Do Forecasters Add Value to Machine Learning Algorithms of Cloud-to-Ground Lightning?

CARA GREGG*

National Weather Center Research Experiences for Undergraduates Program
Norman, Oklahoma

KRISTIN CALHOUN AND TIFFANY MEYER

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and NOAA/National Severe Storms Laboratory
Norman, Oklahoma

ABSTRACT

Cloud-to-ground lightning is an extremely dangerous weather phenomenon resulting in 28 deaths annually over the last decade; currently, there are no requirements for National Weather Service to communicate lightning dangers or hazards to public. A probabilistic algorithm was developed at the National Severe Storms Laboratory using machine learning to create an automated system that generates objects around areas where it predicts cloud-to-ground (CG) lightning will occur. In spring 2017, nine forecasters from the National Weather Service tested a Probabilistic Hazard Information prototype in the Hazardous Weather Testbed in which they used the guidance of the automated system, modified these objects from the system, and created their own objects to ideally create better forecasts of CG lightning. These forecaster and automated objects were verified and aspects of their performance, such as the probability of detection, were compared to see if the forecasters added value to the automated system. Forecasters added value to the system by adding discussion to the objects and through modifying the size, severity, duration, and probability of the lightning storms. However, forecasters found the task particularly tedious to complete. The areas where the forecasters are adding the most value could be used to improve the automated system's performance at predicting CG lightning, further reducing forecaster workload.

1. Background

Cloud-to-ground (CG) lightning is a threatening phenomenon not only to property and electrical grids, but to human life as well (Cummins and Murphy 2009; Mosier et al. 2011). Over the last ten years, the average number of deaths by lightning in the United States is 27 casualties per year (weather.gov cited 2018). In 2008, lightning was the second most common weather-related source of fatality in the United States with 58 fatalities that year (Mosier et al. 2011). At present, there is no formal warning system in place for CG lightning from the National Weather Service (NWS).

The threat of life and property by CG lightning has led researchers at the National Severe Storms Laboratory (NSSL) to develop an automated system for forecasting CG lightning (Meyer et al. 2017). Historically, Doppler weather radars and other radar systems, more commonly used for the detection of thunderstorms, have been used to predict CG lightning with mixed results (Hondl and Eilts 1994; Gremillion and Orville 1999). Gremillion and Orville (1999) found that a 40-dBZ echo at the temperature height of -10°C served as a solid indicator of the start of CG action, revealing a relationship between radar reflectivity in the mixed phase of ice and supercooled liquid water and the prediction of CG lightning. Indicators of lightning and storm electrification may help by providing evidence of intense storm parameters and from there attempt to predict where the lightning itself would strike; but radar features provide no guarantee that they alone signify CG lightning has occurred or may occur. Currently, the NWS is not required to provide information regarding CG lightning to the public. Instead, the public mostly relies on hearing thunder or seeing a lightning flash before taking shelter. The large number of deaths and injuries that occurs annually from CG lightning indicates that this alone is not sufficient.

In most storms, the first in-cloud (IC) flash occurred before the first CG flash (Cummins and Murphy 2009). Studies, such as MacGorman et al. (2011), have shown

**Corresponding author address:* Cara Gregg, The Ohio State University, 1036 Derby Hall, 154 North Oval Mall, Columbus, OH 43210-1361
E-mail: gregg.160@buckeyemail.osu.edu

the first IC flash typically precedes the first CG flashes by tens of minutes, providing lead time to CG activity. However, it was not guaranteed that the first IC flash was correctly identified, nor that the timing was consistent between storms. Additionally, it appeared the lead time was partially dependent on the region. For example, the timing in the high plains in the United States was much longer than the time between flashes for north Texas or Oklahoma due to higher cloud bases of storms (MacGorman et al. 2011). This added a whole other layer of complexity to using the times between flashes to predict when CG lightning struck on its own, thus, leading toward the need for a more detailed system. Through the early 2000s no systematic-automated lightning forecasting algorithms have been routinely utilized by government or private sector (Mosier et al. 2011). However, algorithms are under development and there is hope to put this system into operational use in future years (Cartier cited 2017).

Machine Learning (ML) combines a large amount of different types of data into one model in order to train a system to detect patterns and relationships of a certain phenomenon (Gagne et al. 2017). Meyer et al. (2017) used a Random Forest, which utilized a trained dataset that included a variety of data to develop a CG lightning probabilistic algorithm. The model was trained using a variety of merged radar, environmental and lightning data across the Contiguous United States over a one-year period. The properties of more than 1 million storm samples were recorded and subsets of the data were used to create multiple decision trees providing a "yes" or "no" answer if the storm produced CG lightning (Meyer et al. 2017). Data that is similar to each other should fall into the same nodes of the trees in a random forest, which provides an indication of which variables were important for prediction, training the model (Liaw and Wiener 2001). Meyer et al. (2017) found that lightning data (i.e., if a storm is already producing lightning) to be the top variables for prediction followed by radar data such as maximum reflectivity values in the mixed phase (between 0 and -40°C, Fig. 2). The 300 trees in the Meyer et al. (2017) algorithm are collected into one large dataset (the Random Forest) providing the probabilistic likelihood of any storm to produce CG lightning by taking the number of yes's and dividing it by the total number of trees.

