

Cause-Effect Modeling and Spatial-Temporal Simulation of Power Distribution Fault Events

Yixin Cai, *Student Member, IEEE*, and Mo-Yuen Chow, *Fellow, IEEE*

Abstract—Modeling and simulation are important tools in the study of power distribution faults due to the limited amount of actual data and the high cost of experimentation. Although a number of software packages are available to simulate the electrical signals, approaches for simulating fault events in different environments have not been well developed. In this paper, we propose a framework for modeling and simulating fault events in power distribution systems based on environmental factors and the cause-effect relationships among them. The spatial and temporal aspects of significant environmental factors leading to various faults are modeled as raster maps and probability functions, respectively. The cause-effect relationships are expressed as fuzzy rules and a hierarchical fuzzy inference system is built to infer the probability of faults in the simulated environments. A test case simulating a part of a typical city's power distribution systems demonstrates the effectiveness of the framework in generating realistic distribution faults. This work is helpful in fault diagnosis for different local systems and provides a configurable data source to other researchers and engineers in similar areas as well.

Index Terms—Discrete event simulation, fault diagnosis, fuzzy systems, power distribution faults, power system simulation, system modeling.

I. INTRODUCTION

DISTRIBUTION faults are the major source of customer reliability problems in power systems and have been drawing more and more attention in recent decades [1]. In addition to studying the electrical characteristics of faults, such as fault location based on monitored voltages and currents [2], [3] and the experimental characterization of electrical signals generated during tree contact [4] and lightning strikes [5], many researchers have treated distribution faults as discrete events and established relationships between environments and the root cause of the faults. To mention some, Xu *et al.* [6] and Sahai and Pahwa [7] characterized and modeled environmental factors related to tree-caused and animal-caused faults by logistic regression and Bayesian networks; Xu and Chow *et al.* [8], [9] used biologically inspired algorithms, including

artificial neural networks (ANN), fuzzy systems, and artificial immune recognition systems (AIRS), to identify the fault's root cause based on environmental properties.

Due to the limited amount of actual data and the high cost of experimentation, modeling and simulation are essential tools in studying power distribution faults. Although a number of software packages are available to simulate the electrical signals, the study of faults as discrete events is still limited by the lack of modeling and simulation approaches. The stochastic nature of faults, spatial heterogeneity of distribution systems, and measurement noises make it difficult for one single dataset to represent all systems. Thus, researchers often find that an algorithm performing well on one set of fault records may not achieve the same performance on another fault dataset. With limited access to the same fault dataset, comparing work done by different research groups can be similar to comparing apples and oranges.

In this paper, we propose a framework for modeling and simulating fault events in power distribution systems based on environmental factors and the cause-effect relationships among them. This framework provides a configurable data source capable of representing different local systems in various environments. This data source would help in developing and evaluating distribution fault diagnosis algorithms. Moreover, repeated simulations under different conditions may help engineers to identify regions and time periods with an unusually high probability of faults so that investments can be prioritized and crew deployments can be planned.

Cause-effect relationships have been investigated to help diagnose faults in distribution systems. Deterministic cause-effect relationships are used to locate the fault section. By discovering the causal relationship between the fault section and the corresponding sequence of protective actions, the actual fault section can be identified based on the recorded protective actions. To cope with the complex scenarios of a large distribution system and the inherent noises and uncertainties in the recorded protective actions, Chen *et al.* [10] used a cause-effect network and fuzzy sets, while Zhang *et al.* [11] and Hor *et al.* [12] proposed using Rough Set theory to discover the actual relationships. Reliability-based modeling of equipment failure is also well studied. The residual life of a device before its next failure is modeled as various probability distributions. Techniques to calibrate and validate the model parameters are developed as well [1], [13]. In this type of model, faults are more likely to occur as time advances. Taking the environment into consideration, Kuntz *et al.* [14] and Radmer *et al.* [15] included yearly average temperature and tree growth status in the reliability model via regression and ANN.

These techniques provide good tools for fault analysis and serve their purpose quite well. However, they are not adequate

Manuscript received January 22, 2010; revised January 22, 2010, April 11, 2010, and May 25, 2010; accepted June 22, 2010. Date of publication August 09, 2010; date of current version April 22, 2011. This work was supported by the National Science Foundation (NSF) through Grant No. ECCS-0653017 (Small World Stratification for Power System Fault Diagnosis with Causality). Paper no. TPWRS-00042-2010.

The authors are with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: ycai2@ncsu.edu; chow@ncsu.edu).

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

Digital Object Identifier 10.1109/TPWRS.2010.2055899

for the cause-effect simulation of fault events. First, distribution system faults are essentially stochastic events without any deterministic causal relationships. The cause-effect model can only represent the approximate trend of faults with large uncertainties. Second, the fault characteristics are location-dependent and vary over time. Faults in a new metropolitan distribution system should behave quite differently from those in a 40-year-old single-phase long feeder in a rural area. Thus, both spatial and temporal information need to be considered in the model. Last but not least, environmental factors, most importantly weather conditions, need to be modeled and simulated with a time resolution of minutes or hours instead of a monthly or yearly average.

All these requirements make the cause-effect modeling and spatial-temporal simulation of distribution fault events a challenging task. In the following sections, we lay out our solution. We first identify significant environmental factors leading to distribution faults, find appropriate models describing these natural phenomena, and simulate the environments at different times and locations. Then the causal relationships between these factors and various fault types are investigated. Knowledge gained from field experience and the analysis of historical data is represented as rules and modeled using fuzzy inference systems (FIS) [16]. As the causal relationships apply to all systems and environmental factors vary across space and time, the generated fault events can represent realistic characteristics under a wide range of conditions.

