JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD038715

High-Accuracy Classification of Radiation Waveforms of Lightning Return Strokes

Ting Wu[1], **Daohong Wang**[1], **and Nobuyuki Takagi**[1]

[1]Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu, Japan

Key Points:

- A machine-learning classifier for negative return strokes (RSs) is built using a large data set with 3-D location information
- Both an accuracy and an efficiency of about 98.8% are achieved and the accuracy-efficiency tradeoff can be easily controlled
- Some RSs and intracloud discharges produce special waveforms that are fundamentally difficult to classify without 3-D location results

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

T. Wu,
wu.ting.x4@f.gifu-u.ac.jp

Citation:

Wu, T., Wang, D., & Takagi, N. (2023). High-accuracy classification of radiation waveforms of lightning return strokes. *Journal of Geophysical Research: Atmospheres*, *128*, e2023JD038715. https://doi.org/10.1029/2023JD038715

Received 16 FEB 2023
Accepted 12 JUL 2023

Abstract A machine-learning classifier for radiation waveforms of negative return strokes (RSs) is built and tested based on the Random Forest classifier using a large data set consisting of 14,898 negative RSs and 159,277 intracloud (IC) pulses with 3-D location information. Eleven simple parameters including three parameters related with pulse characteristics and eight parameters related with the relative strength of pulses are defined to build the classifier. Two parameters for the evaluation of the classifier performance are also defined, including the classification accuracy, which is the percentage of true RSs in all classified RSs, and the identification efficiency, which is the percentage of correctly classified RSs in all true RSs. The tradeoff between the accuracy and the efficiency is examined and simple methods to tune the tradeoff are developed. The classifier achieved the best overall performance with an accuracy of 98.84% and an efficiency of 98.81%. With the same technique, the classifier for positive RSs is also built and tested using a data set consisting of 8,700 positive RSs. The classifier has an accuracy of 99.04% and an efficiency of 98.37%. By examining misclassified waveforms, we show evidence that some RSs and IC discharges produce special radiation waveforms that are almost impossible to correctly classify without 3-D location information, resulting in a fundamental difficulty to achieve very high accuracy and efficiency in the classification of lightning radiation waveforms.

Plain Language Summary Lightning location systems are required to classify return strokes (RSs) from intracloud (IC) discharges accurately and efficiently because the RS is the main discharge component that poses direct threats to the human society. In this paper, we report a machine-learning classifier for negative RSs built using a large data set with accurate 3-D location information. The classifier has an accuracy of 98.84% (98.84% of classified RSs are correct classifications) and an efficiency of 98.81% (98.81% of RSs can be correctly classified). With the same technique, we also built a classifier for positive RSs with similarly high accuracy and efficiency. Our classifiers only require some simple waveform parameters so the same technique can be used by various lightning location systems relatively easily. A sample Python script to use the classifier is provided and readers are encouraged to test the classifier using their own data set. We also demonstrate that some RSs and IC discharges produce abnormal waveforms, so 100% accuracy or efficiency is fundamentally difficult to realize using only waveform information.

1. Introduction

Ground-based lightning location systems (LLSs) are widely used to monitor lightning activities. A prominent feature of ground-based LLSs is that lightning activities in a wide area can be monitored in real time with only a limited number of sensors. Some famous national and continental LLSs include the National Lightning Detection Network (NLDN) covering the continental United States (e.g., Cummins & Murphy, 2009), the European Cooperation for Lightning Detection network (EUCLID) covering the European continent (e.g., Schulz et al., 2016), and the Earth Networks Total Lightning Network (ENTLN) (e.g., Zhu et al., 2022) with the aim of a global coverage.

It is a basic requirement for LLSs to automatically and efficiently classify cloud-to-ground (CG) lightning flashes from intracloud (IC) flashes as the former consist of discharges with direct connections to the ground and thus pose a much larger threat to the human society. The fundamental difference between a CG flash and an IC flash is that a CG flash contains one or more return strokes (RSs), so the classification of CG flashes is basically realized by classifying RSs. Further, it is well known that RSs produce characteristic electric field radiation waveforms that are largely different from those of IC discharges (e.g., Lin et al., 1979), so most LLSs classify RSs based on their waveform characteristics.

However, RSs actually can produce radiation waveforms with a variety of special features under some special conditions. For example, some RSs in winter thunderstorms are known to produce abnormal radiation waveforms, some of which could not be correctly classified by LLSs (Wu, Wang, Huang, & Takagi, 2021; Wu, Wang, & Takagi, 2021). It is also well known that RSs striking tall objects produce much narrower radiation waveforms (Pavanello et al., 2007; Zhu et al., 2018). On the other hand, IC discharges include various discharge processes such as narrow bipolar events and recoil leaders, some of which may produce radiation waveforms with certain similar features as RS waveforms. As a result, for most LLSs, it is basically very difficult to achieve a very high classification accuracy of RSs. For example, Zhu et al. (2016) reported that out of 339 RSs in Florida in 2014 that were also recorded by the NLDN, 312 (92%) were correctly classified as RSs by the NLDN. Kohlmann et al. (2017) reported that the classification accuracy of EUCLID for RSs were generally around 90% based on ground-truth data in various regions of Europe. For some particular thunderstorms or some special types of discharges, misclassifications by LLSs can be more common. For example, Fleenor et al. (2009) found that 204 out of 376 (54%) of RSs reported by the NLDN during a field campaign in 2005 were actually IC discharges. Leal et al. (2019) found that compact intracloud discharges with estimated peak currents larger than 50 kA were all falsely classified as RSs by both NLDN and ENTLN. Paul et al. (2020) reported that out of 40 RSs detected at the Peissenberg Tower, 12 (30%) were falsely classified as IC discharges.

In order to overcome the uncertainties in classifications based only on radiation waveforms, Betz et al. (2004) proposed a pseudo 3-D technique to assist the discrimination of RSs and IC discharges based on the fact that the elevation of IC discharges would have some contributions to the time delay. However, this technique also has some limitations. For example, IC discharges need to have significant elevations, the baseline of the LLS cannot be too long, and lightning discharges first need to be located accurately in 2-D. These limitations prevented the wide implementation of this technique.

In recent years, machine-learning techniques have been developing rapidly, and these techniques seem to be promising in significantly increasing the classification accuracy of lightning radiation waveforms. Wang et al. (2020) developed a convolutional neural network to classify radiation waveforms of lightning discharges recorded by the Advanced Direction-time Lightning Detection System in China. They reported an accuracy of over 99%. However, they apparently did not have the height information of lightning discharges and thus could not unambiguously differentiate RSs and IC discharges, so the accuracy remains questionable. Zhu et al. (2021) used the Support Vector Machines (SVM) model to classify CG and IC flashes recorded by the Cordoba Marx Meter Array. The lightning data were in 3-D, so they could employ the discharge height information to build a data set with accurate discharge types. They reported an overall accuracy of 97%. However, their proposed method requires full waveform information, while most LLSs only retrieve a few parameters of electric field waveforms of lightning discharges. Apart from classifying RS waveforms, machine-learning techniques have also been used to classify characteristic waveforms related with terrestrial gamma-ray flashes (Pu et al., 2023).

In this paper, we report a simple yet high-accuracy machine-learning technique based on the Random Forest classifier to classify RSs. We will use a large data set containing about 15,000 negative RSs and many more IC discharges with accurate 3-D location information to train and test the classifier. As will be described in this paper, many of the recorded RSs and IC discharges produced atypical radiation waveforms that were challenging to be correctly classified. However, the accuracy of our classifier is close to 99% demonstrated by evaluations in various respects. Our classifier requires only some simple parameters of lightning radiation waveforms, so the same technique can be adopted by most LLSs relatively easily.

2. Observation and Data

During the summer of 2017, we set up a low-frequency lightning mapping system called Fast Antenna Lightning Mapping Array (FALMA) in central Japan. The FALMA consisted of 12 sites covering an area of about 80×80 km^2. Locations of these 12 sites are shown as red squares in Figure 1a. At every site, a fast antenna working in the frequency band of 500 Hz to 500 kHz was used to receive radiation signals from lightning discharges. The signals were recorded with a sampling rate of 25 MS/s. As described by Wu et al. (2018a), thanks to improvements made in both the hardware and the software, we realized high-quality 3-D lightning mapping with the FALMA. As can be seen from examples of lightning flashes in Wu et al. (2018a, 2019), 3-D mapping results of FALMA have similar quality to those of very-high-frequency (VHF) systems such as the Lightning Mapping Array (Rison et al., 1999).