Progressing toward a more developed warning system, Probabilistic Hazard Information (PHI) is currently being developed and utilized in order to provide warnings of severe events that follow along with the event, rather than staying stagnant around one area (Karstens et al. 2015). A prototype PHI tool has been tested annually by NWS forecasters in the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) since 2015 (Karstens et al. 2015). Forecasters evaluate the use of PHI during real time events in which they created probabilistic forecasts for events such as wind, hail, and tornadoes (Karstens et al. 2015). The output from the forecasters, PHI grids provide more information about storm hazards sooner with more updates than the current severe thunderstorm and tornado warning paradigm. This offers a mixture of both human elements and automated machine elements in a system. When running through cases, the forecasters were given the guidance of the automated objects generated by the ML algorithm as well as additional information such as radar (Karstens et al. 2018). In the 2017 experiment, there were four levels of automation that could be issued for the objects by the forecasters: (1) completely forecaster made, (2) partially automated with a forecaster geometry, (3) partially automated with the automated geometry, or (4) completely automated (Karstens et al. 2018). Karstens et al. (2018) found the prototype performed relatively well, however, noted a need for better verification methods in order to get a better idea of how the automated system was performing.

Emergency managers and other end-users were able to review the resulting PHI grids from forecasters within an experimental display called the Enhanced Data Display (EDD) (Karstens et al. 2018). The end-users made decisions based on the cases and PHI grids, providing insight of their opinions of the PHI objects and the system overall. The opinions of these users is important because if the PHI is planned to be used operationally in the future, it must be usable by those beyond the forecasters.

The PHI prototype is part of a larger project: the Forecasting a Continuum of Environmental Threats (FACETs), currently being developed by the NSSL. This project is working toward providing clear and informative data to the public regarding warnings of severe events, such as tornados and hail, and non-severe events like CG lightning (National Severe Storms Laboratory cited 2018). FACETs will make use of polygons and grid-based probabilities of a severe event as created by the PHI tool (National Severe Storms Laboratory cited 2018). This will allow for the warning to be specific to a certain area and for the public to have a better idea of their chance of receiving severe weather.

The purpose of this paper is to review the displaced real-time lightning case from the 2017 HWT experiment that was run by all forecasters within the PHI tool Calhoun et al. (2018) and determine whether the forecasters added value to the automated system alone. The verification of both the automated system's and forecasters run through a case study allowed for an examination of the systems ability to successfully forecast CG lightning, something that previous studies have lacked. It will explain whether the forecasters added value to the automated system and if so, which areas the forecasters added the most value to the system. By doing so, these areas of value could be picked out and used to enhance the automated system for potential operational use.

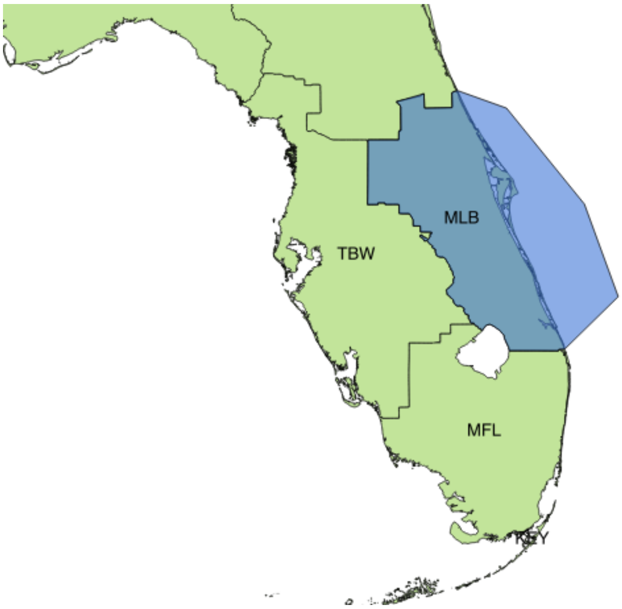

FIG. 1. Melbourne County Warning Area, Florida

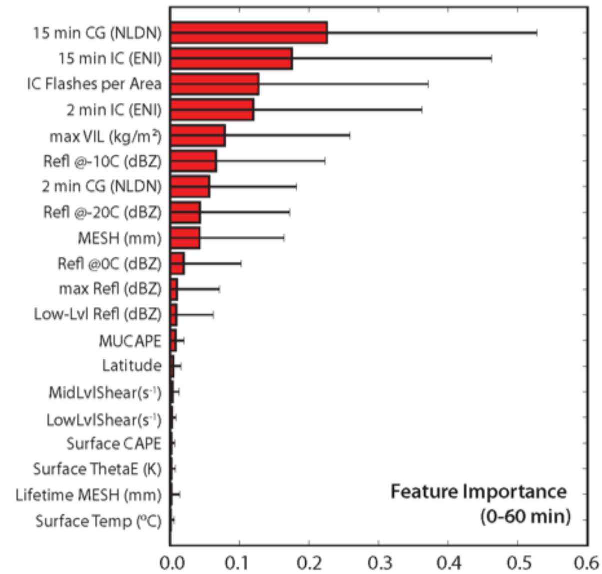

FIG. 2. Important variables in the ML algorithm (Meyer et al. 2017)

2. Data and Methods

a. Event Background: 1 Sept 2016

All forecasters in the 2017 participated in the same study using the same radar, lightning data and automated objects from the Meyer et al. (2017) probabilistic algorithm. This case chosen was from 1 Sept 2016 in the Melbourne (MLB) County Warning Area in Florida as Hurricane Hermine made landfall on the west coast of Florida (Fig. 1). Hurricane Hermine was a category 1 hurricane that moved across western Florida, Georgia, South Carolina, and North Carolina. Hermine became a tropical storm by 0600 UTC on 31 Aug and intensified over the warm waters of eastern Gulf of Mexico as it moved north-northeastward and northeastward. By 1 Sept, an eye was seen in visible imagery and Hermine reached hurricane intensity around 1800 UTC south-southwest of Apalachicola, Florida with maximum winds of 65 kt (Berg 2017). This period of near peak intensity just prior to landfall was encapsulated in the case study period, which took place from 1755 to 1955 UTC. Hermine produced a total of ten tornadoes with two EF-0 tornadoes on the evening of September 1st within the MLB county warning area: one in Winter Garden in Orange County and the other in Lake County (Berg 2017). A large number of CG lightning flashes also struck during this event. This case was found to be particularly difficult by the forecaster participants; Hermine likely assisted in making it challenging for the forecasters and automated system as land-falling Tropical Cyclones are rare events.

