

Article

New Method for Single-Site Cloud-to-Ground Lightning Location Based on Tri-Pre Processing

Bingzhe Dai ¹, Qilin Zhang ^{1,*}, Jie Li ¹, Yi Liu ^{2,3} and Minhong Zhao ¹

- ¹ Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Key Laboratory of Meteorological Disaster, Ministry of Education (KLME), Research Institute of Intelligent-Sensing and Disaster Prevention for Extreme Weather, Nanjing University of Information Science and Technology, Nanjing 210044, China; 20201106001@nuist.edu.cn (B.D.); 20221106001@nuist.edu.cn (J.L.); 202412492016@nuist.edu.cn (M.Z.)
- ² Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing 210044, China; rk226404@reading.ac.uk
- ³ Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK;
- * Correspondence: qlzhang@nuist.edu.cn

Abstract: The single-site lightning detection system can provide timely and effective information on lightning activity in areas where a multi-site lightning network cannot be built. Using deep learning, the single-site lightning detection achieves better performance than traditional methods, but it is highly dependent on the quality of the training dataset. To address this, this paper proposes a method called Tri-Pre to improve dataset quality and thereby enhance the performance of single-site cloud-to-ground lightning detection based on deep learning. After using the Tri-Pre method, the location model's distance estimation error decreases by 36.08%. For lightning with propagation distances greater than 1000 km, the average relative error of the results from the built model based on the Tri-Pre method is 3.78%. When verified using additional measured data, the model also shows satisfactory accuracy, particularly for lightning with propagation distances beyond 1000 km. Specifically, for lightning with propagation distances between 1500 and 1600 km, the average relative location error is approximately 5.46%.

Keywords: single-site lightning location; preprocessing method; lightning waveform; deep learning



Academic Editor: Marios Anagnostou

Received: 17 April 2025

Revised: 14 May 2025

Accepted: 16 May 2025

Published: 19 May 2025

Citation: Dai, B.; Zhang, Q.; Li, J.; Liu, Y.; Zhao, M. New Method for Single-Site Cloud-to-Ground Lightning Location Based on Tri-Pre Processing. *Remote Sens.* **2025**, *17*, 1766. <https://doi.org/10.3390/rs17101766>

Copyright: © 2025 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) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lightning occurs continuously on Earth, emitting a large amount of electromagnetic energy with a frequency range from below 1 Hz to nearly 300 MHz [1]. Among these, the VLF (Very Low Frequency) LEMP (Lightning Electromagnetic Pulse) signal, ranging from 3 to 30 kHz, can propagate thousands of kilometers in the Earth-ionosphere waveguide [2,3]. Real-time lightning location provides important information for mitigating the hazards caused by lightning [4] and related severe weather [5,6]. Lightning location results provided by detection systems make it possible to effectively nowcast thunderstorms [7,8], thereby helping to reduce disaster-related losses. This highlights the importance of developing effective lightning location methods.

Although multi-site lightning location networks using the TDOA (Time Difference of Arrival) method, such as NLDN (National Lightning Detection Network) [9] and GLD360 (Global Lightning Detection Network) [10], have high location accuracy, these networks require a large number of stations with identical configurations (≥ 4 stations), and each

station requires a high-precision timing system and a low-latency network environment. This makes construction challenging and operational cost high, making them unsuitable for certain regions. However, single-site location does not require as many stations as the multi-site lightning location, nor does it rely on high-precision timing systems or an ideal network environment, thereby compensating for the limitations of multi-site lightning location networks.

The single-site lightning location can be divided into two components: direction and distance estimation. Direction finding relies on the principle of the MDF (Magnetic Direction-Finding) method [11–13]. In traditional approaches, a rigorous algorithm for identifying the reliable peak values is also required. However, due to the influence of environmental noise, the non-ideal orthogonality of the antenna, deviations in antenna orientation, and variations in antenna circuits, the recorded waveform peaks may be offset in both amplitude and position, making the design of a reliable algorithm a significant challenge. Meanwhile, there is currently no deep learning model specifically designed for lightning direction estimation.

In terms of distance estimation, there are also some limitations in its application. The method proposed by Koochak and Fraser-Smith [14], based on the different arrival times of ELF (Extremely Low Frequency) and VLF radiation components, requires the simultaneous recording of both. Although the distance estimation accuracy is better than 6.7%, this method is only applicable to lightning with a propagation distance of more than 1000 km. The surface impedance method proposed by Chen et al. [15] is only effective in locating close-range ground flashes within the range of 15–60 km in the VLF/LF (Very Low Frequency/Low Frequency) frequency band. The Kharkov method proposed by Rafalsky [16,17], using phase spectra, requires two crossed magnetic antennas and an electric antenna sensitive to 1.8–3.2 kHz for observation. However, due to the need for multiple calculations and fitting, its real-time performance is poor. With the further development of AI (Artificial Intelligence), relatively efficient and accessible AI methods have been introduced into distance estimation. Compared with traditional non-AI methods that rely on manually constructed equations and expert-defined relationships, data-driven artificial intelligence approaches can automatically learn complex patterns from the data. Typically, as long as there is a sufficient amount of high-quality data, AI models can effectively learn the underlying patterns and produce reliable results. This eliminates the time-consuming and intricate process of manually accounting for numerous influencing factors and their interactions when constructing equations. As a result, development time is significantly reduced, and the method becomes more convenient and easier to implement. de Sá et al. [18] used machine learning to estimate the distance of lightning within a propagation range of 100–700 km in the LF (Low Frequency) band. However, it still necessitates manual feature selection and design, which poses a significant challenge in the field of lightning detection. Wang et al. [19] further employed deep learning to predict lightning distance in the VLF/LF band, mitigating the problem and achieving significant results within a range of approximately 1000 km.

Although AI methods are more convenient and easier to implement, their accuracy is highly dependent on the quality of the dataset [20]. If the dataset contains label errors, the prediction accuracy of the model will be significantly reduced. For detection instruments with wide coverage, electromagnetic waves generated by lightning at different locations may be received simultaneously within a short period of time. This implies that, when performing multi-site lightning location, some of the waveforms used may actually originate from different lightning events. Consequently, if the multi-site location results (used as labels) and the observed waveforms from one of the stations in the multi-site are used to form the dataset for training single-site lightning location model, mismatches may occur

between the waveforms and their corresponding labels—these are referred to as label errors. However, effective methods for eliminating these errors and improving the quality of the dataset are still lacking.

Therefore, developing effective data preprocessing techniques to construct high-quality datasets is a crucial challenge when applying AI methods to single-site lightning location. Due to the diversity of lightning sources and the nonlinear variation of lightning waveforms with propagation distance, this problem is substantially different from those involving artificial signals with known intensities and frequencies. In such cases, it is difficult to manually define suitable features for filtering. Moreover, the large volume of data necessitates the development of automated methods.

To address these challenges, this paper proposes a method called Tri-Pre to perform effective preprocessing of single-site lightning location datasets. In addition, deep learning is employed for direction estimation component in the lightning location model, thereby avoiding the need to design complex peak-search algorithms. Finally, we evaluate the impact of the Tri-Pre method on the model outputs and verify the effectiveness of the proposed model when combined with the Tri-Pre method.

2. Instrumentation and Data Source

The data used in this study come from stations with complete single-site lightning detection equipment within the VLF-LLN (Very Low Frequency Long-range Lightning Location Network) [21], as well as the location results provided by the VLF-LLN. These stations, used for single-site lightning location, are equipped with electric field sensors with whip antennas, magnetic field sensors with orthogonal magnetic loop antennas, and data recording devices. The detected electric field is used to determine the polarity of lightning, which helps eliminate ambiguity when performing direction-finding based on magnetic measurements. In this study, the station located in Nanjing, China, was chosen. Since the instrument has a detection range of 3000 km, the area covered by the study includes most of China. The signal frequency band recorded is in the VLF range. The level triggering method is employed to record the horizontal magnetic field component signals and the vertical electric field signals simultaneously. The analog-to-digital conversion has a resolution of 16 bits, and the sampling frequency is 1 MSPS (Million Samples Per Second), with a sampling time of 1 ms and a pre-trigger time of 300 μ s. Figure 1 presents the recorded electric and magnetic field waveforms of a cloud-to-ground lightning stroke.

