity of lightning. Compared with this method, random forests are often used to assess the degree of lightning risk. It is an ensemble learning algorithm based on supervised learning. By constructing multiple decision trees and combining their prediction results through voting or averaging, it achieves the classification assessment of lightning risk and outputs the continuous weight values of each feature's contribution to the risk. Random forests are beneficial to some extent in preventing overfitting, reducing prediction errors [26], and effectively improving the accuracy of model output weights.

However, machine learning models are typically regarded as “black boxes”, with their internal decision-making mechanisms being complex and difficult to explain [27]. This limits the credibility of the models' application in scenarios such as index construction and insurance pricing that require clear attribution. To address this issue, SHAP values calculate the marginal contribution of features under different combinations based on game theory and output the contributions allocated to each feature. This method enhances the interpretability of the model and helps salespeople and decision-makers understand its working principle [15,28].

2.2.4. Fuzzy Mathematics Theory

Fuzzy mathematical theory can also be used in the construction of the lightning risk index, with greater accuracy and interpretability. The constructed index system using fuzzy mathematical theory receives input variables, which are several indicators used to determine the acceptability of risks. These indicators are usually different indices constructed by the fuzzy mathematics method described above. After the system's operation, it outputs a fuzzy number as the result of the index construction. Depending on the different application scenarios and requirements, the operation method and the meaning of the output result may vary. Luis [20] constructed a system that outputs a value named “Risk Acceptability” through fuzzification and rule reasoning. Guo et al. [29] obtained the assessment value of regional risk by calculating the fuzzy matrix, which is used to confirm the lightning strike risk level of the region. Since the fuzzy logic system can combine numerical input and fuzzy input through appropriate language interpretation methods and suitable rule libraries, it has been proposed as a suitable tool for estimating the acceptability of risks.

2.3. Application of Parametric Insurance and Extreme Value Theory

Natural disasters are characterized by their suddenness and high destructiveness. It is difficult to completely avoid the losses caused by them solely through technical means. To avoid and hedge against various natural risks, including lightning risks, insurance mechanisms have been widely adopted. However, traditional insurance has certain limitations in its operation, such as high management costs, long claim processing times, and difficulties in effectively addressing adverse selection and moral hazard issues. To overcome these limitations, index insurance, as an innovative risk management tool, has been introduced. Index insurance is a special financial contract, and its uniqueness lies in the fact that the triggering conditions for compensation depend on a pre-defined index rather than the losses suffered by the policyholder [30]. Since indices cannot be predicted or manipulated by individuals or groups, this compensation mechanism is feasible in reducing information asymmetry problems [31]. Therefore, index insurance is receiving increasing attention and is rapidly developing in various industries.

At present, the research on index insurance for lightning disasters is still relatively limited, and the relevant reference literature is scarce. Therefore, it is necessary to transfer knowledge from the relatively mature field of weather index insurance research and draw on its modeling ideas and methods to provide theoretical support for the construction of lightning index insurance. In the price of weather index insurance, the costs are usually decomposed into risk costs, management costs, costs for timely obtaining funds in a timely manner, etc. [32]. The compensation amount generated by the index insurance contract has a multistep form, and this loss cost can be explained by the relatively typical positive correlation between agricultural output and indicators such as rainfall [30].

When designing insurance contracts based on a single index, precisely determining the trigger threshold is a crucial consideration. A simple method for determining the trigger threshold is to set it as the historical average index [33,34]. Although this method is relatively simplified, it has the advantages of being intuitive and highly operational and not relying on loss data. Considering the need to balance the feasibility of the research and the accuracy of the model, other studies have set the disaster level at one to two standard deviations below the average index [35–37]. In practical applications, insurance products should target catastrophic events rather than average losses to enhance demand and affordability [38]. A precise method is to construct a peaks over threshold (POT) threshold model based on the generalized Pareto distribution. When it is challenging to determine whether the small and large values in the sample can be well described by a single simple distribution, or whether the peaks are independent of each other, the POT model can be used to introduce a relatively high threshold to avoid these uncertainties [39,40].

3. System Works

This paper presents a design methodology for a lightning risk index insurance product based on multi-source remote sensing data and machine learning. As shown in Figure 1, the methodology consists mainly of three core modules: the data pre-processing and feature extraction module, the lightning risk indices construction module, and the index insurance product design module. First, based on the features of the data, the main variation information is extracted through principal component analysis and dimensionality reduction and combined with the analysis of the forest model and SHAP values. Composite feature weights showing the intensity of lightning activity is constructed in this way. Second, fuzzy mathematical theory is used to combine the weight of features with the actual distribution of data, achieving the calculation and classification of the lightning risk indices. Finally, based on the risk level thresholds, the trigger conditions, compensation amounts and premium standards for index insurance products are designed. And the insurance

compensation records and calculations are completed through a daily determination and aggregation approach.

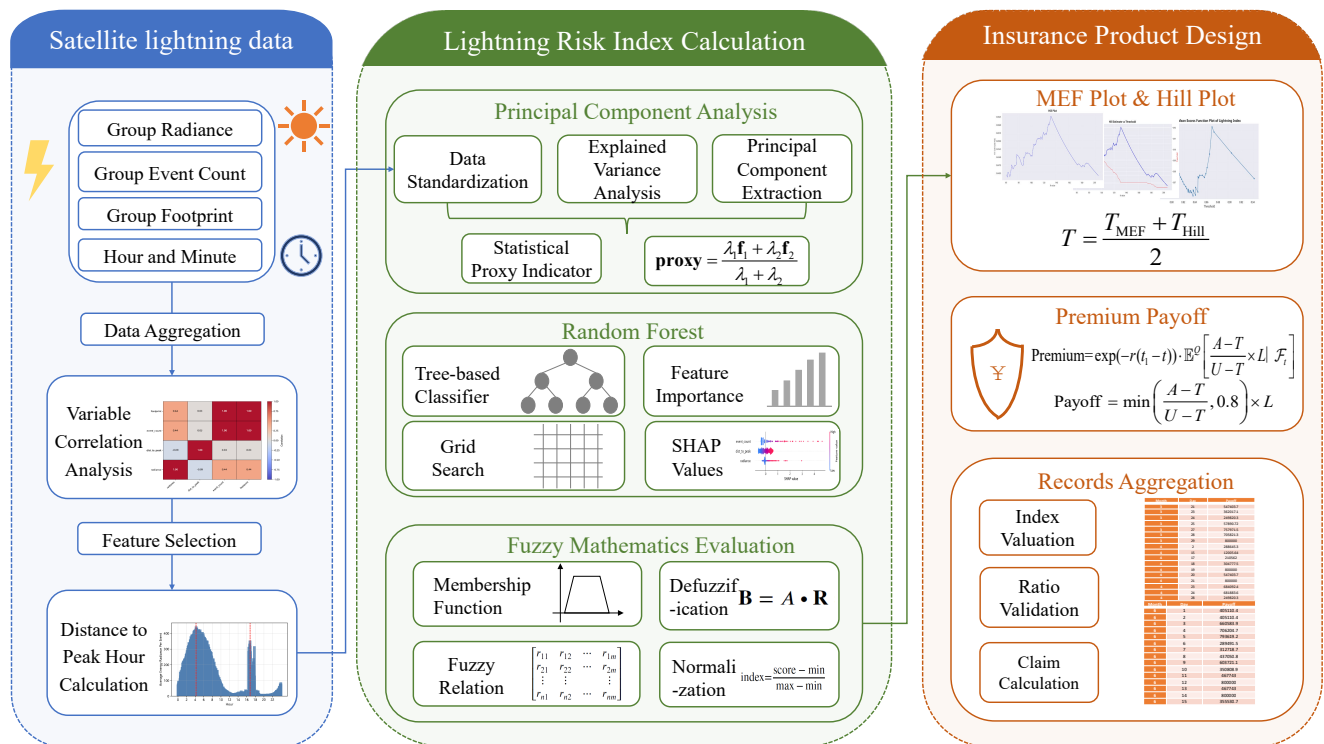


Figure 1. Framework structure flowchart.

Principal component analysis applies a linear orthogonal transformation to compress the high-dimensional correlated variables into several composite components. While preserving the major information structure in the raw data, PCA effectively filters out random noise and redundant information, thereby providing a statistically sound linear baseline for subsequent modeling. The contribution of PCA does not lie in defining risk itself but in distilling the high-dimensional monitoring information into a high-quality training label for the forest model.

PCA is only capable of capturing linear covariation among variables, so it inevitably carries bias from the linearity assumption and cannot directly interpret the loading coefficients as the risk weights of each feature. Through the splitting mechanism based on ensemble decision trees the random forests can autonomously detect nonlinear relationships between features and the target value, thus making a correction to the linear baseline. During the fitting process, the feature importance evaluation mechanism within it naturally fulfills the function of weight assignment: features that are tightly coupled with the thunderstorm physical mechanisms receive higher weights, whereas those weakly related to intensity or purely noisy are automatically suppressed.

Although the corrected score given by the forests is mathematically accurate, its output is a continuous value. Once the score lies near grade boundaries, slight input fluctuations or measurement errors may alter the final judgment. Fuzzy mathematical theory is placed at the end of pipeline, adjusting the physical direction of the feature and constructing membership functions to map the lightning data onto several risk grades. When the score varies around the threshold, the final risk determination is gradual rather than abrupt, and this flexible grading approach better suits the application needs of the insurance business. The output of each stage in the three modules constitutes the input for the next stage, and all three are indispensable.

3.1. Construction of Lightning Risk Indices

3.1.1. Construction of Training Labels Using Principal Component Analysis

Principal component analysis is a method of shifting multiple indicators into a few composite indicators for statistical analysis of observations containing multiple interrelated cause variables. Data should be standardized first, as the results are sensitive to the variance in data. For each sample, the standardized value is

$$x_{i,j}^{(std)} = \frac{x_{i,j} - \bar{x}_j}{\sigma_j}, \quad (2)$$

where $\bar{x}_j$ is the average of all data for the indicator and σ_j is the standard deviation for all data for the indicator.