b. Machine Learning and Creation of Automated Objects

The ML algorithm was able to determine which variables and conditions tended to precede CG lightning. The features that were the most important in the ML algorithm were the 15 minute CG lightning strike data from the National Lightning Detection Network (NLDN), 15 minute IC lightning data from Earth Networks, Inc (ENI), IC flashes per area, and 2 minute IC lightning data from ENI respectively (Fig. 2). In real-time, when the system detected that the conditions were occurring, the algorithm predicted whether or not CG lightning would strike from the storms. The automated objects were generated from this algorithm, producing polygons around the areas where it predicted CG lightning would occur for specific case studies.

c. Creation of Forecaster Objects

Nine NWS forecasters of varying skill and expertise all participated in the 1 Sept 2016 case study in order to determine if the forecasters added value to the automated system. All forecasters had the automated objects from the random forest algorithm as a ‘first guess’ as they utilized the PHI tool. Similar to Karstens et al. (2018), the forecasters had choices to either leave the automated objects that were previously created by the system, modify the automated objects, or manually create their own objects around areas where they predicted lightning would occur that the automated system might have missed, corresponding to the levels of modification provided by the PHI tool. In addition to adjusting and creating objects, the forecasters also had the ability to add discussion regarding what they saw in the model, such as providing information about the specific locations of CG lightning from the National Lightning Detection Network (NLDN) and trends in the lightning activity associated with an object. They also included a percentage to each storm which gave their confidence of that storm's ability to produce CG lightning.

d. National Lightning Detection Network (NLDN)

The NLDN lightning data was used to determine the specific CG lightning locations for forecast verification (Vaisala cited 2018). NLDN data was collected through a system of antennas that were located across the United States. (Hondl and Eilts 1994). This data gave the location of CG lightning strikes around the United States. The CG flashes that fell within the MLB warning area and during the time of the case were recorded. The automated and forecaster objects were compared to the lightning data in order to determine how accurately they forecasted where lightning struck during the simulation.

e. Object Verification

As the automated objects were generated by the ML algorithm and additional objects were created by the forecasters, verification of the objects is necessary to determine how accurate the automated system and forecasters were at predicting CG lightning. By using the NLDN data to determine where CG lightning flashes occurred during the time of the case, a program was developed to count the number of flashes that were successfully forecasted by the objects.

The objects that were included in the verification were those that fell between 1755 and 1955 Coordinated Universal Time (UTC), the period of forecaster created objects in the case study. Only the storms within the MLB area including a buffer with an average of around 45 miles off the Atlantic coast from the warning area were considered. All forecasters began with the automated objects as a 'first guess' to coverage and probabilistic likelihood of CG activity. Each of the following were included in calculating the verification statistics for each forecaster (1) automated objects that forecasters chose not to modify (and did not block from end-users), (2) automated objects that forecasters modified, and (3) newly created objects by forecasters. If forecasters modified an automated object, the original automated objects were removed from the verification dataset by matching the IDs and start time of the modified objects as some objects were modified for only part of that objects existence. The valid CG flashes for each object were determined as occurring between the start time of that object and one minute after the start time. All datasets were converted to epoch time to assist with the validation process. The data was examined in two ways: storm verification and lightning verification. Storm verification was whether the storms created successfully forecasted the occurrence of CG lightning. Lightning verification was if the CG flashes was within a storm object (hit) or outside any storm objects (miss).

For the storm verification, four categories were used: Hit, False Alarm, Correct Null, and Miss (Wilks 2005). Each object, for its duration, were sorted into these four categories at each threshold from 0 % to 100 % in 10 % increments. The duration of the automated objects were 60 minutes and those of the forecaster objects were dependent on forecaster choice. A storm was recorded as a Hit if it had at least one CG flash and the probability of the storm was above the given threshold. A storm was sorted as a False Alarm if it had no CG flashes but had a probability above the threshold. A Correct Null was recorded when the storm had no CG flashes and the probability of the storm was below the threshold. Finally, a Miss was recorded when there was at least one CG flash inside the storm, but the storms probability was below the threshold.

With the lightning verification, CG flashes were scored during the duration of each object using the same thresholds as in the storm verification. Each flash was organized into one of two categories: Hit or Miss. If the CG flash was inside a storm and the storms probability was above the threshold, a hit was recorded. If the CG flash was not inside a storm it was recorded as Miss 1. A second type of miss was defined when the CG flash was inside a storm but the storms probability was below the threshold: Miss 2.

These verification values were organized using a 2x2 Contingency Table in two sections: event forecasted and event observed (Fig. 3). Each section was further broken down into yes or no, which created four positions for the data to be organized into (Roebber 2009). A Hit was in the position which referred to when an event was both observed and forecasted, a False Alarm was located where an event was forecasted but not observed, a Miss was in the position when the event was observed and not forecasted, and a Correct Null was in the spot which referred to an event that was both not forecasted and not observed. For storm verification, the table was viewed by starting at the event observed section and the lightning verification was approached from the event observed section.