According to the precise location results of the VLF-LLN, each waveform can be labeled with distance and direction information to create the single-site lightning location database. We collected the data from June to July 2024 for this study. This is because summer is the peak period for lightning activity in China, which can provide sufficient valid data. Meanwhile, considering computer performance and the limited time for data collection, only the data from these two months were used. Both the raw electric and magnetic field waveforms were processed using a zero-phase band-pass filter with a frequency range of 2–30 kHz to improve the signal-to-noise ratio. Additionally, all electric field waveforms were normalized by their maximum absolute value to reduce the influence of varying intensities of different lightning events with the same propagation distance. The two magnetic field waveforms were not normalized, as direction finding relies on the amplitude ratio between them. Among the data used, the number of LEMP waveform samples with a propagation distance of 100–2800 km accounts for approximately 96.5% of the total, indicating that the effective detection and location range of the LEMP system in this study exceeds 1000 km.

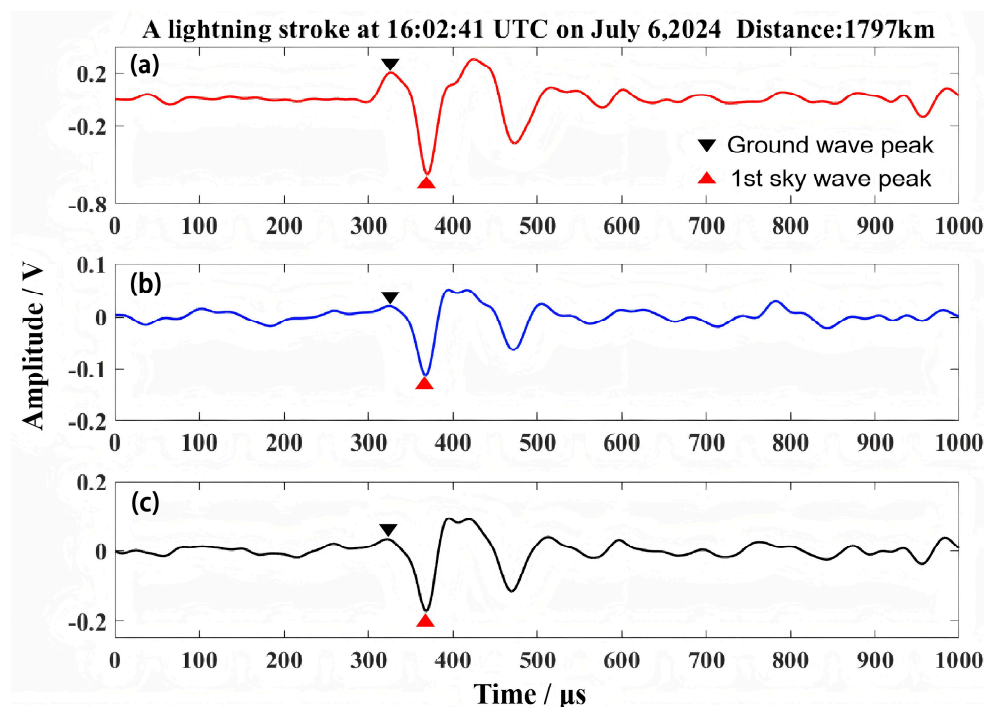


Figure 1. The electric and magnetic field signals measured of a return stroke. (a) Electric field signal, (b) North-south magnetic signal, and (c) East-west magnetic signal.

3. Methods

3.1. Tri-Pre Method

As mentioned in the introduction, there is a need for a method that can preprocess the dataset to improve its quality and enhance the performance of the single-site lightning location model. Considering the complex, nonlinear variations of lightning signals with propagation distance [3], filtering out erroneous data based on similarity becomes a more effective solution. However, the commonly used cross-correlation is a similarity measurement method based on linear operations. As a result, it cannot capture nonlinear relationships and is easily affected by noise, which limits its ability to measure the similarity of lightning signals. In contrast, the triplet network [22] can capture nonlinear similarity relationships and learn feature representations that effectively separate different sample categories. The waveform features it automatically extracts can be directly used to distinguish different categories. Moreover, the network can be effectively trained with only a small number of labeled samples. Therefore, this study adopts the triplet network for similarity-based processing.

The implementation process of the Tri-Pre method is illustrated in Figure 2. The core of this method lies in obtaining a waveform feature extractor and deriving typical features for each category. First, a few-shot dataset comprising 30 categories is constructed by selecting waveforms from observational data, guided by the simulated waveform bank and propagation velocity constraints. The few-shot dataset contains only a small number of samples per category, which are used for training models with limited data. Then, it is used to train a triplet network. Subsequently, the feature extraction module is separated from the well-trained network. Using this feature extractor, the typical features of each category are computed by averaging the extracted features within the few-shot dataset. These typical features are used to judge whether new data meet the required criteria.

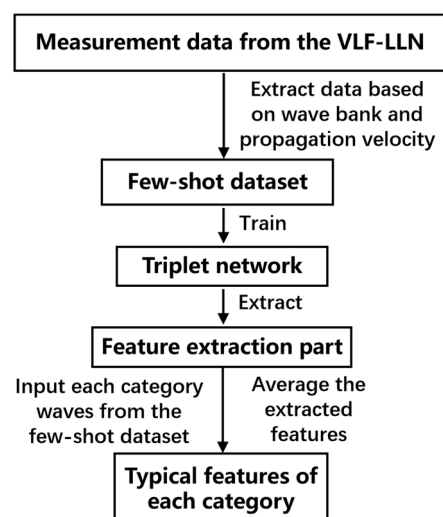


Figure 2. The implementation process of the Tri-Pre method.

To construct the few-shot dataset, waveforms from the simulated waveform bank and propagation speed constraints are used to ensure the accuracy and reliability of the extracted observational data. The waveform bank is based on the method proposed by Hou [23], which has been shown to effectively simulate lightning waveforms at various distances. The waveform bank contains 60 waveforms (30 for daytime and 30 for nighttime), corresponding to propagation distances from 0 to 3000 km in 100 km intervals. The main distinguishing elements between waveforms are the relative arrival timing and amplitudes of the ground wave, the first sky wave, and the second sky wave, as indicated by the red, black, and blue triangular markers, respectively, in Figure 3. When lightning occurs, the resulting electromagnetic waves radiate in all directions. The wave that propagates along the Earth's surface is referred to as the ground wave, while the wave that travels through the Earth-ionosphere waveguide is called the sky wave. The first sky wave is reflected once by the ionosphere, and the second sky wave is reflected twice. These waves follow different propagation paths, with the second sky wave traveling the longest distance and arriving last among the three. When lightning occurs near the observation station, the differences in propagation distances among the three waves are significant, resulting in large time intervals between their arrivals. As the distance between the lightning and the observation station increases, the lengths of the three propagation paths gradually converge, resulting in smaller time differences between them. Meanwhile, considering that the operating frequency band of the VLF-LLN instrument is essentially the same as that of the WWLLN (World Wide Lightning Location Network), and that waveforms within the same frequency band correspond to approximately the same group velocity, together with the proven effectiveness of the group velocity used by WWLLN for lightning localization, it is reasonable to assume that the propagation velocities provided by VLF-LLN falling within the WWLLN-adopted group velocity range ($0.9914c$ – $0.993c$, where c is the speed of light) correspond to valid lightning waveforms. Therefore, observation waveforms are selected if their main distinguishing elements align with those of simulated waveforms at the corresponding distances (i.e., the relative position deviation between the ground wave and the first sky wave is less than 10%, and the peak amplitude difference is also less than 10%) and their propagation speeds given by the VLF-LLN fall within the range of $0.9914c$ to $0.993c$.

