Lightning Distance Estimation Using LF Lightning Radio Signals via Analytical and Machine-Learned Models

Andre L. Antunes de Sá, *Student Member, IEEE*, and Robert A. Marshall, *Member, IEEE*

***Abstract*— Lightning geolocation is useful in a variety of applications, ranging from weather nowcasting to a better understanding of thunderstorm evolution processes and remote sensing of the ionosphere. Lightning-generated radio signals can be used in range estimation of lightning return strokes, for which the most commonly employed technique is the time difference of arrival in lightning detection networks. Though these instrument networks provide the most reliability and best accuracy, users without access to them can instead benefit from lightning geolocation using a standalone instrument. In this article, we present the framework for training fast models capable of estimating negative cloud-to-ground lightning location from single-instrument observations of very low frequency/low frequency (VLF/LF, 3–300 kHz) radio pulses or "sferics," without knowledge of the ionosphere's D-region state. The models are generated using an analytical method, based on the delay between ground and skywave, and a machine learning method. The training framework is applied to three different data sets to assess model accuracy. Validation of the machine-learned models for these data sets, which include both simulated and observed sferics, confirms this technique as a promising solution for lightning distance estimation using a single receiver. Distance estimates using a machine-learned model for observed sferics in Kansas yield an RMSE of 53 km with 68% of them being within 9.8 km. Estimates using the analytical method are found to have an RMSE of 54 km with 68% of them being within 32 km. Limitations of our methodology and potential improvements to be investigated are also discussed.**

***Index Terms*— Distance measurement, inverse problems, ionosphere, lightning, low-frequency, multilayer perceptrons (MLPs), neural networks, radio remote sensing, single station.**

I. Introduction

LIGHTNING return stroke geolocation has brought advantages to many industrial fields and commercial groups such as electric power utilities, aviation, forestry, and weather forecasting/nowcasting [1], through the use of time difference of arrival (TDOA) real-time detection networks such as the National Lightning Detection Network (NLDN) [2], the World Wide Lightning Location Network (WWLLN) [3], the Global Lightning Dataset (GLD360) [4], the British Meteorological Office's lightning geolocation network [5], and the University of Frankfurt's global lightning triangulation using extremely low frequency (ELF) (0.3–3 kHz) waves [6]. Real-time lightning detection and geolocation provides critical information on mitigating hazards posed by lightning [7] and the severe weather associated with it [8], i.e., heavy precipitation, hail, strong convection, and high wind shear. Lightning geolocation information has been instrumental to recent developments in lightning physics, including observations of terrestrial gamma-ray flashes [9], [10]. Also, lightning geolocation has a significant influence in the study of prevailing scientific questions surrounding the climatology of thunderstorms, and lightning's relationship to higher altitude geophysical systems, such as the variation of electron density in the D-region of the ionosphere [11].

A network data connection is an essential requirement for the aforementioned real-time detection services, greatly reducing their suitability for users at more remote locations with expensive or nonexistent data channels. Commercial and general aviation, remote scientific operations, the maritime sector, and even local community events planning can benefit from a standalone device capable of detecting and locating lightning events either for lightning and severe weather avoidance or for decision-making on mission operations involving lightning research.

The self-contained lightning location device suggested in this article operates on similar lightning radio observations, "sferics," as that of the current global lightning networks; these sferics can be acquired by compact, low-cost magnetic loop antennas. Alongside the antenna and data processing unit, the portable instrument requires a model for deducing lightning location from the observations. Determining lightning direction can be accomplished through magnetic direction finding (MDF) [12]–[14], where azimuthal data are derived from sferics acquired by two orthogonal antennas. On the other hand, solutions for range measurements of lightning are more diverse, each with a different set of limitations, and, though currently accurate enough for lightning early detection, these systems will become better suited for navigation, meteorology, and space research if their fidelity can match more closely that of multiple-site networks.

Among the single-site methods for estimating lightning distance, the most prominent approach is to treat the Earth–Ionosphere waveguide as the duct between two circular

Manuscript received May 29, 2019; revised September 30, 2019 and January 12, 2020; accepted January 29, 2020. Date of publication February 19, 2020; date of current version July 22, 2020. This work was supported in part by the Integrated Remote and In Situ Sensing (IRISS) Program through the University of Colorado Boulder and in part by NSF under Grant AGS 1661726. The work of Andre L. Antunes de Sá was supported by the NASA Earth and Space Science Fellowship under Grant 80NSSC17K0392. *(Corresponding author: Andre L. Antunes de Sá.)*

The authors are with the Smead Aerospace Engineering Sciences Department, University of Colorado Boulder, Boulder, CO 80309 USA (e-mail: andre.antunesdesa@colorado.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2972153

shells and assume perfect ionosphere reflections, in order to infer the return stroke distance from the delay between the main sferic and its ionospheric reflections. The biggest limitation of this approach is the required knowledge of the D-region ionosphere. This method was described by Smith *et al.* [15] for a class of in-cloud lightning known as narrow bipolar events, and employed by Zhang *et al.* [16] and Cummer *et al.* [17]. Similarly, the approach is also commonly used to estimate the reflection height of the D-region ionosphere when the return stroke distance is known [18], [19].

A similar and arguably more developed approach for long-distance lightning, beyond 2000 km, involves using the delay between the start of the very low frequency (VLF) "sferic" and ELF "slow-tail" [20], where knowledge of the ionosphere is still assumed and an ELF/VLF antenna is required. The best implementation of this approach is described in [21] and [22], with a model fractional error mean of 0.76% and a standard deviation of 9.22% for daytime ionospheres and fractional error mean of 2.79% and standard deviation of 8.52% for nighttime.

Yet another approach, known as the "Kharkov" method, uses the phase spectrum of the transverse electric (TE) first-order waveguide mode to estimate lightning distance [23], [24], which requires two crossed magnetic antennas and one electric antenna sensitive to 1.8–3.2 kHz. The Kharkov method was originally claimed to have a model uncertainty of 5% for distance estimation depending on the ionospheric model used. This approach was validated by Brundell *et al*. [25], who showed that for a data set of summer nighttime sferics within 2000 km from the receiver, a model error standard deviation of 72–73 km was found depending on the ionospheric model used. This method requires many computations of the first-order phase spectrum for a least-squares fit to the data, which can be computationally expensive, especially for a real-time estimate, even when using the simplest parallel-plate ionospheric model.

Commercial standalone avionics also exist, such as the *Strike Finder* [26] and *Stormscope* [27] instruments. Their performance characteristics are not readily available, but they detect lightning within a range of about 200 km, with decreasing precision as return stroke distance increases.

This article aims to provide a proof of concept and preliminary model error performance estimates for a machine-learned model capable of estimating range from lightning sferics, which could be used in a standard, low-cost, VLF/low-frequency (VLF/LF, 3–300 kHz) receiver. Additionally, a detailed analysis of the best case implementation of an analytical range-finding method, based on the delay between the main sferic and its ionospheric reflections, is presented in Section II and later compared to the machine-learned models. Both methods are examined in the context of existing networks and their performance, such as NLDN's 300-m accuracy [28] and GLD360's ∼2 km accuracy [4], as well as to the single-site accuracy values referenced above. The use of higher frequency sferics allows for simpler and more compact design of the antennas, which requires fewer turns and smaller cross-section area for the desired sensitivity and bandwidth [29], while the major drawback of higher attenuation in these frequencies does not affect our chosen region of interest, 100–700 km.

A. Background

The large currents produced in the return strokes of lightning generate broadband electromagnetic pulses. In the VLF/LF range, these pulses are known as sferics (short for "radio atmospherics"), and they contain information about the generating lightning return strokes (e.g., 30). These sferics can be observed at hundreds to thousands of kilometers from the source as the VLF components propagate without major attenuation in the Earth–Ionosphere waveguide [31], whereas the LF components decay after ∼1000 km; as such the LF signals are most suitable for lightning location within approximately 1000 km from the observer.

Lightning sferics can be obtained with a magnetic flux sensor, such as a wire loop antenna, where two orthogonal antennas can be used for direction finding. Polarization errors in this approach are described by Yamashita and Sao [32], [33], and improvements over the basic approach are to only use the initial part of the sferic for direction finding [12], [13]. Note that if the signal polarity is not known, such as when differentiating between positive and negative return strokes, there is a 180° ambiguity in the calculated direction. A vertical electric field antenna can be added to the receiver to resolve the 180° ambiguity and decrease errors due to polarization through the computation of the Poynting vector [23]. A ferromagnetic core can be used to increase the magnetic permeability of the loop antennas, then referred to as search coils, at the expense of a nonlinear frequency response but allowing for a smaller antenna with the same gain. The transfer function nonlinearity of this antenna type can be corrected or addressed and still provide a size advantage over the air-core antenna at the cost of more complex manufacturing [29].

A simulated cloud-to-ground (CG) return stroke sferic, from the data set described in Section I-C1, is shown in Fig. 1, with a diagram illustrating the geometry of the system. Some major features observed in this sferic, which are directly related to properties of the originating return stroke, include the peak B-field; the ground wave, direct from the source to the receiver; and one or more sky waves, either reflected by the ionosphere directly or reflected by the ground and the ionosphere (for in-cloud lightning). While there is a closed-form solution for extracting lightning range information for some of these observation features, it is only valid with ELF/VLF signals [ELF "Slow Tail" 20]. Currently, single-site lightning location in the time-domain relies on some of these waveform physical features to approximate an inverse model capable of recovering the range to the causative lightning return stroke, e.g., the time delay between the ground wave and the sky wave(s), the ELF "Slow Tail," or even the peak B-field [e.g., 14, 35].