2x2 Contingency Table		Event Observed	
		Above Threshold	Below Threshold
Event Forecast	Above Threshold	Hit	False Alarm
	Below Threshold	Miss	Correct Null

FIG. 3. 2x2 Contingency Table for Storm and Lightning Verification Values

f. Skill Score

Using the values from the storm and CG lightning verifications, the skill scores of each forecaster and the automated system were calculated. The skill scores included the False Alarm Ratio (FAR), Critical Success Index (CSI), and Probability of Detection (POD). The FAR provides the proportion of forecasted storms that failed to actually produce lightning (Gremillion and Orville 1999). The equation used to calculate FAR was:

$$FAR = \frac{FalseAlarm}{Hit + FalseAlarm}$$

The CSI, sometimes referred to as threat score, provides the correctly forecasted storms divided by the sum of the total number of forecasted storms and the number of incorrectly forecasted storms that did not actually occur (Gremillion and Orville 1999). The equation used to calculate CSI was:

$$CSI = \frac{Hit}{Hit + FalseAlarm + Miss}$$

The POD provides the chance that the CG lightning flashes would be forecasted if the CG flashes actually occurred (Gremillion and Orville 1999). The equation used to calculate POD was:

$$POD = \frac{Hit}{Hit + Miss}$$

Each of these values were calculated for each threshold, 0 % to 100 %, in 10 % intervals. The FAR and CSI were calculated using the storm verification and the POD used the lightning verification values. The use of these calculations helped to determine, overall, how accurate the forecasting model actually was (Gagne et al. 2017) for the automated system and the forecasters.

3. Results

FAR, CSI, and POD are plotted together for each of the forecasters and the automated system to provide a comparison across all forecasters and the automation for the event (Fig. 4). For each individual, the threshold was on the x-axis at 10 % intervals and the skill scores were plotted on the y-axis with values from 0 to 1. All of the forecasters were able to increase the POD above the automated system alone. Out of the forecasters at the 0 % threshold level, Forecaster 1 had the lowest POD and Forecaster 4 had the highest POD. The FAR and CSI values were more consistent between the automated system and the forecasters.

Bar graphs were constructed to show the variability of the Hit, False Alarm, Correct Null, and Miss categories for the storm verification for each forecaster and automation. The threshold was plotted on the x-axis, similarly to the line graphs, and the number of storms for each category was plotted on the y-axis. Four bars, representing each of the categories, were created for each threshold level. At the 0 % threshold level, Forecaster 1 had the most hits and Forecaster 6 had the least amount of hits. Only forecasters 1 and 4 had more hits than the automated system, however, the forecasters all had fewer false alarms than the automated system.

Bar Graphs were also generated to show the variability of Hits and Misses for the lightning verification. Graphs showing the misses overall and graphs splitting the misses into the two different options for misses were created, revealing how many Miss 1' s and Miss 2' s were present. Overall, all the forecasters had more hits than the automated system. At the 0 % threshold level, Forecaster 4 had the most hits and Forecaster 1 had the least amount of hits. The automated system had the most of Miss 1, and Forecaster 1 followed as the forecaster with the most of Miss 1. The 20 % threshold level is the first threshold level where Miss 2 was present, for which, Forecaster 6 had the most.

A performance diagram, also known as a Roebber diagram, compared how successful the automated system and each forecaster were at predicting CG lightning relative to each other. This performance diagram plots the POD, FAR, and CSI as well as the bias of the verification, all four values geometrically related to each other (Roebber 2009). The POD is plotted along the y-axis and the Frequency of Hits (FOH), otherwise known as the Success Ratio, is plotted on the x-axis. The value of the FOH is 1-FAR (Roebber 2009). CSI values are plotted as curves on the diagram from 0.1 to 0.4. The biases are each plotted as straight, dashed lines on the diagram, from 0.1 to 10. Each object was plotted as a point based on the FAR and POD of that object. The 0 % threshold was used so that all of the objects would be valid for each individual. Those located in the middle diagonal, where the bias is equal to 1, were unbiased and as they approached the upper-right corner of the diagram the forecasts were more accurate (Roebber 2009).

The performance diagram generated for the automated system and forecasters reveals that the forecasters did add value to the automated system (Fig. 5). The ranges of the axes for this figure are from 0.6 to 1 for the FOH and 0.8 to 1 for the POD to make it easier to see where the automation and forecasters fall within the diagram. All of the forecasters improved the POD of the automated system and most of the forecasters also decreased the overall FAR. Forecaster 1 was located closest to the upper-right corner of the performance diagram and thus was revealed to have the most successful POD and FOH performance out of all of the individuals; however, this does not account for areal coverage of the storm objects.

A reliability diagram was also constructed in order to view the automated system and forecasters abilities at