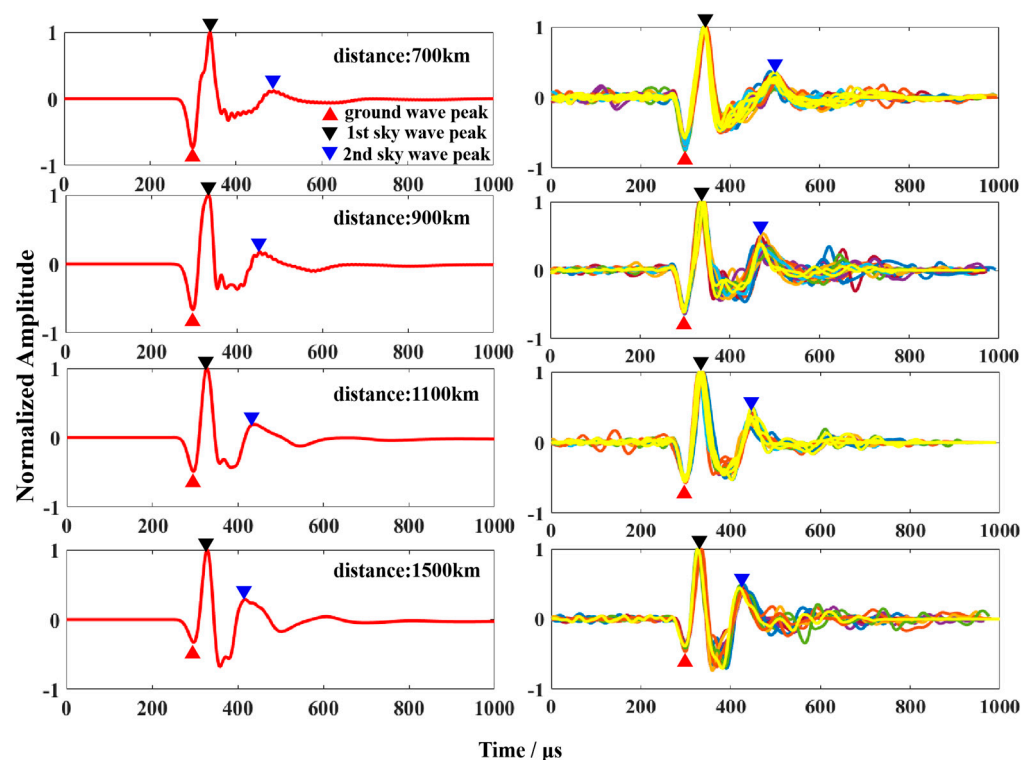


Figure 3. Examples of both the simulated waveforms in the wave bank and the observed waveforms selected according to selection criteria are shown. The left side displays waveforms from the wave-bank at specific distances in red, while the right side of each row shows overlays of 10 observed waveforms selected within ± 50 km of the corresponding distance on the left in different colors.

Figure 3 presents the simulated and observational waveforms selected according to the criteria mentioned above. The left side displays waveforms from the wave bank at specific distances, while the right side of each row shows overlays of 10 selected observational waveforms within ± 50 km of the distance on the left in different colors. The observed similarity among waveforms within the same distance range, along with the clear differences between different ranges, confirms the validity of the selection process. For each of the 30 categories, we selected 30 waveforms. To ensure polarity balance, all selected waveforms were augmented by polarity inversion, allowing the model to handle both positive and negative cloud-to-ground lightning. The few-shot dataset comprises 1800 waveforms, with 1200 for training and 600 for testing.

The process of training the triplet network using the constructed few-shot dataset is shown in Figure 4. The network is trained using triplets as input. Each triplet consists of an anchor and a positive sample from the same category and a negative sample from a different category. All three are passed through a shared feature extraction module to compute the triplet loss, which guides the network parameter updates. This process is repeated for each epoch until all waveforms are used to generate triplets. Table 1 presents the structure of the triplet feature extraction network, which is adapted from the feature extraction module used by Wang [24] with the addition of a dense layer at the end. The original module used by Wang has been shown to effectively extract lightning waveform features, and the added dense layer is designed to better suit the current application. In the network, the 1D convolutional layers extract local features from waveform data using sliding windows. The pooling layers are used for reducing computational complexity. Max pooling captures the most salient features by selecting local maxima, enhancing robustness to positional shifts, while average pooling computes local means, preserving overall feature distribution and smoothing noise. The activation layers apply nonlinear functions (e.g.,

ReLU function) to increase the network's expressiveness for modeling complex patterns. Dense (fully connected) layers aggregate global features via high-dimensional nonlinear combinations, typically used for final classification or regression outputs.

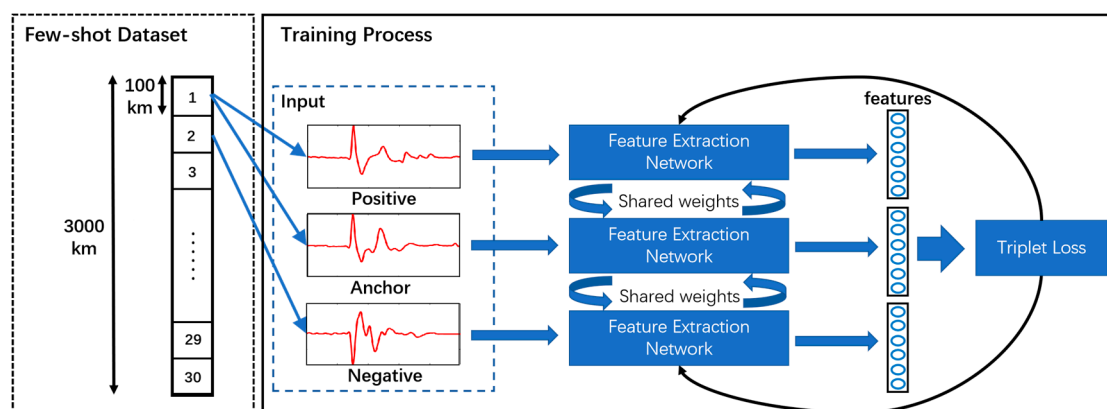


Figure 4. The training procedure of the triplet network. The few-shot dataset consists of 30 distance-based categories, with waveforms extracted from real observations according to the judgement criteria. Training is performed using triplets as input. Each triplet includes an anchor, a positive, and a negative: the anchor and positive are waveforms from the same category, while the negative is from a different category.

Table 1. Triplet feature extraction module composition.

	Layer	Filter Number	Kernel Size	Pooling Window Size	Padding	Stride	Activation Function	Output Shape
1	Input	⁵ /	/	/	/	/	/	(1000, 1)
2	¹ 1D-Conv	16	32	/	⁶ √	/	⁸ ReLU	(1000, 16)
3	1D-Conv	16	32	/	√	/	ReLU	(1000, 16)
4	² Max-Pooling	/	/	2	⁷ ×	2	/	(500, 16)
5	1D-Conv	32	32	/	√	/	ReLU	(500, 32)
6	1D-Conv	32	32	/	√	/	ReLU	(500, 32)
7	Max-Pooling	/	/	2	×	2	/	(250, 32)
8	1D-Conv	64	16	/	√	/	ReLU	(250, 64)
9	1D-Conv	64	16	/	√	/	ReLU	(250, 64)
10	Max-Pooling	/	/	2	×	2	/	(125, 64)
11	1D-Conv	128	8	/	√	/	ReLU	(125, 128)
12	1D-Conv	128	8	/	√	/	ReLU	(125, 128)
13	Max-Pooling	/	/	5	×	5	/	(25, 128)
14	1D-Conv	256	3	/	√	/	ReLU	(25, 256)
15	1D-Conv	256	3	/	√	/	ReLU	(25, 256)
16	³ Mean-Pooling	/	/	5	×	5	/	256
17	⁴ Dense	/	/	/	/	/	/	16

¹ 1D-Conv: one-dimensional convolutional layer; ² Max-Pooling: max pooling layer; ³ Mean-Pooling: average pooling layer; ⁴ Dense: fully connected layer; ⁵ / indicates not applicable; ⁶ √ / ⁷ × indicate whether the component is enabled; ⁸ ReLU: activation layer with ReLU activation function.

After training, the feature extractor can be used independently to obtain the 16-dimensional feature vector $(x_1, x_2, \dots, x_{16})$ for any given waveform. By averaging the extracted feature vectors of waveforms in each category of the few-shot dataset, a representative feature vector $(X_{n1}, X_{n2}, \dots, X_{n16})$ is computed for each of the 30 categories.

In practical use of the Tri-Pre method to verify whether a waveform correctly matches its label and removes incorrect data, the feature extractor is used to extract its feature vector $(y_1, y_2, \dots, y_{16})$, which is then compared with the representative features of each category

using Euclidean distance. The distance is calculated using Equation (1), where n denotes the category index and D represents the Euclidean distance.

$$D = \sqrt{\sum_{i=1}^{16} (X_{ni} - y_i)^2} \quad (1)$$

If the minimum value is below 1, the waveform is assigned to the category (distance range) corresponding to that value. If the assigned category does not match the labeled distance, the waveform is considered mismatched and is removed from the data. Figure 5 illustrates the complete verification procedure.