B. Sferic Forward Model and Distance Estimation

The lightning-generated sferics and their attenuation through the Earth–Ionosphere waveguide define the overall

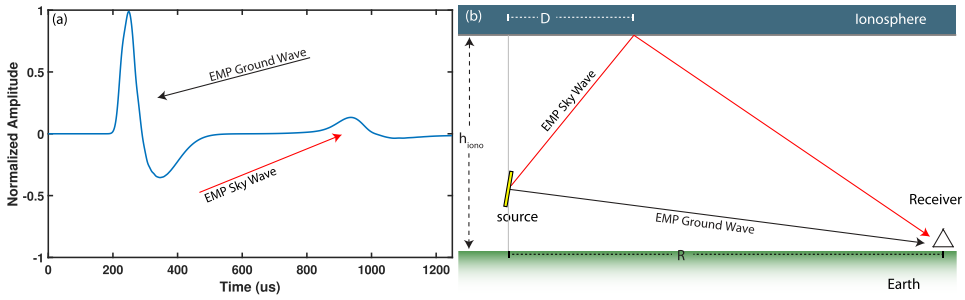

Fig. 1. (a) Simulated sferic for a 100-km distant CG strike. (b) Illustration of the lightning electrogmagnetic pulse (EMP) propagation path toward the receiver, adapted from [34].

physical system and can be represented in the standard observation equation

$$y = \mathcal{H}(x) \tag{1}$$

where the properties of the generating lightning return stroke, which affect the observations, are captured by x, and the forward model $\mathcal{H}$ maps these states into the observed sferic, y. The properties of the generating return stroke, x, are diverse and include the distance to the observer, d; the discharge type, e.g., CG or in-cloud; the peak current and current waveform; and the state of the D-region ionosphere, which is often parameterized by a reflection height h' and a sharpness parameter β [36]. The observation, y, is a time series, usually spanning less than a few milliseconds, and contains the propagated waveform signature from the lightning return stroke at the observer's location, i.e., the sferic. Note that there are many other processes that also affect the sferics and ranging estimates but are neglected in this study due to their second-order nature and a small contribution to the current model errors. These processes include the difference in eastward versus westward propagation of sferics due to the anisotropy of the Earth–Ionosphere waveguide, and distance-dependent polarization given the dispersive nature of the Earth–Ionosphere waveguide [31].

The forward model $\mathcal{H}$ has been successfully employed before using finite-difference time-domain (FDTD) models of the lightning–ionosphere interaction [37]. In this scenario, the forward model can be described through a collection of physical models including Maxwell's equations combined with the Langevin equation describing the motion of the cold, collisional magnetized plasma of the lower ionosphere (e.g., [37]). As such, the direct inversion of this forward model is extremely difficult.

In order to extract the distance d between the receiver and causative discharge, we need to determine the inverse model, $\mathcal{G}_d$

$$d = \mathcal{G}_d(y). \tag{2}$$

However, given that the observation model is not invertible, we must approximate $\mathcal{G}_d$ through other means, such as using the physical waveform features to inform an inverse model, with a possible methodology described in Section II. Alternatively, in this article, we describe an inverse model that uses the complete time-series of the sferic in a machine-learned neural network model. The method has the advantage that it does not require the identification of sferic features, such as a skywave; it provides a distance estimate and an uncertainty for each sferic; it requires little processing power once the model is trained; it is trained to be independent of the ionospheric D-region state; and it is readily applicable in a real-time system.

C. Data

Three distinct data sets are used to assess the viability of the two methods presented in this article for lightning distance estimation. These data sets have different noise characteristics, and without some instrument model to generalize them, any empirical model for one data set cannot be used with a different data set, without significant estimate errors. In order to circumvent this limitation, an operational approach should include a calibration routine that will fit or train a model for each single-site instrument, eliminating the need for a general model for every instrument installation.

The magnitude of the lightning radio signal observed by two orthogonal antennas is normalized through standardization, i.e., mapped to have zero mean and unit variance. Only negative return strokes are used in this research due to the lack of enough positive strokes for proper model training, as discussed in Section III-C. The signal-to-noise ratio (SNR) for the observed data sets below is computed from the peak of each waveform divided by the root-mean-squared value of the first one hundred points, where the ground wave peak is centered around the 207th point, before standardization and truncation of the data. The data sets include both daytime and nighttime observations, except for the Udall data set.

1) Simulated Sferics: A simulated sferic data set is generated specifically for assessing the error performance of a lightning distance estimation model in the absence of any random and systematic error that might be present in empirical data sets. Using an FDTD model [37], 800 simulations of 100-kA CG return strokes are analyzed between 100 and 700 km from the source at 1-km intervals. The discretization time step is 0.1 μs, and the source current is low-pass filtered to 500 kHz. The source is in Huntsville, Alabama (Lat 44.6464°, Lon −67.2811°), and decay is given by the modified transmission line model with linear current decay with height (MTLL) [38]. Each simulation uses a randomly sampled realization of seven input variables from uniform

TABLE I

SEVEN RANDOM INPUTS TO THE SIMULATIONS AND THE DISTRIBUTIONS FROM WHICH THEIR VALUES ARE SAMPLED

τ_r	$\sim U(5, 20)$ µs
τ_f	$\sim U(30, 100)$ µs
l_{ch}	$\sim U(2, 10)$ km
h'	$\sim \begin{cases} U(72, 76) \text{ day} \\ U(83, 88) \text{ night} \end{cases}$
β	$\sim \begin{cases} U(0.3, 0.5) \text{ day} \\ U(0.5, 0.9) \text{ night} \end{cases}$
V_{rs}	$\sim U(0.3c, 0.8c)$

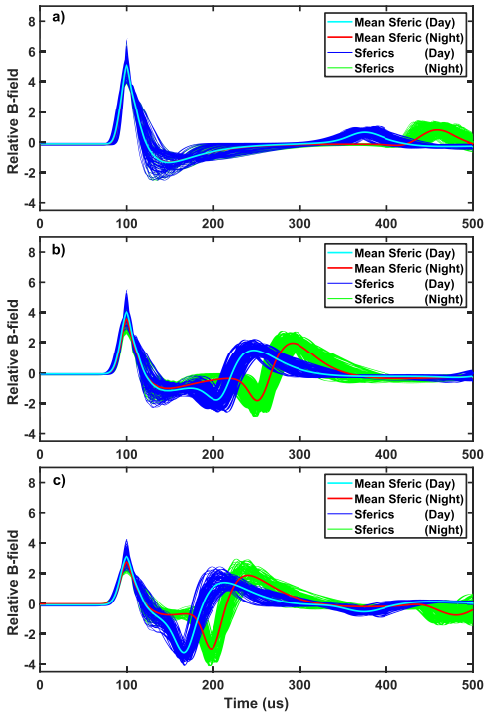

Fig. 2. Plots of sferics at three different distances from the observer for all 800 simulations, and their corresponding mean. (a) Simulated sferics at 100 ± 3 km. (b) Simulated sferics at 300 ± 3 km. (c) Simulated sferics at 500 ± 3 km.

distributions that are representative of all probable states of the local environment (Table I), where the variables are: pulse rise time, τ_r; fall time, τ_f; channel length for MTLL, l_{ch}; ionospheric parameters, h' and β; and the return stroke propagation speed, V_{rs}. Although these uniform distributions are not realistic, e.g., climate system properties and seasonal variations skew some of these nonuniformly [39], [40], they cover the span of all possible states uniformly, which is critical for the generalization of a machine-learned model (see Section III-A). Of the 800 simulations, half of them use daytime ionosphere parameters, $h' = 72$–76 km and $\beta = 0.3$–0.5, while the other half use nighttime parameters, $h' = 83$–88 km and $\beta = 0.5$–0.9.

The simulated data set is characterized in Fig. 2, where a subset of sferic data is plotted for three particular distances from the source. Note that the distance distribution for this set is uniform, and that while all return strokes have the same peak current, adding more simulations with different return stroke currents is not expected to change the standardized sferics in any significant way, since the distant field is linearly proportional to the peak current. The predominant groundwave and skywave features are present in the sferics. There is significant variation in the sferics depending on the sampled random inputs, but the mean is representative of a canonical sferic at different distances for the given random input distributions. The average decrease in peak B-field with distance is seen even after standardization, given some correlation between the two peak heights and the variance, however, the true peak B-field value for each sferic, constant for all simulations at each 1-km distance, is lost in the standardization and would not, by itself, reliably inform a model for lightning distance estimation.

2) Udall Data: The next data set is composed of sferics extracted from radio observations in Udall, Kansas, using a version of the AWESOME receiver described by Cohen *et al.* [41], modified for LF frequencies (3–300 kHz), during four nonconsecutive days of significant lightning activity in early August 2013, between 22:00 and 05:00 local time. Sferics are extracted from the raw data with the use of a matched filter, and as with any other single-site method available, overlapping sferics cannot be processed individually and become a single sferic event. Lightning event data from the NLDN is matched to the sferic observations providing information on the observed sferics' originating lightning return stroke, such as the true distance, necessary for training and validation, and other useful parameters such as return stroke peak current. Lightning return strokes with NLDN-observed peak current smaller than -10 kA and distances between 90 and 710 km from the receiver are included in this data set, which consists of 128 461 sferic observations, and of these 55% are from intracloud (IC) lightning and 45% are from CG lightning, as reported in the NLDN data.

The Udall data set is characterized in Fig. 3, where a subset of sferic data is plotted for three particular distances from the source. Similar observations can be made compared to the simulated sferics, i.e., ground and sky wave peaks can be identified with a clear relationship between the time delay and the originating return stroke's distance, but also present is the receiver and environment noise, which can include other lightning-generated signals not associated with each other. Fig. 4 shows distributions of the distance, peak current, and SNR, plotted individually and in color maps of event distribution, as well as the SNR distribution with respect to peak current and distance. Most relevant to the model error performance below, the distance distribution is nonuniform, with a large peak at around 600 km, a valley at around 300 km, and a more uniform region between 100 and 300 km. Fig. 5 displays the location density of return strokes associated with this data set. The range-azimuth distribution is not uniform and there are many gaps in location coverage, which is expected to affect the model training. The mostly uniform topology of the region around the receiver relaxes the coverage requirement to some extent. As expected, the majority of events (Fig. 4) are of low peak current and low SNR. Naturally, peak current and SNR seem to correlate well, but also note that for distance-peak current combinations with a lower

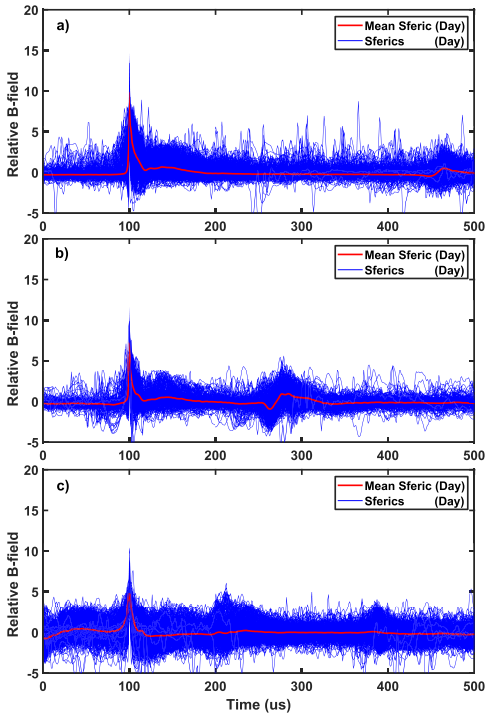

Fig. 3. Plots of around 1000 sample sferics from the Udall data set at different distances from the observer and their corresponding mean. (a) Udall sferics at 100 ± 3 km. (b) Udall sferics at 300 ± 3 km. (c) Udall sferics at 500 ± 3 km.

density of events, the average SNR is higher than would be otherwise.