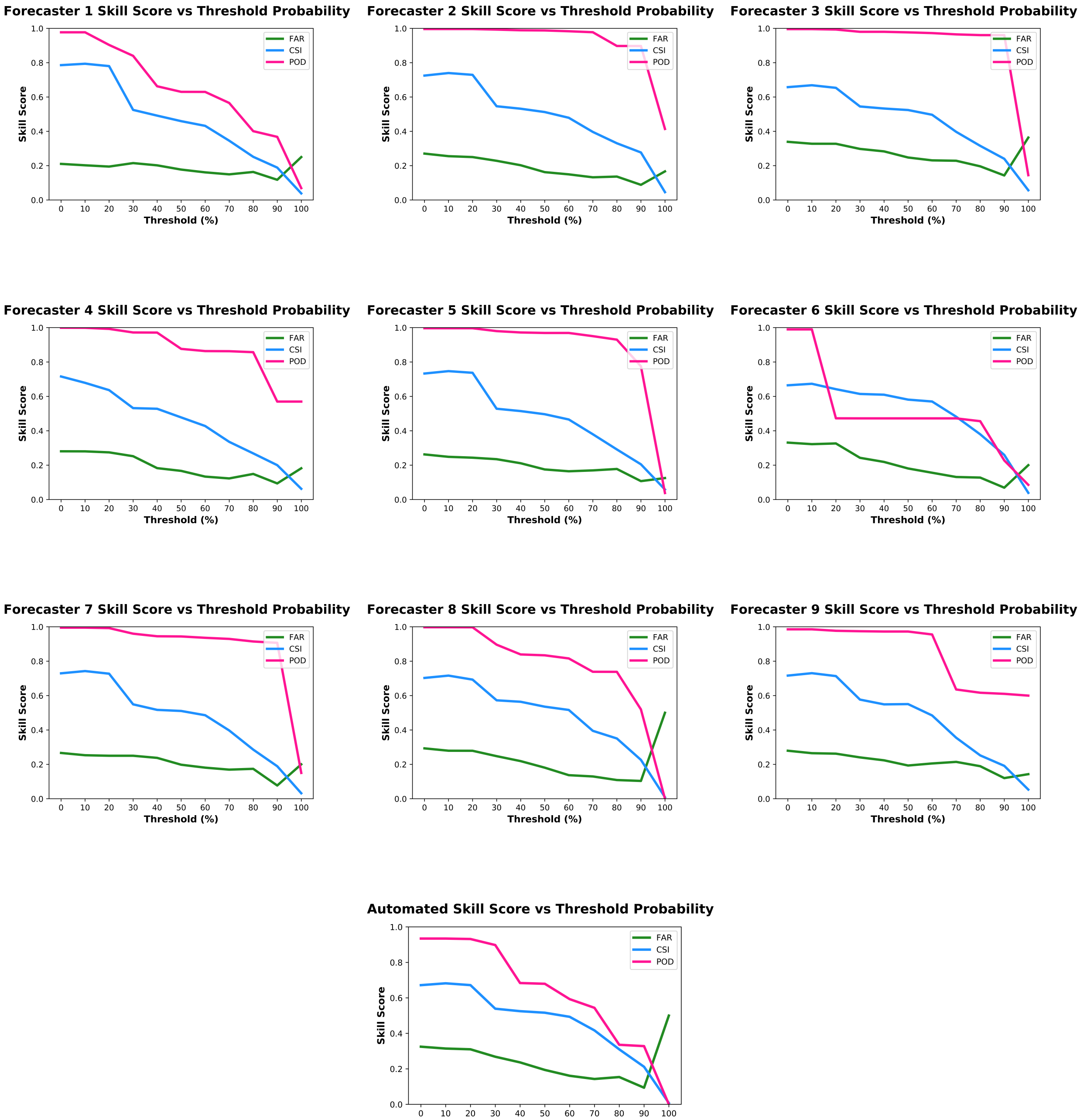

FIG. 4. Skill Score values of FAR (blue line), CSI (green line), and POD (pink line) of the nine forecasters and the automation

forecasting CG lightning across the entire probabilistic range (Fig. 6). The dashed grey line represents an idealized perfectly reliable forecast; anything above this line is under-forecasted and anything below the line is over-forecasted. The thick, black line represents the automated system and rainbow colors, the forecasters. The automated system and forecasters all under-forecasted for most of the thresholds, but tended to over-forecast around the 50 % and 90 % thresholds. The forecasters improved on the reliability at the 50 % and 70 % thresholds, but many also had greater difficulty than the automation at lower probabilities, under-forecasting more often.

A bar graph, separated by each forecaster and the automated system, was also plotted in the bottom right of the diagram in order to show how many objects were forecasted at each threshold. Overall, this shows that the