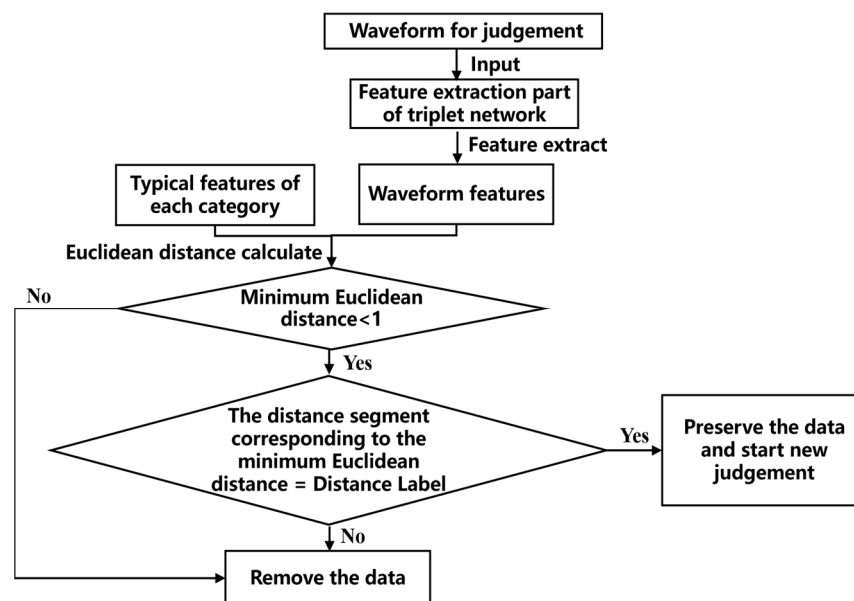


Figure 5. The mismatch judgement process of the Tri-Pre method.

3.2. Single-Site Lightning Location Model

The variation of lightning electromagnetic waveforms with propagation distance is a key factor enabling single-site lightning localization using deep learning. However, waveform characteristics also differ due to factors such as lightning type (intra-cloud or cloud-to-ground) and the diurnal variation of the ionospheric D-region [25]. Intra-cloud lightning generally contains more high-frequency components and exhibits a distinct bipolar symmetric structure, which is significantly different from the cloud-to-ground lightning. Additionally, due to the influence of solar radiation, the ionospheric D-region exists during the day but is nearly absent at night, resulting in VLF LEMPs that exhibit different characteristics between daytime and nighttime after propagating through the Earth-ionosphere waveguide. These influences cannot be directly captured by the feature extraction network through waveform analysis alone. Therefore, based on domain knowledge in lightning physics, they must be manually incorporated into the model as additional features or addressed through constraints applied before inputting to the model. We adopt the Tri-Pre method described in Section 3.1 to ensure that the data used for location corresponds to cloud-to-ground lightning, and we use arrival time as an additional input feature to represent the diurnal variation of the ionosphere. The arrival time is expressed as the number of seconds elapsed from 00:00 on the day to the signal trigger time, ranging from 0 to 86,400 s.

The single-site lightning location model is composed of the direction and distance estimation part. Both consist of a feature extraction network and a multi-feature fusion

network. The feature extraction networks they used have the same structure. Each layer of the networks may include various types of neural network structures, such as Fully Connected Networks, CNN (Convolutional Neural Networks) [26], RNN (Recurrent Neural Networks) [27], LSTM (Long Short-Term Memory networks) [28], ResNet (Residual Networks) [29], and others. The workflow of the direction and distance estimation components is as follows. First, the feature extraction network extracts features from each input waveform separately. Then, the extracted features, along with the manually added features based on lightning knowledge mentioned earlier, are integrated into the multi-feature fusion network using a 'branch-fusion' connection method to produce the output.

3.2.1. Feature Extraction Network

The structure of the feature extraction network is illustrated in Figure 6. The input to the network is either an electric field or a magnetic field waveform with a size of 1000×1 . Based on this structure, 16 features are extracted from each input waveform. The same network architecture is employed for both direction and distance estimation.

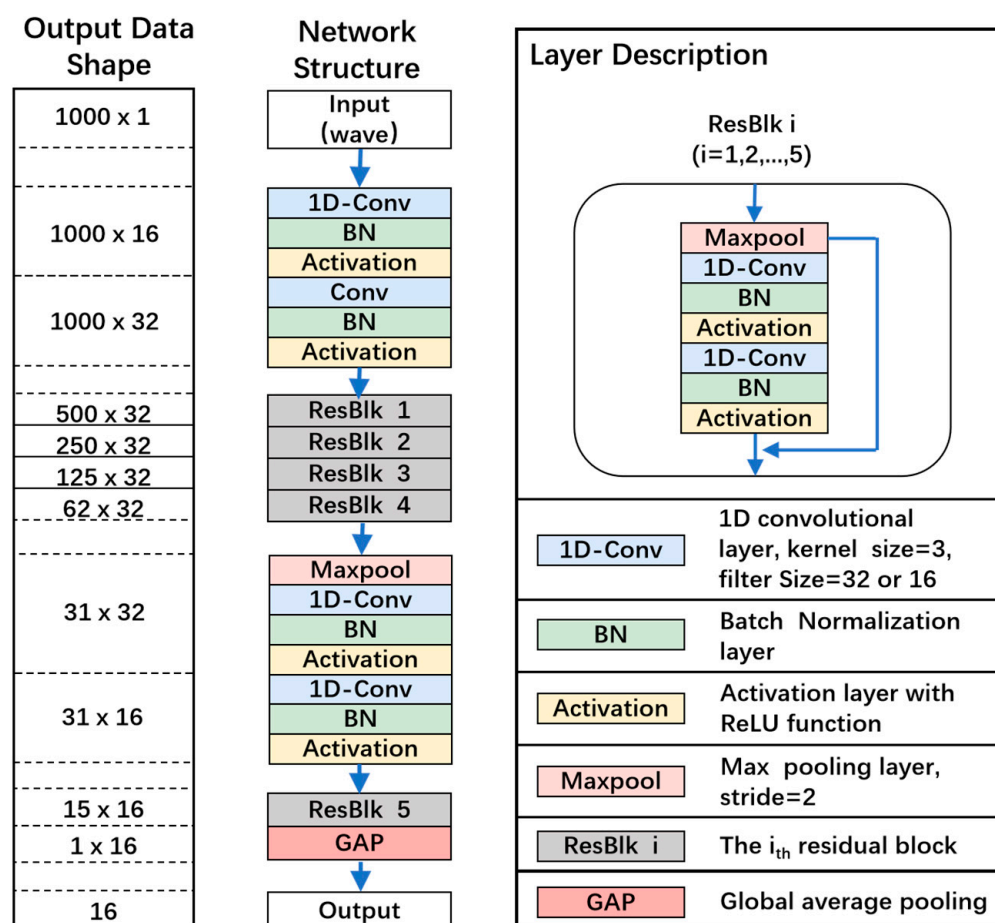


Figure 6. The details of the feature extraction network.

In the network, the BN (Batch Normalization) layers stabilize activation distributions, accelerate convergence, suppress gradient issues, and improve generalization. The residual blocks use skip connections to add inputs to convolution outputs element-wise, mitigating gradient vanishing/explosion and facilitating deep network training by learning residual mappings. Global average pooling reduces parameters and enhances spatial translation robustness by averaging each channel across its spatial dimensions.

3.2.2. Multi-Feature Fusion Network for Direction Estimation

The input to the multi-feature fusion network for direction estimation consists of features extracted from the electric field waveform, the north–south magnetic field waveform, and the east–west magnetic field waveform by the feature extraction network, along with the waveform arrival time. These inputs are fused into a 49×1 feature vector in the concatenation layer, which is then sequentially processed by six residual layers, one fully connected layer, and an activation layer to produce the direction estimation output. Figure 7 illustrates the structure of the multi-feature fusion network. To strictly constrain the output azimuth range to $(0, 360)$, the output values are the sine and cosine of the azimuth. By applying an inverse tangent conversion, the azimuth can subsequently be obtained. Additionally, a tanh activation layer is used before the output, and Leaky ReLU [30] is used as the activation function in the residual block, because it is more suitable than ReLU for tasks involving negative output values.