If a dataset contains I observations and is described by J indicators, the dataset can be expressed in the $I \times J$ -size matrix $\mathbf{X}$. Matrix $\mathbf{X}$ allows the following singular value decomposition [41]:

$$\mathbf{X} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T, \quad (3)$$

where $\mathbf{P}$ is the $I \times L$ left singular vectors matrix; $\mathbf{Q}$ is the $J \times L$ right singular vectors matrix, called a loading matrix; and $\mathbf{\Delta}$ is a diagonal matrix containing singular values. At the same time, the $I \times L$ size factor score matrix $\mathbf{F}$ can be derived from

$$\mathbf{F} = \mathbf{P}\mathbf{\Delta} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^T\mathbf{Q} = \mathbf{X}\mathbf{Q}, \quad (4)$$

where column i of $\mathbf{F}$ is the score vector $\mathbf{f}_i$ for the principal component i . By Equation (4), it is known that the loading matrix $\mathbf{Q}$ provides a coefficient of linear combinations used to calculate scores for each principal component. The element in the load matrix $\mathbf{Q}$ is referred to as the loading, indicating the linear combination coefficient of the original variable in the construction of the main components. In practical applications, loadings are also commonly associated coefficients between the principal component and the original variable, whose absolute value reflects the importance of the variable in the principal component.

Abdi et al. [13] gives the method of calculating the explanation deviation for the principal component i :

$$\mathbf{F}^T\mathbf{F} = \mathbf{Q}^T\mathbf{X}^T\mathbf{X}\mathbf{Q} = \mathbf{Q}^T\mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T\mathbf{Q} = \mathbf{\Lambda}, \quad (5)$$

where the diagonal matrix $\mathbf{\Lambda}$ contains non-zero eigenvalues for the principal components. After sorting the eigenvalues in descending order of their absolute values, the explained variance of the i -th principal component is equal to the corresponding eigenvalue λ_i , where the interpretation variance ratio is

$$\frac{\lambda_i}{\sum_{k=1}^L \lambda_k}, \quad (6)$$

where L is the matrix $\mathbf{X}$'s rank. The higher the interpretation variance of a principal component, the greater the information in the data matrix $\mathbf{X}$ is reflected in the principal component. The magnitude of the lightning data information and the complexity of the relationships between the indicators shows that it is not a good way to substitute the original data with the first principal component. Therefore, the first and second principal components extracted after data dimension reduction were used as weights, based on their respective interpretation variance ratio. And the statistical training label is weighted by the score vector of the first and second principal components as

$$\text{proxy} = \frac{\lambda_1 \mathbf{f}_1 + \lambda_2 \mathbf{f}_2}{\lambda_1 + \lambda_2}. \quad (7)$$

In the composite proxy indicator vector, each element proxy_i is the training label corresponding to each sample. The indicator vector is the target label of the random forests model.

3.1.2. Training Weights Using the Random Forests Model

The principal component analysis method has succeeded in condensing the main variation information of multiple data features into the weighted scores of the first and second principal components. However, the score represents only the largest difference in data and does not have a direct physical or risk connotation. Therefore, using it as a statistical training label, this paper uses the random forest model to learn, thereby generating objective importance weights of different features.

Random forest is an improved parallel ensemble learning algorithm based on Bagging integration, where decision trees are used as base learners. In this algorithm, a random attribute selection mechanism is introduced during the training of decision trees. It consists of a series of tree classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, 2, \dots\}$, of which $\mathbf{x}$ is the input vector, $\{\Theta_k\}$ is a set of mutually independent and identically distributed random vectors, and each tree classifier votes for the most frequent categories at input vector $\mathbf{x}$ [14].

Random forest processes are designed to be simple, with a small scale and high performance levels, as well as lower generalization errors. Moreover, as the number of tree classifiers increases, the random forests model is less likely to suffer from overfitting. Lightning data are often distributed at a thick end and presented by complex relationships, while random forests models have a strong robustness for abnormalities and noise and are therefore more suitable for training this type of data than other common algorithms, such as Adaboost [42].

While random forest models can output the importance of features, they do not reveal the positive or negative direction of the features to predict results. Therefore, SHAP values are introduced to decompose and interpret model predictions. SHAP values are an additive explanation method for models. In this approach, model predictions are broken down into the sum of contributions to each feature, the projection expectations in the absence of any feature are used as baseline values, and the predicted values under the different sequenced feature combinations are calculated on average and then derived from SHAP values. SHAP values are calculated by

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)], \quad (8)$$

where ϕ_i is the i -th SHAP value, S is the set of non-zero indices in $\mathbf{z}' \in \{0, 1\}^M$. M is the quantity of input features. $\mathbf{z}' = 1$ indicates that the feature is observed, while $\mathbf{z}' = 0$ indicates that the feature is unknown or is covered when calculating. N is the set of all input features [15].

This paper is based on the scikit-learn library of Python for the practical realization of random forest models in research. Specific application processes are as follows:

1. Divide the data set into training and test sets at 8:2. When training the random forests, each decision tree is constructed using bootstrap sampling from the original dataset. Data not extracted are not learned by any tree learner as out-of-bag (OOB) data, so OOB data can be used directly as a test set to assess general errors in evaluating the generalization error of the overall model, and no additional data need to be reserved [14]. Nevertheless, the amount of data used in this study is sufficient to avoid the need to use OOB estimates to save training samples. To ensure the rigor of model training, this paper still divides training and testing sets. Considering the limitations of computer memory and computational cost, in this paper, a 5% subset was extracted

from the training set using the stratified sampling method for lightweight training. This approach effectively reduces the training cost of the model while ensuring the sufficiency of the training data and maintaining the same data features of the training set. However, full test data are still used for model testing.

2. Use GridSearchCV class to optimize the parameters of the random forest model. Specifically, by setting the parameter $cv = 5$, the training set is divided into five folds. The parameter combinations are traversed through grid search, and the average performance obtained from cross-validation is used as the evaluation norm. After determining the optimal parameters, in order to fully utilize the data information, the final model is retrained using the entire training set instead of a single fold's training subset.
3. Extract the importance weights of the features from the trained model.
4. Calculate the SHAP values for each feature, and use the SHAP method to explain the model's prediction results. This is used to reveal the positive or negative contribution directions and influence degrees of the features on the prediction results. Sum and normalize the extracted feature importance weights to form the weight vector for the subsequent fuzzy mathematics comprehensive evaluation.

3.1.3. Constructing Risk Indices by Fuzzy Mathematics

After extracting feature weights from the random forest model, this paper introduces fuzzy mathematics for integrated risk assessment. Because of the vagueness of the lightning risk, the information on the lightning monitoring data does not provide a clear picture of the extent of the lightning risk. Fuzzy integrated evaluations can effectively address such uncertainties [29].

Three basic elements required before a fuzzy integrated evaluation are described below:

1. Factor set $U = \{u_1, u_2, \dots, u_n\}$. The factor set includes the levels for evaluating the severity of lightning strike risks and the features to be considered when calculating the lightning strike indices. Three main features are used to measure the level of the thunder strike risk, which are the number of lightning strikes contained in each lightning cluster, the radiation intensity of the lightning cluster, and the difference between the detection time of the lightning cluster and the time of the radiation peak during the study. The peak radiation time is defined as the time when the radiation of each lightning event reaches its maximum value over an average day. The time of lightning occurrence is converted to the margin of peak time in order to capture more precisely the relationship between the magnitude of the lightning risk and the time of lightning occurrence. A lightning event takes place at a time of t , and there are a total of n radiation peaks per day at $p_1, p_2, \dots, p_n$. When lightning occurs, it is converted to a circular distance from the peak time as

$$d(t) = \min_{i \in \{1, \dots, n\}} (|t - p_i|, 24 - |t - p_i|), \quad (9)$$

where lightning time t is calculated as

$$t = \text{hour} + \frac{\text{minute}}{60}. \quad (10)$$

2. Evaluation set $V = \{v_1, v_2, \dots, v_n\}$. The set includes different levels of judgment. The risk of lightning is classified into low-risk, medium-risk and high-risk categories, so the evaluation set $V = \{\text{low risk, medium risk, high risk}\}$.
3. Weight vector $\mathbf{A} = \{a_1, a_1, \dots, a_n\}$. The weight vector includes the weight of each factor in the outcome. The weight vector used in this paper has been output from the model in Section 3.1.2.

Having obtained the above three elements, this paper presents a fuzzy comprehensive evaluation as follows:

1. Determine the membership function $A(x)$. The membership function depicts the degree of conformity of each data point to different levels of the evaluation set in each feature. The commonly used methods for determining the membership functions include the fuzzy statistics method, the assignment method, and the binary comparison ranking method [43]. This paper uses the assignment method to determine the function as a trapezoidal distribution, which is used as the basic form of the function. We also determine the specific parameters according to the kurtosis and measure of skewness of the different features. The functions attached to each feature are defined here as shown in Table 1, where p_n indicates the n -th quantile of the corresponding feature. In the data we use, the skewness of the lightning strike quantity and the radiation intensity are 3.09 and 9.64, in severe right-skew pattern; the kurtosis of those are 14.67 and 162.42, with extremely thick tails. The selection of the quantiles by the conventional method would result in a large number of conventional lightning events being incorrectly classified as medium or even high risk, seriously reducing the interpretability of the model. Thus, the parameters for lightning strike quantity and radiation intensity are selected as generally high percentiles. However, the peak time difference has a skewness and kurtosis of -0.13 and -1.15 , which are basically symmetrically distributed, so we select parameters for it as more conventional quantiles.

Table 1. Membership functions of different features and level.