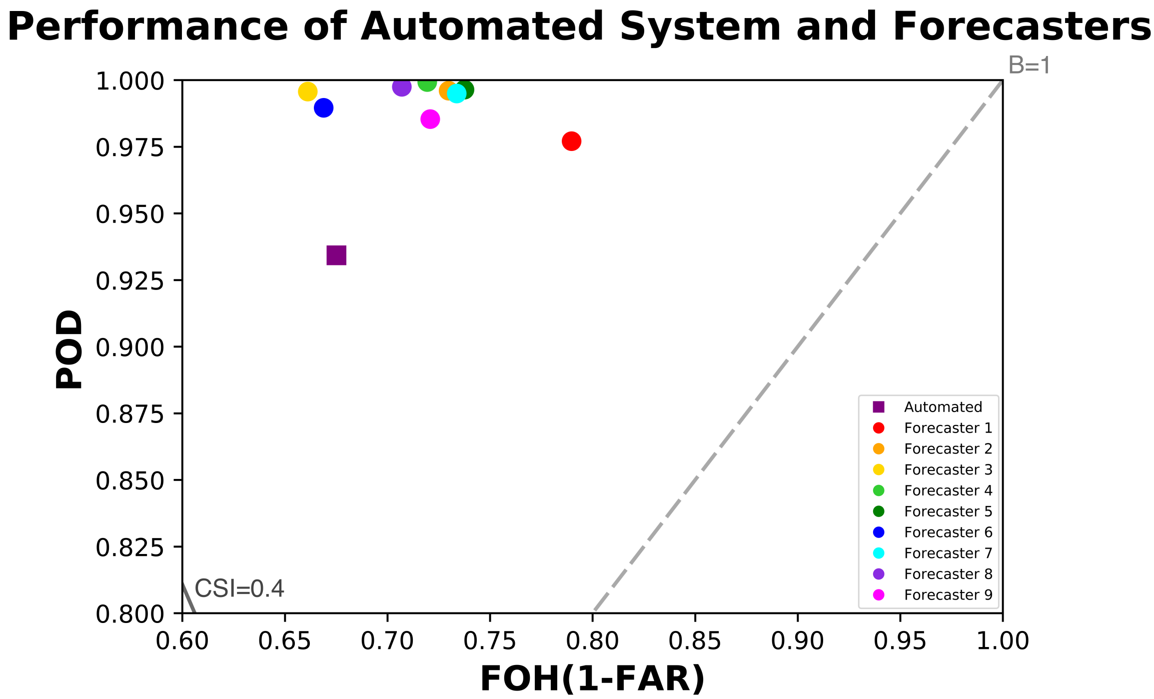

FIG. 5. Performance Diagram of Automation and Forecasters

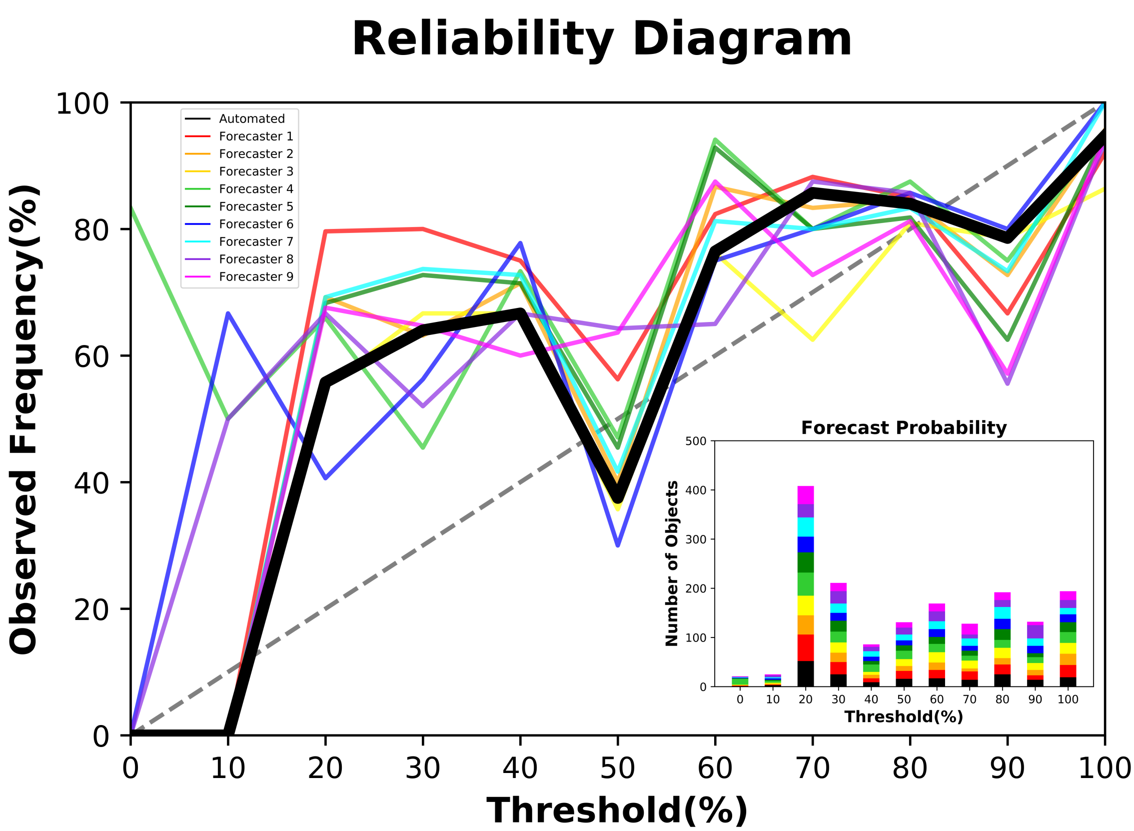

FIG. 6. Reliability Diagram of Forecasters and Automation

forecasters forecasted a similar number of objects at each threshold as the automated system, however, Forecaster 2 and Forecaster 5 appear to have a smaller number of objects at most of the thresholds than the other forecasters and automation does and could be considered outliers compared to the others. The 30 % threshold has a much greater number of objects than the other thresholds. The automated system forecasted the most objects at this threshold and all of the forecasters likely followed suit when using the automation as a guidance, thus allowing this threshold to have the most objects overall. Currently there is no definite answer to why the 30 % threshold was forecasted at so much, but a possible explanation could have to do with the lack of experience of the system to forecast tropical events.

In order to determine how the forecasters added value to the automated system, the data from each forecaster object was examined. Forecasters could modify many aspects of the automated objects including: Speed, Direction, Discussion, Severity (the frequency of lightning), the object shape, and probability. Looking at each characteristic, the number of modified automated objects that were modified for an individual characteristic were tallied (Fig. 7). The type with the most modifications among the automated objects was the discussion of the objects, forecasters often provided details on individual cities and if they were expecting the storm to strengthen or weaken here. Other characteristics that were most frequently modified was the duration of the storms CG lightning, the probability of the CG lightning, and the severity (or frequency) of CG lightning. The qualities that were not modified as often was the object itself, the warning threshold, impacts, confidence, actions, and warning type of the CG lightning.

Modifications of the characteristics were binned by forecasters to determine if certain characteristics were only changed by a few of the forecasters or if it was a quality that all the forecasters felt was important to adjust (Fig. 7). The discussion, duration, probability, and severity of the CG lightning of the storms were adjusted by all the forecasters while the alert level, warning threshold, impacts, confidence, and actions of the objects were only modified by four of the nine forecasters. Forecaster 8 made the largest number of modifications of the automated objects (with 343 changes) and Forecaster 3 made the fewest modifications (with only 12 modifications across the entire event).