3) Boulder Data: The third data set consists of sferics from radio observations in Boulder, CO, using a similar receiver as the one at Udall, during 25 days in June and July 2017. The data are of lower quality than the Udall data set due to a combination of hardware issues and environment noise. This lower-quality data set is intentionally used here to explore the use of the neural network model training for noisier sferics, i.e., higher noise floor and a larger number of spurious signals. Again, the data are matched with NLDN truth data, and 63% of sferics are IC and 37% are CG. Due to the sensitivity of the instrument and the lower SNR data, far fewer sferics could be extracted at any one time; the total data set consists of 21 881 sferics paired with their corresponding NLDN data, of which 6250 are between 8:00 and 18:00 local time, and 15 631 are between 20:00 and 06:00 local time.

The Boulder data set is characterized in Fig. 6, where a subset of sferic data is plotted for three distance ranges. Fig. 7 shows distributions of the distance, peak current, and SNR plotted individually and in color maps of event distribution, and the SNR distribution with respect to peak current and distance. Most importantly, the SNR values are far lower than in the Udall data set. As with the Udall data, the distance distribution is nonuniform, with few events less than 300 km and few events at around 550 km. On the location map of the sferics, Fig. 8, while it can be seen that locations are significantly not balanced, at least there is better coverage of the region than in the Udall data set. Also note that the topology around the Boulder receiver is less uniform than around the Udall receiver, with the Rocky Mountains to the west.

II. Skywave Delay Method and Performance

An approximate surrogate model of $\mathcal{G}_d$ for estimating lightning distance can be analytically obtained from the delay between the direct groundwave and the reflected skywave, a major physical feature of sferics, as discussed in the introduction. This geometric approach is illustrated in Fig. 1(b); most importantly, the method requires some knowledge of the D-region ionosphere, providing the altitude and refraction angle of sferics. In this section, the skywave delay approach for lightning distance estimation is presented along with its performance error when applied to different data sets. Regardless of the refraction model used, the dependence on the state of the ionosphere will be significant, and while the approach can be made more complex to minimize errors due to the influence of the D-region ionosphere, this dependence is intrinsic and a limiting systematic error. Identification of skywaves in the sferics is another limiting factor, as sferics without a discernible skywave must be discarded, lowering the overall detection efficiency.

The simulated lightning data set of CG sferics with varying ionospheric states, described in Section I-C1, is used in assessing the best possible error performance for the skywave delay model, for noiseless data with discernible skywaves and a range of realistic ionospheric states. The Udall data set (Section I-C2) is also used in assessing the error performance of the skywave delay method, to validate the most important results of the study with the simulated data set, such as the characterization of the method's limitations.

The ground and skywave are identified as the two tallest peaks in each sferic's observation window, separated by at least 50 μs, and the time difference between them is recorded as the skywave delay. These peaks are shown for a few sample sferics in Fig. 9(a) and (b), for the nighttime simulated and Udall data sets, respectively. The computed delays form skywave-distance curves for each of the ionospheric states represented in the data set, seen in Fig. 9(c) and (d). The nighttime simulated data set presents complete curves of 400 realizations of ionospheric states, since for every realization there are observations at all distances from the source. The Udall data set includes much fewer distance observations for certain ionospheric states, and so only partial curves for the theoretical relationship is visible. The Udall data set also contains a large number of incorrect estimations of the skywave delay, mainly caused by the skywave peak being lower than the noise floor or the presence of any spurious signals with a taller peak than the skywave, including secondary reflections of the signal.

Although the simple computation method described above can likely be improved to better handle the incorrect estimations of the skywave delay, none of these possible improvements would decrease the uncertainty band of the skywave-distance curves since they are caused by the dependence on the ionospheric D-region state. In order to focus on the main limitation of the skywave delay method, which is uncertainty in the ionospheric state, sferics from the Udall

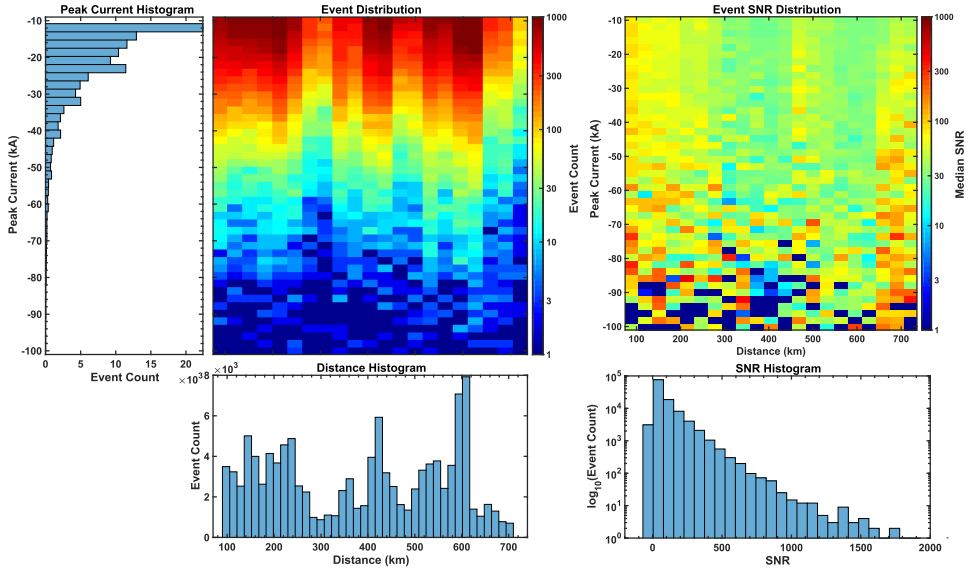

Fig. 4. Plots of the distance, peak current, and SNR individual distributions, and color maps of event distribution and SNR distribution with respect to peak current and distance, for the Udall data.

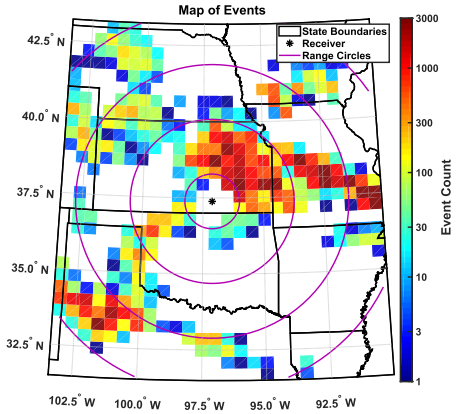

Fig. 5. Location map for lightning associated with the sferics in the Udall data set. The receiver location, and range circles of 100, 300, 500, and 700 km from the receiver are shown for reference.

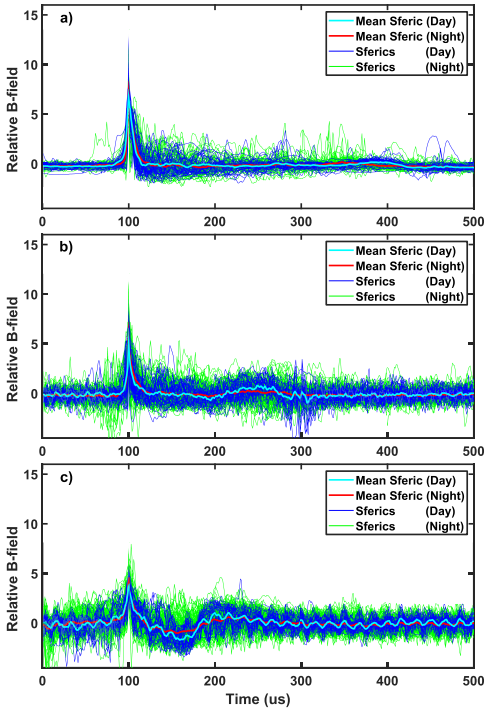

Fig. 6. Plots of around a hundred sample sferics from the Boulder data set at different distances from the observer and their mean. (a) Boulder sferics at 100 ± 3 km. (b) Boulder sferics at 300 ± 3 km. (c) Boulder sferics at 500 ± 3 km.

data set are filtered to include only the correctly identified skywave delays as given by the empirical spread of the skywave-distance curves, bypassing any error caused by the process of identifying the skywave delay on real data. The final error performance discussed is thus a projection of the best possible implementation of the skywave delay method, where the identification of skywave peaks is near perfect, and might actually not be possible to implement. The data filtering is accomplished using an envelope defined by top and bottom skywave-distance curve boundaries, which are spread out from a center skywave-distance curve. The envelope center and spread are determined by the distribution of skywave delays at 6-km binned distances, where 1) the delay distribution peak for each distance determines the envelope center and 2) two averages of the peak's full-width at half-maximum, for distances before and after 470 km, define the spread of the envelope. Given the significant spread in these curves, a large error is expected when estimating lightning distance using the skywave delay alone and without knowledge of the D-region ionosphere. Although there is general agreement between the skywave delay curves for simulated and real data sets, the spread is smaller for the real data set, likely due to a more limited set of ionospheric states captured in the four nights of sferic data for this data set.

A fifth-order polynomial is fit to the skywave delay for different lightning distances and its error performance histogram (true distance minus estimated) is shown in Fig. 9(e) and (f).

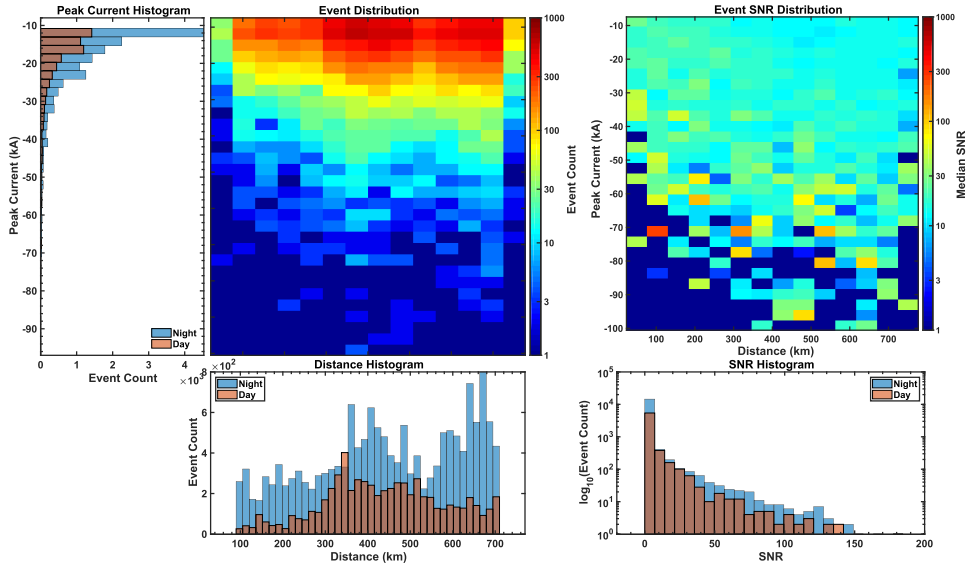

Fig. 7. Plots of the distance, peak current, and SNR individual distributions, and color maps of event distribution and SNR distribution with respect to peak current and distance, for the Boulder data.