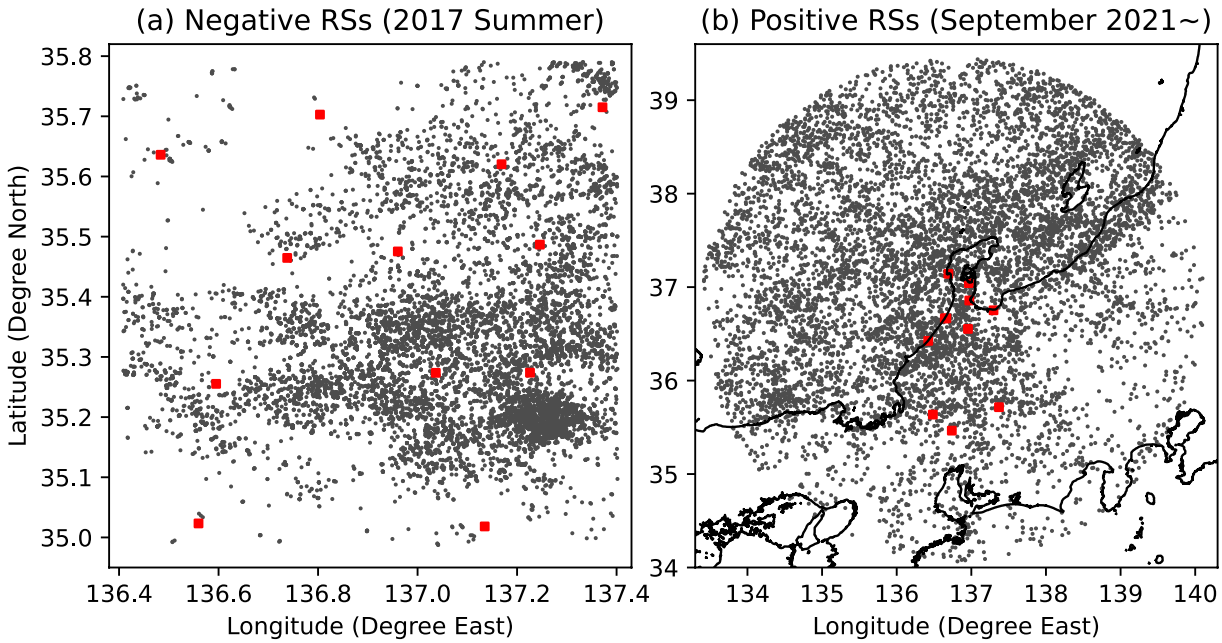

Figure 1. (a) Negative return strokes (RSs) (black dots) observed from 19 July to 26 August 2017. (b) Positive RSs observed from 26 September 2021 to 3 September 2022. Red squares represent observation sites of Fast Antenna Lightning Mapping Array.

Data obtained from 19 July to 26 August are used in this study for building and testing the classifier for negative RSs. All data are reprocessed for this study. As a negative RS produces a positive radiation pulse (as in the atmospheric electricity sign convention), in order to classify negative RSs, we need to examine all positive pulses. A positive pulse here simply means a positive electric field change whose amplitude exceeds a pre-defined threshold, and we do not set any restrictions on other characteristics of the pulse to ensure the simplicity of this method and thus its easy implementation in real observations. The largest positive pulse in each 20-ms window is located in 3-D, and discharges located in the region shown in Figure 1a, a 90×90 km^2 area over the FALMA network, are used in order to ensure reliable 3-D locating. Pulses with source heights lower than 500 m are treated as candidates of RSs. Their waveforms are then confirmed manually, and for some ambiguous pulses, they are further manually located to determine their source heights. In this way, we can unambiguously determine that the selected pulses are truly RSs. The number of IC discharges are much larger than that of RSs, so we cannot manually confirm waveforms of all IC discharges, and we only use pulses with source heights larger than 3000 m as IC pulses. There are 14,898 pulses confirmed as negative RSs and 159,277 pulses as IC discharges. Although the height thresholds of 500 m for RSs and 3000 m for IC discharges are somewhat arbitrary, only 1064 (0.6%) pulses are located between 500 and 3000 m, so neglecting these pulses will have little influence on the results of this study. Locations of all negative RSs are shown as black dots in Figure 1a. It should be noted that we will build a classifier for negative RSs rather than negative CG flashes; a CG flash consists of at least one RS and also many IC discharges, both of which need to be correctly classified.

Using the high-quality data set of 2017 summer, we will establish the technique for building the classifier as will be described in Sections 3.1–3.5. Further, using the same technique, we will also build a classifier for positive RSs as will be described in Section 3.6. However, positive RSs in central Japan in summer are quite rare (Wu et al., 2018b). In order to accumulate a large number of positive RSs, we will use the data collected during a long period, from 26 September 2021 to 3 September 2022. During this period, we set up a FALMA network covering a large area for 2-D locating of both summer and winter lightning. Observation sites are shown as red squares in Figure 1b. A total of 8700 positive RSs observed in an area with a radius of 300 km are identified and will be used for building and testing the classifier for positive RSs. Locations of these positive RSs are shown as black dots in Figure 1b. The procedure for the identification of these positive RSs will be further described in Section 3.6.

Our classifiers will be built and tested mainly based on the Random Forest classifier, which is one of the most widely used machine-learning models for classification tasks. The random forest classifier is an ensemble

classifier consisting of multiple decision trees, with each decision tree being built on a subset of training samples, and the final classification result is based on predictions made by the ensemble of decision trees. It has been shown that ensemble classifiers are more likely to achieve higher accuracy and robustness (e.g., DeFries, 2000; Miao et al., 2011). Moreover, the random forest classifier is well known for its computational efficiency and simple parameter setting as compared with other ensemble classifiers (e.g., Belgiu & Drăguţ, 2016). In Section 3.5, we will also make a brief comparison with the SVM classifier, another popular machine-learning model.

3. Methods and Results

3.1. Method to Evaluate the Performance of a Classifier

Before building the classifier, first we need to define some parameters as indicators of the performance of a classifier. One obvious parameter to evaluate the performance is the classification accuracy, or simply *accuracy*, that is, the percentage of true RSs in the waveforms classified as RSs. However, only this parameter is apparently not enough, as it is always possible to build a classifier with very strict criteria so that it only identifies very typical RS waveforms. Another important parameter is the identification efficiency, or simply *efficiency*, that is, the percentage of correctly classified RSs in all RSs.

Suppose the number of RSs is N_R, and the number of IC discharges is N_I. Of the N_R RSs, N_{Rc} are correctly classified (the subscript c stands for "correct"), and the remaining $N_R - N_{Rc}$ are misclassified as IC discharges. Of the N_I IC discharges, N_{Ic} are correctly classified, and the remaining $N_I - N_{Ic}$ are misclassified as RSs. The accuracy and the efficiency can be calculated as follows.

$$Accuracy = \frac{N_{Rc}}{N_{Rc} + (N_I - N_{Ic})} \tag{1}$$

$$Efficiency = \frac{N_{Rc}}{N_R} \tag{2}$$

During the process to build the classifier, we will experiment and tune various parameters of the classifier to make the accuracy and the efficiency as high as possible.

Normally a data set is split into a larger training set and a smaller test set, with the training set used to train a classifier and the test set used to test or evaluate the performance of the classifier. In this study, we use an improved approach. All RS and IC data are combined, shuffled and then divided into five equal parts. Each part is in turn used as the test set and the remaining four parts combined are used as the training set. In this way, a classifier is built and tested for five times and five results of accuracy and efficiency are calculated. The average values of five tests will be used as the final results. In this way, we can avoid any random biases in the test set. Moreover, as will be described in Section 4, in this way all data can be tested and we can find as many atypical waveforms as possible that are difficult to be correctly classified.

3.2. Waveform Parameterization

We will define some waveform parameters to be used for building the classifier. First we describe the procedure to calculate waveform parameters based on multiple-site records. As waveforms recorded at a close distance contain the electrostatic and induction field components (e.g., Thottappillil et al., 1997) that may significantly distort the waveforms, observation sites within 40 km from a discharge are first excluded. Waveforms recorded by the remaining sites are used to calculate the parameters, and for each parameter, the median value of the results calculated based on these sites are used as the final result of the parameter for the discharge. The median value is used because different sites have different noises which may have some influences to the calculation of parameters, and we believe using the median value is the simplest method to reduce the influences. Figure S1 in Supporting Information S1 also shows the distributions of median distances with observation sites more than 40 km away for all negative RSs and IC discharges used for building the classifier.

