J11.2 USING MULTIPLE LINEAR REGRESSION TO DEVELOP A PLANT DAMAGE MODEL FOR A MAJOR UTILITY COMPANY

Brian J. Cerruti * and Steven G. Decker
Rutgers, the State University of New Jersey, New Brunswick, NJ

1. INTRODUCTION

Public Service Electric and Gas (PSE&G) is New Jersey's largest public utility, providing service to several large urban areas such as Trenton, Newark, New Brunswick, and the Philadelphia suburbs. The weather can have a dramatic impact on workforce operations, as a typical overtime crew can cost PSE&G $2 400 per hour, and a large cleanup effort may cost over $1 000 000 (Cerruti et al. 2009). In an attempt to alleviate these large costs, the present work describes the development of a plant damage model using data pertaining to the four PSE&G service territories shown in Figure 1.

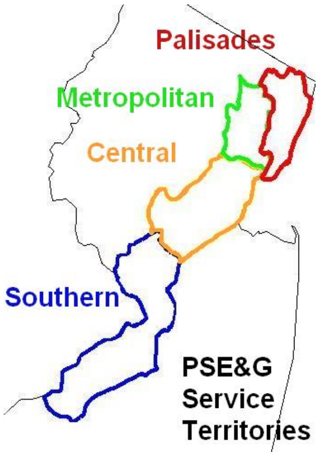
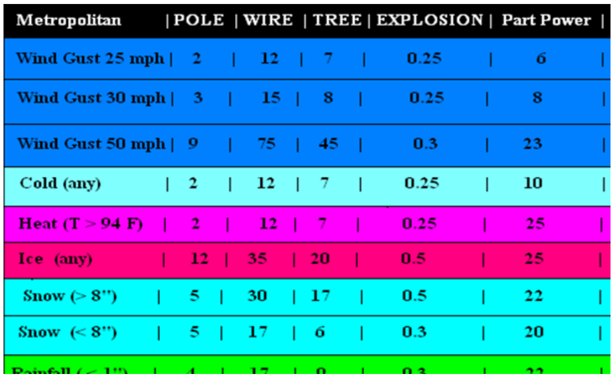


FIG. 1. A map of much of New Jersey showing the four PSE&G service territories.

A plant damage model is a tool forecasters may use to predict the damage to utility equipment. In this case, it will be a statistical relationship using multiple linear regression with weather observations as the predictors and plant damage as the predictand. The model may be used to create a deterministic forecast of plant damage while allowing for uncertainty in the model to be addressed explicitly via statistical analysis of the model results. Presumably, a utility company receiving such data would be more adequately prepared for an oncoming storm.

2. BACKGROUND

Wittman (2006) describes an early attempt at producing a plant damage model, which was based on three years of customer phone call data (18 September 2003–7 September 2006). With temperature, wind gust, and precipitation as the primary predictors, the weather observations were sorted by a "cause and effect" relationship and mainly single variable regression was used. Here, the principal cause for the observed customer calls was subjectively identified by investigating daily surface weather observations for each of the four service territories. The resulting model, an excerpt of which is shown in Table 1, was presented in tabular form for quick reference forecasting, which proved beneficial during pop-up thunderstorms.



Metropolitan	POLE	WIRE	TREE	EXPLOSION	Part Power
Wind Gust 25 mph	2	12	7	0.25	6
Wind Gust 30 mph	3	15	8	0.25	8
Wind Gust 50 mph	9	75	45	0.3	23
Cold (any)	2	12	7	0.25	10
Heat (T > 94 F)	2	12	7	0.25	25
Ice (any)	12	35	20	0.5	25
Snow (> 8")	5	30	17	0.5	22
Snow (< 8")	5	17	6	0.3	20
Rainfall (< 1")	4	17	9	0.3	22

TABLE 1. An excerpt from the Wittman (2006) damage model for Metropolitan division.

The forecaster used the tables by applying their forecast of the weather to obtain a value of expected phone calls, which were then used as a proxy for damage. If multiple events were forecast, the tabular data was summed across all relevant storm modes. The total forecast was then subjectively adjusted according to storm coverage.

* *Corresponding author address:* Brian J. Cerruti, Rutgers Univ., Dept. of Environmental Sci., 14 College Farm Rd., New Brunswick, NJ 08901;
e-mail: bcerruti@eden.rutgers.edu

The first damage model includes several shortcomings that are addressed in the current study. The shortcomings include using call data as a proxy for damage, as one downed pole in a heavily populated area can yield numerous phone calls. In addition, the use of single variable regression yields low correlation coefficients, and the use of subjectivity in the application of the model leads to low confidence forecasts.