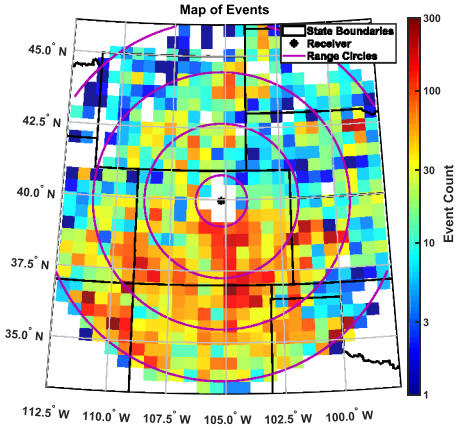

Fig. 8. Location map for lightning associated with the sferics in the Boulder data set. The receiver location, and range circles of 100, 300, 500, and 700 km from the receiver are shown for reference.

Table II further tabulates the performance for later comparison with the models trained using machine learning. The error distribution, which closely resembles a normal distribution for the simulated data sets, has a root-mean-square error of 58.2 km for daytime simulated data, 38.4 km for nighttime simulated data, 79 km for the whole simulated data set, and 54.2 km for the Udall data set (nighttime). The error distribution projected for the Udall data set does not resemble a Gaussian distribution as much as the other distributions, being narrower at the top and wider at the base, and so the percentile edges of 32.12 km (68th percentile), 120 km (95th percentile), and 198 km (99.7th percentile) represent the error performance better. As expected, the ionosphere-driven spread in skywave plots as a function of source distance, shown in Fig. 9(c) and (d), inherently limits the precision of a skywave delay distance estimation method, even if perfect identification of ground and skywave peaks can be achieved. The error on the whole simulated data set is much larger than on day or night alone since the spread in skywave-distance curves is even larger. Also note that the model performance built and validated using the Udall data set has better error performance than the nighttime simulated data set, as predicted from the smaller curve spread for the Udall data set in Fig. 9(d) discussed previously.

In Sections III and IV, a machine learning approach to the lightning distance estimation problem is presented to address the ionospheric knowledge limitation with an improved error performance, while also eliminating the need for identification of skywave peaks.

III. Machine Learning Methodology

A novel approach for estimating an inverse model involves training a surrogate machine-learned model using artificial neural networks (ANNs), which have become increasingly powerful in the last few years [42], [43]. Machine learning grants many advantages in finding an inverse model, including their capacity for generalization, discussed in the next section, and their fast run time once trained, since they use simple arithmetic operations, as opposed to intense computations and/or matrix inversions required by more complex inverse methods. The model learning process in ANNs, through minimization of a cost function, can exploit instructive information that may not be clearly accessible, if accessible at all, through the extraction of known physically relevant features in observations, such as the time delay between the ground and sky waves in the waveform of sferics. Thus, a machine learning approach to inverse modeling can expand the practical observability of a problem. The inverse modeling requires the use of truth data, known as supervised learning, to accurately approximate the desired function.

A. Generalization

An important advantage of machine learning is its intrinsic approach to model generalization for the desired domain,

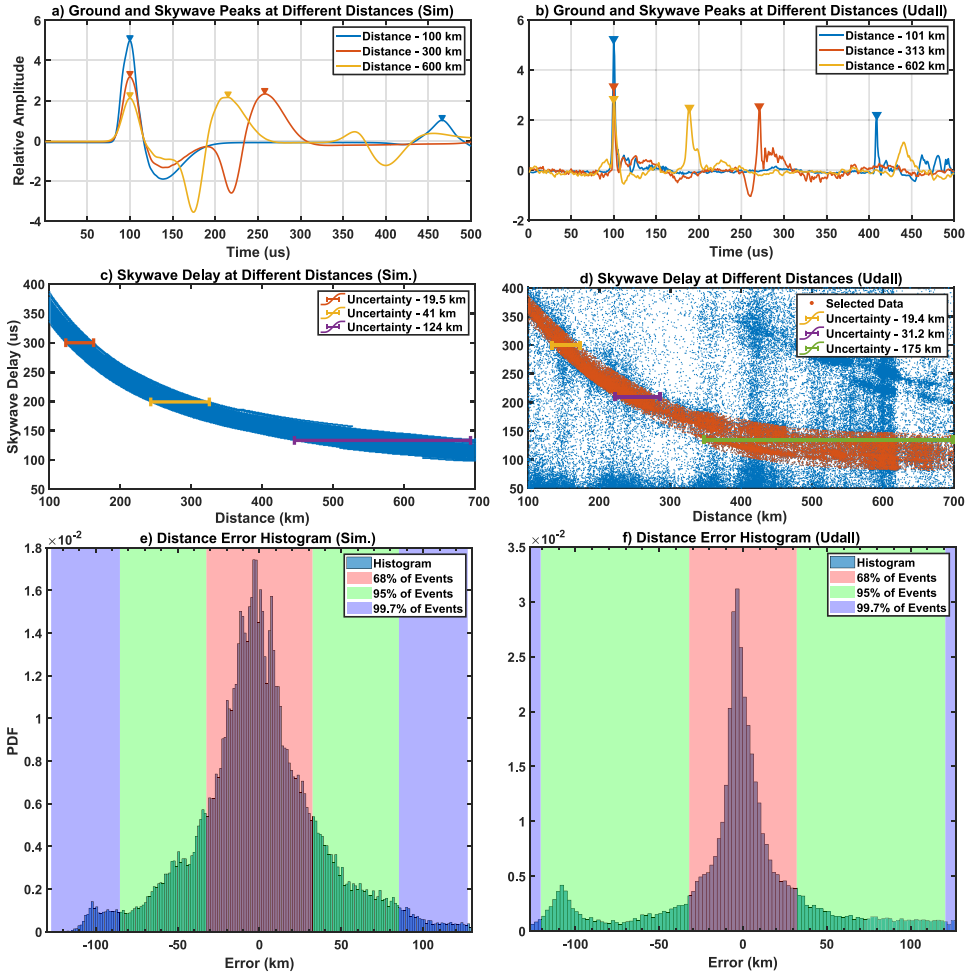

Fig. 9. The top plots, (a) and (b), illustrate the peaks used in finding the delay between ground and skywave, for a few sample sferics at different distances from the observer. Note that the Udall data do not contain lower frequencies present in the simulated data due to the receiver's frequency response. The middle plots, (c) and (d), show the relationship between skywave delay and distance, which can be described as a specific curve for a given ionospheric state. The bottom plots, (e) and (f), display the estimated error distribution accompanied by coverage shadings to indicate the percentage of events that lies in each shaded band. Plots on the left correspond to the nighttime simulated sferics data set, while plots on right to the Udall data set of nighttime sferics. Note that plot (f) corresponds to a projection of the best case implementation of the skywave delay method, with a near-perfect measurement of the skywave delay.

including a domain with noise that could have been intractable through other simpler estimation methods, but where the desired output is still observable to the learning process. Similarly, the model can also be trained to be independent of otherwise critical inputs in the data, as long as they are themselves observable separately from the desired input. For example, in lightning distance estimation using sferics, a neural network model can be trained to be independent of ionospheric parameters, h' and β, as long as there are observations more strongly correlated with the ionospheric parameters than to lightning distance or vice versa. However, some care must be taken in properly preparing the input data for the model pre- and posttraining, while understanding the domain of the trained model in terms of its generalization is critical in its application.

Naturally, there will be cases where generalization of the model for the desired domain cannot be achieved by machine learning, e.g., when there are not enough observations or they are not representative of the desired domain. This is increasingly common for higher complexity models as the density of the training samples is inversely proportional to model complexity. Here, it is best to divide the desired domain wherever possible and train a model for each separate domain, which introduces a new requirement of choosing one of the trained models for use with any given data. This approach is especially useful when working with wildly different environments, such as day/night, land/sea, or positive/negative return stroke, where specifically trained models can be used depending on some selection criteria in real time instead of tasking the learning algorithm with a wide generalization domain.

Another approach that could improve the training of an all-purpose model, i.e., wide domain, without requiring real-time choice of the trained model, would be the use of bootstrap aggregation (bagging) [44]. Bagging takes an ensemble of models trained through bootstrapping (see Section III-D) and uses the ensemble's majority vote on the estimation of a target. This technique can greatly decrease variance in the

TABLE II

SUMMARY OF THE ESTIMATED ERROR PERFORMANCE, IN km, FOR THE MODELS CORRESPONDING TO VARIOUS DATA SETS. THE PERCENTILES GIVEN ARE CENTERED ON EACH MEAN. THE FIRST FOUR RESULTS ARE FOR THE SKYWAVE DELAY METHOD DEVELOPED IN SECTION II

Dataset	**RMS**	Mean	Std. Dev.	**68th %**	95th %	99.7th %
Skywave Delay Models						
Simulated (Day)	**58.2**	0.00	58.2	**51.1**	121	204
Simulated (Night)	**38.4**	0.00	38.4	**32.5**	85.2	127
Simulated (Both)	**79.0**	0.00	79.0	**81.3**	150	206
Projected Udall (Night)	**54.20**	0.00	54.20	**32.12**	120	198
Machine-Learned Models						
Simulated (Day)	**2.35**	-0.50	2.29	**2.27**	4.36	7.61
Simulated (Night)	**2.59**	-0.52	2.54	**2.33**	4.66	9.14
Simulated (Both)	**3.19**	-0.51	3.15	**2.39**	5.02	9.60
Boulder (Day)	**64.6**	-0.24	64.4	**46.2**	154	306
Boulder (Night)	**58.1**	-1.57	58.1	**38.7**	124	328
Boulder (Both)	**71.9**	1.51	71.9	**52.7**	158	351
Udall (Night)	**52.9**	-0.28	52.9	**9.76**	118	345
Udall (Groundwave Only)	**86.6**	0.19	86.6	**61.8**	198	357
Udall (SNR>50)	**42.4**	-3.55	42.3	**7.56**	71.6	336
Udall (SNR>200)	**31.1**	-5.30	30.8	**7.87**	50.0	372
Udall (d <250km)	**36.3**	0.82	36.3	**4.96**	43.8	345

model error estimates, increase generalization, and decrease the number of gross misclassifications when compared to the single model approach.