3.2.1. Parameters Related With Pulse Characteristics

First we define three basic parameters related with pulse characteristics. Definitions of these parameters are illustrated using an RS pulse (a first RS) in Figure 2a and an IC pulse (preliminary breakdown) in Figure 2b (blue parameters).

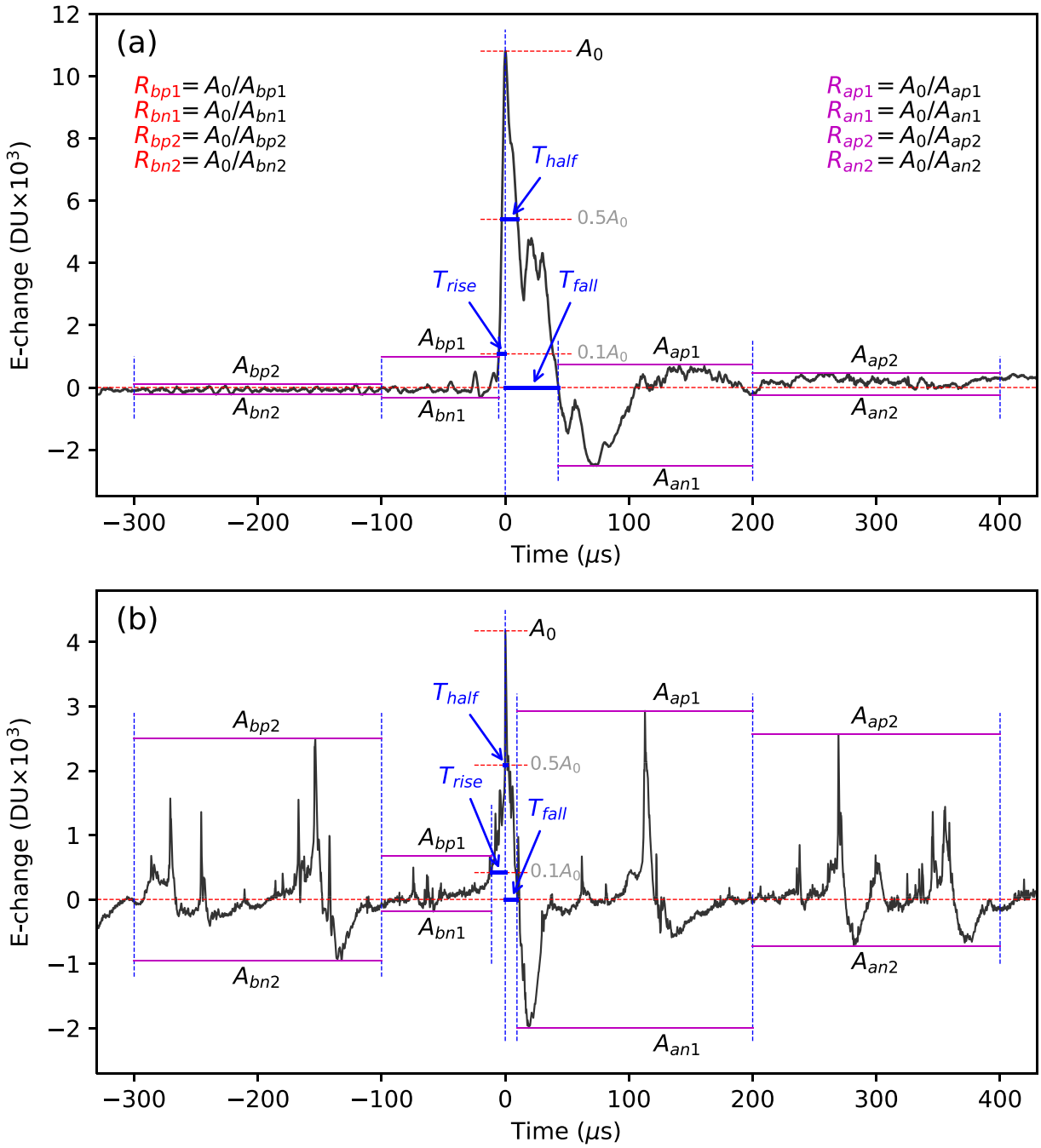

Figure 2. Illustration of waveform parameters using (a) a return stroke pulse and (b) an intracloud pulse.

1. T_{rise}: The rise time of a pulse (10% to peak).
2. T_{fall}: The fall time of a pulse (peak to zero).
3. T_{half}: The pulse width at the half maximum.

With only these three basic parameters, we trained and tested the Random Forest classifier using the negative RS and IC data set obtained in 2017 summer. As described in Section 3.1, the data set is divided into five parts and each part in turn is used as the test set, so the classifier is trained and tested for five times. The accuracy ranges from 72.25% to 73.57% with an average of 72.82%, and the efficiency ranges from 70.80% to 72.81% with an average of 71.59%. We also tried to add two related parameters, including the pulse width, which is the sum of

the rise time and fall time, and the ratio of fall time to rise time, but the result has little difference (the average accuracy is 72.17% and the average efficiency is 70.86%).

Indeed, with only these basic pulse parameters, it is difficult to accurately classify RSs.

3.2.2. Parameters Related With Relative Strength

An important feature of the RS waveform is that pulses right before and after an RS pulse is usually much weaker. The following parameters are defined to employ this feature. These parameters are also illustrated in Figures 2a and 2b.

1. R_{bp1}: The ratio of A_0 to A_{bp1}, in which A_0 is the peak amplitude of the target pulse, and A_{bp1} is the maximum amplitude of pulses right before the target pulse (from −100 μs to 10% peak) as illustrated in Figure 2. The subscript b stands for "before," and the subscript p stands for "positive."
2. R_{bn1}, R_{bp2}, R_{bn2}, R_{ap1}, R_{an1}, R_{ap2}, R_{an2}: These parameters are defined in the same way as R_{bp1}, also illustrated in Figure 2. Note that the subscript a stands for "after," and the subscript n stands for "negative."

As can be seen in Figure 2, 300-μs waveform before and 400-μs waveform after the target pulse are used for calculating these ratios. It should be noted that the lengths are to some extent arbitrary; the point here is that these lengths should be long enough to contain enough information about the strength of discharges both before and after, but at the same time it should not be too long because the recording length, especially the pre-trigger length, for one trigger may be quite short in some systems. Waveforms too long may also contain information that is not consistently related with the target pulse and thus introduce noises.

The three parameters defined in Section 3.2.1 along with the eight new parameters defined above are used to train the Random Forest classifier. The accuracy of five tests ranges from 98.86% to 99.32% with an average of 99.02%, and the efficiency ranges from 98.02% to 98.66% with an average of 98.34%. It is clear that these new parameters representing the relative strength are very effective in the classification of RSs.

3.3. Tradeoff Between Accuracy and Efficiency

From the above result, we can see one feature of the classifier is that the accuracy is always higher than the efficiency. It is obvious that increasing the efficiency usually implies decreasing the accuracy. However, it is desirable if we can control the tradeoff between the accuracy and the efficiency. For example, in some situations, it may be required to identify as many RSs as possible, so a high efficiency is essential while a low accuracy is tolerable. Next we will investigate two factors that influence the tradeoff between the accuracy and the efficiency.

3.3.1. Influence of Sample Size Imbalance

One reason for the higher accuracy in the classifier built in the previous section is a much larger sample of IC discharges compared with the sample of RSs. With such a biased data set, the classifier is more likely to misclassify RSs, as also noted by Zhu et al. (2021). We can simply duplicate the sample of RSs to make the classifier identify more RSs, though at the cost of more misclassifications of IC discharges. Note that the duplication should only be made for the training set.

With the original data set, 247 of 14,898 RSs (1.7%) are misclassified, but only 145 of 159,277 IC pulses (0.091%) are misclassified. If we duplicate the data set of RSs in the training set, the number of misclassified IC pulses increases to 162 while the number of misclassified RSs decreases to 199. We tried to make more duplications and tested the classifier, and the results of the accuracy and the efficiency are shown in Figure 3a. With one duplication of the RS training set, the accuracy decreases from 99.02% to 98.91% but the efficiency increases from 98.34% to 98.66%. With two duplications, the accuracy decreases to 98.84% but the efficiency increases to 98.76%, very close to the accuracy. With further duplications, we can see that both the accuracy and the efficiency are generally very similar, changing between 98.75% and 98.85%, indicating that the sample size imbalance does not have a significant effect any more.

If we use the average value of the accuracy and the efficiency as the indicator of the overall performance of a classifier, we can see from Figure 3a that with four duplications of the RS training set, the classifier has the highest performance with an accuracy of 98.84% and an efficiency of 98.81%. We treat this as the best performance of the classifier for negative RSs and this classifier will be used for further evaluations in the following section.