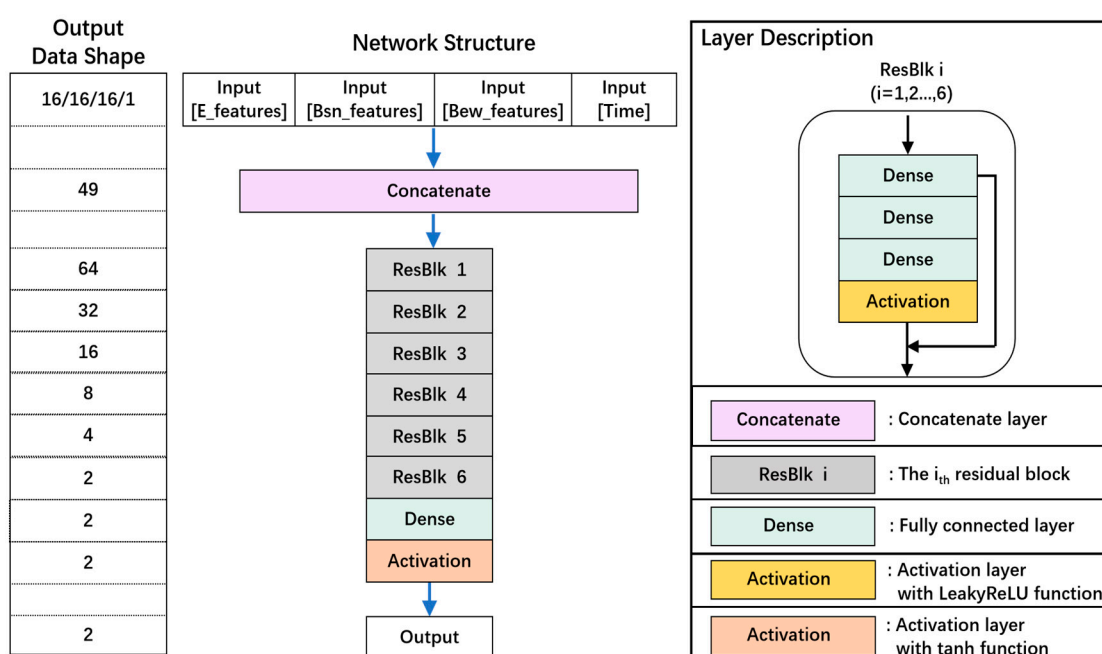


Figure 7. The details of the multi-feature fusion network for direction estimation.

3.2.3. Multi-Feature Fusion Network for Distance Estimation

The input to the multi-feature fusion network for distance estimation consists of the features extracted from the electric field waveforms by the feature extraction network, along with the waveform arrival time. These inputs are fused into a 17×1 feature vector by the concatenation layer, which is then sequentially processed by six residual layers, one fully connected layer, and an activation layer to produce the distance estimation output. Figure 8 illustrates the structure of this multi-feature fusion network. In the distance estimation task, only the normalized electric field signal is input during feature extraction. The reasons for only using the normalized electric field waveform are as follows. First, this approach helps to reduce the computational load and accelerates the training process. Second, although the intensity of the waveform decays with distance, the lightning intensity can vary significantly. Incorporating intensity information into the model may introduce unnecessary interference and result in large deviations in the distance estimation.

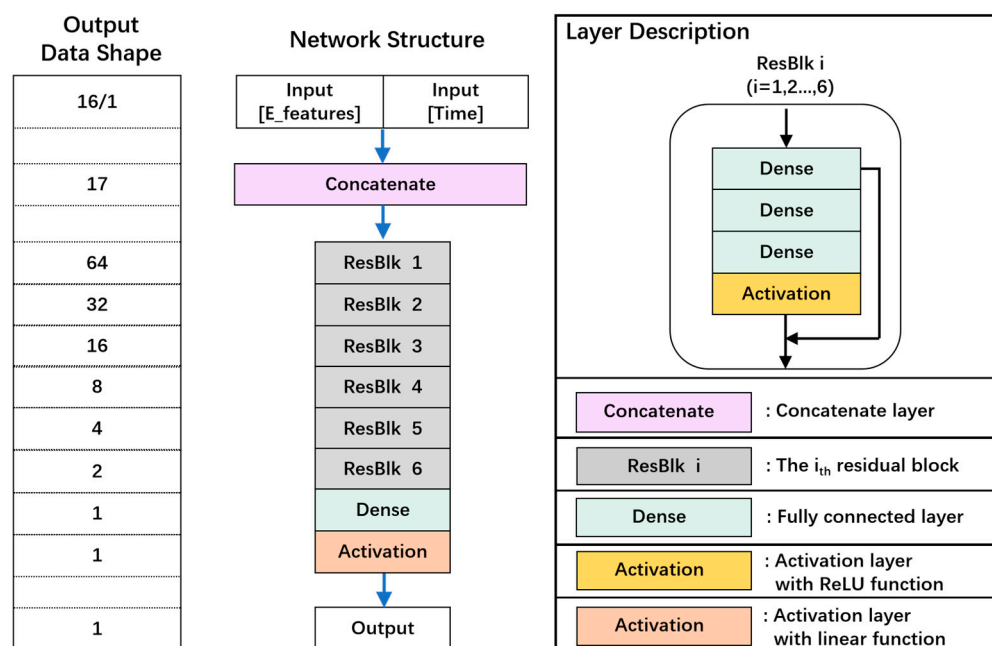


Figure 8. The details of the multi-feature fusion network for distance estimation.

3.3. Training Parameter Settings and Evaluation Metrics

During training, the batch size is set to 128, and the Adam optimizer [31] is used to minimize the loss. The initial learning rates for direction and distance estimation are set to 0.0001 and 0.001, respectively. If the loss on the test set does not decrease after five consecutive epochs, the learning rate will be automatically reduced by a factor of 2. Meanwhile, an early stopping strategy is adopted. When the loss on the test set does not decrease for ten consecutive epochs, training is halted. A training process that utilizes all the data in the training set is referred to as an epoch, and the maximum number of epochs is set to 120.

The performance of the model can be more effectively evaluated using two indicators: relative error (E_R) and absolute error (E_A). For the input data N , E_R , and E_A can be expressed by Equations (2) and (3), respectively, as follows:

$$E_R = \frac{1}{N} \sum_{n=1}^N \frac{|D_{true} - D_{pre}|}{D_{true}} \times 100\% \quad (2)$$

$$E_A = \frac{1}{N} \sum_{n=1}^N |D_{true} - D_{pre}| \quad (3)$$

In Equations (2) and (3), D_{true} is the actual distance from the lightning to the station, calculated based on the station's location and the lightning location determined by the VLF-LLN. D_{pre} is the distance between the lightning and the station estimated by the single-site lightning location model.

4. Results

In Section 4.1, we first demonstrate the effectiveness of the Tri-Pre method. In the subsequent two sections, we further evaluate the performance of the single-site lightning location model based on the Tri-Pre method.

4.1. Effectiveness of Tri-Pre Method

We illustrate the effectiveness of the Tri-Pre method by comparing the training results on the cross-correlation processing, Tri-Pre processing, and unprocessed datasets. The cross-

correlation processing is performed by computing normalized cross-correlation between the measured waveform and each of the waveform in the simulated wave bank (same as the wave bank mentioned in Section 3.1). If the maximum value of the 30 normalized cross-correlation values exceeds 0.6, and the distance corresponding to the maximum value matches the distance label of the waveform, the data are retained. To ensure the validity of the comparison, we ensure that the amount (298,135 in the Tri-Pre processed dataset) and distribution of the data across the three datasets are essentially the same. The training results on the three datasets are shown in Figure 9. It can be observed that the estimation performance of the distance is optimized after the triplet processing. As we can see in Figure 9, the result of triplet processing is the lowest in both the absolute error and the relative error. Compared with the unprocessed case, the distance estimation shows a reduction of 36.08% in absolute error and 59.93% in relative error. These prove the effectiveness of the Tri-Pre method.