Additionally, machine learning algorithms can be guided by prior assumptions on the kind of function they are to learn so as to better generalize the learned model, e.g., the assumption that the sferic data are centered. A common and implicit assumption for the learning process is that the function to be learned is smooth, i.e., it changes little within a small region [43, Ch. 5.11.2] [42]. With this assumption, it is clear that similar to any interpolator, the machine learning algorithm requires a large number of training data that is also representative of the desired generalized domain; otherwise, the smoothness assumption will be broken in regions of the domain for which there is not enough training data. Not only is there a need for proper data coverage of the domain space, but a balanced distribution of data is also required for optimal results in model training [45]. Many resampling techniques have been documented to tackle the class imbalance problem satisfactorily, such as under- or oversampling the training data set to make it more uniform. Simply adding more training data will increase the trained model error performance regardless of class imbalance, however, the learning algorithm suffers the "curse of dimensionality" [42], and as such resampling the data set to minimize class imbalance might be the most practical approach to ensuring representative data coverage over the desired generalization domain.

B. Multilayer Perceptron

While there are many types of neural networks available for inverse modeling, the multilayer perceptron (MLP) class of feed-forward neural networks has been an increasingly common choice among researchers [42], [46]. The MLP has been established as a universal estimator, capable of training a model to approximate virtually any smooth and measurable function [47].

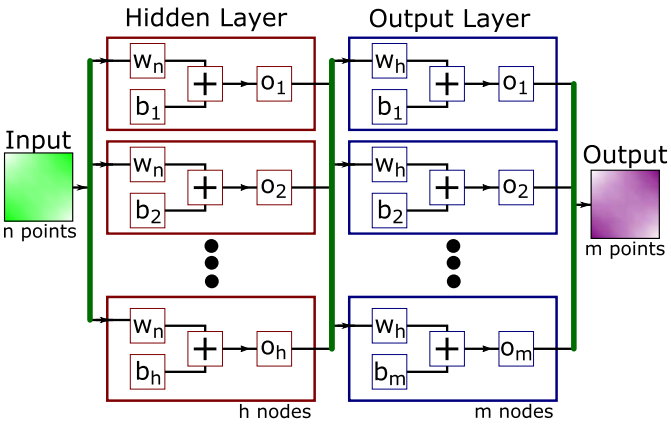

Fig. 10. MLP feedforward neural network example structure. Each node has weights for all of its inputs, which will be found during the training process, as well as independent bias values.

The network, illustrated in Fig. 10, consists of hidden layers, each containing a number of nodes or artificial neurons, that map all of the previous inputs to output values through weights (w_n) and biases (b_n) for each input. The output layer maps the outputs from the last hidden layer either into a continuous-valued estimate using one node for regression networks or into several categorical probability density estimates for classification networks, using multiple nodes for each category. Although the distance estimation problem requires a continuous output, suggesting the use of a regression approach, the main difference between the two approaches at the output layer does not impact the training of the hidden layers, in terms of both speed and accuracy. This feature was verified in this study, and comparable error performance of the trained model

was achieved by using either the classification or regression network. The advantage of the classification network is that a probability density function (PDF) estimated at the output can be used to augment the simple parameter estimate with an error uncertainty based on that output PDF; i.e., for every sferic, we obtain an estimate of its source distance from the receiver, as well as an estimate of the uncertainty in that distance.

The process of training the classification neural network involves computing values for the weights and biases of each neuron in the network. This is accomplished with a backpropagation algorithm, commonly used in neural network implementations, which involves minimizing a cost function iteratively, where the errors calculated at the output are back-propagated and distributed to the neurons in each hidden layer [48]. The performance function, also sometimes referred to as the loss function or cost function, dictates the cost associated with the errors, directly affecting the rate at which weights and biases are adjusted.

It is recommended to preprocess the data for the learning process to improve the efficiency of the learning algorithm. For example, standardization, i.e., mapping the data to have zero mean and unit variance, allows for faster training convergence and better generalization of the trained model [49].

C. Neural Network Configuration

For the lightning distance estimation problem, the learning algorithm must map the input to output (sferic, y and distance, d) with given example pairs, referred to as supervised learning. Thus, true distances, d, for each sferic are needed for the training, as well as for the model validation, where the trained model's distance estimate is compared to the truth. The model outputs an estimated PDF of distance bins, from which an sferic is likely to have originated, and an expected value over the PDF gives a continuous estimate of the sferic range information, essentially recovering the information given by a regression neural network. The final inverse model is described by

$$d = \mathcal{G}_d(y) = \mathbb{E}(\mathcal{G}_{\mathrm{NN}}(y)). \tag{3}$$

where $\mathcal{G}_{\mathrm{NN}}(y)$ is the neural network model and $\mathbb{E}$ is the expected value operator. The observation data, y, comes into the model as a 1-D time-series of B-field values within a $\sim$1 ms "sferic window," sampled at 1 MHz. Although it would also be possible to use this data in the frequency domain or in some other orthogonal basis, this study did not explore such options.

The model training uses scaled conjugate gradient back-propagation, a robust option for memory-intensive training [50], with a cross-entropy performance function [51], implemented using MATLAB's neural network toolbox [52].

Each sferic is shifted in the time dimension so as to center the main ground peak to the same relative time for all data, so that the hidden nodes of the model can train with the assumption that the data are centered. Each sferic is also normalized through standardization, i.e., mapped to have zero mean and unit variance. For the classification scheme, the data are discretized in 10-km distance bins, which was found through a search between 1, 2, 4, 5, 10, 15, and 20-km bin sizes to be the optimal size for the Udall data set, described in Section I-C2. Note that the distance estimate can be more accurate than the 10-km bin size, thanks to the PDF of the distance estimate. The corresponding truth data must also be preprocessed for training, as the algorithm expects the training set truth to be in the same format as the model output, i.e., a PDF of distance estimate bins. For this processing, we employ a one-hot encoding of the truth, where all bins are set to zero except for the one bin corresponding to the correct distance, which is set to one.

Throughout the training process, several improvements were explored and included in the final training framework. These procedures are described below and included optimal parameter selection, data editing, and uncertainty derived from the signal-to-noise ratio.

The final topology of the network, i.e., the number of hidden layers and their nodes, was chosen through a grid search for the trained model with the smallest root-mean-square error for the Udall data set (Section I-C2). The performance was estimated using a K-fold cross-validation scheme [53], and the possible number of nodes ranged from 50 to 800 nodes for the first hidden layer, and from 0 to 800 nodes independently for the second, with a step interval of 50 nodes for both layers. The optimal configuration was found to be 200 and 50 nodes for the first and second layers, respectively. A similar grid search was employed for determining the optimal "sferic window" size and location, i.e., the time duration of each sferic with respect to the ground wave peak originally centered at 207 μs, ranging from 1 to 550 μs for the left boundary and 550–1100 μs for the right boundary with a step interval of 50 μs. The optimal size was found to be 700 μs, i.e., 700 data points, with the peak centered at the 147th point.

Given the different types of lightning return strokes and how they affect the sferic to be observed, especially positive versus negative return stroke types, an all-purpose trained model suffers greatly in error performance due to a generalization difficulty (See Section III-A). The problem can instead be divided into specific models with improved accuracy, for positive and negative return strokes separately. Though this can easily be achieved during training, as polarity is part of the truth data, it is complicated to achieve when applying the model in the real world, as explained in Section I-A, regarding the 180° ambiguity in the calculated direction of the lightning return stroke. Another scheme must be applied first in determining which trained model is suitable for use depending on the incoming sferic's return stroke polarity. In our data sets, the problem is exacerbated with a smaller and less varied population of positive return strokes, and so only the more ubiquitous negative lightning return strokes are considered in this article (Section I-C). Just as with day/night, by using an adequate data set it might be possible to train a single model that is capable of estimating both positive and negative return strokes, with only a marginal loss of accuracy, but this possibility remains unexplored.

Finally, SNR was identified as a large driver of model error performance, and, as a quantity easily measured from the observations, it can be leveraged for better uncertainty

estimation and user warnings in the real-time application of the model. Categorizing the distance estimates in SNR-derived uncertainty regimes is explored in Section IV.

D. Validation

For validation of the trained models in this research, bootstrapping (sampling with replacement) is employed for estimating the trained model error statistics, and the best trained model selected from a pool of different trained models [54]. It is important to note that as an estimator, the model learning is constrained through the bias-variance tradeoff [55], and so both are minimized together in the search for a model with the lowest root-mean-squared error (RMSE), avoiding under- and overfitting associated with minimizing bias or variance alone.

First, the supervised learning data set, observation plus truth data, is randomized and divided into k partitions, where the size of each partition is referred to as the fold size. The training process, using MATLAB's network training function *train* on the MATLAB network object *patternnet*, follows by sampling 80% of the partitions for training, except for the simulated data set where only 10% of the partitions are used for training because of the larger size of that data set. MATLAB's own training function will separate 10% of the training data and use it internally for its own validation of under- versus overfitting and as a convergence stop condition for the training. The trained model error statistics are estimated, as explained in the next paragraph, and the RMSE is used as a measure of the model error, i.e., accuracy and precision. The process is repeated until a model with the best error performance is found and the next one hundred iterations of the training fail to generate a better model.

Once the model is trained, it is independently validated by sampling a fold size sector from the original data set with replacement, and calculating the estimate's error, i.e., the truth minus the model's estimated distances. The validation step is repeated a few hundred times in order to build a statistically meaningful distribution of the estimate errors, from which we can approximate each model's estimate error statistics, a process known as bootstrapping. If the validation data set is large enough, the true error statistics can be calculated and compared with one arrived at by bootstrapping. The true error statistics cannot be computed directly for the Udall and Boulder data sets since the subset of data not used in training is too small to be statistically meaningful. However, the simulated data set does contain a large number of sferics not used in training which can be used for inferring the trained model error statistics. Given the possible approximation errors expected in bootstrapping [53], the bootstrapping approximation of error statistics is compared against the true error statistics for trained models in the simulated data set. A fold size of at least 250 sferics is empirically found to be suitable for approximating the error statistics for the trained models, with less than 0.5 km of error in bias and variance approximation of the trained model error, and the k-value is adjusted to maintain this minimum fold size for the different data sets. If a smaller fold size is desired, the improved bootstrap 632+ technique can be used instead [56].