Factors	Risk Level		
	Low Risk	Medium Risk	High Risk
Number of lightning strikes	$A(x) = \begin{cases} 1, & x \leq p_{90} \\ \frac{p_{95}-x}{p_{95}-p_{90}}, & p_{90} < x < p_{95} \\ 0, & x \geq p_{95} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{90} \\ \frac{x-p_{90}}{p_{95}-p_{90}}, & p_{90} < x < p_{95} \\ 1, & p_{95} \leq x \leq p_{99} \\ \frac{p_{99.9}-x}{p_{99.9}-p_{99.5}}, & p_{99.5} < x < p_{99.9} \\ 0, & x \geq p_{99.9} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{99.99} \\ \frac{x-p_{99}}{p_{99.99}-p_{99}}, & p_{99} < x < p_{99.99} \\ 0, & x \leq p_{99} \end{cases}$
Radiance intensity	$A(x) = \begin{cases} 1, & x \leq p_{95} \\ \frac{p_{95}-x}{p_{99.5}-p_{95}}, & p_{95} < x < p_{99.5} \\ 0, & x \geq p_{99.5} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{99} \\ \frac{x-p_{99}}{p_{99.5}-p_{99}}, & p_{99} < x < p_{99.5} \\ 1, & p_{99.5} \leq x \leq p_{99.9} \\ \frac{p_{99.95}-x}{p_{99.95}-p_{99.9}}, & p_{99.9} < x < p_{99.95} \\ 0, & x \geq p_{99.95} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{99.999} \\ \frac{x-p_{99.9}}{p_{99.999}-p_{99.9}}, & p_{99.9} < x < p_{99.999} \\ 0, & x \leq p_{99.9} \end{cases}$
Peak time difference	$A(x) = \begin{cases} 1, & x \geq p_{90} \\ \frac{x-p_{70}}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \leq p_{70} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{30} \\ \frac{x-p_{30}}{p_{50}-p_{30}}, & p_{30} < x < p_{50} \\ 1, & p_{50} \leq x \leq p_{70} \\ \frac{p_{90}-x}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \geq p_{90} \end{cases}$	$A(x) = \begin{cases} 1, & x \leq p_{10} \\ \frac{p_{50}-x}{p_{50}-p_{10}}, & p_{10} < x < p_{50} \\ 0, & x \geq p_{50} \end{cases}$

2. Build a matrix of fuzzy relationships. For each set of data, the degree of membership is applied to the different levels of the functions by feature, each feature forming a line vector. For the i -th feature, if the corresponding vector is $(r_{i1}, r_{i2}, \dots, r_{im})$, we can generate a fuzzy matrix $\mathbf{R}$ as

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix}. \quad (11)$$

3. Fuzzy synthesis calculation. Perform weighted summation along the feature dimensions for each sample to form combined assessment matrix **B**, as

$$\mathbf{B} = \mathbf{A} \bullet \mathbf{R} = (b_1, b_2, \dots, b_n). \quad (12)$$

In this study, the integrated assessment matrix **B** is $(b_{\text{low}}, b_{\text{medium}}, b_{\text{high}})$.

4. Defuzzification. This paper uses the centroid method to calculate scores for each sample:

$$\text{score} = b_{\text{low}} \times 2 + b_{\text{medium}} \times 5 + b_{\text{high}} \times 8. \quad (13)$$

The weights multiplied by each degree of membership are used to distinguish the severity of different risk classes. The three weights in this study are derived from the five-tier risk scoring system specified in the *Technical Specification for Lightning Disaster Risk Assessment* (QX/T 85-2018) issued by the China Meteorological Administration [44]. This standard assigns membership scores of 1, 3, 5, 7, and 9 to the five ascending risk levels. Since our study reduces lightning risk to only three levels, we cannot directly adopt the five original weights. Instead, we take the arithmetic mean of each pair of adjacent original weights to define our three new weights. Specifically, the low-risk level uses the mean value of 1 and 3, the medium level retains 5, and the high level uses the mean value of 7 and 9. Although this specification is intended for building-specific lightning risk evaluation, its core fuzzy-mathematics-based lightning risk index formula is conceptually transferable.

5. Get risk indices. The scores after min–max normalization are the final lightning risk indices, whose formula is

$$\text{index} = \frac{\text{score} - \min}{\max - \min}, \quad (14)$$

where max and min are the maximum and the minimum original scores.

3.2. Insurance Threshold Selection

The meaning of “risk” includes not only the magnitude of damage caused by a single lightning strike, but also relates to the potential probability of resulting losses. “Index” refers to indicators of the intensity of a given disaster, such as earthquake magnitude, typhoon speed, rainfall amount, etc. The compensation for index insurance is not based on the actual loss of the insured person but on whether or not the predefined disaster intensity reaches the trigger level to determine the amount of compensation [45]. Therefore, the “lightning risk index” not only characterizes the intensity of a single lightning event but also describes the likelihood of it evolving into a lightning disaster, thereby indicating the potential probability of triggering an insurance payoff. The lightning risk index is highly relevant to the loss, and the index calculation data are derived from satellite monitoring and cannot be predicted or manipulated; there has been historically sufficient data to support the pricing of parameterized insurance products. It is therefore reasonable and feasible to design parameterized insurance products for a lightning disaster. As the satellite data selected in this paper cover a very wide range of regions, this paper designs the insurance product as a regional index insurance. For the wide range of insured areas, if the lightning strike index reaches the trigger threshold, the insurance company will provide compensation to the policyholders to cover the possible economic losses caused by lightning strikes.

The most important thing to design is to establish the trigger threshold for insurance. The thunderstorm risk is a catastrophe risk, with a small probability of large losses, which are usually distributed in a thick tail. The POT model, using a generalized Pareto distribution to aggregate the tail of the loss distribution, is more accurate in describing features of loss distributions at the end than other statistical methods [46], and the POT model is more

accurate in the case of small sample data than other methods [47]. This model is therefore used to select the trigger threshold for insurance. In particular, the methodology described in the following two subsections is used.

3.2.1. Mean Excess Function Graph

To draw the mean excess function (MEF) graph, we need to calculate the average difference between each sample and the given threshold u as $e(u) = E[X - u \mid X > u]$. Usually, values are uniformly selected from the end of the data, and then all sample points greater than each threshold are selected. The arithmetic mean is calculated by subtracting the threshold from these sample points. Based on the ascending order of the exceedances, the mean exceedance function graph can be drawn, where the horizontal axis represents the threshold and the vertical axis represents the average exceedance function value [48]. The mean excess of the function chart often remains positive and shows an overall downward trend. Usually a location where the mean excess function begins to rise rapidly is selected as a threshold [49].

3.2.2. Hill Plot

The Hill plot is an important tool for estimating distribution in extreme value theory. First, sort the risk indices in descending order to obtain the order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. For each upper threshold k , take the threshold $X_{(k+1)}$ and calculate Hill's estimate [50]:

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \ln \frac{X_{(i)}}{X_{(k)}}, k = 1, 2, \dots, n-1. \quad (15)$$

In view of the fact that most of the features of the lightning data are in a severe right-skewed pattern, the k values have been selected from 60. In order to prevent excessive sample sizes and to avoid introducing system deviations, the maximum k value has been set at 1‰ of the sample volume, and 200 points have been set equally between the range for observation. The Hill map is based on k values and Hill estimates, where the horizontal and vertical axes represent the k values and corresponding Hill estimates. Usually, the index corresponding to the Hill statistic, where the plot displays a linear trend above this index, is selected as the threshold [51].

3.3. Premium and Claim Design

The selection of the trigger threshold addresses the underlying issue of when to pay, but it is not sufficient to specify the conditions for payment to constitute a complete index insurance scheme. Two core economic parameters need to be further defined for a practical index insurance product: the standard of claim and the premium. The former determines the amount of financial compensation that an insured person can obtain in the event of a disaster, which directly affects the safeguarding function of insurance and the incentive to participate, while the latter relates to the reasonableness of the pricing of insurance products, financial sustainability, and the willingness of the insured to pay. This section sets out the methodology for determining the amount of compensation to be paid and the premium standard separately.

3.3.1. Calculate the Claim

A claim can be calculated using a linear compensation structure. Normally, a base rate calculation structure is [31]

$$r = \frac{A - T}{U - T}, \quad (16)$$

where r is the claim rate, T is the threshold, A is the actual index value, and U is the index ceiling for insurance. In this study, T is determined by the average value of the threshold determined by the two methods described in Section 3.2, and U should make sure the insurance trigger rate no greater than 20%.

However, compensation mechanisms for index insurance are limited to multilayered arrangements for the transfer of catastrophic risks. In catastrophe risk transfer systems, insurance companies spread risk through reinsurance, capital market instruments and massive-disaster funds in which the government participates, with different levels of loss shared by the participating entities. Therefore, insurance companies should have a clear ceiling on their own liability for compensation [52]. Losses above that limit will be taken over by a superior body, such as the reinsurers or the government, and insurance companies will no longer be subject to additional liability. Following the industry practice in property insurance, this article sets the upper limit of the compensation rate at 80%. The formula used for calculating the rate of claim is therefore

$$r = \min\left(\frac{A - T}{U - T}, 0.8\right). \quad (17)$$

According to Equation (17), the formula for calculating compensation can be

$$\text{Payoff} = r \times L, \quad (18)$$

where L is coverage.

3.3.2. Calculation of Premiums

This paper will calculate premiums in such a way as to achieve a theoretical balance of payments:

$$\text{Premium} = \exp(-r(t_1 - t)) \cdot \mathbb{E}^Q \left[\frac{A - T}{U - T} \times L \mid \mathcal{F}_t \right], \quad (19)$$

where $\mathbb{E}^Q \left[\frac{A - T}{U - T} \times L \mid \mathcal{F}_t \right]$ means that in the current time period, compensation is expected to be paid at a risk-neutral measure Q [53], and $\exp(-r(t_1 - t))$ will discount the compensation amount from the time t_1 when the compensation is paid, using the risk-free interest rate, to the time t when the premium is determined.

The lightning risk index could not be stored or traded, so the lightning risk parametric insurance market is an imperfect market. In the imperfect market, due to the existence of multiple risk neutral measurements, the sole theoretical price of the lightning insurance product could not be determined only through no-arbitrage pricing theory [54]. Therefore, this paper limits the theoretical price of insurance products by introducing risk-added factors.