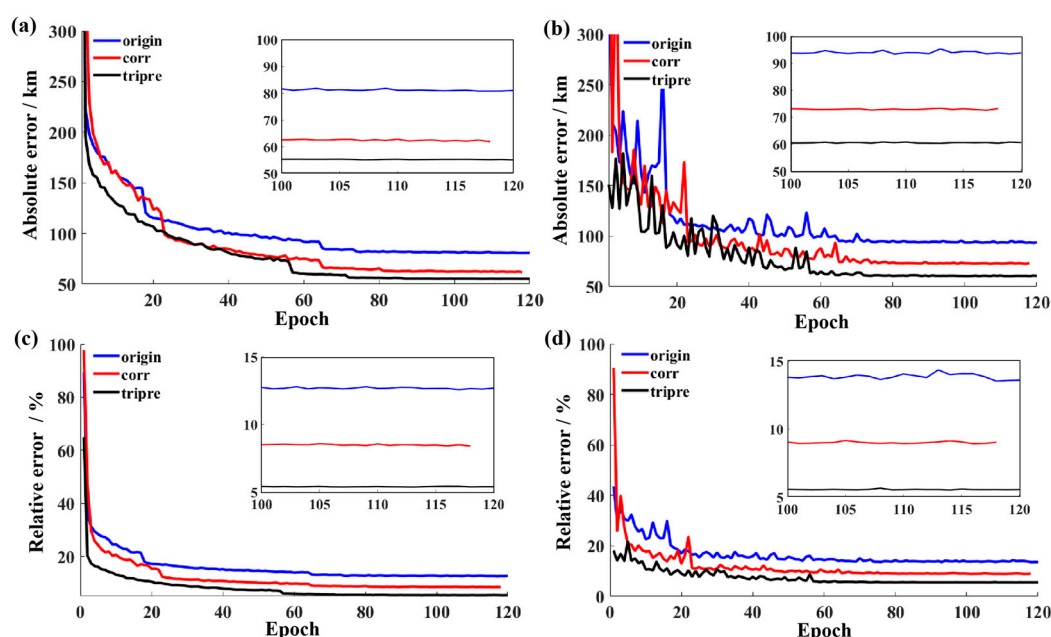


Figure 9. The relative and absolute errors of the models trained by different datasets: (a) absolute errors on the training dataset, (b) absolute errors on the test dataset, (c) relative errors on the training dataset, and (d) relative errors on the test dataset. ‘Origin’ refers to the unprocessed dataset, ‘corr’ refers to the dataset processed by cross-correlation, and ‘tripre’ refers to the dataset processed by Tri-Pre.

4.2. The Performance of Single-Site Lightning Location Model Based on Tri-Pre Method

According to the results in Section 4.1, the single-site lightning location model based on the Tri-Pre method demonstrates significantly improved performance compared to models using cross-correlation processing or unprocessed data. Therefore, starting from this section, we further analyze the performance of the model trained on the Tri-Pre-processed dataset. This analysis is conducted under the assumption that models performing well during training are also likely to perform effectively in real-world applications.

As shown in Figure 9a,b, the relative error of the lightning distance estimation can be controlled to within 5.6% based on the training data, with the optimal relative error being 5.37% (absolute error of 55.09 km). On the test dataset, the model also performed well, with the optimal relative error being 5.54% (absolute error of 60.39 km). The reason for the high absolute error is that, due to the large detection range of the equipment, data with a propagation distance exceeding 1000 km constitute a significant proportion of the entire dataset. As the propagation distance increases, the distance estimation error

inevitably increases, which leads to the overall large absolute error. If only the data with propagation distances below 1000 km are considered, the absolute error on the test set is 36.5 km. Additionally, for the data with propagation distances exceeding 1000 km, the optimal relative error is 3.78%.

For the direction estimation developed in this work, as shown in Figure 10, more than 96.5% of the direction estimation errors are concentrated in the interval of $[-10^\circ, 10^\circ]$. Although the current results are slightly worse than the azimuth results obtained by Wang et al. [19] using the traditional MDF method, our approach avoids the need to design complex algorithms to search for suitable peaks for calculation. Instead, it utilizes a data-driven approach, making the process simpler. Additionally, the current dataset is relatively small, and as more data accumulate, the direction estimation accuracy may improve further, potentially reaching or exceeding the results of the MDF method. Additionally, the proposed method avoids the issue in the MDF method that requires extensive data correction to account for the angle deviations caused by antenna orientation errors during the installation process.

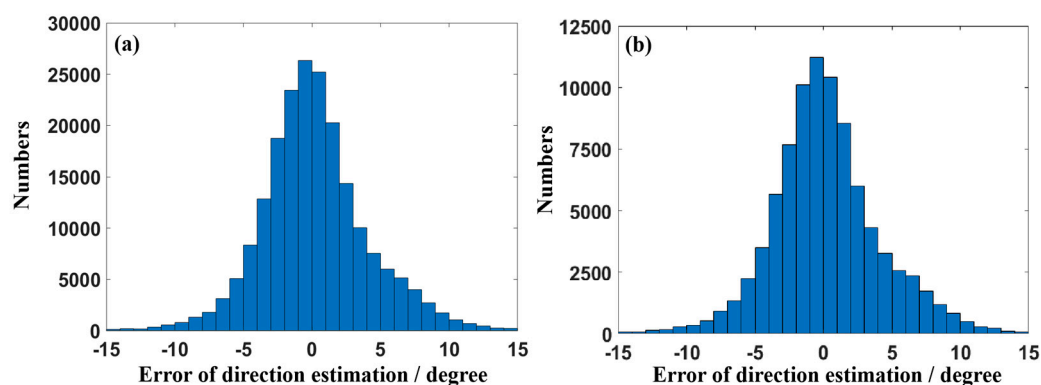


Figure 10. The statistical results of the direction estimation error on training and testing datasets. (a) direction estimation error on the training dataset, (b) direction estimation error on the testing dataset.

4.3. Measured Results

We employed additional data to evaluate the accuracy of the single-site lightning location model combined with the Tri-Pre method in this section. For each new waveform sample, we used the feature extraction network described in Section 3.1 to extract features and then calculate the Euclidean distances to the representative features of all categories. If none of the distances are less than 1, the current sample is discarded. This ensures that all waveform data used belong to cloud-to-ground lightning. The predicted distance and direction results were converted to longitude and latitude coordinates using the Haversine formula and then compared with the coordinates provided by the VLF-LLN. As lightning events typically occur in regions of strong convection with low CTT (Cloud-Top Temperature), we also use the CTT data from the FY4B (Fengyun 4B) satellite to validate the location results. Clearly, most lightning events are more likely to occur in strong convective regions with low CTT. If the lightning locations provided by the system also follow this pattern, it can be considered that the system is capable of effectively locating lightning in thunderstorms.

Figure 11 shows the location results of both the single-site and VLF-LLN overlaid on the CTT image. These lightning events occurred during thunderstorms within three different distance ranges. The colored areas represent the CTT image, where regions with lower values indicate more vigorous convection. The red ‘*’ marks indicate the location results of VLF-LLN, and the blue ‘+’ marks represent the single-site location results. It can be observed that the reference lightning location results provided by the VLF-LLN

are mostly concentrated in regions with lower CTT values, indicating that using these results as the standard is reasonable. Moreover, the single-site lightning location results are also distributed around the reference results, demonstrating that the current single-site lightning location method is effective in locating lightning events. Figure 11a shows the comparison of observed results for thunderstorms within the 300–500 km range. A total of 430 lightning events occurred from 06:00 to 06:15 UTC on 23 June 2024. The average relative error of the estimation provided by the single-site location model is 2.63% (absolute error: 25.97 km). Figure 11b shows the comparison of observed results for 210 lightning events occurring within the 1100–1200 km range from 05:45 to 06:00 UTC on 17 June 2024. The average relative error of the single-site location model is 4.84% (absolute error: 43.95 km). Figure 11c shows the comparative observation results for thunderstorms within the 1500–1600 km range, with a total of 186 lightning events occurring between 14:45 and 15:00 UTC on 14 June 2024, and an average relative error of 5.46% (absolute error: 47.85 km). At the same time, it can be observed that the location results are concentrated in areas with low CTT values, which are more likely to produce lightning. This illustrates the reliability of the results.