II. ENVIRONMENTAL MODELING AND SIMULATION

The fundamental information needed for the cause-effect modeling and spatial-temporal simulation of distribution faults is about the power system itself, such as its power line topologies and switchgear configurations. Some of these properties affect the fault characteristics (e.g., whether the power line is overhead or underground). Others are kept for post-fault analysis.

The first step in environmental modeling is to model the significant environmental factors that lead to various faults. The selection of significant factors and the modeling of the spatial and temporal aspects of these factors are discussed in this section.

A. Significant Environmental Factors

Field experience and the knowledge gained analyzing historical fault data are used to select the significant environmental factors. For example, it is easy to understand that when strong winds blow the tree branches harder, we are more likely to have tree-caused faults. Meanwhile, strong winds reduce animal activities, and historical data tells us that fewer animal-caused faults occur on windy days (as shown in Fig. 1). Therefore, wind speed is chosen as a significant factor. As recommended by experienced distribution engineers, factors identified as significant for fault diagnosis [17], such as season and rainfall intensity, are considered important for environmental modeling in this paper. As environmental characteristics change, more factors may be required to model a particular local system.

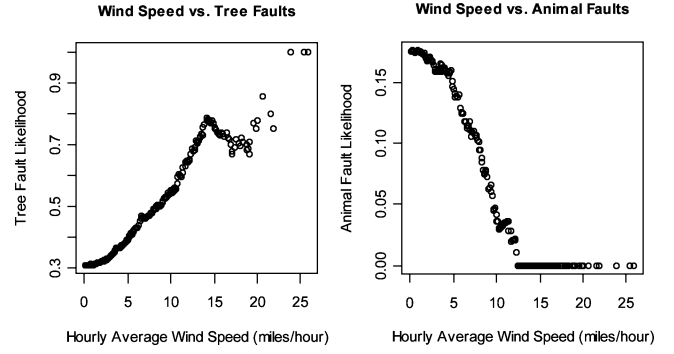


Fig. 1. Likelihood of tree-caused and animal-caused faults under different wind speeds. (Data source: distribution faults in the Garner operation center of Progress Energy Carolinas between 2005 and 2006)

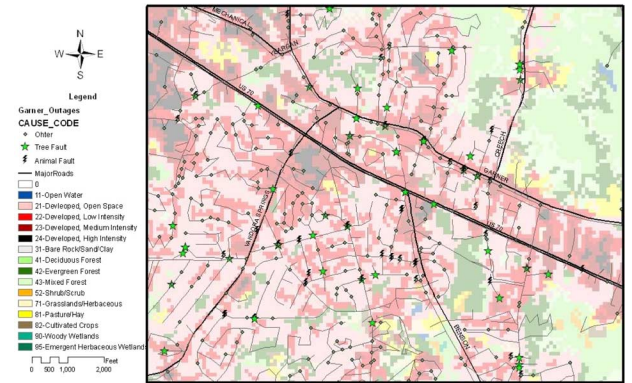


Fig. 2. Example of raster maps showing land use types.

B. Modeling Spatial Information

The spatial properties of environments describe the locations of those geographic features of interest, such as the layout of highways and vegetation. Maps are the everyday tool for representing spatial information and are used in our model. Although power lines are usually represented in vector format in power system software, raster maps are preferred in our model due to their flexibility for modeling various features and their ease of manipulation [18]. A raster map consists of grids of a fixed size, where every grid is given a particular number to represent the value, category, or presence of a certain factor. For example, with other layers containing additional spatial information, land use types are represented as the different integer values of the cells, shown as different colors in Fig. 2.

Spatial information is assumed to be static during the simulation period. Although some of the factors do change over time (e.g., the distribution system may expand, urban development may change roads or land use), these changes usually take a long time to complete. Then, they can be treated as new environments, modeled in updated maps, and simulated in a new study period.

C. Modeling Temporal Information

Theoretically, every aspect of the environment is constantly changing at a different rate. For example, utility companies manage the trees along power lines and carry out equipment maintenance and upgrades every few years; the starting and

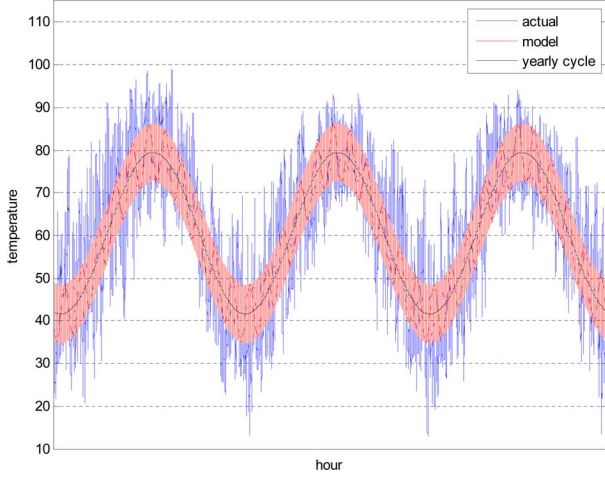


Fig. 3. Three-year hourly average temperature. (Data source: CLAY weather station in Clayton, NC, between 2004 and 2006)

ending dates of a season drift slowly from year to year; precipitation patterns change throughout the year; wind speed changes every instant.

Weather conditions are the most significant temporal information affecting the cause-effect model. Assume the region under simulation is small enough that the weather conditions within it can be regarded as the same. A large region can be divided into smaller areas, modeled, and simulated separately. Three major weather measures, temperature, wind speed, and rainfall intensity, are considered in this paper.