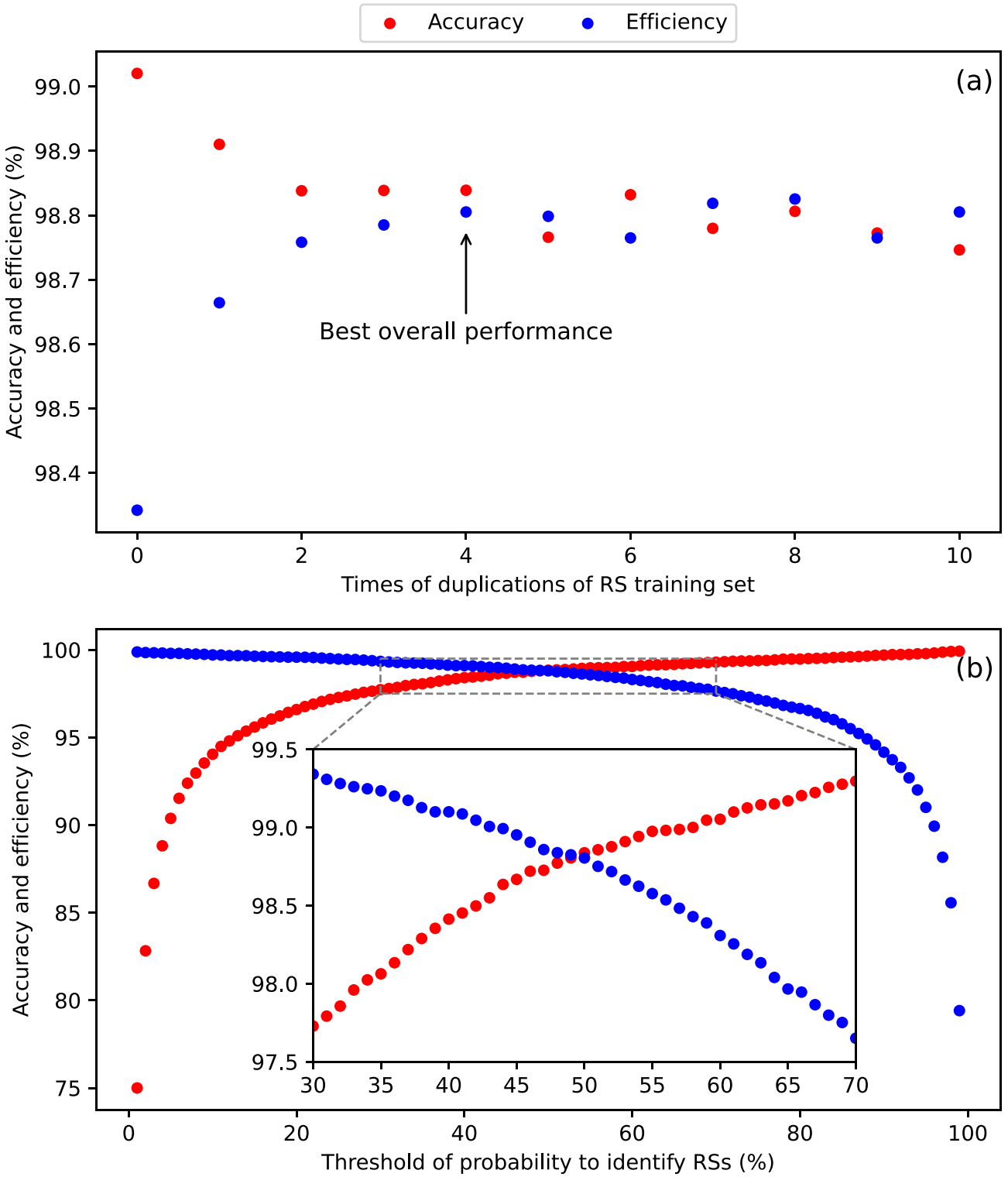

Figure 3. Variations of the accuracy and the efficiency with (a) different times of duplications of return stroke (RS) training set and (b) different thresholds of probability to classify RSs.

3.3.2. Influence of Probability Thresholds

When classifying a pulse, the Random Forest classifier can output the probability that the pulse is a true RS. By default, the classifier determines a pulse as an RS when the probability is larger than 50%. By changing the probability threshold, we can conveniently tune the accuracy-efficiency tradeoff.

Figure 3b shows variations of the accuracy and the efficiency related with the probability threshold. We can see that as the probability threshold increases, the accuracy increases while the efficiency decreases. This is easy to understand; a higher probability threshold represents stricter criteria to classify RSs, so naturally the identified RSs are more likely true RSs (higher accuracy), but at the same time fewer RSs can be identified (lower efficiency). In practice, when using the classifier we can set a customized probability threshold that fits the specific requirements of an application to achieve desired accuracy or efficiency.

3.4. Parameter Importance and Parameter Reduction

The Random Forest classifier outputs a value indicating the relative importance of each parameter in contributing to the performance, from which we can evaluate the effectiveness of each parameter in the classification of RSs. The results are shown in Figure 4a. Values of the importance of all parameters combined equal to 1. We can see that parameters related with the pulse strength relative to previous pulses (red parameters in Figures 2 and 4a) are generally more important than other parameters. This is easy to understand as an RS pulse is preceded by leader pulses which are usually much weaker than the RS pulse. By contrary, an IC pulse is usually preceded by other IC pulses with comparable amplitudes. Therefore, parameters related with the relative strength are very effective in the classification of RSs.

We can also see that parameters related with pulse characteristics (blue parameters) have relatively low importance, which is why the classifier performance is very poor with only these parameters as described in Section 3.2.1. It also indicates that traditional RS classification methods based on pulse characteristics are not very reliable.

Using the information of the relative importance of each parameter, it is possible to simplify the parameterization by removing some parameters with low importance. Based on the model built in Section 3.3.1, we investigated how the accuracy and the efficiency change when removing some parameters with low importance. When removing the parameter R_{ap1}, which has the lowest importance, the accuracy decreases from 98.84% to 98.78%, although the efficiency increases slightly from 98.81% to 98.83%. When further removing R_{ap2}, which has the second lowest importance, the accuracy stays unchanged as 98.78%, but the efficiency decreases to 98.72%. We tried to further remove other parameters in order of importance, from lowest to highest, and the results are shown in Figure 4b. We can see that with fewer parameters, the accuracy and the efficiency generally decrease as expected, but the decreasing is not significant when removing parameters with low importance. Specifically, when removing the six parameters with lowest importance and using only the remaining five parameters to train the classifier, the accuracy is 98.44% and the efficiency is 98.07%, which are still satisfactory results. Therefore, in applications with limited computational resources, parameters with low importance can be removed to make the classifier easier to be implemented.

3.5. Comparison of Different Machine-Learning Models

Apart from the Random Forest classifier, another popular machine-learning model for classification is the SVM classifier, which was used by Zhu et al. (2021) for the classification of lightning pulses. Here we make a brief comparison of the Random Forest and the SVM classifiers. Note that the same classifier may have very different performances in different types of applications, and the comparison here is only of the Random Forest and the SVM classifiers for the classification of lightning radiation waveforms with the parameterization schemes described in this paper. First we use the scheme described in Section 3.2.2 (using 11 parameters illustrated in Figure 2) to train the classifiers, and the results are shown in Table 1 (upper part). We can see the SVM classifier has slightly lower accuracy and efficiency than the Random Forest classifier. Further, we use the scheme described in Section 3.3.1 (duplicating the RS training data set) to train the classifiers, and again, the SVM classifier has slightly lower accuracy and efficiency. Another difference is in the time needed to train a classifier; it takes less than 30 s to train an Random Forest classifier while the time needed to train an SVM classifier is around 100 s. A significantly shorter time to build a classifier is potentially very useful as it would be more convenient to experiment various combinations of parameters in order to boost the performance of the classifier.