According to the Radon–Nikodym derivative, the risk-neutral measure Q and the realistic probability measure P are related as follows [55]:

$$\frac{dQ}{dP} = \frac{\exp(\gamma X)}{\mathbb{E}^P[\exp(\gamma X)]}, \quad (20)$$

where X is the claim and γ is an additional risk factor. Therefore, the expected value of compensation in the risk-neutral measure Q can be indicated in the realistic probability measure P as

$$\begin{aligned} \mathbb{E}^Q[X] &= \mathbb{E}^P \left[X \cdot \frac{dQ}{dP} \right] \\ &= \mathbb{E}^P \left[X \cdot \frac{\exp(\gamma X)}{\mathbb{E}^P[\exp(\gamma X)]} \right] = \frac{\mathbb{E}^P[X \cdot \exp(\gamma X)]}{\mathbb{E}^P[\exp(\gamma X)]}. \end{aligned} \quad (21)$$

This paper replaces the real probability distribution of the sample by the real probability distribution of X , as

$$\mathbb{E}^Q[X] \approx \frac{\frac{1}{N} \sum_{i=1}^N X_i \exp(\gamma X_i)}{\frac{1}{N} \sum_{i=1}^N \exp(\gamma X_i)} = \frac{\sum_{i=1}^N X_i \exp(\gamma X_i)}{\sum_{i=1}^N \exp(\gamma X_i)}. \quad (22)$$

As this insurance product follows a daily activated monthly settlement model, the discount effect is minimal. For the sake of simplicity, $\exp(-r(t_1 - t))$ will be ignored. At the time the premium was set for insurance for the first time, due to the lack of data on compensation for the insurance period, the expectation is multiplied by a certain risk-adding factor to cover the uncertainty of the parameter estimates, the tail risk premium, and the operating and management costs of insurance companies. This is set to 1.2. Upon the accumulation of data on compensation payments, historical premiums are linked to the corresponding amounts and the numerical method is applied to Equation (22) to obtain a significant risk-added factor γ and adjust it.

When determining premiums on the basis of multi-month historical data, projections based on averages are not reliable owing to significant differences in historical data levels. Accordingly, this paper uses the maximum of the monthly claims. After the above steps, the premium is determined.

3.3.3. Claim Record Combination

The core feature of index insurance is that compensation is based on objective indicators rather than actual losses, so that the frequency of index values and aggregation directly influence the outcome of the claim. Satellite monitoring data used in this paper are provided on a daily basis, so the insurance trigger is implemented accordingly. Specific processes are as follows:

1. Calculate the daily indices. The value of the current-day insurance index is calculated daily on the basis of features such as radiation intensity, lightning strike quantity, and so forth. The combination of the indices reflects the intensity of lightning activity in the area monitored on that date and is the basis for determining whether or not the insurance trigger was triggered.
2. Select the maximum value within the day. Given that the intermittent and volatile nature of one-day lightning activity, a single-time monitoring value may not be able to provide a full picture of the overall risk level at that date. For the purpose of conservatively estimating risk exposure and simplifying the processing of claims, the maximum value of the single-time insurance indices is used as the representative value of the date for all monitoring times. As long as the maximum value reaches or exceeds the predefined trigger threshold, the insurance claim is deemed to have been triggered on that date.
3. Converge records of claims on a daily basis. After a day-by-day determination of the trigger result, all dates during the insurance period are recorded and a series of successive claims records is made. Each record contains a separate date and the amount of compensation calculated according to steps 1 and 2.

4. Experiments

To test the actual performance of the lightning risk index insurance products constructed in Section 3, this section conducts a series of experiments based on satellite monitoring data and historical disaster records in the study area. The experiment included a daily calculation of the insurance indices and trigger determination, measurement of

the frequency and amount of payment, comparison of different percentile parameters, various model settings, and the reasonableness test of loss ratios. In the above-mentioned experiment, the aim is to assess the safety and financial sustainability of the product.

4.1. Datasets and Experimental Setup

The study selected L2-class lightning group data, which was formed in minutes, from the Chinese weather satellite FY-4A lightning mapping imager. The dataset used in this study includes the following key information: the time range spans from 1 March 2023 to 31 May 2023; the observation area covers part of China and Australia; and the total sample size is 2,061,376. The variable information is shown in Table 2.

Table 2. Basic variable information of the dataset.

Feature	Unit	Variable Abbreviation	Access
Group Radiance	$\mu\text{J}/\text{m}^2/\text{ster}$	radiance	Directly included in dataset
Hour	None	hour	Extracted from the file name according to the official naming format
Minute	None	minute	Extracted from the file name according to the official naming format
Group Event Count	None	event_count	Directly included in dataset
Group Footprint	km^2	footprint	Directly included in dataset

The lightning mapping imager can detect lightning in China and the surrounding regions, thereby enabling severe convective weather monitoring and tracking, providing early warning of lightning disasters. The LMI used a CCD array that operated at a wavelength of 777.4 nm to capture lightning optical emission. The optical events were filtered and clustered to “Event”, “Group” and “Flash” products. A “Group” is suggested as a lightning discharge. An “Event” is suggested, as an optical event of one single pixel exceeded the background threshold during one frame.

By contrast, lightning events are only instant light detected by individual pixels, which are susceptible to background light and instrument noise, having a low and unstable signal-to-noise ratio. Flashes mix multiple independent discharge processes and lose important process details. Group data allows for both stable signals of high signal-to-noise ratio and the extraction of abundant feature information, such as the duration of discharge, total energy, spatial coverage, etc. These features reflect the strength and scale of the thunderstorm system and are of the most immediate physical significance for assessing flash disaster risk.

The experiments in this study were conducted using Python 3.13.2, which was run on a computer equipped with an AMD Ryzen 9 7945HX processor with Radeon Graphics at 2.50 GHz, a NVIDIA GeForce RTX 4060 graphics card, and 16 GB of memory.

4.2. Feature Engineering

In order to ensure the quality of the subsequent feature construction and the stability of model training, the raw data is preprocessed in this study. We remove samples with null values for the features and those with obvious outliers, mainly including negative or zero values for physical variables such as radiation intensity, energy, etc. A total of 150 samples were removed from the experiment and 2,061,226 data points remained for the experiment.

4.2.1. Calculation of Derivative Features

This section converts the lightning detection time into peak distance time (hereinafter referred to as *dist_to_peak*) according to the methods described in Section 3.1.3. The radiation intensity for each hour is shown in Figure 2.

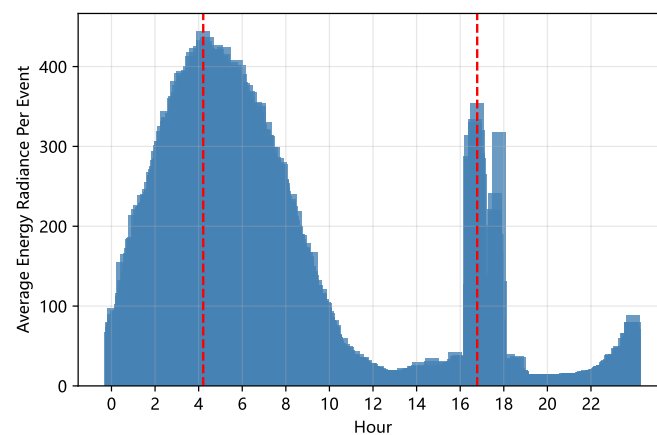


Figure 2. Daily statistics chart of radiation intensity by hour.

Figure 2 and related statistical information indicate that there are two peak hours of daily radiation intensity shown as two red dotted lines, which are 4.21 (approximately 4:13) and 16.78 (approximately 16:47). In the dataset, the time difference between the transformed radiation peak values is [0.003,6.28] h after conversion, with an average time of 3.19 h.

4.2.2. Variable Selection

To study the correlation between variables and avoid introducing redundancies into models, this paper calculates the correlation coefficients between variables and draws the heat map as shown in Figure 3. As can be seen from Figure 3, the correlation between other variables is generally weak, except the full correlation between lightning events and the footprint. It has been observed that all lightning footprints in the data collection are 100 times larger than the number of lightning events. Therefore, we delete the Group Footprint variable and keep the three remaining variable input models to calculate the lightning risk indices.

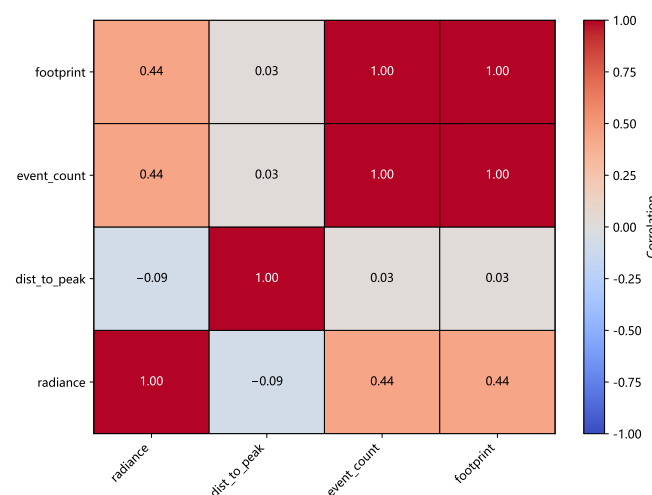


Figure 3. Variable correlation coefficient heatmap.

4.3. Calculation Results of Risk Indices

This experiment starts with principal component analysis. The interpretation variance of the first principal component in this dataset is 48.17%, and that of the second is 33.68%. The sum of the two is 81.85%, indicating that the first and second components include most of the information from the original data. The load matrix is shown in Table 3, where PCA 1 stands for the loading of the first principal component while PCA 2 stands for the loading of the second principal component.

Table 3. Principal component analysis load matrix.

Feature	PCA 1	PCA 2
radiance	0.7100	−0.0568
event_count	0.6980	0.1896
dist_to_peak	−0.0938	0.9802

The first principal component has a high and close positive payload in total radiation energy and the quantity of lightning strikes, while the payload is close to zero at times of peak risk. This indicates that the first component mainly reflects the overall energy output scale of lightning activity and represents a combination of radiation energy and the number of events.

The second principal component has a very high positive payload at the distance and time from the peak hour, while the payload is close to zero in the total radiation energy and quantity of lightning. This indicates that the second principal component primarily captures the temporal features of lightning occurrence as the proximity to the peak period of risk.