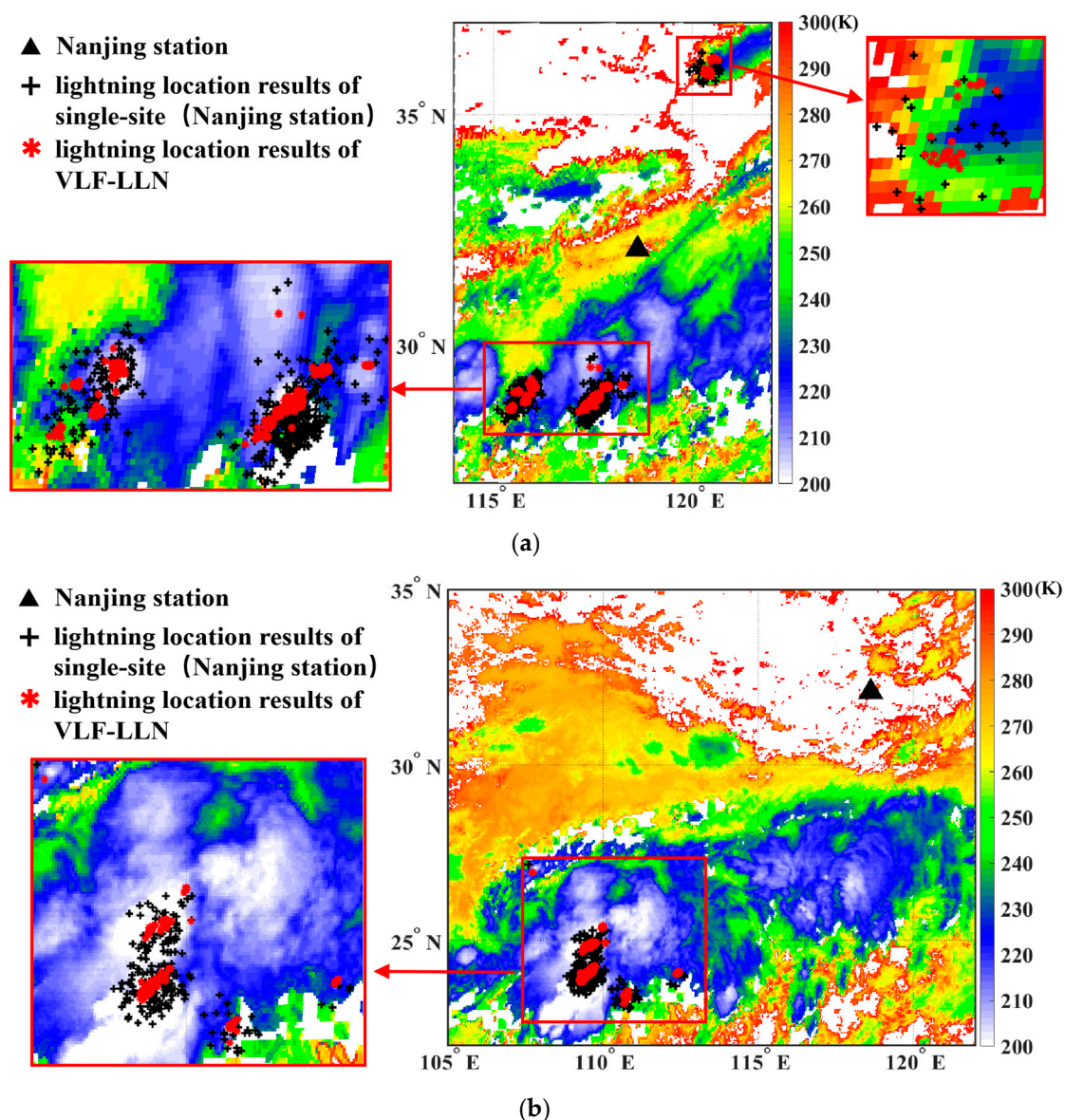


Figure 11. Cont.

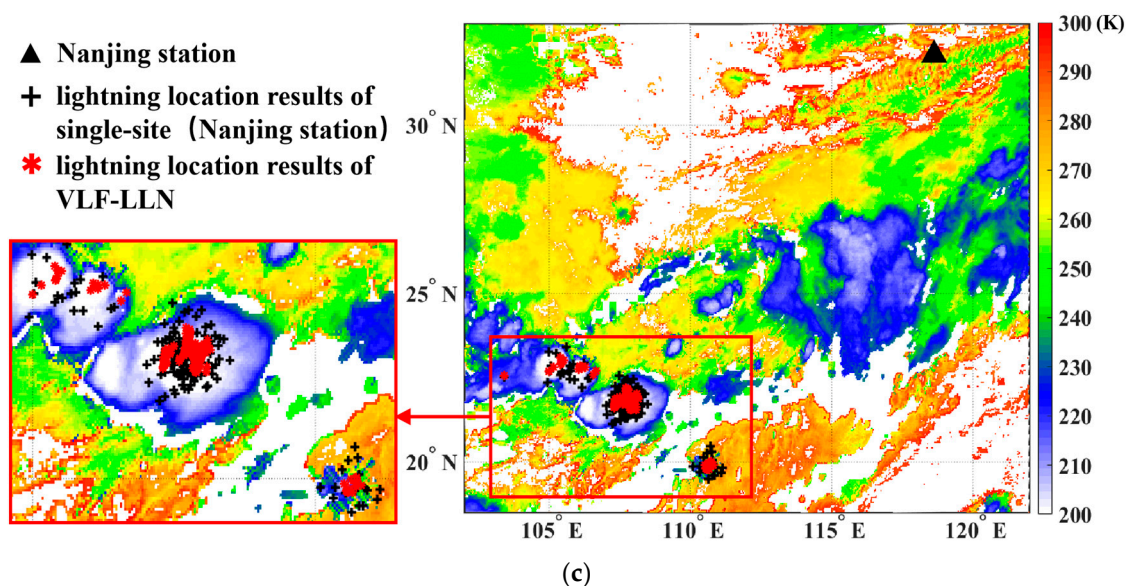


Figure 11. The comparison between the single-site and the VLF-LLN location results is presented. To better illustrate the validity of the results, they are overlaid on the CTT (Cloud-Top Temperature) image. The colored regions represent the CTT image, where lower values indicate more intense convection; lightning is more likely to occur in these areas. A zoomed-in view of the concentrated localization region is also provided. The corresponding time and the distance from the Nanjing station for each image are shown as follows: (a) Location results of lightning strikes during a thunderstorm occurring 300–400 km from Nanjing station between 06:00 and 06:15 UTC on 23 June 2024. (b) Location results of lightning strikes during a thunderstorm occurring 1100–1200 km from Nanjing station between 05:45 and 06:00 UTC on 17 June 2024. (c) Location results of lightning during a thunderstorm occurring 1500–1600 km from Nanjing station between 14:45 and 15:00 UTC on 14 June 2024.

It is evident that the directional error significantly impacts the location error, particularly when the lightning is far from the detection station. This issue could potentially be mitigated by incorporating additional high-quality data, obtained through Tri-Pre, to train the directional estimation model.

5. Discussion

In the process of using deep learning for single-station localization, ensuring the quality of the dataset is crucial for building an effective model. However, due to the potential presence of different source conditions in multi-station localization, mismatches may arise between the provided labels and the measured waveforms, leading to the inability to train a model that provides accurate results. Therefore, this paper proposes a Tri-Pre method to reduce such errors. The method does not require manual feature specification and significantly reduces manual processing costs. Additionally, the direction estimation component based on deep learning in the localization model not only avoids the complex peak-search algorithms commonly used in traditional lightning direction estimation methods but also eliminates the need for correcting inherent biases.

The single-site lightning location model developed in this study takes electric field signals, north-south magnetic field signals, east-west magnetic field signals, and signal arrival times as input features and outputs the propagation distance and direction of lightning relative to the station. The geographic coordinates (latitude and longitude) of the lightning are then determined based on the model output and the station's location. Based on the training data, the absolute error of the location model outputs is reduced by 36.08% after Tri-Pre processing. Meanwhile, the average relative error of distance estimation on

the test dataset is less than 5.6%, and for CG lightning with propagation distances greater than 1000 km, the best relative error is as low as 3.78%. When validated on additional measured data, the relative error for CG (Cloud-to-Ground) lightning with propagation distances between 1500 and 1600 km is only 5.46%. These results demonstrate that the proposed Tri-Pre method effectively cleans mismatched data and enhances the dataset quality and the newly established single-site lightning location model can also accurately locate cloud-to-ground lightning with a propagation distance exceeding 1000 km, which is better than the 6.7% relative error reported by the method proposed by Koochak and Fraser-Smith [14]. The single-site lightning location model, based on the Tri-Pre method, can accurately locate cloud-to-ground lightning even at propagation distances exceeding 1000 km.