Temperature is widely known to have yearly and daily cyclic patterns, which are modeled as

$$T = \bar{T} + A \sin \frac{2\pi d}{365} + B \cos \frac{2\pi d}{365} + a \sin \frac{2\pi h}{24} + b \cos \frac{2\pi h}{24} + N(0, 1) \quad (1)$$

where $\bar{T}$ is the yearly average temperature, d is the day of the year ($1, 2, \dots, 365$), h is the hour of the day ($0, 1, 2, \dots, 23$), A, B, a , and b are parameters reflecting the amplitude and phase of the sinusoidal waves in (1). One example of the observed hourly average temperature during a three-year period is shown in Fig. 3. Note that even though this model does not provide the exact temperature patterns, it gives an extraction of features that is good enough for the simulation.

In the literature, wind speed is usually modeled by a Weibull distribution [19], described as

$$f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (2)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. One example of the actual one-year hourly average wind speed is shown in Fig. 4.

As shown in Fig. 4, the probability of the wind speed being 0 is much greater than that described in the Weibull model. Thus, we first add the constraint that the probability that the wind speeds are greater than 0 and then use a Weibull distribution

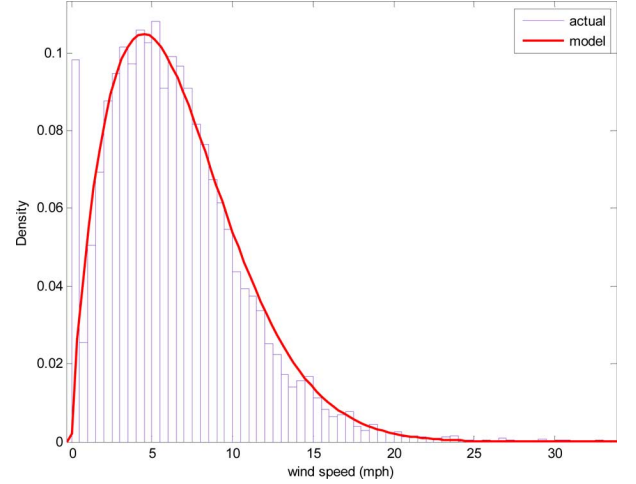


Fig. 4. One-year hourly average wind speed. (Data source: LAKE weather station in Raleigh, NC, in 2003)

to model the speed. Another practical constraint on wind speed is that it cannot be arbitrarily high. Although very large values appear with a tiny probability in the model (2), the upper limit is set to prevent unrealistic wind speeds from being generated. The final wind speed model used in the simulation is as follows:

$$f(x; \lambda, k) = \begin{cases} 0 & x > x_{\max} \\ \frac{kp}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x_{\max} \geq x > 0 \\ \frac{1-p}{1} & x = 0 \\ 0 & x < 0 \end{cases} \quad (3)$$

where x is the wind speed, $x_{\max} > 0$ is the maximum wind speed that is allowed, $k > 0$ is the shape parameter, $\lambda > 0$ is the scale parameter, and $p > 0$ is the probability of the wind speed being greater than 0. Based on our observations, p often takes a value close to 1. Note that λ is not the average wind speed. However, with p and k fixed, λ can be used to represent the average strength of the wind.

Similarly, the intensity of rainfall is modeled as an exponential distribution constrained by the probability of rainfall, as follows:

$$f(x; \lambda) = \begin{cases} 0 & x > x_{\max} \\ \frac{p}{\lambda} e^{-x/\lambda} & x_{\max} \geq x > 0 \\ \frac{1-p}{1} & x = 0 \\ 0 & x < 0 \end{cases} \quad (4)$$

where x is the rainfall intensity, $x_{\max} > 0$ is the maximum rainfall intensity allowed, $\lambda > 0$ is the scale parameter, and $p > 0$ is the probability of rainfall. Unlike wind, rainfall is a relatively rare weather event, so the value of p in (4) is usually set close to 0. One example of the actual one-year hourly rainfall data is shown in Fig. 5. Note that hours with exactly zero rainfall are excluded from the plot.

III. CAUSE-EFFECT MODELING OF FAULT EVENTS

Cause-effect relationships summarize the typical correlation between environmental factors and faults, disregarding the time

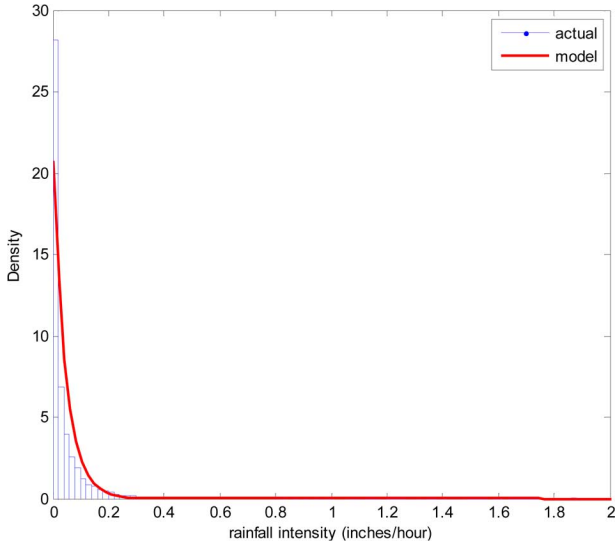


Fig. 5. One-year hourly rainfall intensity excluding the hours without any rainfall. (Data source: RDU weather station in Morrisville, NC, in 2005)

and the location. For example, flourishing trees around power lines during a thunderstorm are very likely to lead to tree-caused faults. In our model, these causal relationships are represented by fuzzy rules and are used to infer the probability of various faults given the environmental conditions.