3. NEW METHODOLOGY

3.1 INPUT DATA

Several improvements on the first model attempt were made. The first was utilizing a unified damage database made available by PSE&G stretching from 1 January 2003 to 31 October 2008. Transformers; poles; trees; and service, secondary, and primary wires were selected from the database as the predictands. Multiple linear regression was preformed relating surface weather observations to plant damage, akin to model output statistics (Glahn et al. 1972). The stations chosen were Newark Liberty International Airport (KEWR) for Metropolitan division, Teterboro Airport (KTEB) for Palisades, Somerset Airport (KSMQ) for Central, and Trenton Mercer County Airport (KTTN) for Southern (Fig 2). Surface data was downloaded from the National Climate Data Center.

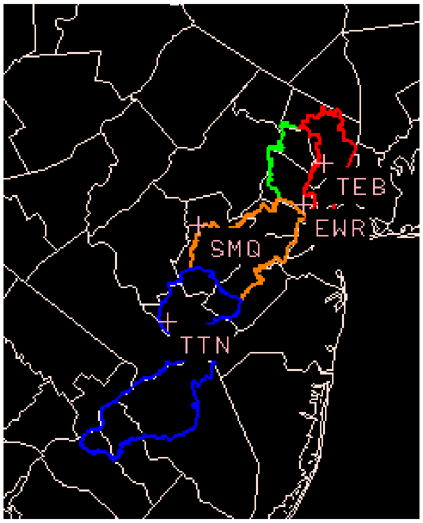


FIG. 2. Weather stations used for the regression.

3.2 STORM TYPE

The weather observations for the selected stations in each territory were analyzed to identify each day as a particular “storm type”. The idea is to determine a regression equation for each plant element for each division for each storm type. The storm types used are thunderstorm, warm cyclone, cold cyclone, mix cyclone, heat wave, and none. Thunderstorm was diagnosed as the storm type if thunderstorms were observed at the station or in the vicinity of the station or if severe weather occurred within the territory as measured by reports via the Storm Prediction Center (SPC 2009). A warm cyclone was diagnosed if the only form of precipitation measured was rain and precipitation accumulated greater than 0.01″. A cold cyclone was diagnosed if any “wintry” precipitation was observed such as snow, freezing rain, or sleet with measured liquid equivalent precipitation of at least 0.01″. A mix cyclone was diagnosed if a combination of warm and cold storm types occur (rain *and* at least one of snow, sleet, or freezing rain) with liquid equivalent precipitation of at least 0.01″. A heat wave was diagnosed if maximum temperatures exceeded 90ºF (32ºC), and measured precipitation was no more than 0.01″. A no weather day was diagnosed if precipitation was measured to be no more than 0.01″, no precipitation was reported, and maximum temperatures were no more than 90ºF.

Type	CEN	MET	PAL	SOU
T-storm	164	131	134	158
Warm	427	560	551	474
Cold	95	91	104	101
Mix	38	43	58	46
Heat	78	106	102	55
None	1158	1167	1146	1186
??	332	163	168	267
Modified	22/9	17/6	30/10	37/13

Table 2. A summary of storm mode by occurrence. The modified days in the bottom row depict the number of total days in which data was modified / total number of events causing lagged damage correction.

A seventh storm type emerged as a result of missing or suspect data or if liquid equivalent precipitation was measured to be greater than 0.01″ while no report of falling precipitation was observed. A summary of the occurrence of each storm type is contained in Table 2.

3.3 POST PROCESSING

The damage database is a collection of field reports compiled on a daily basis with a “day” defined as midnight to midnight LST. The data is collected as it is reported meaning that a significant late-day thunderstorm’s damage is likely to mainly be reported on the next day. For large and powerful severe weather events, it may take several days for all of the damage from that event to be reported and logged. This necessitates a post processing of the plant element damage data to account for any potential lag in reporting during severe weather. A summary of the post processing information is included in Table 2. While the post processing of the input data does act to nudge the data to better fit the weather observations, this method may be accepted if any verification data is post-processed in the same objective manner.