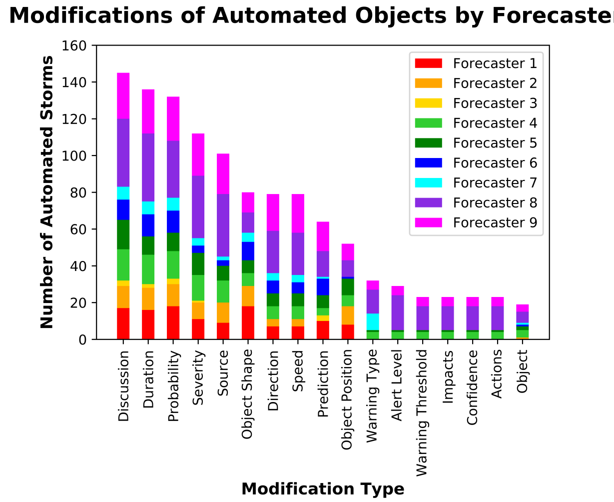

FIG. 7. Graph of modifications made by forecasters separated by forecaster

4. Discussion

The case held a lot of importance due to both the area where it took place and the challenge that the case presented. The MLB area includes many outdoor tourist attractions such as Disney World, Universal Studios, and Sea World in Orlando, Florida. The warning area also contains Cape Canaveral, where the Kennedy Space Center is located, and many beaches along the Atlantic coast of Florida. These locations are typically popular which led to the potential for many people to be outdoors and exposed to CG lightning when the storms hit. The case also took place during Hurricane Hermine which was mostly affecting the Western coast of Florida. As discussed in the event background, the presence of Hermine made this case rather difficult for both the automation (since this is a rare event) and the forecasters. The difficulty of the case, however, was part of the motivation for running the system through it in order to see how the automated system and the forecasters performed with a challenge.

Several limitations occurred over the course of the research. Only nine forecasters came and ran through the case, which is a small sample size. Thus, the data could easily be skewed and it may be difficult to get strong results. Also, in the performance diagram, the only value that wasnt considered when calculating the skill scores was the Correct Null values from the storm verification. However, the failure to include the Correct Null values did not hinder the ability of the diagram to accurately show the performance, but actually had the potential to enhance the diagram by excluding the excess events that were not forecasted and did not occur (Roebber 2009).

Another limitation could be that halfway through the case, Forecaster 1 became overwhelmed by the numerous amount of objects present and began to focus on only the top half of the MLB county warning area. This meant the automated system took over for the bottom half of the region for the remaining half of the case for this forecaster. No other forecaster did this during their case, so this change for Forecaster 1 could affect the results. According to the performance diagram, Forecaster 1 had the most successful performance with predicting CG lightning, which indicates that a human-machine mix in which the forecaster is not doing as much work with the automated objects could have the potential to work well.

Broadcasters and emergency managers were brought in to run a simulation of the case with the EDD, which used the PHI objects from a forecasters case. The forecasters provided useful information for these end users, such as the added discussion to the objects, that the automated system alone would not have. This information gave the end users a better idea of what was occurring with the objects and could help them pick out which objects may be more important to pay attention to, especially if there was a multitude of objects at one time during the case. Many of the broadcasters tended to think that the lightning objects were confusing and claimed they likely would not share them on air due to lower level of importance in comparison to the threat of tornadoes or severe storms. However, by having the discussion and other qualities about the objects, they had some guidance of when to warn certain areas about CG lightning. The emergency managers gave a positive response to the system and claimed that they wanted the product in operational use as soon as possible (Meyer et al. 2017). The end users provided some insight on how forecasters were adding value to the automated system, pointing out the discussion as an important component that the forecasters supply.

When looking at the modifications that were made by the forecasters, it was interesting to view which forecasters made the most and least modifications and how those forecasters performed relative to each other. Forecaster 8 made the most modifications of the automated objects while Forecaster 3 made the least amount of modifications. When looking at the performance diagram, Forecaster 8 had a slightly higher POD and a higher FOH than Forecaster 3, indicating that the modifications did make an impact in helping the success of predicting CG lightning.

An automated system alone may not be completely trustworthy to a forecaster because of possible detection errors or misuse of the system that may have occurred in the past (Karstens et al. 2018). However, a system in which the forecaster has to manually create and issue all the objects themselves is unreasonable; this would waste a lot of time, often which the forecaster does not have. The forecasters who ran through the PHI prototype found that task of modifying and creating polygons based off the automated system was already quite tedious. They felt as though they were spending too much time with the objects and if they had to do the task along with their regular work, especially during a time with a lot of severe events occurring, there would be too much for them to be in charge of alone. This is where determining where the forecasters added value to the automated system came in handy.

All of the forecasters improved the automated system in terms of performance (Fig. 5). The bar graph of the modifications revealed that the forecasters provided the most value with the added discussion of the objects as well as with changing the duration, probability, and severity of the CG lightning polygons. An option for the forecasters could be to let the automated system be the baseline and create the polygons in an area, and they could come in and add discussion for the objects and quickly edit these qualities that were modified the most during the case. This would allow the forecasters to spend less time with the objects, but still let them edit the objects so they are not relying on only the automated system, thus providing the mix of both human and machine elements into one system.

Another solution would be to continue to improve the automated systems ability at forecasting CG lightning. For example, adding a buffer to the automated objects to increase the area the objects cover would increase the POD of the automated system. The objects would likely catch more of the CG lightning strikes and thus miss less of the CG flashes. However, increasing the area of the objects would potentially make the model more difficult to use operationally. The area could look messy with many large objects and become confusing for end users viewing the objects. Some broadcasters viewed many large objects from individual forecasters, making the PHI system appear cluttered and not truly discriminating among areas with this highest threat. This served as a motive for why some chose not to use the lightning information at all when running through the simulation. There becomes a tradeoff between having a higher POD value and better performing system and having a more complicated system for the end users. However, with more research there could be other enhancements that would improve the performance of the automated system.