A. Fuzzy Environmental Factors

To apply the fuzzy rule-based inference, significant environmental factors are first converted to fuzzy sets. Membership functions are defined to map the values generated by the environmental models into linguistic descriptions.

For categorical values, the linguistic categories are already defined. Fuzzy sets mainly represent the uncertainty of the input value. The number of fuzzy sets used to represent different categories is the same as the number of levels in the input variable. The membership function for these categorical variables has multiple peaks where the location and peak value can be specified. The uncertainty of the input being one category is described by varying the height of the peaks. For each fuzzy set, the highest peak represents the corresponding input level, while the lower peaks at other values represent the uncertainty. One example uses nine fuzzy sets to represent the nine land use types defined by the USGS Land Cover Institute [20]. Fig. 6 shows the fuzzy membership function used to describe land use type *forest*. For categories that are clearly separable from *forest*, such as *water*, the height of the lower peak is close to 0. For categories that are not so easy to differentiate, such as *non-natural woods*, the lower peak could be of a height similar to the highest one.

Continuous values are more conventional as input to fuzzy systems. Therefore, a large collection of standard membership functions, including triangular, trapezoid, and Gaussian, is readily available in the Matlab Fuzzy Logic Toolbox [21]. Based on the nature of the input variables, fuzzy sets with corresponding membership functions are defined. For example, the distance to the nearest trees and to the nearest roads can be represented by three fuzzy sets, *near*, *med*, and *far*. Intuitively, a tree or a road is near a power line if it is within a couple

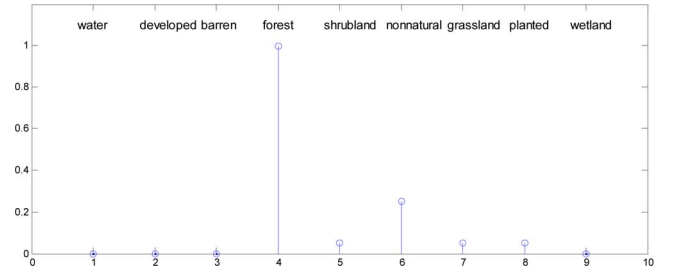


Fig. 6. Example of membership functions for categorical variables.

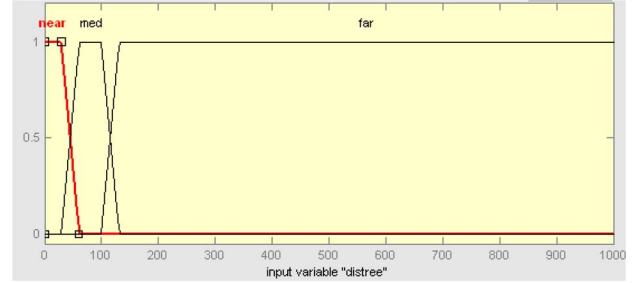


Fig. 7. Example of membership functions for continuous variables.

of meters. Considering that the common spatial resolution of raster maps is 30 meters, distances within 30 to 60 meters are defined as *near*, those of more than 100 meters are considered as *far*, and distances in-between are regarded as *med*. As an example, we used the standard trapezoid membership function, as shown in Fig. 7.

Fuzzy sets and membership functions are defined to reflect most of the common situations under study. We may need to change a membership function or use a different number of fuzzy sets when modeling different environments. For example, two fuzzy membership functions describing the *dry* and *wet* seasons in southern California may better reflect reality than the four fuzzy sets of *spring*, *summer*, *fall*, and *winter*, which are more suitable in eastern North Carolina.

B. Hierarchical Rule Base

If we use n input variables and each variable has m fuzzy sets, the total number of rules needed in a single FIS is m^n . This number grows exponentially with n . This so-called *curse of dimensionality* makes it very difficult to design and implement a complex FIS. This problem has been addressed extensively in fuzzy system literature, and one widely used solution is the hierarchical fuzzy system (HFS) [22], [23]. Breaking the system down into several subsystems can substantially reduce the number of rules needed while the HFS retains the properties of the original FIS [24]. The HFS shown in Fig. 8 (which is used in the illustrative example in Section IV) has 122 rules, which is less than 0.1% of the number required by a single-layer FIS.

As an example, the knowledge regarding underground or overhead lines in the first subsystem is that underground systems are less prone to external interference, thus being less likely to have tree-caused or animal-caused faults. In the form of rules, this is stated as “if the system is underground, then the probability of tree-caused faults is low and the probability of animal-caused faults is low.” The following subsystem takes

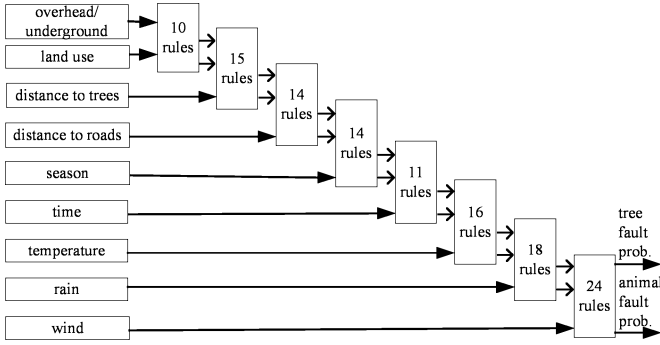


Fig. 8. Structure of a hierarchical fuzzy system (HFS).