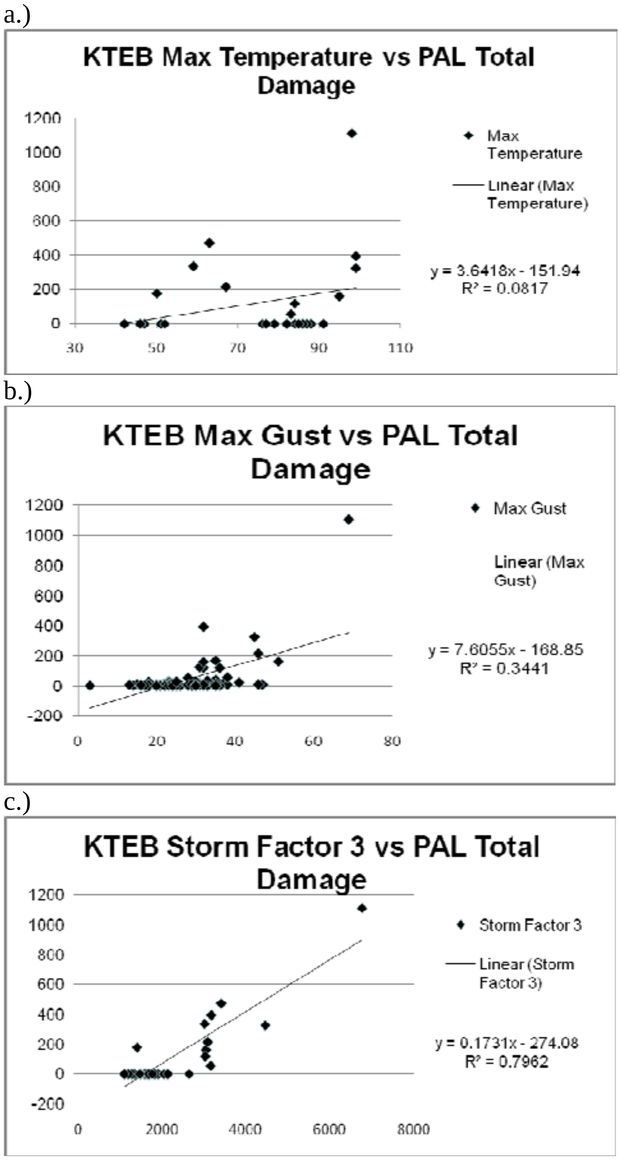


FIG. 3. Example of data investigation to find input variables for regression analysis. The regression equations will calculate each plant element individually, but storm total damage was used to investigate variables to inspect which variables would be of use for all elements.

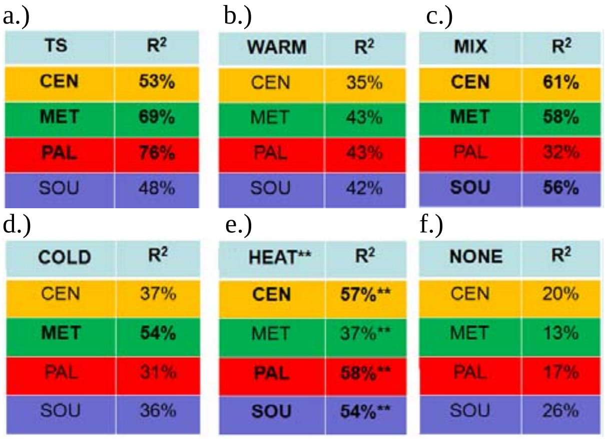
3.4 VARIABLES

Improving on the previous attempt, multiple variables will be used for every regression. The regression will include all relevant variables as a first pass, and then proceed by backwards elimination to create an operational model (Wilks 2006).

The variables chosen that were included at times in the previous attempt are maximum wind gust, maximum temperature, and liquid equivalent precipitation. Additional variables previously not included are the ten-day accumulated precipitation, three-day maximum temperature sum, and the number of severe weather reports in a given division. The ten-day accumulated precipitation is the sum of the station-measured precipitation from one to ten days prior to the forecast day. The ten-day accumulated precipitation is intended to serve as a proxy for the amount of moisture in the top layer of the soil, which is thought to be a contributor to downed trees (Wittman 2006).

An additional variable, the three-day maximum temperature sum, was considered as a proxy for the cumulative heat equipment has been exposed to and is calculated as the sum of the maximum temperature for the previous two days and the forecast day.