We train the random forest model after calculating training labels. In the experiment, we divide the data into a total of 82,449 samples for training and 412,246 samples for testing. The experiment optimized the key parameters using grid search combined with a five-fold cross-validation. The search space is set as follows:

1. Number of decision trees ($n_estimators$): 100, 200. This parameter controls the number of base learners in ensemble learning. If the value is too small, it may lead to under-fitting. If it is too large, it will increase the computational cost and the benefits will decrease.
2. Maximum depth (max_depth): 10, 20, None, where None indicates that tree nodes will continue to split until all leaf nodes are pure or reach minimum sample limits, allowing models to learn complex non-linear relationships.

The evaluation index for model performance is the determination coefficient (R^2), which is chosen to ensure the robustness of parameter selection. Finally, the parameter group that yields the highest average R^2 on the validation set is selected as the optimal model configuration. After searching, the best parameters determined by the grid search are $n_estimators = 100$ and $max_depth = 20$. The R^2 of this model under this configuration on the test set is 0.99926.

In the random forest model, a high determination coefficient is expected and reasonable. The main reason for this is that the learning objective of the model is the training label constructed through the PCA, which is a linear combination of original features. Random forests, with powerful nonlinear fitting capabilities, can closely approximate this deterministic mapping relationship with extremely high accuracy. It is important to emphasize that the core output of the random forest model in this study is not the predicted value but the feature importance based on the prediction process. This importance is used to measure the contribution of lightning features to the training label and provide objective weights for the subsequent fuzzy comprehensive evaluations. Therefore, the high accuracy of the model in fitting does not affect the rationality and interpretability of the final weight results.

The feature weights provided by the model for group event count, group radiance and peak distance time are 0.6019, 0.2276 and 0.1705. The feature importance of random forest output indicates that the number of lightning events is the most contributing risk factor, followed by total radiation energy and distance time from the peak period. This finding shows that the frequency of lightning activity has the most significant impact on risk; moreover, total energy and time series features are also important.

The statistical chart of the impact of SHAP values on model output is shown in Figure 4. The SHAP values of the number of flash events and total radiation energy are distributed on both sides of the zero axis, with negative values concentrated near the zero axis, while positive values are more dispersed and spread over a large range, suggesting that few high-energy and large-scale lightning events can have a significant boost to risk. Most ordinary flashes contribute less, with negative values attached to the zero axis, conforming to the rule that extreme events are rare. The distribution pattern of the two on the positive and negative sides is similar, suggesting that the risk of a high number of lightning strikes is often accompanied by high energy and multiple events. The SHAP values of the distance time from the peak risk period are consistent with their physical expectations as a feature of time, which means the closer to the peak, the higher the risk; the farther away, the lower the risk. The magnitudes of positive and negative effects are similar. These features above indicate that the risk judgment mechanisms identified by the model are consistent with physical perception and that the feature importance results are reasonable.

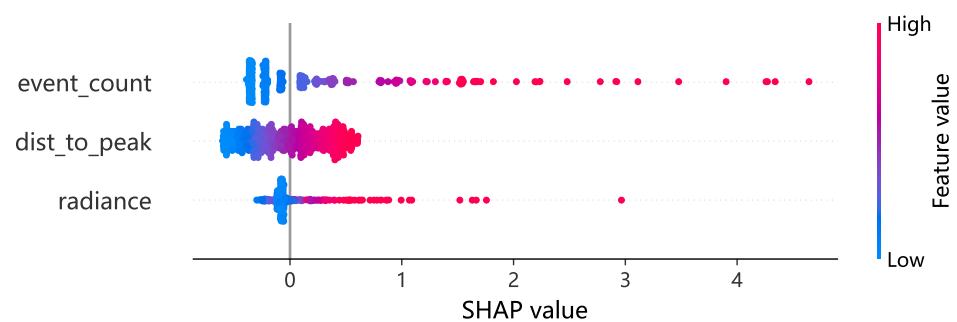


Figure 4. Statistical analysis of the influence of SHAP values on model outputs.

SHAP values also have direct practical value in insurance operations. In practice, when the index is triggered but the policyholder believes that no actual loss has occurred or vice versa, basis risk disputes appear to arise. SHAP value analysis is capable of providing a feature-by-feature contribution decomposition for each risk determination. Such attributional information can help insurance institutions explain to policyholders the basis for each trigger decision, thereby enhancing the transparency and credibility of index-based insurance and reducing claim disputes technically.

After obtaining the feature material weight vector, this paper uses fuzzy mathematic methods to calculate the lightning risk indices. The statistics of data risk levels are shown in Table 4, where we still adopt the lightning risk grading criteria from reference [44]. Table 4 shows that the distribution of sample risk levels is significantly skewed to the right. Low-risk samples absolutely dominate, while medium-risk samples account for about 4% and high-risk samples account for only 0.03%. This distribution structure is highly consistent with the natural attribute that long periods of silence are followed by short bursts. This indicates that the model has a good ability to distinguish between normal conditions and extreme events.

Table 4. Risk level statistics situation.

Risk Level	Risk Index Range	Amount	Percentage
Low risk	[0, 0.4]	1,971,044	95.62%
Medium Risk	(0.4, 0.8]	89,605	4.35%
High risk	(0.8, 1]	577	0.03%

The number and percentage distribution of samples in each score range are shown in Figure 5. Figure 5 shows a highly concentrated single-peak pattern of risk score distribution,

with peaks in the range [0.1, 0.2), with a total of 93% of samples concentrated in the lower range [0.1, 0.3). As the scores rises, the percentage of samples falls sharply, cumulatively at 1.5% for more than 0.5 and 70 samples for more than 0.9, indicating that the model had a strong screening capability for extremely high-risk samples.

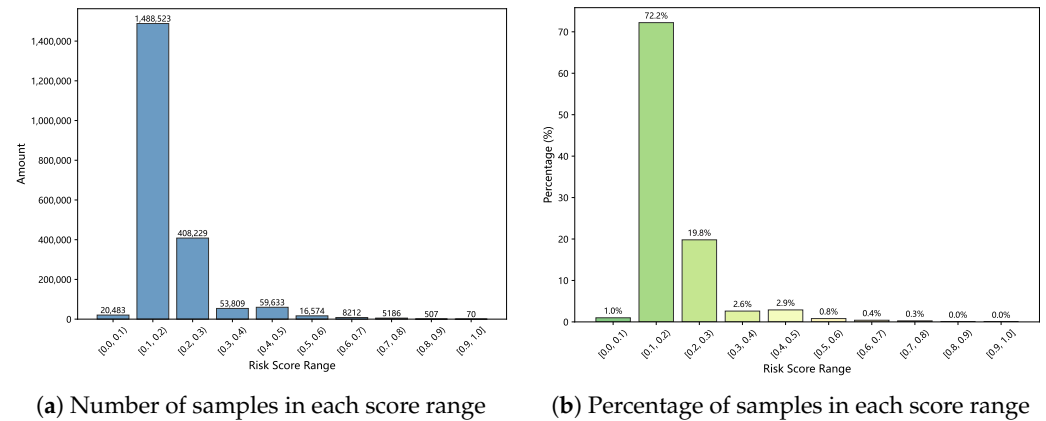


Figure 5. Histogram of sample quantities and proportion distribution for each score range

4.4. Insurance Premium and Claim Calculation Results

According to the methodology described in Section 3.2.1, the MEF and Hill plots are drawn as Figures 6 and 7. The mean excess rises very rapidly when the threshold in Figure 6 is 0.8120. Thus, the MEF plot sets the trigger threshold at 0.8120. In Figure 7, when the K value is 183, there is a significant and continuous increase, at which time the corresponding threshold is 0.8324. Combining the two methods above, we ultimately set the trigger threshold as an average of 0.8222. Considering that the study area covers two very large countries, China and Australia, the experiment assumed that the limit for compensation per accident is 800,000 yuan. Following the methodology described in Section 3.3, this experiment calculates the theoretical compensation and insurance premiums for all data and consolidates them into the records of the claims shown in Table 5. Due to the limitations of the space available, only records of claims greater than zero are kept. Table 5 shows a total of 64 events triggered in the experimental data during the three-month period, of which 55 were partial and nine were capped at 800,000 yuan. Taking into account the positioning of this study as a regional insurance product, the nine events reflect nine extremely intense convective weather events occurring in a wide area covering China and Australia, with a density of only about 0.42 per million square kilometers per month. The frequency is completely in line with the climatic characteristics of thunderstorm activities in the regions of both countries. This result is therefore evidence that the model can effectively identify regional extreme disaster events.

Table 5. Experimental aggregation of non-zero claim records.

Month	Day	Payoff	Month	Day	Payoff
3	21	673,153.1	4	27	147,926.6
3	22	131,403.6	4	28	458,250.4
3	23	539,274.5	4	29	800,000
3	24	458,250.4	5	2	351,147.5
3	25	319,646.6	5	3	169,748.5
3	26	195,682.8	5	4	509,348.7
3	27	800,000	5	5	800,000
3	28	787,556	5	6	539,369.5
3	29	800,000	5	7	800,000
3	31	221,041.1	5	8	800,000

Table 5. Cont.