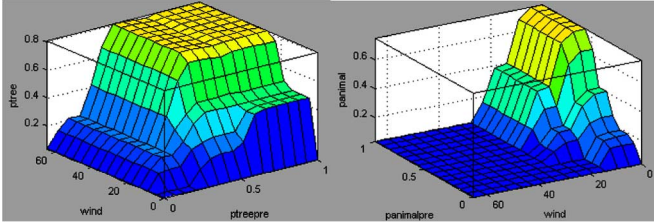


Fig. 9. Sample response surfaces of the input wind speed.

the outputs of the probability of tree-caused or animal-caused faults from the previous subsystem and an additional environmental factor as inputs, and infers new probabilities. Sample response surfaces of the last subsystem regarding wind speed are shown in Fig. 9. As we can see, when the wind speed grows higher, the probability of tree-caused faults increases and the probability of animal-caused faults decreases, reflecting the trends we showed in Fig. 1.

C. Fault Event Generation

All the available environmental factors are evaluated at every map cell where there are valid distribution system components. For each cell, the probability of faults comes from the HFS; fault events are generated according to these probabilities; and generated events with the corresponding properties are returned and stored in a file.

IV. ILLUSTRATIVE EXAMPLE

A. Simulator Structure and GUI

To demonstrate the cause-effect modeling and spatial-temporal simulation framework, we have developed a power distribution fault simulator with Matlab [25]. The fault simulator consists of three major parts: input module, fault generator, and graphic user interface (GUI), as shown in Fig. 10.

The input module takes user settings in the form of parameters and maps and converts them to environmental factors. These settings describe the layout and properties of the distribution system under study, the major geographical features, such as land use, vegetation, and roads, and local weather characteristics. The fault generator converts the environmental factors into fuzzy representations, infers the probability of tree-caused and

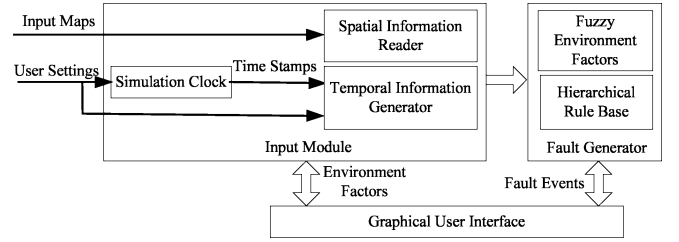


Fig. 10. Overall structure of the fault simulator.

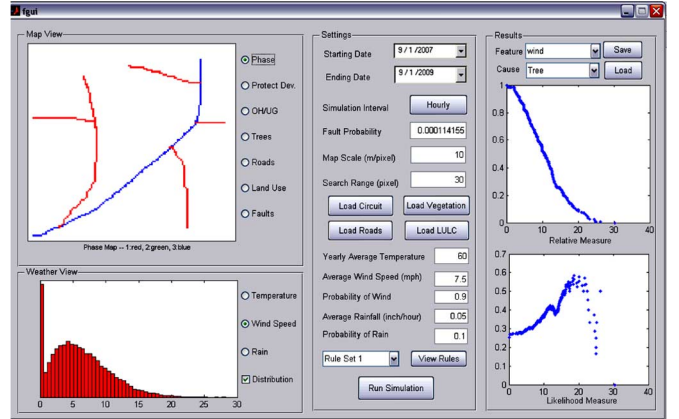


Fig. 11. Sample GUI for the fault simulator.

animal-caused faults using the hierarchical fuzzy rules, and generates and records fault events. Users interact with the simulator through the GUI (Fig. 11).

B. Environmental Factors

Four maps regarding circuits, vegetation, roads, and land use types are required for this simulator. The circuit map describes the distribution system layout with properties encoded in three layers, including the number of phases, the devices protecting the section, and whether it is overhead or underground. The vegetation map records where the tall trees, low bushes, and grasses are. Each category is encoded by different integer values in one layer. The road map represents the layout of major roads in the area under simulation. The land use map encodes the land use type by different integer values in one layer. Fig. 12 shows the spatial information loaded from a set of sample maps. With dense distribution lines, scattered trees, and most land in developed status, this is a typical distribution system for a city region.

By fitting the model (1) to the actual data shown in Fig. 3, the following temperature model is used in our simulation:

$$T = 60.5 - 5.0 \sin \frac{2\pi d}{365} - 18.3 \cos \frac{2\pi d}{365} - 6.4 \sin \frac{2\pi h}{24} - 2.2 \cos \frac{2\pi h}{24}. \quad (5)$$

In the wind speed model (3), k is set to be 1.7. $\lambda = 7.46$ and $p = 0.95$ are estimated from the data shown in Fig. 4. Similarly, λ equals to 0.48 and p equals to 0.11 in the rainfall model (4), as estimated from the data shown in Fig. 5. These models are then used to generate weather conditions during a

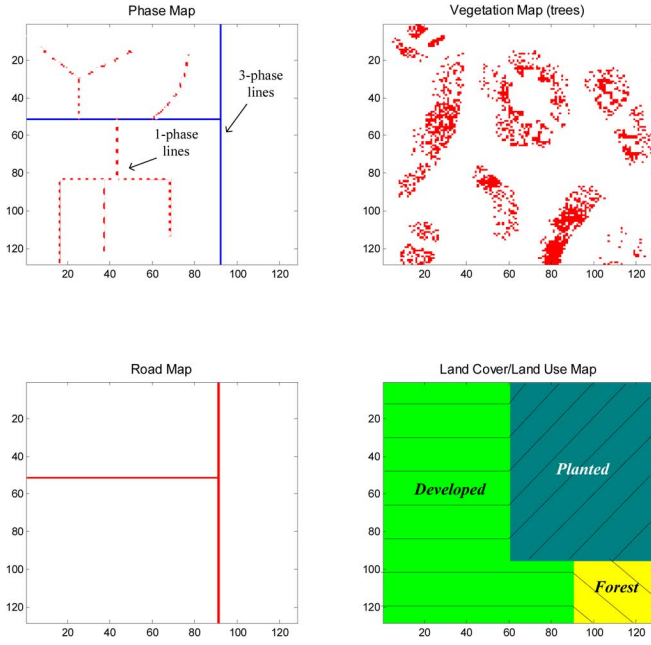


Fig. 12. Example of spatial information loaded from maps.

two-year period. The histogram of generated wind speed can be found in the bottom-left corner of Fig. 11.