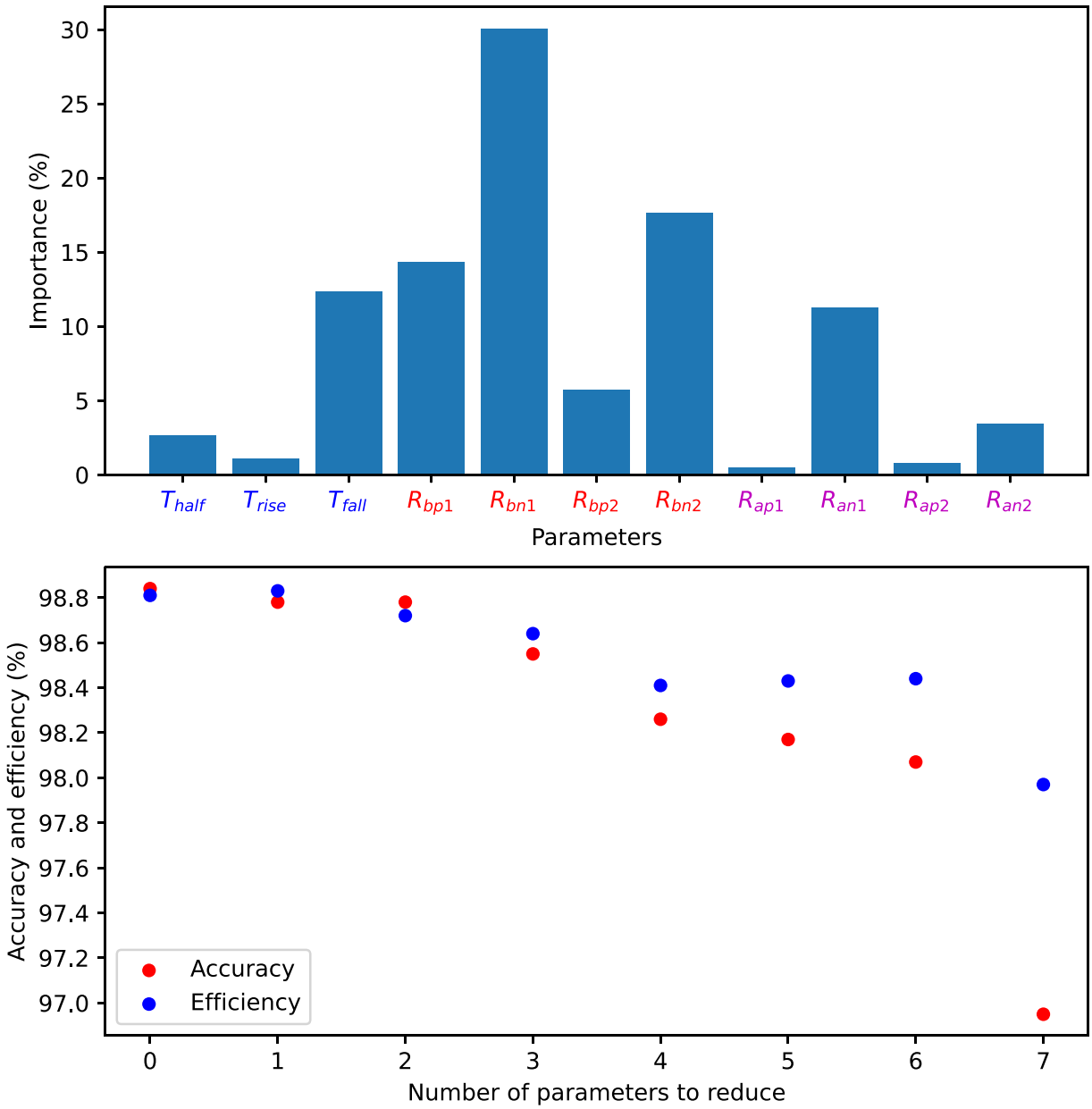

Figure 4. (a) Relative importance of waveform parameters. Definitions of these parameters are illustrated in Figure 2. (b) Variations of accuracy and efficiency when reducing the number of parameters. Parameters are reduced in order of importance, from low to high, that is, R_{ap1}, R_{ap2}, T_{rise}, T_{half}, R_{an2}, R_{bp2}, and R_{an1}.

Table 1
Comparison of the Random Forest Classifier and the Support Vector Machines (SVM) Classifier

Classifier	Accuracy (%)	Efficiency (%)	Time cost (s)
Eleven parameters (Section 3.2.2)			
Random forest	99.02	98.34	20
SVM	98.43	97.42	96
Duplicating return stroke training set (Section 3.3.1)			
Random forest	98.84	98.81	28
SVM	97.98	98.09	108

3.6. Classification of Positive Return Strokes

The methods described above can also be used to build a classifier for the classification of positive RSs. However, positive CG flashes are very rare in summer thunderstorms in central Japan. As reported by Wu et al. (2018b), only 46 positive CG flashes consisting of 53 positive RSs were observed and could be located in 3-D during the summer observation of 2017. Therefore, here we also include the data obtained in other periods. First we use the 690 positive RSs observed during the winter of 2018 (Wu et al., 2022) to build a preliminary classifier for the identification of positive RSs. Then we use this classifier to search the data recorded in about 1 year from September of 2021 for possible positive RSs. As described in Section 2, during this period, we

set up a FALMA network with long baselines for 2-D locating of both summer and winter lightning. Waveforms of the identified positive RSs by the preliminary classifier are manually confirmed to exclude obvious false classifications. Indeed, the preliminary classifier identified many pulses that were clearly IC pulses and we painstakingly excluded all apparent IC pulses by manual inspections. In this way, we collected the data of 8700 positive RSs, locations of which are shown in Figure 1b. Note that there is no height information for these positive RSs, so this data set is not as accurate as the negative RS data set in 2017 summer used in previous sections.

For IC data, we also use the data of summer observation of 2017 as these data have accurate 3-D location results. However, different from the IC data set for the negative RS classifier, IC pulses for building positive RS classifier should have the same polarity as positive RSs. So we located IC pulses having the same polarity as positive RSs and selected those with heights larger than 3 km, the same treatment as that in building the negative RS classifier. On the other hand, as the size of positive RS data set is relatively small, we do not need too many IC data, so for simplicity, we only located one IC pulse in every 50-ms window. Finally, we collected a total of 113,922 IC pulses.

Using these data sets, and with the same scheme for building negative RS classifier described in Section 3.3.1, we built and tested the classifier for positive RSs. It is found that with the RS training set duplicated for one time, the classifier has the best overall performance with an accuracy of 99.04% and an efficiency of 98.37%. This result demonstrated that as long as there are enough data of positive RSs, we can also build a high-accuracy classifier for positive RSs in the same way as building the negative RS classifier.

Although the data set of positive RSs does not have 3-D location information, and the usage of the preliminary classifier to assist the selection of positive RSs may also introduce some biases, as positive RSs are much rarer than negative RSs and it is very difficult to collect a large and reliable sample, we believe our classifier is very valuable for future observations and researches. Moreover, as all waveforms of identified positive RSs have been manually confirmed, the classifier likely has an accuracy similar to that of the manual classification.

4. Atypical Intracloud and Return Stroke Waveforms

As described in Section 3.1, the whole data set of 2017 summer is divided into five parts, with each part in turn used as the testing set and the remaining four parts combined used as the training set. In this way, all pulses can be tested and we can identify as many pulses as possible that are potentially difficult to classify. Using the classifier built in Section 3.3.1 with the RS training set duplicated for four times, all pulses are classified. Of the 14,898 RSs, 178 were misclassified as IC discharges, and of the 159,277 IC pulses, 173 were misclassified as RSs. Waveform figures of all these misclassified pulses are provided in the data repository.

There are several common reasons for misclassifications of RS pulses as IC pulses. Waveforms of four examples are shown in Figures 5a–5d, all of which are misclassified as IC discharges and whose source heights have been confirmed to be close to the ground. First, it is well known that RSs striking tall grounded objects usually produce very narrow pulses (Araki et al., 2018; Cai et al., 2022; Pavanello et al., 2007; Zhu et al., 2018), making it easy to misclassify them as IC discharges. One example is shown in Figure 5a. This pulse is located near a transmission tower, and its pulse width is only about 9 μs, indicating that it is likely produced by an RS striking the tower. Second, two RSs sometimes occur sequentially with a very small time difference of a few tens of microseconds, and if the second RS has a larger peak than the first one, the second RS may be misclassified as an IC pulse. One example is shown in Figure 5b. Such RSs are likely the so-called "multiple-termination strokes" (Kong et al., 2009; Sun et al., 2016) or "forked strokes" (Ballarotti et al., 2005), with two RSs induced by two branches of the same leader. Third, an RS may occur almost simultaneously with IC discharges of other lightning flashes, resulting in a peculiar waveform and thus misclassified as an IC discharge. One example is shown in Figure 5c. While the positive pulse is confirmed to be produced by an RS, the preceding negative pulse labeled as "IC" is produced by IC discharges in an independent lightning flash and is located about 87 km away from the RS. Such large negative pulse is normally not seen right before negative RSs, and as a result the overall waveform is misclassified as an IC discharge. Finally, some RSs apparently produce waveforms that are largely different from typical RS waveforms but the reason is not yet clear. One example is shown in Figure 5d. The pulse has a rise time of about 18 μs while its fall time is only about 9 μs.