An investigation of these variables revealed a stronger relationship if the product of some variables were considered. Known as Storm Factors (SF), the product of the wind gust and liquid equivalent precipitation (SF1), the product of the wind gust and ten day accumulated precipitation (SF2), and the product of the wind gust and maximum temperature (SF3) have an important effect in the regression for most storm types. An example may be found in Figure 3. In Figure 3a, the observed maximum temperature for KTEB is plotted against the total observed damage for Palisades territory where the storm mode was diagnosed to be “Thunderstorm”, yielding an R^2 value of ~0.08. Figure 3b shows the total damage for Palisades plotted with the maximum wind gust, with an R^2 of ~0.34. The product of the maximum temperature and wind gust (SF3) is plotted in Fig. 3c against the total Palisades damage and the R^2 increases dramatically to ~.80, demonstrating the potential usefulness of the Storm Factors.



a.)

TS	R^2
CEN	53%
MET	69%
PAL	76%
SOU	48%

b.)

WARM	R^2
CEN	35%
MET	43%
PAL	43%
SOU	42%

c.)

MIX	R^2
CEN	61%
MET	58%
PAL	32%
SOU	56%

d.)

COLD	R^2
CEN	37%
MET	54%
PAL	31%
SOU	36%

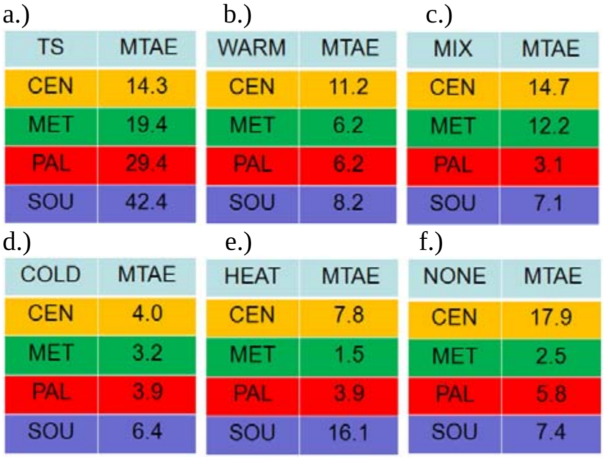
e.)

HEAT**	R^2
CEN	57%**
MET	37%**
PAL	58%**
SOU	54%**

f.)

NONE	R^2
CEN	20%
MET	13%
PAL	17%
SOU	26%

TABLE 3. Summary of regression models for each storm type and division. The R^2 value shown is an average of the R^2 values from each of the six plant element regressions. For storm type “Heat”, only the R^2 value for the transformer regression is shown, as denoted by asterisks.



a.)

TS	MTAE
CEN	14.3
MET	19.4
PAL	29.4
SOU	42.4

b.)

WARM	MTAE
CEN	11.2
MET	6.2
PAL	6.2
SOU	8.2

c.)

MIX	MTAE
CEN	14.7
MET	12.2
PAL	3.1
SOU	7.1

d.)

COLD	MTAE
CEN	4.0
MET	3.2
PAL	3.9
SOU	6.4

e.)

HEAT	MTAE
CEN	7.8
MET	1.5
PAL	3.9
SOU	16.1

f.)

NONE	MTAE
CEN	17.9
MET	2.5
PAL	5.8
SOU	7.4

TABLE 4. Summary of error results from validation data. Mean Total Absolute Error (MTAE) values are shown for each division and each storm type.

4. DATA

4.1 REGRESSION RESULTS

An overview of the regression results can be found in Table 3. The baseline storm type, “None”, represents the situations where meteorological conditions are assumed unfavorable for causing plant damage, and this type displays the lowest R^2 values.

The “Thunderstorm” and “Mix” days have the highest R^2 values. Thunderstorms can be expected to have a strong correlation between plant damage and observed weather. Mix days presumably have very high correlation coefficients because cyclones which bring multiple precipitation types to the same area tend to have higher precipitation totals and stronger winds according to the data observed in this study.

4.2 VALIDATION

An independent data set was obtained from PSE&G for the period 1 November 2008–15 November 2009. The daily weather observations were obtained for the appropriate weather stations and the damage data was post processed according to the methods in Section 3.3.

A summary of the validation data for each storm type can be found in Table 4. Here, the error score is the mean total absolute error (MTAE), which is calculated by taking the absolute error for each of the six plant elements, summing the daily errors, then lastly taking the average over all days. This was performed for each storm type in each division.

Despite having the highest R^2 coefficients, the “Thunderstorm” mode is subject to the largest mean absolute error. This may be due to the station data failing to capture the small scale of the convection within the PSE&G territories. Additionally, the variance in the plant damage is greatest during the thunderstorm mode than for any other storm type.