Independently, the PDF distance bin estimate given by the model to an input sferic allows us to infer some uncertainty information for each lightning return stroke distance estimate; however, the PDF was found to be a poor estimator of uncertainty when compared to the actual model statistics, either approximated by the bootstrapping or the true model statistics in the simulated data set. An adequate map may be found for sferic uncertainty based on the trained model's output PDF, but in this study, there was no attempt to find this map. A classification approach is still chosen so that future implementation might take advantage of it.

IV. Machine-Learned Model Performance

In this section, the trained models' error performances for the three data sets described above are presented and analyzed. The model error is computed from the truth minus estimated distance, and the computed root-mean-square error (RMSE) is used as the primary error performance measure for the models. Given that some error distributions have nonzero positive kurtosis and deviate from a normal distribution, the RMSE, and standard deviation are inflated estimates of model uncertainty [57], and so percentile measures are given as well for more accurate comparisons. A summary of the error performance of different models described in this section and in Section II is found in Table II.

A. Simulated Sferics

The trained models based on the simulated data set give information about the best performance that can be achieved with our methodology, as it is a highly controlled data set of noise-free sferics with a balanced distribution of data covering the desired domain space, i.e., vast data of uniform distance and ionospheric parameter distributions.

The performance of the simulated data set model is shown in Fig. 11. The distance error histogram shows that the estimated error distribution is close to Gaussian. The small bias is inevitable given that the model was selected according to RMSE, and the variance would likely suffer if bias would have been minimized instead.

The confusion matrix shows no misclassifications beyond one 10-km bin, and this indicates that improvements in the error performance, i.e., rms estimate error, could still be seen by decreasing the 10-km bin size, provided that the data set size is increased accordingly. Also, note the slight increase in estimate error near the distance boundaries. This error is likely caused by the hard boundary for the categorical distance estimate, which can only under- or overestimate in these regions. Since the expected value of the predicted distance is used in determining the final estimate, an even larger systematic effect is expected. If this systematic effect is not balanced perfectly between the left and right boundaries, there will be a contribution to the approximated model error mean, but in this model this contribution is insignificant and the bias is due to a number of overestimate outliers for sferics in the region between 100 and 200 km. Finally, the model is also more accurate when the training/validation data are separated between day and night, as expected in

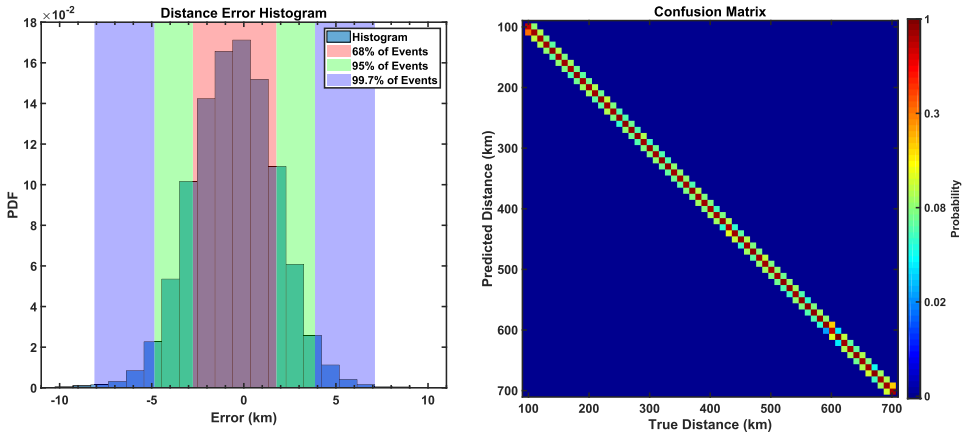

Fig. 11. (Left) Histogram of the error in estimated distances for the simulated sferics, accompanied by coverage shadings to indicate the percentage of events that lie in each shaded band. (Right) Confusion matrix plot shows the inferred probabilities of the model predicted distances for a given true distance, with each true distance row normalized.

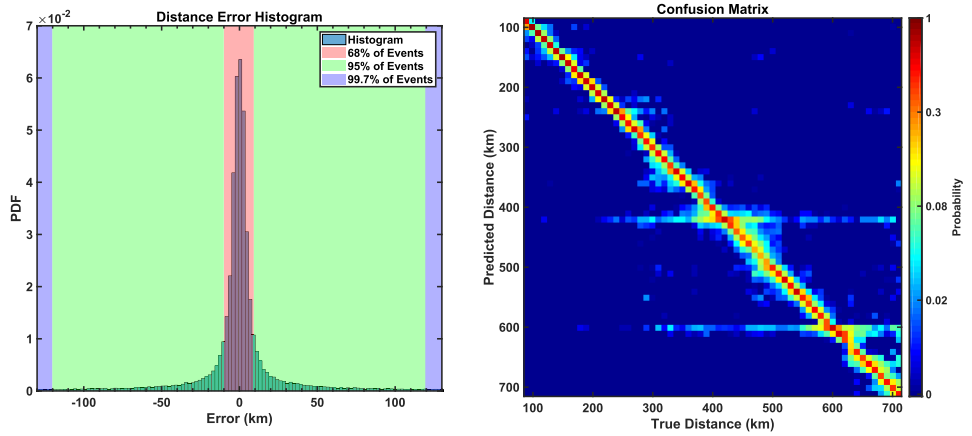

Fig. 12. (Left) Histogram of the error in estimated distances for the Udall data set, accompanied by coverage shadings to indicate the percentage of events that lie in each shaded band. The tails of the distribution continue beyond the error axis. (Right) Confusion matrix plot shows the inferred probabilities of the model predicted distances for a given true distance, with each true distance column normalized.

terms of generalization requirements. This is expected given the increase in complexity for the model that requires a larger domain space, and thus more data for proper training. Given a large amount of data available for training here, and the very small increase in RMSE, the worse performance of the more complex model is likely to be a limitation of this approach, and different network topology (number of neurons and bin size) might be required for any possible improvement.

B. Udall Data Set

The model trained on the Udall data set, mostly affected by class imbalance, has a promising accuracy for real-time application, with an RMSE of 53 km, in light of possible improvements discussed in Section IV-D.

The error performance of the model trained on the Udall data set, shown in Fig. 12, suffers primarily from the nonuniformity of the data distance distribution. On the histogram plot (left), the error estimate has a distribution much narrower at the top and wider at the base compared to a normal distribution. The gross misclassifications, larger than 50 km, though small in number, continue beyond 100 km and greatly affect the model accuracy. As a measure of precision, the standard deviation and RMSE seem to be conservative; more illustrative are the error bounds of 9.76 km for the 68th percentile and 117.77 km for the 95th percentile. The large positive kurtosis

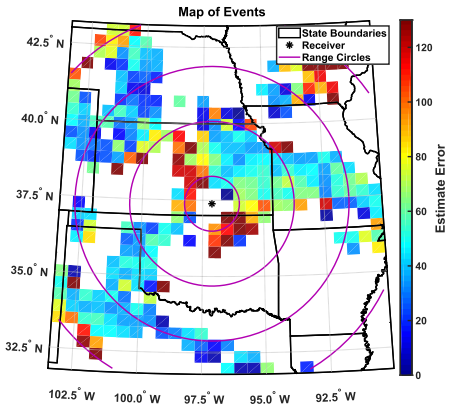

Fig. 13. Map of the estimate RMSE for sferics in the Udall validation data set. The receiver location, and range circle of 100, 300, 500, and 700 km from the receiver are shown for reference.

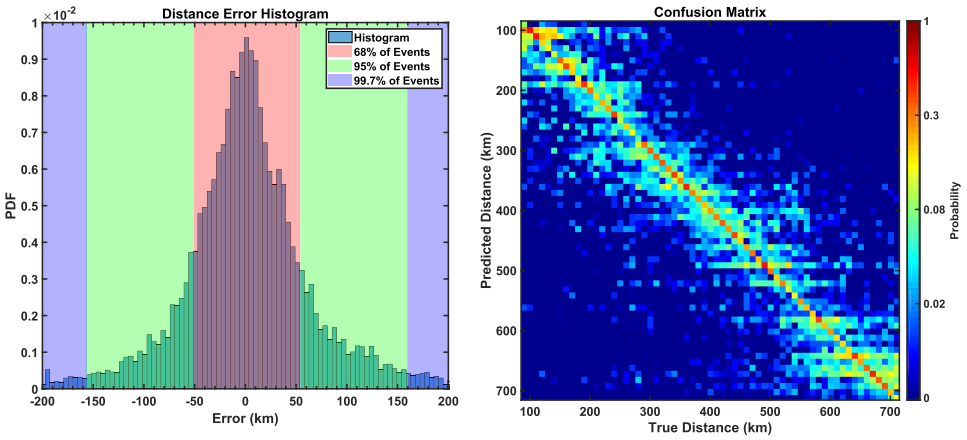

Fig. 14. (Left) Histogram of the error in estimated distances for the Boulder data set, accompanied by coverage shadings to indicate the percentage of events that lie in each shaded band. The tails of the distribution continue beyond the error axis. (Right) Confusion matrix plot shows the inferred probabilities of the model predicted distances for a given true distance, with each true distance column normalized.

of the error distribution indicates a large number of outliers. It is possible that careful data selection will allow for better performance on some of the grossly misclassified sferics, larger than 300 km, as these might arise from improperly classified or overlapping sferics, as well as the use of bagging, as explained in Section III-A.

The confusion matrix shows a satisfactory performance for sferics closer than 300 km. The overall error performance correlates directly with the number of events for the given distances, displayed in Fig. 4, from 1) the large uniform number of sferics closer than 300 km, 2) a peak of events at 420 km, 3) the lack of sferics in the 300–600 km region, to 4) a peak of events at 600 km. The class imbalance is seen to be so problematic in this case, that the model tends to predict distances of 420 and 600 km, which had a vast number of training sferics associated with them, and that is especially the case for sferics which belong to underrepresented distance bins. The weakness of the model predictions cannot be solely attributed to the lack of sferics at specific regions, as the intrinsically lower SNR values for more distant sferics and the fact that some of them are below the noise floor, effectively reducing data coverage, also contributes to the model inaccuracies.