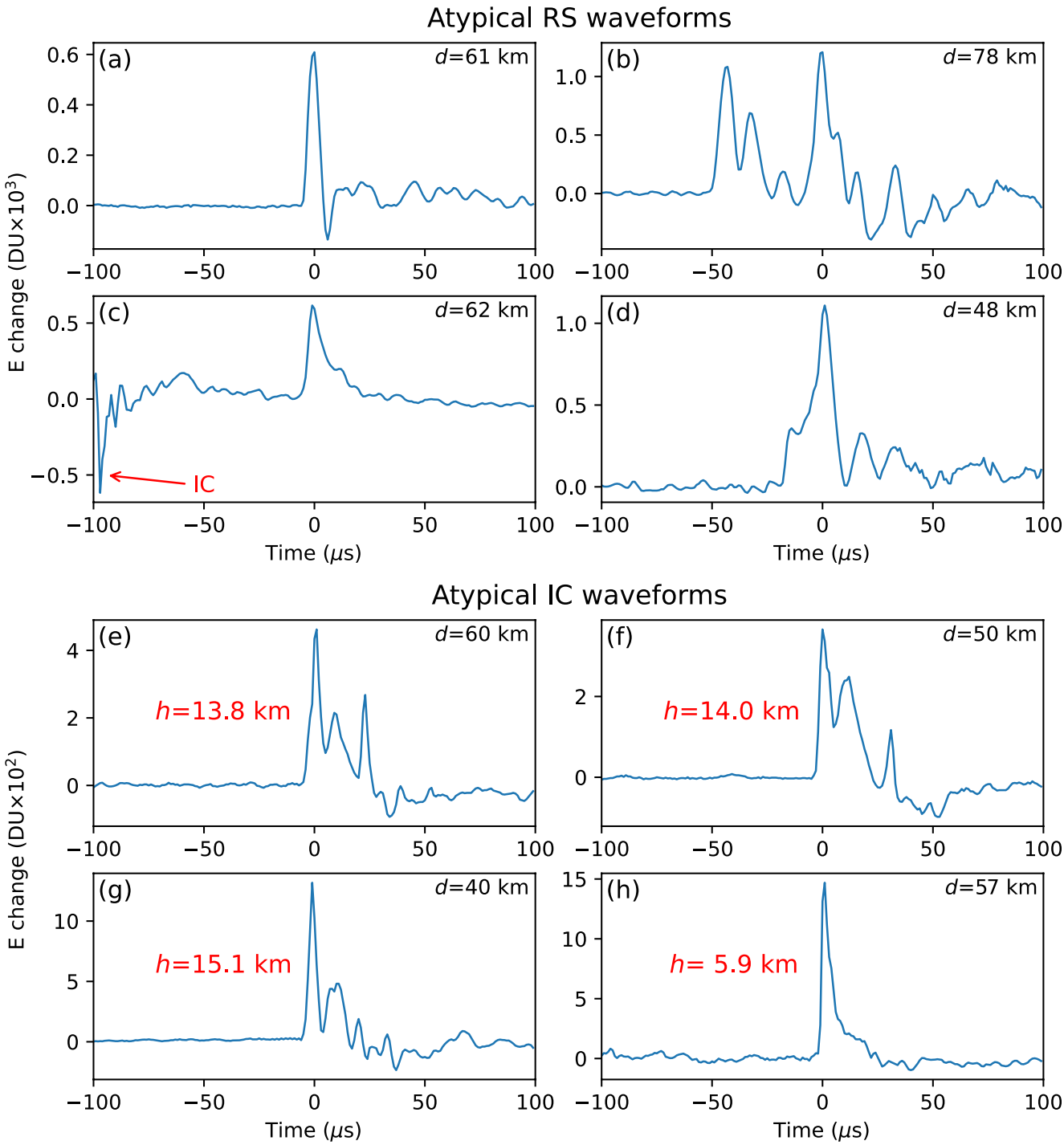

Figure 5. (a)–(d) Atypical E-change waveforms produced by return strokes (RSs) but misclassified as intracloud (IC) discharges. (e)–(h) Atypical E-change waveforms produced by IC discharges but misclassified as RSs. The value of *d* represents the distance between the discharge and the observation site recording the waveform. The value of *h* represents the source height of the IC discharge.

Another example of an RS producing abnormal waveform is shown in Figure 6 along with location results of the preceding leader. This RS is a subsequent RS. We can see from the location results in Figure 6a a dart leader with a speed of about 4×10^6 m/s preceding the RS, and the RS is located very close to the ground as indicated by the cross sign. From Figure 6c, we can see details of the RS waveform. It contains two peaks with the second peak much larger than the first one, resulting in a much larger rise time than the fall time. Without the 3-D location

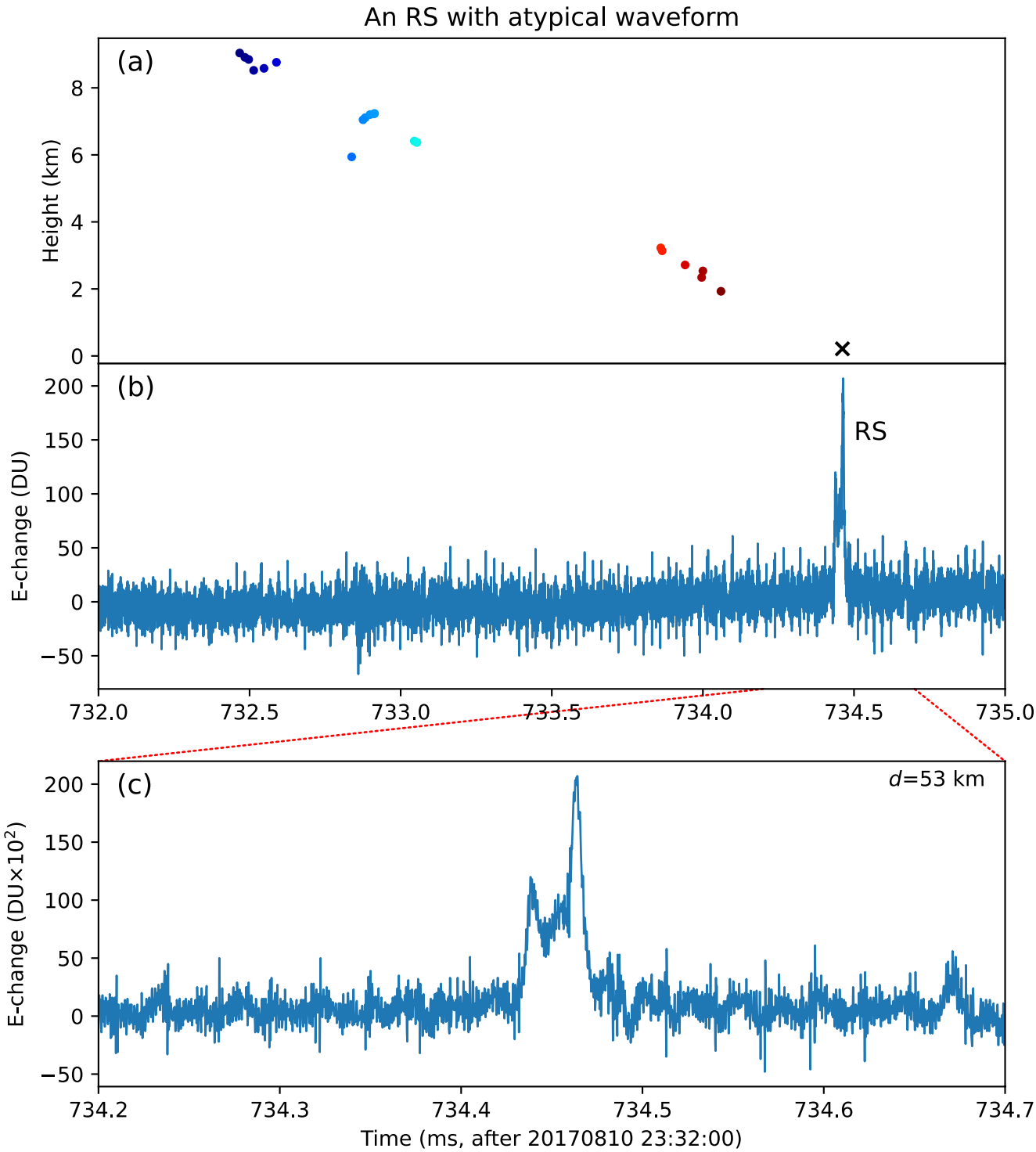

Figure 6. Location result and E-change waveforms of a return stroke (RS) misclassified as an intracloud discharge. (a) Height-time location results of the dart leader preceding the RS. The cross sign represents the RS. (b) E-change waveform of the RS and preceding discharges. (c) E-change waveform of the RS. The value of *d* represents the distance between the RS and the observation site recording the waveform.

results, it is very difficult to determine that the waveform is produced by an RS. Waveforms of this RS recorded by all sites are shown in Figure S2 in Supporting Information S1.