Month	Day	Payoff	Month	Day	Payoff
4	1	192,512.5	5	9	673,153.1
4	2	486,288.3	5	10	139,213.5
4	3	171,804.7	5	11	59,769.32
4	4	262,012.3	5	12	131,403.6
4	5	188,335.2	5	13	51,604.82
4	6	267,317	5	14	601,771.3
4	7	107,525.5	5	15	171,804.7
4	10	131,403.6	5	17	131,403.6
4	11	35,891.24	5	18	577,640.8
4	12	242,760.1	5	19	256,707.5
4	13	114,078.1	5	20	800,000
4	14	192,941.2	5	21	267,317
4	15	286,510.3	5	22	246,316.4
4	17	429,899.6	5	23	219,560.9
4	18	497,938.3	5	24	628,834.8
4	19	800,000	5	25	336,093.9
4	20	673,153.1	5	26	493,345.8
4	21	800,000	5	27	374,538.4
4	22	219,560.9	5	28	195,682.8
4	23	771,864.2	5	29	147,926.6
4	24	770,269.1	5	30	651,102.5
4	25	192,353.7	5	31	509,348.7

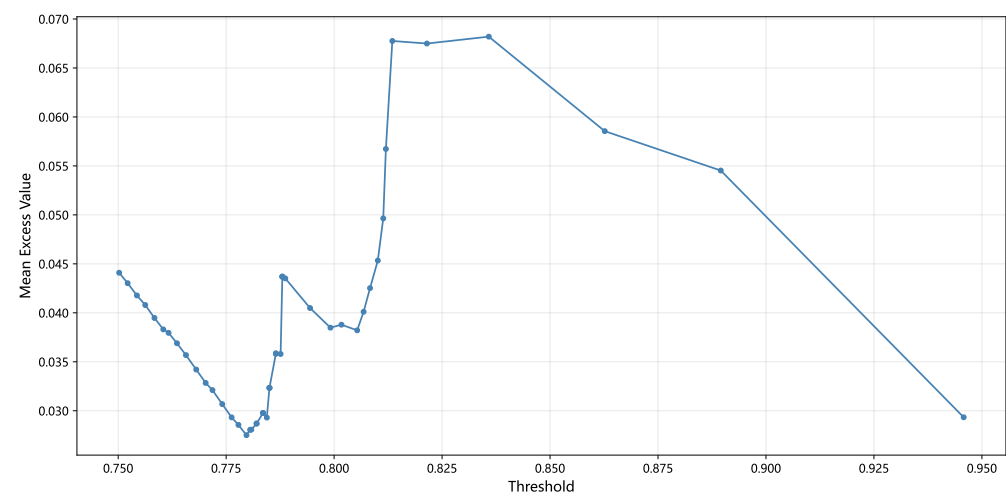


Figure 6. Mean excess function plot.

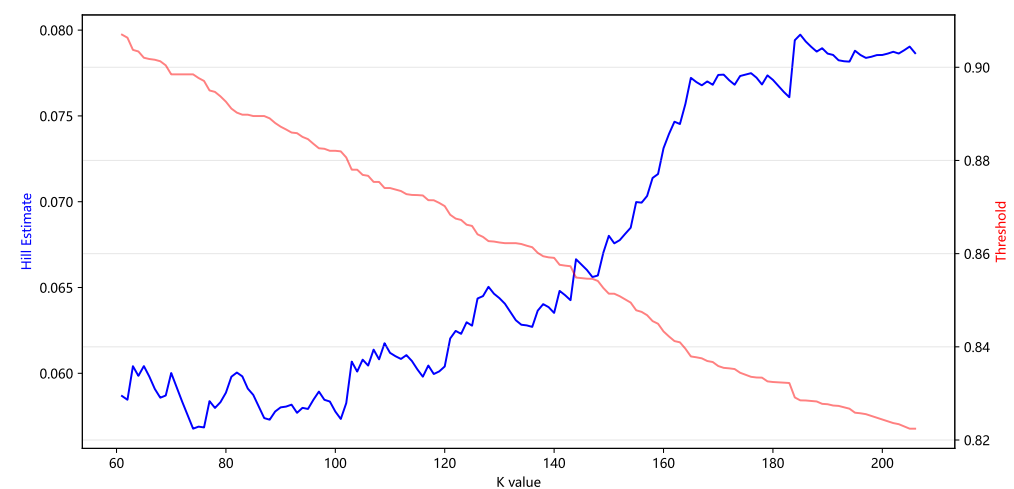


Figure 7. Hill plot.

Ultimately, due to the fact that the premium is the highest in May, the experiment resulted in determining the final premium of 13,960,984.95 yuan.

4.5. Verification of Loss Ratio

This subsection uses monitoring data generated by the same satellite from 1 June 2023 to 30 June 2023 to simulate compensation records to validate model stability by calculating loss ratio. The loss ratio is calculated as

$$\text{Loss ratio} = \frac{\text{Payoff}}{\text{Premium}} \times 100\%. \quad (23)$$

The June claims records are shown in Table 6. Table 6 shows that 30 days in June all triggered compensation, and there is no zero-compensation day. This might be because the study area was in the peak summer period when thunderstorm activities were at their strongest, resulting in a significant increase in the frequency and intensity of severe convective weather. Payoffs on a single-day basis were concentrated between 486,899.4 yuan and 800,000 yuan, of which the maximum was reached on 5, 12 and 14 June. The overall distribution of payments was more even and not extremely volatile.

Table 6. Payment records in June.

Month	Day	Payoff	Month	Day	Payoff
6	1	570,394.7	6	16	532,858.6
6	2	570,394.7	6	17	595,320.4
6	3	754,887.4	6	18	779,075.7
6	4	787,832.9	6	19	649,430.7
6	5	800,000	6	20	634,758.1
6	6	486,899.4	6	21	570,394.7
6	7	503,673.1	6	22	652,098.9
6	8	593,460.8	6	23	531,748.6
6	9	713,823.3	6	24	615,625.4
6	10	531,180.3	6	25	486,899.4
6	11	615,625.4	6	26	486,899.4
6	12	800,000	6	27	486,899.4
6	13	615,625.4	6	28	684,746
6	14	800,000	6	29	771,339.2
6	15	534,590.3	6	30	593,460.8

According to the provisions on the mandatory stress scenarios for stress testing in *Insurance Company Solvency Supervision Rules (II)* issued by the China Banking and Insurance Regulatory Commission, a comprehensive loss ratio increases to 120% of the base scenario within the next quarter constitutes a stress scenario that requires attention. Given that the test window for this experiment covers only a single month, the coverage area is a wide range of China and Australia, the volatility of the compensation rate is magnified by the shorter observation time window, and therefore the compensation ratio is considered reasonable between 0.5 and 1.5. The total amount actually claimed in June is calculated at 18,749,943.00 yuan, with the ratio of the premium based on the data of the previous three months is 1.343, indicating that the model did not show systematic pricing deviations in the off-sample data. The loss ratio passes the test, and the model was robust.

4.6. Sensitivity Analysis

Lightning strikes are rare events, but when they occur, they often cause severe consequences. This is reflected in the lightning data by extremely high kurtosis and skewness. Therefore, the percentile thresholds in the membership functions cannot be set to the conventional values commonly used in fuzzy mathematics, such as p_{10} , p_{30} , p_{50} and p_{70} .

To justify the parameter selection, we respectively modify the thresholds for the number of lightning strikes and radiance intensity to the conventional percentiles, as shown in Tables 7 and 8, and compare the resulting loss ratios across different parameter settings. The peak time difference exhibits relatively normal skewness and kurtosis and thus remains unchanged. The experimental results are presented in Table 9.

Table 7. Membership functions with different percentiles for number of lightning strikes.

Factors	Risk Level		
	Low Risk	Medium Risk	High Risk
Number of lightning strikes	$A(x) = \begin{cases} 1, & x \leq p_{10} \\ \frac{p_{30}-x}{p_{30}-p_{10}}, & p_{10} < x < p_{30} \\ 0, & x \geq p_{30} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{30} \\ \frac{x-p_{30}}{p_{50}-p_{30}}, & p_{30} < x < p_{50} \\ 1, & p_{50} \leq x \leq p_{70} \\ \frac{p_{90}-x}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \geq p_{90} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{90} \\ \frac{x-p_{70}}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \leq p_{70} \end{cases}$
Radiance intensity	$A(x) = \begin{cases} 1, & x \leq p_{95} \\ \frac{p_{95}-x}{p_{99.5}-p_{95}}, & p_{95} < x < p_{99.5} \\ 0, & x \geq p_{99.5} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{99} \\ \frac{x-p_{99}}{p_{99.5}-p_{99}}, & p_{99} < x < p_{99.5} \\ 1, & p_{99.5} \leq x \leq p_{99.9} \\ \frac{p_{99.95}-x}{p_{99.95}-p_{99.9}}, & p_{99.9} < x < p_{99.95} \\ 0, & x \geq p_{99.95} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{99.999} \\ \frac{x-p_{99.9}}{p_{99.999}-p_{99.9}}, & p_{99.9} < x < p_{99.999} \\ 0, & x \leq p_{99.9} \end{cases}$
Peak time difference	$A(x) = \begin{cases} 1, & x \geq p_{90} \\ \frac{x-p_{70}}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \leq p_{70} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{30} \\ \frac{x-p_{30}}{p_{50}-p_{30}}, & p_{30} < x < p_{50} \\ 1, & p_{50} \leq x \leq p_{70} \\ \frac{p_{90}-x}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \geq p_{90} \end{cases}$	$A(x) = \begin{cases} 1, & x \leq p_{10} \\ \frac{p_{50}-x}{p_{50}-p_{10}}, & p_{10} < x < p_{50} \\ 0, & x \geq p_{50} \end{cases}$

Table 8. Membership functions with different percentiles for radiance intensity.

Factors	Risk Level		
	Low Risk	Medium Risk	High Risk
Number of lightning strikes	$A(x) = \begin{cases} 1, & x \leq p_{90} \\ \frac{p_{95}-x}{p_{95}-p_{90}}, & p_{90} < x < p_{95} \\ 0, & x \geq p_{95} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{90} \\ \frac{x-p_{90}}{p_{95}-p_{90}}, & p_{90} < x < p_{95} \\ 1, & p_{95} \leq x \leq p_{99} \\ \frac{p_{99.9}-x}{p_{99.9}-p_{99.5}}, & p_{99.5} < x < p_{99.9} \\ 0, & x \geq p_{99.9} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{99.99} \\ \frac{x-p_{99}}{p_{99.99}-p_{99}}, & p_{99} < x < p_{99.99} \\ 0, & x \leq p_{99} \end{cases}$
Radiance intensity	$A(x) = \begin{cases} 1, & x \leq p_{10} \\ \frac{p_{30}-x}{p_{30}-p_{10}}, & p_{10} < x < p_{30} \\ 0, & x \geq p_{30} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{30} \\ \frac{x-p_{30}}{p_{50}-p_{30}}, & p_{30} < x < p_{50} \\ 1, & p_{50} \leq x \leq p_{70} \\ \frac{p_{90}-x}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \geq p_{90} \end{cases}$	$A(x) = \begin{cases} 1, & x \geq p_{90} \\ \frac{x-p_{70}}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \leq p_{70} \end{cases}$
Peak time difference	$A(x) = \begin{cases} 1, & x \geq p_{90} \\ \frac{x-p_{70}}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \leq p_{70} \end{cases}$	$A(x) = \begin{cases} 0, & x \leq p_{30} \\ \frac{x-p_{30}}{p_{50}-p_{30}}, & p_{30} < x < p_{50} \\ 1, & p_{50} \leq x \leq p_{70} \\ \frac{p_{90}-x}{p_{90}-p_{70}}, & p_{70} < x < p_{90} \\ 0, & x \geq p_{90} \end{cases}$	$A(x) = \begin{cases} 1, & x \leq p_{10} \\ \frac{p_{50}-x}{p_{50}-p_{10}}, & p_{10} < x < p_{50} \\ 0, & x \geq p_{50} \end{cases}$

Table 9. Comparison of Percentile Thresholds.