Fig. 13 shows a map of RMSE for the sferics of the validation data set. The azimuthal dependence of the model and its contribution to model accuracy are considerably less significant than the other dependencies described above. Even with the sparsity of azimuthal variety at 500 km, where there is a large number of sferics to the southwest and east, for sferics whose azimuths correspond to the northwest the error is indistinguishable from the prediction error for events to the southwest or east.

C. Boulder Data Set

The model trained on the Boulder data set, with low SNR and various parasitic signals, has much lower accuracy, RMSE of 72 km and 68th percentile error of 52.7 km, but still performs reasonably well in estimating distances given the poor data quality. Thus, the methodology presented here can

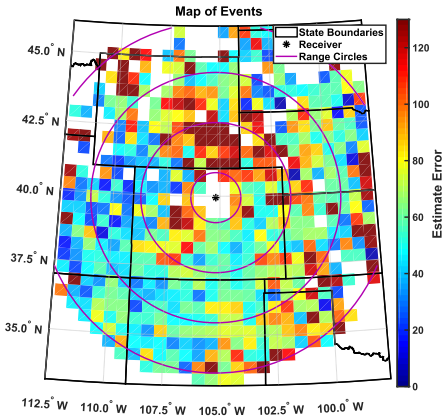

Fig. 15. Map of the estimate RMSE for sferics in the Boulder validation data set. The receiver location, and range circle of 100, 300, 500, and 700 km from the receiver are shown for reference.

be used even in a noisier environment at the cost of larger uncertainties.

The error performance of the trained model for the Boulder data set, as seen in Fig. 14, is poor and falls far from what can be achieved with the simulated sferics. The lower SNR for this data set is the primary cause of the model's deficiency. Again nonuniformity in the sferics' distance distribution leads to a model with corresponding distance regions of better performance (300–500 km) and of worse performance (100–300 km). Thus, it seems that the method overemphasizes training on more numerous sferics at a given distance. The estimation at farther distances suffers from poor error performance due to a different mechanism explained by the SNR decrease with distance, which has been seen on the Udall data set. Additionally, the models perform better, in terms of estimate error, when only day or night data are used separately just like with the simulated data set, only here the difference is significant. The reason for the significant difference between using the day and night data together or separately is likely due to the larger domain model needing more data to be properly trained. This unmet requirement is especially true on

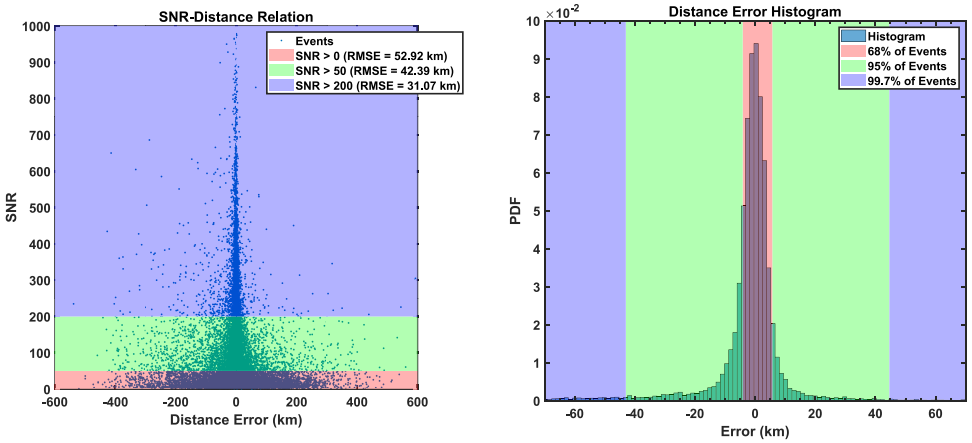

Fig. 16. (Left) Plot of the SNR distance error relation discovered in the Udall model performance. (Right) Distance error histogram for only events predicted to be closer than 250 km.

the dayside of the domain, which has a significantly lower number of sferics than its night counterpart. Additionally, as opposed to the Udall data set, this set shows more azimuthal dependence, where a large number of sferics to the southwest, with relatively small errors, does not translate to good accuracy for the same distances in other directions, as seen in Fig. 15. It would be important to distinguish if this is truly an azimuthal dependence caused by the topology in the region, or because of a lack of representation of the ionosphere states given the time of day.

D. Model Performance Considerations

As the best model performance for real data is still much worse than the trained model from simulated sferics, some limitations and adjustments are considered in this section. As a properly calibrated low-noise instrument is expected to generate observations with quality similar to that of the Udall data set, only that data set is studied here.

Although a sensitivity analysis was not performed as part of this study, it is possible to learn whether there is any usable information on stroke distance from the observer present in the groundwave portion of a sferic alone. For this experiment, a model was trained for the Udall data set with a very narrow sferic "window size," with the right boundary only 10 points (10 μs) after the groundwave peak, essentially removing the skywave from the input data. The results for this method, shown in Table II, demonstrate that this method can estimate distances, though they are worse than the results using the skywave. These results indicate that there is some observability of the distance in the sferic groundwave. This confirms that the skywave delay is not the only feature that can be exploited in a trained neural network model, which as seen in Section II, is not enough for arriving at the highest precision estimates with the simulated or Udall data sets.

Regarding the uncertainty associated with the model's output, note that there is a strong relationship between SNR and the estimated distance error, as shown explicitly in Fig. 16. This information can be used at the application level for better informing the expected uncertainty for a given sferic. As an illustration of such an application, three SNR bands were delimited spanning SNR values from $\geq$0 (original result from above), $\geq$50, and $\geq$200, where standard deviations and percentiles present the statistics for each band, as given in Table II.

Additionally, it is clear that the model error performance is greatest for closer distances. By limiting the scope of the estimation to sferics that were estimated to be closer than 250 km, a better error performance from the original model can be expected, RMSE of 36 km, as shown in Fig. 16. Note that in this Udall data set, there is a large local uniform distribution of sferics in this regime of sferics closer than 250 km from the receiver, which is also necessary for the performance observed in the first place. By limiting the estimates on predicted distances as opposed to true distances, this technique can be employed at the application level, and still allows for estimation of more distant events. With the prerequisites of enough data uniformly distributed at least for closer sferics, which is easier than farther sferics because of the smaller area delineated, the improvement in error performance with closer sferics is expected to be true for any data set. This improvement in error performance is expected to be limited, however, as the skywave becomes relatively weak compared to the ground wave for closer sferics. Note also that with a more adequate data set uniformly covering all of the distances and hours of the night and with enough sferics, we are led to conclude that the accuracy seen here for closer sferics is attainable for all of the desired distances.

V. Conclusion

In this preliminary study on the use of machine-learned models for distance estimation of lightning through its emitted radio signatures, a promising methodology and associated software framework have been developed and documented, including thorough validation techniques and deliberation on improvements and critical elements. A simpler and more physically natural analytical method, based on a sferic's skywave delay, can be used as a simplified, first approach for estimating lightning distance. The analytical method provides some

advantages over competing methods, such as our machine learning method, including a much smaller requirement on the amount of data needed to produce a model, and allowing for a simple, direct improvement to the distance estimate if the information on the D-region ionosphere is available. The limitations of this analytical method were also described, providing context to the disadvantages to be overcome by any competing methods, including its intrinsic dependence on the ionospheric D-region state, and a required algorithm capable of robustly measuring the skywave delay on any sferic. The machine-learned model's accuracy presented here helps to illustrate the level of performance that can be achieved for different types of data sets, and elucidates the data set qualities that strongly correlate with the model error performance. The model accuracy achieved in this study with real data is comparable to that reported on other articles using single-site distance estimation, but it is likely that an error performance improvement can be achieved with better data coverage used in training, use of resampling techniques, and implementation of bootstrap aggregating. The model error performance is far worse than that achieved in detection networks, except for the model trained with simulated data, which informs us of the type of accuracies possible when using this technique with comprehensive and noiseless data. We confirm that as the observed sferics depend on system parameters, e.g., distance and ionospheric states, enough comprehensive data sampling is required for the training of accurate models. Additionally, as the desired domain increases, so does the model complexity and the amount of data required to train a model accurately; i.e., training for both daytime and nighttime sferics produces a model with a significant increase in error due to a small and imbalanced data set, but even with a large amount of training data there will still be a marginal accuracy cost to the more complex model, as seen with the vast and uniform simulated data set. The observations are not as sensitive to other parameters, such as azimuth and topology, which are shown to have a smaller effect in data coverage for training. Different approaches for determining the uncertainty at the application level were discussed, which could include error estimates based on SNR or computed distance. Further research must be done to improve the methodology and data preprocessing of sferics for better performance, as well as to better understand the tradeoffs of training different models for different environment generalizations.

ACKNOWLEDGMENT

The authors would like to thank Dr. Patrick Blaes for sharing his initial insights into this research, and Mr. Noah Holland-Moritz for his contributions to the sferic extracting algorithm.

REFERENCES

[1] A. Nag, M. J. Murphy, W. Schulz, and K. L. Cummins, "Lightning locating systems: Insights on characteristics and validation techniques," *Earth Space Sci.*, vol. 2, no. 4, pp. 65–93, Apr. 2015.

[2] K. Cummins, E. Krider, and M. Malone, "The US national lightning detection network and applications of cloud-to-ground lightning data by electric power utilities," *IEEE Trans. Electromagn. Compat.*, vol. 40, no. 4, pp. 465–480, Nov. 1998.

[3] C. J. Rodger, J. B. Brundell, R. L. Dowden, and N. R. Thomson, "Location accuracy of long distance VLF lightning locationnetwork," *Ann. Geophys.*, vol. 22, no. 3, pp. 747–758, Jun. 2010.

[4] R. Said and M. Murphy, "GLD360 upgrade: Performance analysis and applications," in *Proc. 24th Int. Lightning Detection Conf.*, San Diego, CA, USA: Vaisala, 2016, pp. 1–8.

[5] A. C. L. Lee, "An operational system for the remote location of lightning flashes using a VLF arrival time difference technique," *J. Atmos. Ocean. Technol.*, vol. 3, no. 4, pp. 630–642, Dec. 1986.

[6] M. Füllekrug and S. Constable, "Global triangulation of intense lightning discharges," *Geophys. Res. Lett.*, vol. 27, no. 3, pp. 333–336, Feb. 2000.

[7] S. A. Changnon, "Damaging thunderstorm activity in the United States," *Bull. Amer. Meteor. Soc.*, vol. 82, no. 4, pp. 597–608, Apr. 2001.

[8] C. J. Schultz, W. A. Petersen, and L. D. Carey, "Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends," *Wea. Forecasting*, vol. 26, no. 5, pp. 744–755, Oct. 2011.