C. Fuzzy Rules

Nine environmental factors are considered to be significant in the simulation, including overhead or underground power lines, land use, distance to trees, distance to major roads, season, time of day, temperature, wind speed, and rainfall intensity. They are organized as an HFS, as shown in Fig. 8.

The two categorical factors, underground or overhead and land use, are fuzzified based on the membership functions shown in Fig. 6. Fuzzy sets for the distance to trees and major roads are as shown in Fig. 7.

The definition of season and time of day is inherently fuzzy. Although spring begins generally in March in North America, the exact date varies in different locations and from year to year. Similarly, in different places, different seasons, and for different purposes, morning could start as early as 5:00 am, or as late as 8:00 am. Considering the cyclic property of the hours of the day and the days of the year, a multi-modal trapezoid membership function is used for *evening*, *night*, and *winter*, as shown in Fig. 13.

Weather factors are fuzzified based on our daily description of the weather. Temperature is represented by five fuzzy sets, *extrecold*, *cold*, *normal*, *hot*, and *extrehot*. The normal temperature ranges from about 55°F to 85°F. Temperatures above 100°F are extremely hot and those below 20°F are extremely cold. The intensity of rainfall is usually described in six levels [26]. We adopt the scale and simplify it to *norain*, *light*, *med*, and *heavy*. Similarly, wind speed is represented by *nowind*, *light*, *med*, and *strong*, according to the commonly used Beaufort scale [27]. All these fuzzy sets use the standard trapezoid membership functions similar to those in Fig. 7.

As shown in Fig. 14, the same trapezoid membership functions are used for the probability of tree-caused and animal-

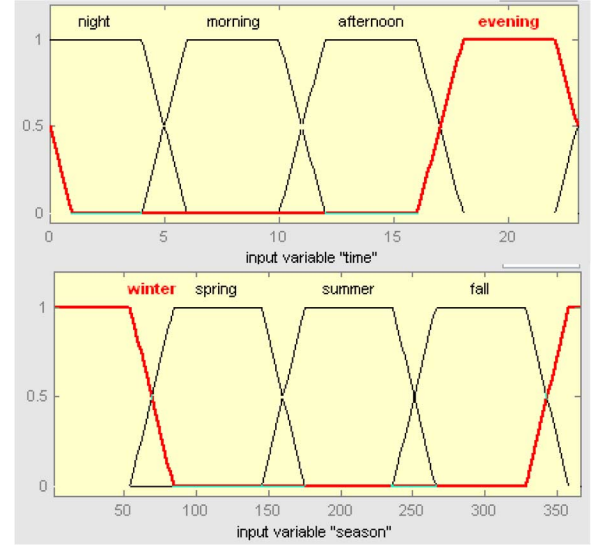


Fig. 13. Membership functions for the season and time of day.

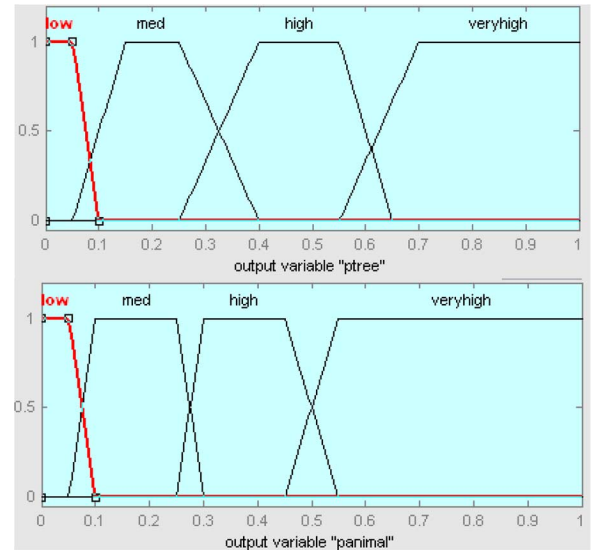


Fig. 14. Membership functions for the probability of tree-caused and animal-caused faults.

caused faults, both in the final output and all the intermediate results. The probability of animal-caused faults at *med*, *high*, and *veryhigh* level is slightly lower than those of tree-caused faults, as observed in the actual data.

D. Test Case

The simulation settings for the test case are summarized in Table I. It is to simulate a part of the actual distribution systems in the town of Garner, North Carolina.

The actual fault data are the power outage records in the distribution system of Progress Energy Carolinas within Garner operation center between 2005 and 2006. We have integrated real-time weather measurements and other geographical information with the outage records using spatial and spatial-temporal relationships [28], so the actual measurements of the environmental factors while the outages occurred are available.