5. Conclusion

The automated system, while not perfect on its own, provided a baseline to the forecasters when forecasting CG lightning. The forecasters added value to the automated system and improved the automated objects through modifying the qualities such as the discussion and duration of the storms. This project provided an example of performance for this automated system if it were put into operational use right now. It shows which aspects of the objects the forecasters would likely need to focus on when using the automated system as well as which areas they would not need to concentrate on so they do not spend too much time with the system, but still produce accurate forecasts of CG lightning.

We suggest that additional improvements be made to the automated system in order to enhance its success of forecasting CG lightning following the guidance of where forecasters added value. This could allow the system to do a better job on its own, thus lessening the need of the forecasters to focus on the automated system and allowing them to continue their work where they feel is the most important. Eventually the automated system could be put into operational use and CG lightning could officially be predicted by forecasters, particularly as the NWS stresses the importance of Impact Decision Support Services (IDSS). The information could also potentially work towards the development of a CG lightning warning system in which physical warnings are issued to areas regarding if they have a high risk of lightning that could affect their area.

Acknowledgments. The corresponding author thanks Dr. Daphne LaDue and the National Weather Center Research Experience for Undergraduates. Thanks are also extended to Melanie Schroers and Jacqueline Waters for their help and support through this experience and Patrick Campbell for his assistance with coding.

This work was prepared by the authors with funding provided by National Science Foundation Grant No. AGS-1560419, and NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NOAA, or the U.S. Department of Commerce.

References

Berg, R., 2017: Hurricane Hermine. *National Hurricane Center Tropical Cyclone Report*, 1–63.

Calhoun, K., and Coauthors, 2018: Cloud-to-Ground Lightning Probabilities and Warnings within an Integrated Warning Team. *American Meteorological Society*, Austin, Texas, Amer. Meteor. Soc., [Available online at https://ams.confex.com/ams/98Annual/meetingapp.cgi/Paper/329888].

Cartier, K. M., cited 2017: New Model Predicts Lightning Strikes; Alert System to Follow. [Available online at https://doi.org/10.1029/2017EO088591].

Cummins, K. L., and M. J. Murphy, 2009: An Overview of Lightning Locating Systems: History, Techniques, and Data U ses, With an In-Depth Look at the U.S. NLDN. *IEEE Transactions on Electromagnetic Compatibility*, **51 (3)**, 499–518, doi:10.1109/TEMC.2009.2023450.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, **32 (5)**, 1819–1840, doi:10.1175/WAF-D-17-0010.1.

Gremillion, M. S., and R. E. Orville, 1999: Thunderstorm Characteristics of Cloud-to-Ground Lightning at the Kennedy Space Center, Florida: A Study of Lightning Initiation Signatures as Indicated by the WSR-88D. *Weather and Forecasting*, **14 (5)**, 640–649, doi:10.1175/1520-434(1999)014⟨0640:TCOCTG⟩2.0.CO;2.

Hondl, K. D., and M. D. Eilts, 1994: Doppler Radar Signatures of Developing Thunderstorms and Their Potential to Indicate the Onset of Cloud-to-Ground Lightning. *Monthly Weather Review*, **122 (8)**, 1818–1836, doi:10.1175/1520-0493(1994)122⟨1818:DRSODT⟩2.0.CO;2.

Karstens, C. D., and Coauthors, 2015: Evaluation of a Probabilistic Forecasting Methodology for Severe Convective Weather in the 2014 Hazardous Weather Testbed. *Weather and Forecasting*, **30 (6)**, 1551–1570, doi:10.1175/WAF-D-14-00163.1.

Karstens, C. D., and Coauthors, 2018: Development of a HumanMachine Mix for Forecasting Severe Convective Events. *Weather and Forecasting*, **33 (3)**, 715–737, doi:10.1175/WAF-D-17-0188.1.

Liaw, A., and M. Wiener, 2001: Classification and Regression by RandomForest. **23**.

MacGorman, D. R., I. R. Apostolakopoulos, N. R. Lund, N. W. S. Demetriades, M. J. Murphy, and P. R. Krehbiel, 2011: The Timing of Cloud-to-Ground Lightning Relative to Total Lightning Activity. *Monthly Weather Review*, **139 (12)**, 3871–3886, doi:10.1175/MWR-D-11-00047.1.

Meyer, T., K. Calhoun, D. Kingfield, and C. Karstens, 2017: Using Random Forest to Generate Cloud-to-Ground Lightning Probabilities. *American Meteorological Society*, Austin, Texas, Amer. Meteor. Soc.

Mosier, R. M., C. Schumacher, R. E. Orville, and L. D. Carey, 2011: Radar Nowcasting of Cloud-to-Ground Lightning over Houston, Texas. *Weather and Forecasting*, **26 (2)**, 199–212, doi:10.1175/2010WAF2222431.1.

National Severe Storms Laboratory, cited 2018: Forecasting a Continuum of Environmental Threats. [Available online at https://www.nssl.noaa.gov/projects/facets/].

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Weather and Forecasting*, **24 (2)**, 601–608, doi:10.1175/2008WAF2222159.1.

Vaisala, cited 2018: National Lightning Detection Network NLDN. [Available online at https://www.vaisala.com/en/products/data-subscriptions-and-reports/data-sets/nldn].

weather.gov, cited 2018: U.S Lightning Deaths in 2018: 15**. [Available online at https://www.weather.gov/safety/lightning-fatalities].

Wilks, D., 2005: *Statistical Methods in Atmospheric Sciences, Volume 100, 2nd Edition*. Academic Press, 648 pp.