Membership Function	Thresholds of T and U	Premium in May	Payoff in June	Loss Ratio
As shown in Table 1	$T = 0.8222, U = 1$	13,960,984.95	18,749,943.00	1.343
As shown in Table 7	$T = 0.8952, U = 1$	6,845,081.17	11,278,631.41	1.648
As shown in Table 8	$T = 0.9805, U = 1$	2,347,249.87	1,387,588.74	0.591

The results show that the loss ratios in both cases deviate more significantly from 1 compared to the original model. This is mainly because the traditional percentile method misclassifies a large number of low-risk lightning strike data as medium-risk or even high-risk, thereby introducing excessive noise that prevents the model from accurately identifying the lightning hazards that are truly likely to occur. Therefore, this experiment demonstrates that percentile parameters set according to the actual data distribution, rather than based on empirical rules, are more suitable for operational application scenarios than traditional percentile parameters.

4.7. Model Comparison

To quantify the contribution of each component in the proposed framework, this subsection compares the loss ratios of the original model, its reduced variants, and a variant where forests is replaced by XGBoost. Since only the fuzzy logic-based method follows a conventional normalization practice for deriving the lightning risk index, the outputs of the other models were kept in their original unscaled form to ensure a fair comparison. For each model, the trigger threshold was selected using the same procedure, and the corresponding premiums and payoffs were calculated accordingly. The comparison results are presented in Table 10.

Table 10. Model Comparison.

Model	Thresholds of T and U	Premium in May	Payoff in June	Loss Ratio
PCA + RF + Fuzzy Mathematics	$T = 0.8222, U = 1$	13,960,984.95	18,749,943.00	1.343
PCA	$T = 16.2324, U = 20.5840$	1,852,675.76	0	0
PCA + RF	$T = 15.0340, U = 15.7896$	4,677,587.70	1,761,882.13	0.3766
PCA + XGBoost + Fuzzy Mathematics	$T = 0.8514, U = 1$	10,520,024.19	16,657,199.63	1.583

It is observed that when using only the PCA score for the payout simulation, no claim was triggered throughout the entire month of June. However, there were publicly reported lightning disaster events in China alone during that period, indicating that the model fails to capture actual lightning occurrences and would therefore be unacceptable to policyholders in practice. For the PCA + RF model, the MEF plot did not show a clear upward trend in the mean excess values, while the Hill plot produced stable and interpretable results. Therefore, the trigger threshold for this model was selected solely based on the Hill plot. This also suggests that the raw RF output alone may not be fully suitable for the operational scenario considered in this study. Among all model variants, the original framework yields a loss ratio closest to 1, indicating that it achieves the best overall performance and that each of its components makes a positive contribution to the model's effectiveness.

5. Conclusions

This paper presents a design scheme for a lightning risk index insurance product based on multi-source remote sensing data and machine learning combined with fuzzy mathematics theory. First, using PCA to score the principal components from the radiation intensity, the number of lightning events, etc., and combining random forests regression with the analysis of SHAP values, the importance of the lightning activity features is constructed. Second, the calculation of the lightning risk index was achieved by combining the feature weights with fuzzy mathematics theory. Based on this, the trigger threshold, compensation amount and premium standard for the index insurance were designed, forming a complete product framework. This work provides a transferable index insurance design approach for regions that require real-time multi-source remote sensing monitoring data or lack historical

disaster damage data. It integrates multidisciplinary approaches and has some outreach value in other natural disaster areas such as fires, typhoons and rainstorms.

The experiment verified the insurance scheme based on historical satellite monitoring data. The results showed that the constructed risk index could effectively capture the temporal and spatial characteristics of lightning activities. The random forests model had a low fitting error for the training labels, and the feature importance weights revealed by SHAP values were consistent with physical cognition. Through daily determination and aggregated claim records, the insurance scheme had reasonable frequency and amount distribution of payouts under different trigger thresholds, and the premium pricing took into account both actuarial fairness and market acceptability. Compared with the traditional fixed threshold method, this method has obvious advantages in risk discrimination and payout stability.

The framework proposed in this study demonstrates certain feasibility for deployment in actual insurance systems. Taking the data source used in this study, the FY-4A satellite, as an example, it can observe data on a minute-by-minute basis and generate data within approximately one minute, with the day's observational data made publicly available on the official website within 24 h. Insurance companies may consider establishing operational collaborations with meteorological centers to obtain customized data in a more timely and accurate manner, thereby integrating it into their business processes and reducing the workload of model data preprocessing. Relevant experiments have shown that in model comparisons processing financial data at large scales, it takes random forests 47.3 s to train, while the prediction latency remains at a reasonable 12.4 milliseconds. This makes it suitable for batch processing applications [56]. This satisfies the computational efficiency requirements for insurance applications that settle claims on a monthly basis.

However, due to the limitations of the temporal and spatial resolution of satellite data and the impact of base risk, future research will integrate summer season data, incorporate higher-resolution observational data and combine reinsurance transfer mechanisms and surveys of policyholders' willingness to pay to optimize the pricing strategies and promotion plans for insurance products. Furthermore, due to the temporal granularity of publicly available data, it is not yet possible to correlate the monthly lightning-related losses, insurance claims, and other real-world impact indicators for the corresponding region with the risk index. In future work, this study will attempt to collect or reasonably simulate disaster loss data so as to further validate the rationality of the lightning risk index. Moreover, with the development of blockchain and smart contract technologies, implementing a claim settlement model based on daily claims may be considered in the future [57].

Author Contributions: Conceptualization, G.H., Y.C. and S.J.; methodology, G.H., M.X., Y.C. and S.J.; software, G.H.; validation, Y.C., S.J. and M.X.; formal analysis, Y.C. and M.X.; investigation, G.H.; resources, S.J.; data curation, G.H.; writing—original draft preparation, G.H.; writing—review and editing, S.J. and M.X.; visualization, G.H.; supervision, M.X.; project administration, S.J.; funding acquisition, S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data and code used in this work are available at https://github.com/GuanhuaH/Lightning_risk_insurance_2026, accessed on 14 May 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principal Component Analysis
RF	Random Forests

SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
OOB	Out-of-Bag
POT	Peaks Over Threshold
MEF	Mean Excess Function

References

- China Meteorological Administration. *China Meteorological Disaster Yearbook (2024)*; China Meteorological Press: Beijing, China, 2025.
- Tadesse, M.A.; Shiferaw, B.A.; Erenstein, O. Weather index insurance for managing drought risk in smallholder agriculture: Lessons and policy implications for sub-Saharan Africa. *Agric. Food Econ.* **2015**, *3*, 26. [\[CrossRef\]](#)
- Banerjee, C.; Berg, E. Efficiency of wind indexed typhoon insurance for rice. In Proceedings of the EAAE 2011 Congress: Change and Uncertainty—Challenges for Agriculture, Food and Natural Resources, Zurich, Switzerland, 30 August–1 September 2011. [\[CrossRef\]](#)
- Thomas, M.; Tellman, E.; Osgood, D.E.; DeVries, B.; Islam, A.S.; Steckler, M.S.; Goodman, M.; Billah, M. A Framework to Assess Remote Sensing Algorithms for Satellite-Based Flood Index Insurance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 2589–2604. [\[CrossRef\]](#)
- Khalil, A.F.; Kwon, H.; Lall, U.; Miranda, M.J.; Skees, J. El Niño–Southern Oscillation–based index insurance for floods: Statistical risk analyses and application to Peru. *Water Resour. Res.* **2007**, *43*, 2006WR005281. [\[CrossRef\]](#)
- Yu, S.; Ren, Y. Research on the lightning risk assessment method for Chongqing based on fuzzy mathematics. In Proceedings of the 2014 International Conference on Lightning Protection (ICLP), Shanghai, China, 11–18 October 2014; pp. 1054–1057. [\[CrossRef\]](#)
- Liu, S.; Zhu, C.; Yin, H.; Qin, K.; Lin, H.; Huang, J.; Xia, M.; Weng, L. GLMamba: A Global–Local Mamba Network for Efficient Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 11344–11360. [\[CrossRef\]](#)
- Ni, Y.; Liu, S.; Guo, T.; Xia, M. TiBT-Net: A High-Resolution Remote Sensing Image Change Detection Network Integrating Bi-Temporal Space Enhancement and Token Interaction. *Remote Sens.* **2026**, *18*, 805. [\[CrossRef\]](#)
- Yin, H.; Wang, J.; Liu, S.; Wang, Y.; Liu, Y.; Guo, T.; Xia, M. MISA-Net: Multi-Scale Interaction and Supervised Attention Network for Remote-Sensing Image Change Detection. *Remote Sens.* **2026**, *18*, 376. [\[CrossRef\]](#)
- Ren, Z.; Weng, L.; Xia, M.; Lin, H. MCINet: Multi-attentive cross-level interaction network for cloud and snow segmentation. *J. Appl. Remote Sens.* **2026**, *20*, 021404. [\[CrossRef\]](#)
- Lu, A.; Wang, J.; Guo, T.; Wang, Z.; Xia, M. LECloud: Efficient Cloud and Cloud-Shadow Segmentation Based on Windowed State Space Model and Lightweight Attention Mechanism. *Remote Sens.* **2026**, *18*, 1341. [\[CrossRef\]](#)
- Wang, Y.; Zhang, Y.; Weng, L.; Lin, H.; Xia, M. CFR-Net: A Coarse-to-Fine Concatenated Dual Decoder Based on Frequency-Space-Structure Fusion for Robust Remote Sensing Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 19442–19458. [\[CrossRef\]](#)
- Abdi, H.; Williams, L.J. hPrincipal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [\[CrossRef\]](#)
- Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
- Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 1–10.
- Lundberg, S.M.; Lee, S.I. Consistent Feature Attribution for Tree Ensembles. In Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017), Sydney, NSW, Australia, 10 August 2017.
- Marcilio, W.E.; Eler, D.M. From explanations to feature selection: Assessing SHAP values as feature selection mechanism. In Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Recife/Porto de Galinhas, Brazil, 7–10 November 2020; pp. 340–347. [\[CrossRef\]](#)
- Parhamfar, M. Lightning Risk Assessment Software Design for Photovoltaic Plants in Accordance with IEC 62305-2. *Energy Syst. Res.* **2022**, *5*, 34–54. [\[CrossRef\]](#)
- Han, B.; Ming, Z.; Zhao, Y.; Wen, T.; Xie, M. Comprehensive risk assessment of transmission lines affected by multi-meteorological disasters based on fuzzy analytic hierarchy process. *Int. J. Electr. Power Energy Syst.* **2021**, *133*, 107190. [\[CrossRef\]](#)
- Gallejo, L.E.; Duarte, O.; Torres, H.; Vargas, M.; Montaña, J.; Pérez, E.; Herrera, J.; Younes, C. Lightning risk assessment using fuzzy logic. *J. Electrostat.* **2004**, *60*, 233–239. [\[CrossRef\]](#)
- Murphy, K.M.; Bruning, E.C.; Schultz, C.J.; Vanos, J.K. A Spatiotemporal Lightning Risk Assessment Using Lightning Mapping Data. *Weather. Clim. Soc.* **2021**, *13*, 571–589. [\[CrossRef\]](#)
- Mahdariza, F. The Determination of Lightning Disaster Hazard Index Using Analytical Hierarchy Process. *Elkawanie* **2017**, *3*, 233–238. [\[CrossRef\]](#)
- Montanya, J.; Bergas, J.; Hermoso, B. Electric field measurements at ground level as a basis for lightning hazard warning. *J. Electrostat.* **2004**, *60*, 241–246. [\[CrossRef\]](#)