TABLE I
SIMULATION SETTINGS FOR THE TEST CASE

Simulation period	01/01/2005~12/31/2006
Time step	1 hour
Maps	as shown in Fig. 12
Map resolution	30 meters
Map extent	3840 meters by 3840 meters
Temperature parameters	$\bar{T} = 60.5, A = -5.0, B = -18.3, a = -6.4, b = -2.2$
Wind parameters	$k = 1.7, \lambda = 7.46, p = 0.95$
Rainfall parameters	$\lambda = 0.48, p = 0.11$
Environmental factors	Overhead/underground lines, distance to trees, distance to major roads, season, time, temperature, rainfall intensity, wind speed
No. of subsystems in the FIS	8
No. of rules in the FIS	122

V. CONCLUSION

In this paper, a framework of cause-effect modeling and spatial-temporal simulation of power distribution fault events is proposed to cope with the data availability problem faced by many researchers in the field of distribution fault diagnosis.

Significant environmental factors leading to distribution faults are first identified based on the understanding of distribution faults and the analysis of historical fault data. The spatial information of these factors is modeled as raster maps. The temporal information of weather conditions is modeled as probability functions. The causal relationships among these factors and various types of faults are investigated and expressed in the form of fuzzy rules. A hierarchical fuzzy inference system is built to infer the probability of tree-caused and animal-caused faults in the simulated environments. Fault events are generated and recorded accordingly. The test case demonstrates the effectiveness of the simulation in generating realistic distribution faults.

By changing the environmental settings, we are able to generate fault events for different local systems during different time periods. This framework would help develop and evaluate fault diagnosis algorithms and predict regions with high probability of faults during a certain time period. It also provides a configurable data source as a useful tool for other researchers and engineers in similar areas.

ACKNOWLEDGMENT

The authors would like to thank J. W. Gajda and G. C. Lampley from Progress Energy Carolinas Inc. for their support on data and field experience.

REFERENCES

- [1] R. E. Brown, *Electric Power Distribution Reliability*. New York: Marcel Dekker, 2002.
- [2] S.-J. Lee, M.-S. Choi, S.-H. Kang, B.-G. Jin, D.-S. Lee, B.-S. Ahn, N.-S. Yoon, H.-Y. Kim, and S.-B. Wee, "An intelligent and efficient fault location and diagnosis scheme for radial distribution systems," *IEEE Trans. Power Del.*, vol. 19, no. 2, pp. 524–532, Apr. 2004.
- [3] A. Borghetti, M. Bosetti, M. D. Silvestro, C. A. Nucci, and M. Paolone, "Continuous-wavelet transform for fault location in distribution power networks: Definition of mother wavelets inferred from fault originated transients," *IEEE Trans. Power Syst.*, vol. 23, no. 2, pp. 380–388, May 2008.
- [4] J. A. Wischkaemper, C. L. Benner, and B. D. Russell, "Electrical characterization of vegetation contacts with distribution conductors—Investigation of progressive fault behavior," in *Proc. IEEE PES Transmission and Distribution Conf. Expo.*, Chicago, IL, 2008.
- [5] T. Miyazaki and S. Okabe, "A detailed field study of lightning stroke effects on distribution lines," *IEEE Trans. Power Del.*, vol. 24, no. 1, pp. 352–359, Jan. 2009.
- [6] L. Xu, M.-Y. Chow, and L. S. Taylor, "Analysis of tree-caused faults in power distribution systems," presented at the 35th North American Power Symp., Rolla, MO, 2003.
- [7] S. Sahai and A. Pahwa, "A probabilistic approach for animal-caused outages in overhead distribution systems," presented at the 9th Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS), Stockholm, Sweden, 2006.
- [8] L. Xu, M.-Y. Chow, and L. S. Taylor, "Power distribution fault cause identification with imbalanced data using the data mining-based fuzzy classification E-Algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 164–171, Feb. 2007.

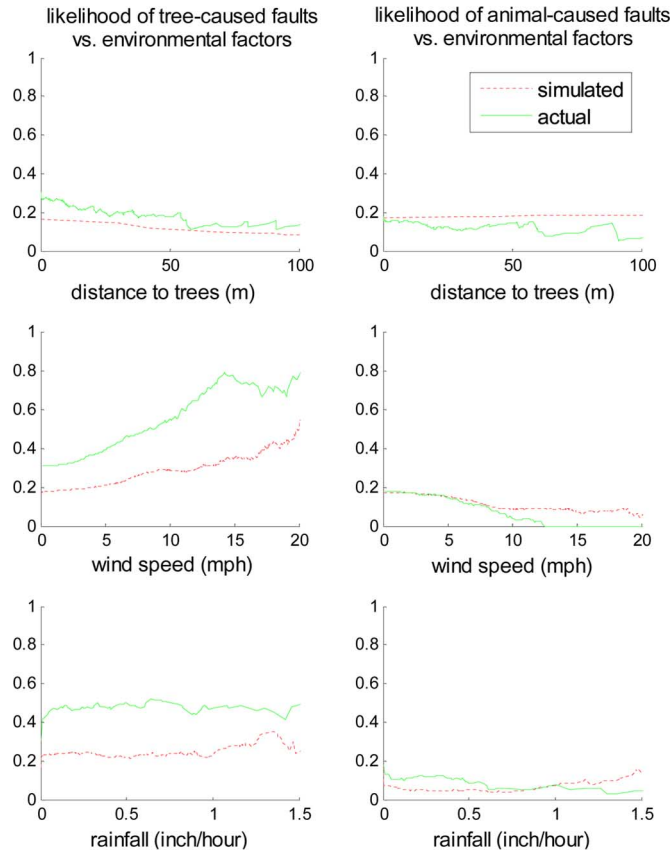


Fig. 15. Likelihood measures of selected environmental factors.