[9] G. J. Fishman *et al.*, "Discovery of intense gamma-ray flashes of atmospheric origin," *Science*, vol. 264, no. 5163, pp. 1313–1316, May 1994.

[10] F. Lyu, S. A. Cummer, and L. McTague, "Insights into high peak current in-cloud lightning events during thunderstorms," *Geophys. Res. Lett.*, vol. 42, no. 16, pp. 6836–6843, Aug. 2015.

[11] S. A. Cummer, U. S. Inan, and T. F. Bell, "IonosphericDregion remote sensing using VLF radio atmospherics," *Radio Sci.*, vol. 33, no. 6, pp. 1781–1792, Nov. 1998.

[12] F. Adcock and C. Clarke, "The location of thunderstorms by radio direction-finding," *J. Inst. Electr. Engineers-III, Radio Commun. Eng.*, vol. 94, no. 28, pp. 118–125, Mar. 1947.

[13] E. P. Krider, R. C. Noggle, and M. A. Uman, "A gated, wideband magnetic direction finder for lightning return strokes," *J. Appl. Meteor.*, vol. 15, no. 3, pp. 301–306, Mar. 1976.

[14] I. I. Kononov, I. A. Petrenko, and V. S. Snegurov, "Radiotechnical techniques for locating thunderstorms," (in Russian), *Hidrometeoizdat*, p. 222, 1986, doi: 10.1016/0021-9169(95)00011-P.

[15] D. A. Smith *et al.*, "A method for determining intracloud lightning and ionospheric heights from VLF/LF electric field records," *Radio Sci.*, vol. 39, no. 1, Feb. 2004, Art. no. RS1010.

[16] H. Zhang *et al.*, "Locating narrow bipolar events with single-station measurement of low-frequency magnetic fields," *J. Atmos. Solar-Terr. Phys.*, vols. 143–144, pp. 88–101, Jun. 2016.

[17] S. A. Cummer *et al.*, "The source altitude, electric current, and intrinsic brightness of terrestrial gamma ray flashes," *Geophys. Res. Lett.*, vol. 41, no. 23, pp. 8586–8593, Dec. 2014.

[18] E. H. Lay and X.-M. Shao, "High temporal and spatial-resolution detection of D-layer fluctuations by using time-domain lightning waveforms," *J. Geophys. Res.*, vol. 116, no. A1, Jan. 2011, Art. no. A01317.

[19] V. B. Somu, V. A. Rakov, M. A. Haddad, and S. A. Cummer, "A study of changes in apparent ionospheric reflection height within individual lightning flashes," *J. Atmos. Solar-Terr. Phys.*, vol. 136, pp. 66–79, Dec. 2015.

[20] J. R. Wait, "On the theory of the slow-tail portion of atmospheric waveforms," *J. Geophys. Res.*, vol. 65, no. 7, pp. 1939–1946, Jul. 1960.

[21] C. Mackay and A. C. Fraser-Smith, "Lightning location using the slow tails of sferics," *Radio Sci.*, vol. 45, no. 5, pp. 1–10, Oct. 2010.

[22] C. Mackay and A. C. Fraser-Smith, "World coverage for single station lightning detection," *Radio Sci.*, vol. 46, no. 3, pp. 1–9, Jun. 2011.

[23] V. A. Rafalsky, A. P. Nickolaenko, A. V. Shvets, and M. Hayakawa, "Location of lightning discharges from a single station," *J. Geophys. Res.*, vol. 100, no. D10, p. 20829, Feb. 2004.

[24] V. Rafalsky, A. Shvets, and M. Hayakawa, "One-site distance-finding technique for locating lightning discharges," *J. Atmos. Terr. Phys.*, vol. 57, no. 11, pp. 1255–1261, Sep. 1995.

[25] J. B. Brundell, C. J. Rodger, and R. L. Dowden, "Validation of single-station lightning location technique," *RadioSci.*, vol. 37, no. 4, pp. 12-1–12-9, Aug. 2002.

[26] I. Avionics. (1997). *Strike Finder Lightning Detection*. Accessed: Apr. 10, 2019. [Online]. Available: http://www.insightavionics.com/strikefinder.htm

[27] L. C. Avionics. (1998). *Stormscope*. Accessed: Apr. 10, 2019. [Online]. Available: https://www.l3commercialaviation.com/avionics/products/stormscope/

[28] S. Mallick *et al.*, "Performance characteristics of the NLDN for return strokes and pulses superimposed on steady currents, based on rocket-triggered lightning data acquired in Florida in 2004-2012: Evaluation of NLDN Performance," *J. Geophys. Res. Atmos.*, vol. 119, no. 7, pp. 3825–3856, Apr. 2014.

[29] S. Tumanski, "Induction coil sensors—A review," *Meas. Sci. Technol.*, vol. 18, no. 3, p. R31, 2007.
[30] U. S. Inan, S. A. Cummer, and R. A. Marshall, "A survey of ELF and VLF research on lightning-ionosphere interactions and causative discharges," *J. Geophys. Res.*, vol. 115, no. A6, Jun. 2010, Art. no. A00E36.
[31] K. P. Spies and J. R. Wait, *Mode Calculations for VLF Propagation in the Earth-Ionosphere Waveguide*, U.S. Dept. Commerce, Office Tech. Services, Washington, DC, USA, 1961, no. 114.
[32] M. Yamashita and K. Sao, "Some considerations of the polarization error in direction finding of atmospherics—I. Effect of the Earth's magnetic field," *J. Atmos. Terr. Phys.*, vol. 36, no. 10, pp. 1623–1632, Oct. 1974.
[33] M. Yamashita and K. Sao, "Some considerations of the polarization error in direction finding of atmospherics-II. Effect of the inclined electric dipole," *J. Atmos. Terr. Phys.*, vol. 36, no. 10, pp. 1633–1641, Oct. 1974.
[34] R. A. Marshall, C. L. Da Silva, and V. P. Pasko, "Elve doublets and compact intracloud discharges," *Geophys. Res. Lett.*, vol. 42, no. 14, pp. 6112–6119, Jul. 2015.
[35] F. Horner, "The design and use of instruments for counting local lightning flashes," in *Proc. IEE—B, Electron. Commun. Eng.*, vol. 107, no. 34, pp. 321–330, Jul. 1960.
[36] J. R. Wait and K. P. Spies, *Characteristics Earth-Ionosphere Waveguide for VLF Radio Waves*. U.S. Dept. Commerce, Nat. Bureau Standards, Washington, DC, USA, 1964, no. 300.
[37] R. A. Marshall, "An improved model of the lightning electromagnetic field interaction with the D-region ionosphere," *J. Geophys. Res.*, vol. 117, no. A3, Mar. 2012, Art. no. A03316.
[38] V. Rakov and A. Dulzon, "A modified transmission line model for lightning return stroke field calculations," in *Proc. 9th Int. Symp. Electromagn. Compat.*, 1991, pp. 229–235.
[39] D. A. Chrissan and A. C. Fraser-Smith, "Seasonal variations of globally measured ELF/VLF radio noise," *Radio Sci.*, vol. 31, no. 5, pp. 1141–1152, Sep. 1996.
[40] D. A. Chrissan and A. C. Fraser-Smith, "A comparison of low-frequency radio noise amplitude probability distribution models," *Radio Sci.*, vol. 35, no. 1, pp. 195–208, Jan. 2000.
[41] M. Cohen, U. Inan, and E. Paschal, "Sensitive broadband ELF/VLF radio reception with the AWESOME instrument," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 3–17, Jan. 2010.
[42] M. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, Aug. 1998.
[43] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[44] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
[45] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Oct. 2002.
[46] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.
[47] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.
[48] C. M. Bishop *et al.*, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[49] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Proc. Neural Netw., Tricks Trade*. London, U.K.: Springer-Verlag, 1998, pp. 9–50. [Online]. Available: http://dl.acm.org/citation.cfm?id=645754.668382
[50] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, no. 4, pp. 525–533, Jan. 1993.
[51] D. M. Kline and V. L. Berardi, "Revisiting squared-error and cross-entropy functions for training neural network classifiers," *Neural Comput. Appl.*, vol. 14, no. 4, pp. 310–318, Dec. 2005.
[52] H. Demuth, M. Beale, and M. Hagan, "Deep learning toolbox, version 13," MathWorks, Natick, MA, USA, 2008, pp. 37–55.
[53] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. IJCAI*, Montreal, QC, Canada, vol. 14, no. 2, 1995, pp. 1137–1145.
[54] R. Tibshirani, "A comparison of some error estimates for neural network models," *Neural Comput.*, vol. 8, no. 1, pp. 152–163, Jan. 1996.
[55] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, Jan. 1992.
[56] B. Efron and R. Tibshirani, "Improvements on cross-validation: The 632+ bootstrap method," *J. Amer. Stat. Assoc.*, vol. 92, no. 438, pp. 548–560, Jun. 1997.
[57] P. H. Westfall, "Kurtosis as peakedness, 1905–2014.R.I.P," *Amer. Statistician*, vol. 68, no. 3, pp. 191–195, Jul. 2014.

Andre L. Antunes de Sá (Student Member, IEEE) received the B.A. degree in physics and astronomy from Amherst College, Amherst, MA, USA, in 2014, and the M.S. degree in aerospace engineering sciences from the University of Colorado Boulder, Boulder, CO, USA, in 2016, where he is currently pursuing the Ph.D. degree in aerospace engineering sciences.

His research interests include radio instrument design for lightning remote sensing, thunderstorm electrification processes, and the relation between lightning and severe weather phenomena.

Mr. Antunes de Sá was a recipient of the NASA Earth and Space Science Fellowship from 2017 to 2020.

Robert A. Marshall (Member, IEEE) received the B.S. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2004 and 2009, respectively.

He held research positions with the Center for Space Physics, Boston University, Boston, MA, USA, and the Department of Aeronautics and Astronautics, Stanford University, Stanford, CA, USA. He is currently an Assistant Professor with the Ann and H. J. Smead Department of Aerospace Engineering Sciences, University of Colorado Boulder, Boulder, CO, USA. He has coauthored the book *Numerical Electromagnetics: The FDTD Method* (Cambridge University Press, 2011) with U. S. Inan. His research activities include the study of lightning–ionosphere interactions, radiation belt precipitation and atmospheric effects, numerical modeling and applications in space physics, and small satellite instrument development.

Dr. Marshall received the First Place in the 2007 International Radio Science Union (URSI) Student Paper Competition. He was a recipient of the URSI Young Scientist Award in 2011.