24. Thomas, A.M.; Noble, S. A physics-based ensemble machine-learning approach to identifying a relationship between lightning indices and binary lightning hazard. *Front. Earth Sci.* **2024**, *12*, 1376605. [[CrossRef](#)]
25. Sheng, J.; Xu, M.; Han, J.; Deng, X. A Lightning Disaster Risk Assessment Model Based on SVM. *J. Big Data* **2021**, *3*, 183–190. [[CrossRef](#)]
26. Zhao, C.; Peng, R.; Wu, D. Bagging and Boosting Fine-Tuning for Ensemble Learning. *IEEE Trans. Artif. Intell.* **2024**, *5*, 1728–1742. [[CrossRef](#)]
27. Li, J.; Zhu, C.; Dong, Y.; Xia, M. Fault Prediction Method of Boost Converter Based on Multi-Modal Components and Temporal Convolutional Networks. *Energies* **2026**, *19*, 1974. [[CrossRef](#)]
28. Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; Wortman Vaughan, J. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14. [[CrossRef](#)]
29. Guo, Z.; Lin, G.; He, Q.; Chang, Y.; Zhu, Y.; Wang, Z.; Xu, Y.; Cao, J. Regional lightning risk assessment based on fuzzy comprehensive evaluation method. In Proceedings of the 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 10–12 August 2010; pp. 1340–1343. [[CrossRef](#)]
30. Assa, H.; Liu, P.; Wang, S. (Eds.) *Quantitative Risk Management in Agricultural Business*; Springer Actuarial; Springer Nature: Cham, Switzerland, 2025. [[CrossRef](#)]
31. Jerry S.; Jason H.; Anne M. *Designing Agricultural Index Insurance in Developing Countries: A GlobalAgRisk Market Development Model Handbook for Policy and Decision Makers*; Global AgRisk: Moose Jaw, SK, Canada, 2009.
32. Collier, B.; Skees, J.; Barnett, B. Weather Index Insurance and Climate Change: Opportunities and Challenges in Lower Income Countries. *Geneva Pap. Risk Insur.-Issues Pract.* **2009**, *34*, 401–424. [[CrossRef](#)]
33. Erec Heimfarth, L.; Musshoff, O. Weather index-based insurances for farmers in the North China Plain: An analysis of risk reduction potential and basis risk. *Agric. Financ. Rev.* **2011**, *71*, 218–239. [[CrossRef](#)]
34. Erec Heimfarth, L.; Finger, R.; Musshoff, O. Hedging weather risk on aggregated and individual farm-level: Pitfalls of aggregation biases on the evaluation of weather index-based insurance. *Agric. Financ. Rev.* **2012**, *72*, 471–487. [[CrossRef](#)]
35. Shibabaw, A.; Berhane, T.; Awgichew, G.; Walelgn, A.; Muhamed, A.A. Hedging the Effect of Climate Change on Crop Yields by Pricing Weather Index Insurance Based on Temperature. *Earth Syst. Environ.* **2023**, *7*, 211–221. [[CrossRef](#)]
36. Chen, W.; Hohl, R.; Tiong, L.K. Rainfall index insurance for corn farmers in Shandong based on high-resolution weather and yield data. *Agric. Financ. Rev.* **2017**, *77*, 337–354. [[CrossRef](#)]
37. Hohl, R.; Jiang, Z.; Tue Vu, M.; Vijayaraghavan, S.; Liong, S.Y. Using a regional climate model to develop index-based drought insurance for sovereign disaster risk transfer. *Agric. Financ. Rev.* **2021**, *81*, 151–168. [[CrossRef](#)]
38. Bokusheva, R. Using copulas for rating weather index insurance contracts. *J. Appl. Stat.* **2018**, *45*, 2328–2356. [[CrossRef](#)]
39. Wang, Q. The POT model described by the generalized Pareto distribution with Poisson arrival rate. *J. Hydrol.* **1991**, *129*, 263–280. [[CrossRef](#)]
40. Falk, M. On testing the extreme value index via the pot-method. *Ann. Stat.* **1995**, *23*, 2013–2035. [[CrossRef](#)]
41. Takane, Y. Relationships among Various Kinds of Eigenvalue and Singular Value Decompositions. In *New Developments in Psychometrics*; Yanai, H., Okada, A., Shigemasu, K., Kano, Y., Meulman, J.J., Eds.; Springer: Tokyo, Japan, 2003; pp. 45–56. [[CrossRef](#)]
42. Li, H.; Xu, X.; Xia, Q.; Xia, M. Multiple-constraints exploration and prototype-guided noise identification for semi-supervised medical image segmentation. *Biomed. Signal Process. Control* **2026**, *120*, 110004. [[CrossRef](#)]
43. Běhounek, L.; Cintula, P. Fuzzy class theory. *Fuzzy Sets Syst.* **2005**, *154*, 34–55. [[CrossRef](#)]
44. QX/T 85-2018; Technical Specification for Lightning Disaster Risk Assessment. China Meteorological Administration Press: Beijing, China, 2018.
45. Wijesena, S.; Pradhan, B. Advancements in Weather Index Insurance: A Review of Data-Driven Approaches to Design, Pricing and Risk Management. *Earth Syst. Environ.* **2025**, *9*, 2355–2379. [[CrossRef](#)]
46. Van Montfort, M.A.J.; Witter, J.V. The Generalized Pareto distribution applied to rainfall depths. *Hydrol. Sci. J.* **1986**, *31*, 151–162. [[CrossRef](#)]
47. Tan, C.S.; Gupta, A.; Ong, Y.S.; Pratama, M.; Tan, P.S.; Lam, S.K. Pareto optimization with small data by learning across common objective spaces. *Sci. Rep.* **2023**, *13*, 7842. [[CrossRef](#)] [[PubMed](#)]
48. Chukwudum, Q.C.; Mwita, P.; Mung'atu, J.K. Optimal threshold determination based on the mean excess plot. *Commun. Stat.-Theory Methods* **2020**, *49*, 5948–5963. [[CrossRef](#)]
49. Smith, R.L. Statistics of Extremes, with Applications in Environment, Insurance, and Finance. In *Extreme Values in Finance, Telecommunications, and the Environment*, 1st ed.; Chapman and Hall/CRC: New York, NY, USA, 2003; pp. 20–97.
50. Drees, H.; Resnick, S.; De Haan, L. How to make a Hill plot. *Ann. Stat.* **2000**, *28*, 254–274. [[CrossRef](#)]
51. Mosala, R.; Rachuene, K.A.; Shongwe, S.C. Most suitable threshold method for extremes in financial data with different volatility levels. *ITM Web Conf.* **2024**, *67*, 01033. [[CrossRef](#)]

52. Hochrainer-Stigler, S.; Reiter, K. Risk-Layering for Indirect Effects. *Int. J. Disaster Risk Sci.* **2021**, *12*, 770–778. [[CrossRef](#)]
53. Benth, F.E.; Šaltytė-Benth, J. Stochastic Modelling of Temperature Variations with a View Towards Weather Derivatives. *Appl. Math. Financ.* **2005**, *12*, 53–85. [[CrossRef](#)]
54. Shreve, S.E. *Stochastic Calculus for Finance. 2: Continuous-Time Models*, corr. print ed.; Springer Finance Textbook; Springer: New York, NY, USA, 2004; Volume 2.
55. Xanthopoulos, S. Martingale Pricing and Single Index Models: Unified Approach with Esscher and Minimal Relative Entropy Measures. *J. Risk Financ. Manag.* **2024**, *17*, 446. [[CrossRef](#)]
56. Xu, S.; Zhou, Y. Machine Learning Applications in Financial Statement Fraud Detection: A Comparative Analysis. *Ann. Appl. Sci.* **2025**, *6*, 1–23.
57. Li, J.; Peng, Q.; Wu, D.; Sun, Y.; Zhao, W. Lightning Insurance: A Fast Claim, High Accuracy Insurance Platform Based on Blockchain Technology and NASNET Algorithm. In Proceedings of the 2021 International Conference on Artificial Intelligence and Blockchain Technology (AIBT), Beijing, China, 4–6 December 2021; pp. 101–108. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.