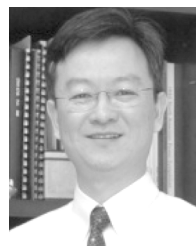
The simulation ran for around 270 min on a laptop with a 2.0-GHz dual-core processor and 2 GB of RAM, and generated 131 tree-caused and 137 animal-caused fault events. As a tool to represent fault characteristics, the likelihood of tree-caused and animal-caused faults under selected environmental factors is shown in Fig. 15.

As we can see in Fig. 15, the trend of simulated fault likelihood matches the actual situation in general. Note that we have a relatively small sample of faults with both the simulated and the actual data. This leads to large variations when estimating the fault likelihood, so an exact match of fault likelihood is hardly achievable.

- [9] L. Xu, M.-Y. Chow, J. Timmis, and L. S. Taylor, "Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 198–204, Feb. 2007.
- [10] W.-H. Chen, C.-W. Liu, and M.-S. Tsai, "On-line fault diagnosis of distribution substations using hybrid cause-effect network and fuzzy rule-based method," *IEEE Trans. Power Del.*, vol. 15, no. 2, pp. 710–717, Apr. 2000.
- [11] Q. Zhang, Z. Han, and F. Wen, "A new approach for fault diagnosis in power systems based on rough set theory," presented at the 4th Int. Conf. Advances in Power System Control, Operation and Management, Hong Kong, China, 1997.
- [12] C.-L. Hor, P. A. Crossley, and S. J. Watson, "Building knowledge for substation-based decision support using rough sets," *IEEE Trans. Power Del.*, vol. 22, no. 3, pp. 1372–1379, Jul. 2007.
- [13] R. E. Brown, G. Frimpong, and H. L. Willis, "Failure rate modeling using equipment inspection data," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 782–787, May 2004.
- [14] P. A. Kuntz, R. D. Christie, and S. S. Venkata, "Optimal vegetation maintenance scheduling of overhead electric power distribution systems," *IEEE Trans. Power Del.*, vol. 17, no. 4, pp. 1164–1169, Oct. 2002.
- [15] D. T. Radmer, P. A. Kuntz, R. D. Christie, S. S. Venkata, and R. H. Fletcher, "Predicting vegetation-related failure rates for overhead distribution feeders," *IEEE Trans. Power Del.*, vol. 17, no. 4, pp. 1170–1175, Oct. 2002.
- [16] T. J. Ross, *Fuzzy Logic With Engineering Applications*, 2nd ed. New York: Wiley, 2004.
- [17] Y. Cai, M.-Y. Chow, W. Lu, and L. Li, "Statistical feature selection from massive data in distribution fault diagnosis," *IEEE Trans. Power Syst.*, vol. 25, no. 2, pp. 642–648, May 2010.
- [18] P. A. Longley, M. F. Goodchild, D. J. Maguire, and D. W. Rhind, *Geographic Information Systems and Science*. Chichester, U.K.: Wiley, 2001.
- [19] Wind Power. [Online]. Available: http://en.wikipedia.org/wiki/Wind_power.
- [20] USGS Land Cover Institute, NLCD Land Cover Class Definitions. [Online]. Available: <http://landcover.usgs.gov/classes.php>.
- [21] Mathworks, Fuzzy Logic Toolbox User's Guide. [Online]. Available: <http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy/>.
- [22] R. R. Yager, "On the construction of hierarchical fuzzy systems models," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp. 55–66, Feb. 1998.
- [23] M.-L. Lee, H.-Y. Chuang, and F.-M. Yu, "Modeling of hierarchical fuzzy systems," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 343–361, Sep. 2003.
- [24] L.-X. Wang, "Analysis and design of hierarchical Fuzzy systems," *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 5, pp. 617–624, Oct. 1999.
- [25] Y. Cai and M.-Y. Chow, "Cause-effect modeling and simulation of power distribution fault events," in *Proc. IEEE Power & Energy Soc. General Meeting*, Minneapolis, MN, 2010.
- [26] Rain. [Online]. Available: <http://en.wikipedia.org/wiki/Rain>.
- [27] Wind. [Online]. Available: <http://en.wikipedia.org/wiki/Wind>.
- [28] Y. Cai and M.-Y. Chow, "Exploratory analysis of massive data for distribution fault diagnosis in Smart Grids," in *Proc. IEEE Power & Energy Soc. General Meeting*, Calgary, AB, Canada, 2009.



power system faults.



Yixin Cai (S'06) received the B.S. and M.S. degrees in control science and engineering from Tsinghua University, Beijing, China, in 2003 and 2006, respectively. He is currently pursuing the Ph.D. degree with the Advanced Diagnosis Automation and Control Laboratory at North Carolina State University, Raleigh.

His research interests include intelligent fault management and fault diagnosis of power distribution systems, geographic information systems in power systems, and spatial-temporal modeling of

Mo-Yuen Chow (S'81–M'82–SM'93–F'07) received the B.S. degree in electrical and computer engineering from the University of Wisconsin-Madison in 1982 and M.Eng. and Ph.D. degrees from Cornell University, Ithaca, NY, in 1983 and 1987, respectively.

He joined the Department of Electrical and Computer Engineering at North Carolina State University, Raleigh, as an Assistant Professor in 1989. He became an Associate Professor in 1993 and has been a Professor since 1999. His research focuses on fault diagnosis and prognosis, distributed control, and computational intelligence. He has been applying his research to areas including mechatronics, power distribution systems, distributed generation, motors, and robotics. He has established the Advanced Diagnosis and Control Laboratory at North Carolina State University. He has published one book, several book chapters, and over 100 journal and conference articles related to his research work.

He is the Editor-in-Chief of the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS and an Associate Editor of the IEEE TRANSACTIONS ON MECHATRONICS.