The major reason for IC discharges misclassified as RSs is that waveforms of some IC discharges have some similar features as those of RSs. Four examples are shown in Figures 5e–5h. All of these waveforms appear very similar to those of RSs. However, their source heights range from 5.9 to 15.1 km, indicating that they are produced by IC discharges. We also manually located these pulses to make sure that there were no large errors in the source height results. We can see that these pulses have relatively short rise times and much longer fall times. Pulses in Figures 5e–5g also have fine structures superimposed on the falling part, similar to waveforms of first RSs, and the pulse in Figure 5h resembles the waveform of a subsequent RS. These similar features as RS waveforms make it almost impossible to correctly classify them as IC discharges without the 3-D location results.

Another example of an IC pulse appearing similar to RS pulses is shown in Figure 7 along with location results of preceding discharges. From the height-time location results in Figure 7a, we can see that a leader first propagated above 6.5 km and then descended to a height of about 5.5 km, and then the large IC pulse is produced, represented by the cross sign. From the E-change waveform in Figure 7c, we can see that the large IC pulse is very similar to an RS pulse, with preceding pulses resembling stepped leader pulses. With the help of the 3-D location results, we can be sure that this RS-like pulse is produced by IC discharges. Waveforms of this IC discharge recorded by all sites are shown in Figure S3 in Supporting Information S1. We are not aware of any study reporting such RS-like IC pulses, and the reason for their abnormal waveforms is not immediately clear even by looking at their 3-D mapping results. In our future studies, we will explore the mechanism responsible for these special IC pulses.

These examples of special RS and IC waveforms illustrate the fact that some RSs and IC discharges produce atypical radiation waveforms from which the discharge types cannot be accurately determined, resulting in a fundamental difficulty to achieve very high accuracy and efficiency using only waveform information. This result also illustrates the importance of accurate 3-D location results in scientific investigations of lightning phenomena.

5. Conclusions

Using a large data set with 3-D location results, we built a classifier for radiation waveforms of negative RSs based on the Random Forest classifier. Eleven simple parameters are defined for building the classifier, including three parameters related with pulse characteristics and eight parameters related with relative strength of pulses. A classification accuracy of 98.84% and an identification efficiency of 98.81% are achieved. We also demonstrated methods to tune the tradeoff between the accuracy and the efficiency so the classifier can be used in applications with different requirements of the accuracy or the efficiency. The importance of the 11 parameters was analyzed and we demonstrated that it is possible to reduce the number of parameters with relatively low importance and at the same time keep the high performance of the classifier. With the same methods, we also built a classifier for positive RSs which has similarly high accuracy and efficiency as the classifier for negative RSs.

Misclassified RS and IC waveforms are examined and some common reasons for misclassifications are analyzed. We demonstrated that RSs sometimes produce radiation waveforms that are largely different from normal RS waveforms, and IC discharges sometimes produce waveforms that appear very similar to RS waveforms. Therefore, some RS and IC waveforms are fundamentally difficult to be correctly classified without 3-D location information, and it is likely that such misclassifications commonly exist in most LLSs. The results also imply the importance of 3-D location results in detailed analyses of lightning phenomena.

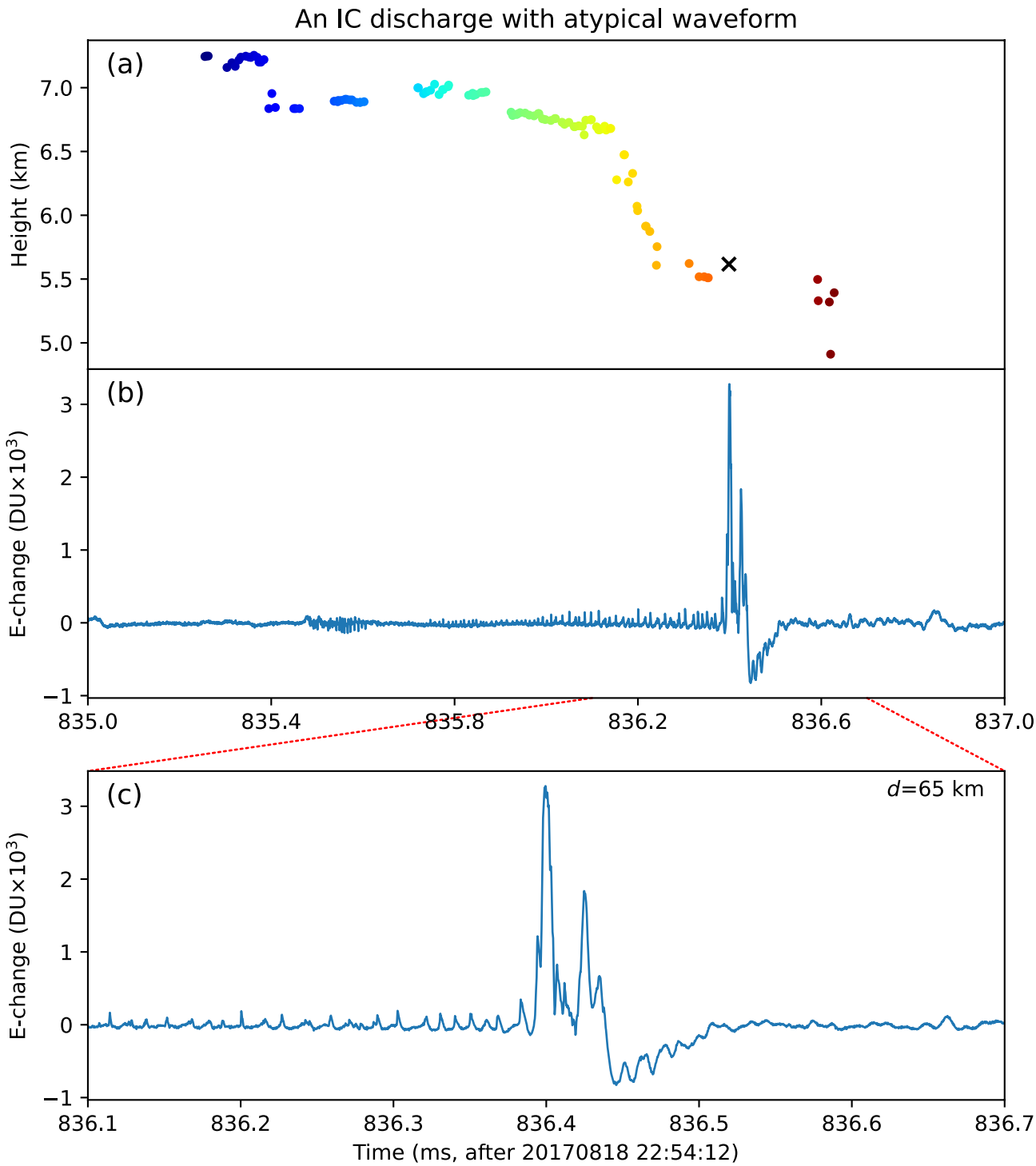

Figure 7. Location result and E-change waveforms of an intracloud (IC) pulse misclassified as a return stroke pulse. (a) Height-time location results of the IC pulse and preceding discharges. The cross sign represents the location of the IC pulse. (b) E-change waveform of the IC pulse and preceding discharges. (c) E-change waveform of the IC pulse. The value of *d* represents the distance between the IC discharge and the observation site recording the waveform.

Data Availability Statement

Data sets for building and testing the classifiers, waveform figures of all positive and negative RSs, data of the trained classifiers and a simple Python script demonstrating the usage of the classifiers can be found at https://doi.org/10.5281/zenodo.7900171 (Wu, 2023).

Acknowledgments
This study was supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (Grants 20H02129 and 21K03681).