Considering that the current Tri-Pre method is based on the wave bank divided at 100 km intervals, there is a possibility that the performance of the Tri-Pre method could be further improved when using a finer division, such as training with waveforms at 50 km intervals.

In future work, we will collect more data to further improve the effectiveness of both the Tri-Pre method and the location model. The impact of expanded data on improving location accuracy will also be evaluated. Furthermore, we plan to explore the use of transfer learning to achieve model generalization.

Author Contributions: B.D.: Conceptualization, Methodology, Software, Writing—original draft. Q.Z.: Resources, Validation, Formal analysis, Supervision, Writing—review and editing. J.L.: Software, Data curation. Y.L.: Software, Funding acquisition. M.Z.: Data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key R&D Program of China (Grant No. 2017YFC1501505), and in part by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 42405148).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We want to thank all personnel and meteorological departments involved in the construction of the lightning location site, lightning data collection, and processing.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

BN	batch normalization
CG	cloud-to-ground
CNN	convolutional neural networks
CTT	cloud-top temperature
E_A	absolute error
ELF	extremely low frequency
E_R	relative error
FY4B	fengyun 4B
GLD360	global lightning detection network
LEMP	lightning electromagnetic pulse
LF	low frequency
LSTM	long short-term memory
MDF	magnetic direction-finding
MSPS	million samples per second

NLDN	national lightning detection network
ResNet	residual networks
RNN	recurrent neural networks
TDOA	time difference of arrival
VLF	very low frequency
VLF-LLN	very low frequency long-range lightning location network
WWLLN	world wide lightning location network

References

1. Rakov, V.A. Electromagnetic methods of lightning detection. *Surv. Geophys.* **2013**, *34*, 731–753. [\[CrossRef\]](#)
2. Dowden, R.L.; Brundell, J.B.; Rodger, C.J. VLF lightning location by time of group arrival (TOGA) at multiple sites. *J. Atmos. Sol.-Terr. Phys.* **2002**, *64*, 817–830. [\[CrossRef\]](#)
3. Said, R.K.; Inan, U.S.; Cummins, K.L. Long-range lightning geolocation using a VLF radio atmospheric waveform bank. *J. Geophys. Res. Atmos.* **2010**, *115*. [\[CrossRef\]](#)
4. Changnon, S.A. Damaging thunderstorm activity in the United States. *Bull. Am. Meteorol. Soc.* **2001**, *82*, 597–608. [\[CrossRef\]](#)
5. Schultz, C.J.; Petersen, W.A.; Carey, L.D. Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends. *Weather. Forecast.* **2011**, *26*, 744–755. [\[CrossRef\]](#)
6. Liu, Y.; Wang, J.; Song, Y.; Liang, S.; Xia, M.; Zhang, Q. Lightning nowcasting based on high-density area and extrapolation utilizing long-range lightning location data. *Atmos. Res.* **2025**, *321*, 108070. [\[CrossRef\]](#)
7. Srivastava, A.; Liu, D.; Xu, C.; Yuan, S.; Wang, D.; Babalola, O.; Sun, Z.; Chen, Z.; Zhang, H. Lightning nowcasting with an algorithm of thunderstorm tracking based on lightning location data over the Beijing area. *Adv. Atmos. Sci.* **2022**, *39*, 178–188. [\[CrossRef\]](#)
8. Zhang, L.; Li, Q.; Zhou, Z.; Yang, K. A lightning augmented recurrent nowcasting model based on self-supervised learning and multi-modal fusion method. *Atmos. Res.* **2025**, *321*, 108089. [\[CrossRef\]](#)
9. Orville, R.E. Development of the national lightning detection network. *Bull. Am. Meteorol. Soc.* **2008**, *89*, 180–190. [\[CrossRef\]](#)
10. Pohjola, H.; Mäkelä, A. The comparison of GLD360 and EUCLID lightning location systems in Europe. *Atmos. Res.* **2013**, *123*, 117–128. [\[CrossRef\]](#)
11. Cummins, K.L.; Murphy, M.J.; Bardo, E.A.; Hiscox, W.L.; Pyle, R.B.; Pifer, A.E. A combined TOA/MDF technology upgrade of the US National Lightning Detection Network. *J. Geophys. Res. Atmos.* **1998**, *103*, 9035–9044. [\[CrossRef\]](#)
12. Krider, E.P.; Noggle, R.C.; Uman, M.A. A gated, wideband magnetic direction finder for lightning return strokes. *J. Appl. Meteorol.* **1976**, *15*, 301–306. [\[CrossRef\]](#)
13. Füllekrug, M. Introduction to lightning detection. *Weather* **2017**, *72*, 32–35. [\[CrossRef\]](#)
14. Koochak, Z.; Fraser-Smith, A. Single-station lightning location using azimuth and time of arrival of sferics. *Radio Sci.* **2020**, *55*, 1–13. [\[CrossRef\]](#)
15. Chen, M.; Lu, T.; Du, Y. An improved wave impedance approach for locating close lightning stroke from single station observation and its validation. *J. Atmos. Sol.-Terr. Phys.* **2015**, *122*, 1–8. [\[CrossRef\]](#)
16. Rafalsky, V.A.; Shvets, A.V.; Hayakawa, M. One-site distance-finding technique for locating lightning discharges. *J. Atmos. Terr. Phys.* **1995**, *57*, 1255–1261. [\[CrossRef\]](#)
17. Rafalsky, V.A.; Nickolaenko, A.P.; Shvets, A.V.; Hayakawa, M. Location of lightning discharges from a single station. *J. Geophys. Res. Atmos.* **1995**, *100*, 20829–20838. [\[CrossRef\]](#)
18. de Sá, A.L.A.; Marshall, R.A. Lightning distance estimation using LF lightning Radio signals via analytical and machine-learned models. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5892–5907. [\[CrossRef\]](#)
19. Wang, J.; Xiao, F.; Yuan, S.; Song, J.; Ma, Q.; Zhou, X. A novel method for ground-based VLF/LF single-site lightning location. *Measurement* **2022**, *196*, 111208. [\[CrossRef\]](#)
20. Jordan, M.I. Artificial intelligence—The revolution hasn’t happened yet. *Harv. Data Sci. Rev.* **2019**, *1*, 1–9.
21. Li, J.; Dai, B.; Zhou, J.; Zhang, J.; Zhang, Q.; Yang, J.; Wang, Y.; Gu, J.; Hou, W.; Zou, B.; et al. Preliminary application of long-range lightning location network with equivalent propagation velocity in China. *Remote Sens.* **2022**, *14*, 560. [\[CrossRef\]](#)
22. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12–14, 2015. Proceedings 3*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015.
23. Hou, W.; Azadifar, M.; Rubinstein, M.; Zhang, Q.; Rachidi, F. An efficient FDTD method to calculate lightning electromagnetic fields over irregular terrain adopting the moving computational domain technique. *IEEE Trans. Electromagn. Compat.* **2019**, *62*, 976–980. [\[CrossRef\]](#)

24. Wang, J.; Huang, Q.; Ma, Q.; Chang, S.; He, J.; Wang, H.; Zhou, X.; Xiao, F.; Gao, C. Classification of VLF/LF lightning signals using sensors and deep learning methods. *Sensors* **2020**, *20*, 1030. [[CrossRef](#)] [[PubMed](#)]
25. Zhou, X.; Wang, J.; Ma, Q.; Huang, Q.; Xiao, F. A method for determining D region ionosphere reflection height from lightning skywaves. *J. Atmos. Sol.-Terr. Phys.* **2021**, *221*, 105692. [[CrossRef](#)]
26. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)]
27. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [[CrossRef](#)]
28. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [[CrossRef](#)]
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
30. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier nonlinearities improve neural network acoustic models. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 17–19 June 2013; Volume 30.
31. Kinga, D.; Adam, J.B. A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, ICLR 7, San Diego, CA, USA, 7–9 May 2015.

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.