References

Araki, S., Nasu, Y., Baba, Y., Rakov, V. A., Saito, M., & Miki, T. (2018). 3-D finite difference time domain simulation of lightning strikes to the 634-m Tokyo skytree. *Geophysical Research Letters*, *45*(17), 9267–9274. https://doi.org/10.1029/2018gl078214

Ballarotti, M. G., Saba, M. M. F., & Pinto, O. (2005). High-speed camera observations of negative ground flashes on a millisecond-scale. *Geophysical Research Letters*, *32*(23), L23802. https://doi.org/10.1029/2005gl023889

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Betz, H.-D., Schmidt, K., Oettinger, P., & Wirz, M. (2004). Lightning detection with 3-D discrimination of intracloud and cloud-to-ground discharges. *Geophysical Research Letters*, *31*(11), n/a. https://doi.org/10.1029/2004gl019821

Cai, L., Liu, W., Zhou, M., Wang, J., Yan, R., Tian, R., & Fan, Y. (2022). Differences of electric field parameters for lightning strikes on tall towers and nonelevated objects. *IEEE Transactions on Electromagnetic Compatibility*, *64*(6), 2113–2121. https://doi.org/10.1109/temc.2022.3207237

Cummins, K. L., & Murphy, M. J. (2009). An overview of lightning locating systems: History, techniques, and data uses, with an in-depth look at the U.S. NLDN. *IEEE Transactions on Electromagnetic Compatibility*, *51*(3), 499–518. https://doi.org/10.1109/temc.2009.2023450

DeFries, R. (2000). Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, *74*(3), 503–515. https://doi.org/10.1016/s0034-4257(00)00142-5

Fleenor, S. A., Biagi, C. J., Cummins, K. L., Krider, E. P., & Shao, X.-M. (2009). Characteristics of cloud-to-ground lightning in warm-season thunderstorms in the Central Great Plains. *Atmospheric Research*, *91*(2–4), 333–352. https://doi.org/10.1016/j.atmosres.2008.08.011

Kohlmann, H., Schulz, W., & Pedeboy, S. (2017). Evaluation of EUCLID IC/CG classification performance based on ground-truth data. In *2017 International symposium on lightning protection (XIV SIPDA)*. IEEE. https://doi.org/10.1109/sipda.2017.8116896

Kong, X., Qie, X., Zhao, Y., & Zhang, T. (2009). Characteristics of negative lightning flashes presenting multiple-ground terminations on a millisecond-scale. *Atmospheric Research*, *91*(2–4), 381–386. https://doi.org/10.1016/j.atmosres.2008.03.025

Leal, A. F., Rakov, V. A., & Rocha, B. R. (2019). Compact intracloud discharges: New classification of field waveforms and identification by lightning locating systems. *Electric Power Systems Research*, *173*, 251–262. https://doi.org/10.1016/j.epsr.2019.04.016

Lin, Y. T., Uman, M. A., Tiller, J. A., Brantley, R. D., Beasley, W. H., Krider, E. P., & Weidman, C. D. (1979). Characterization of lightning return stroke electric and magnetic fields from simultaneous two-station measurements. *Journal of Geophysical Research*, *84*(C10), 6307. https://doi.org/10.1029/jc084ic10p06307

Miao, X., Heaton, J. S., Zheng, S., Charlet, D. A., & Liu, H. (2011). Applying tree-based ensemble algorithms to the classification of ecological zones using multi-temporal multi-source remote-sensing data. *International Journal of Remote Sensing*, *33*(6), 1823–1849. https://doi.org/10.1080/01431161.2011.602651

Paul, C., Heidler, F. H., & Schulz, W. (2020). Performance of the European lightning detection network EUCLID in case of various types of current pulses from upward lightning measured at the Peissenberg tower. *IEEE Transactions on Electromagnetic Compatibility*, *62*(1), 116–123. https://doi.org/10.1109/temc.2019.2891898

Pavanello, D., Rachidi, F., Janischewskyj, W., Rubinstein, M., Hussein, A. M., Petrache, E., et al. (2007). On return stroke currents and remote electromagnetic fields associated with lightning strikes to tall structures: 2. Experiment and model validation. *Journal of Geophysical Research*, *112*(D13), D13122. https://doi.org/10.1029/2006jd007959

Pu, Y., Cummer, S. A., Lyu, F., Zheng, Y., Briggs, M. S., Lesage, S., et al. (2023). Unsupervised clustering and supervised machine learning for lightning classification: Application to identifying EIPs for ground-based TGF detection. *Journal of Geophysical Research: Atmospheres*, *128*(9), e2022JD038369. https://doi.org/10.1029/2022jd038369

Rison, W., Thomas, R. J., Krehbiel, P. R., Hamlin, T., & Harlin, J. (1999). A GPS-based three-dimensional lightning mapping system: Initial observations in central New Mexico. *Geophysical Research Letters*, *26*(23), 3573–3576. https://doi.org/10.1029/1999gl010856

Schulz, W., Diendorfer, G., Pedeboy, S., & Poelman, D. R. (2016). The European lightning location system EUCLID – Part 1: Performance analysis and validation. *Natural Hazards and Earth System Sciences*, *16*(2), 595–605. https://doi.org/10.5194/nhess-16-595-2016

Sun, Z., Qie, X., Liu, M., Jiang, R., Wang, Z., & Zhang, H. (2016). Characteristics of a negative lightning with multiple-ground terminations observed by a VHF lightning location system. *Journal of Geophysical Research: Atmospheres*, *121*(1), 413–426. https://doi.org/10.1002/2015jd023702

Thottappillil, R., Rakov, V. A., & Uman, M. A. (1997). Distribution of charge along the lightning channel: Relation to remote electric and magnetic fields and to return-stroke models. *Journal of Geophysical Research*, *102*(D6), 6987–7006. https://doi.org/10.1029/96jd03344

Wang, J., Huang, Q., Ma, Q., Chang, S., He, J., Wang, H., et al. (2020). Classification of VLF/LF lightning signals using sensors and deep learning methods. *Sensors*, *20*(4), 1030. https://doi.org/10.3390/s20041030

Wu, T. (2023). High-accuracy classification of radiation waveforms of lightning return strokes (version 2) [Dataset]. Zenodo. https://doi.org/10.5281/ZENODO.7900171

Wu, T., Wang, D., Huang, H., & Takagi, N. (2021). The strongest negative lightning strokes in winter thunderstorms in Japan. *Geophysical Research Letters*, *48*(21), e2021GL095525. https://doi.org/10.1029/2021gl095525

Wu, T., Wang, D., & Takagi, N. (2018a). Lightning mapping with an array of fast antennas. *Geophysical Research Letters*, *45*(8), 3698–3705. https://doi.org/10.1002/2018gl077628

Wu, T., Wang, D., & Takagi, N. (2018b). Locating preliminary breakdown pulses in positive cloud-to-ground lightning. *Journal of Geophysical Research: Atmospheres*, *123*(15), 7989–7998. https://doi.org/10.1029/2018jd028716

Wu, T., Wang, D., & Takagi, N. (2019). Velocities of positive leaders in intracloud and negative cloud-to-ground lightning flashes. *Journal of Geophysical Research: Atmospheres*, *124*(17–18), 9983–9995. https://doi.org/10.1029/2019jd030783

Wu, T., Wang, D., & Takagi, N. (2021). Compact lightning strokes in winter thunderstorms. *Journal of Geophysical Research: Atmospheres*, *126*(15), e2021JD034932. https://doi.org/10.1029/2021jd034932

Wu, T., Wang, D., & Takagi, N. (2022). On the intensity of first return strokes in positive cloud-to-ground lightning in winter. *Journal of Geophysical Research: Atmospheres*, *127*(22), e2022JD037282. https://doi.org/10.1029/2022jd037282

Zhu, Y., Bitzer, P., Rakov, V., & Ding, Z. (2021). A machine-learning approach to classify cloud-to-ground and intracloud lightning. *Geophysical Research Letters*, *48*(1), e2020GL091148. https://doi.org/10.1029/2020gl091148

Zhu, Y., Rakov, V. A., Tran, M. D., Lyu, W., & Micu, D. D. (2018). A modeling study of narrow electric field signatures produced by lightning strikes to tall towers. *Journal of Geophysical Research: Atmospheres*, *123*(18), 10,260–10,277. https://doi.org/10.1029/2018jd028916

Zhu, Y., Rakov, V. A., Tran, M. D., & Nag, A. (2016). A study of national lightning detection network responses to natural lightning based on ground truth data acquired at LOG with emphasis on cloud discharge activity. *Journal of Geophysical Research: Atmospheres*, *121*(24), 14651–14660. https://doi.org/10.1002/2016jd025574

Zhu, Y., Stock, M., Lapierre, J., & DiGangi, E. (2022). Upgrades of the Earth networks total lightning network in 2021. *Remote Sensing*, *14*(9), 2209. https://doi.org/10.3390/rs14092209