5. CASE STUDIES

The following is a series of two case studies intended to show the functionality of the plant damage model on a division basis for two of the storm types with the highest regression R^2 values. The model results will be compared with observed damage per division and a discussion will follow.

5.1 9 June 2009: “THUNDERSTORM”

Thunderstorms affected the PSE&G service territories in the early morning hours with frequent lightning, wind gusts of 16 to 31 mph (7−14 m s^{-1}), rainfall of 0.2 to 1.05″ (5−27 mm), and a report of severe hail in the Southern division (SPC 2009). The regression equations were applied for each territory using the "Thunderstorm" mode to the surface observations from each weather station. The results are summed graphically in Fig. 4, and the complete results are in Table 5.

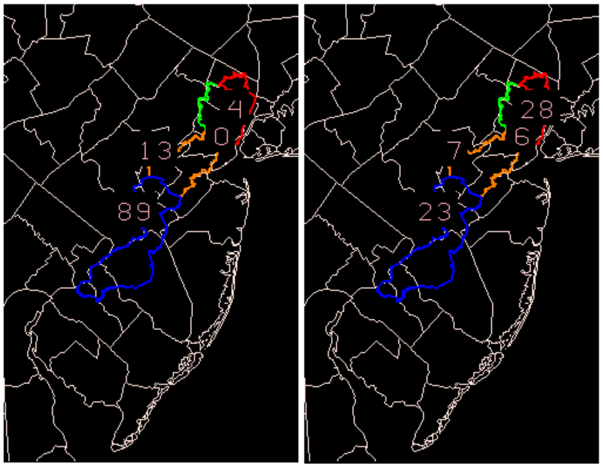


FIG. 4. (left) PSE&G plant damage observations and (right) "Thunderstorm" model prediction for 9 June 2009.

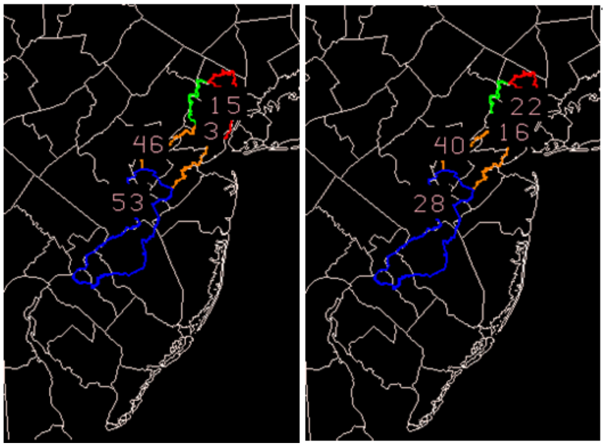


FIG. 5. Same as Fig. 4, but for 11 September 2009 using the "Warm" model.

The model performed well in Central by correctly predicting a low damage total (Table 5a). The model overestimated the damage to Palisades and underestimated damage to the Southern division (Table 5c–d). The damage in Metropolitan was measured to be zero, but the model predicted six elements to break (Table 5b). These errors may be due to the anomalous timing of the convection crossing New Jersey, as temperatures likely would have risen higher had the storms arrived in the late afternoon allowing for a higher model estimated damage forecast due to

a.)

CEN	**O**	**M**
Trans	9	2
Pole	0	1
Service	0	1
Second	0	0
Primary	2	1
Tree	**2**	**2**

b.)

MET	**O**	**M**
Trans	0	1
Pole	0	1
Service	0	2
Second	0	1
Primary	0	1
Tree	**0**	**0**

c.)

PAL	**O**	**M**
Trans	2	5
Pole	2	5
Service	0	11
Second	0	2
Primary	0	3
Tree	0	2

d.)

SOU	**O**	**M**
Trans	18	4
Pole	1	3
Service	**6**	**6**
Second	12	1
Primary	27	5
Tree	25	4

TABLE 5. Complete data for 9 June 2009 Case Study. M = Modeled data (according to "Thunderstorm" storm type); O = Observed data. The values in bold are where model error is no more than one when damage was observed.

a.)

CEN	**O**	**M**
Trans	**3**	**5**
Pole	10	4
Service	**13**	**12**
Second	**1**	**2**
Primary	**6**	**5**
Tree	**13**	**12**

b.)

MET	**O**	**M**
Trans	0	1
Pole	0	1
Service	0	6
Second	0	1
Primary	0	2
Tree	3	5

c.)

PAL	**O**	**M**
Trans	**4**	**2**
Pole	**3**	**3**
Service	7	10
Second	0	1
Primary	**1**	**3**
Tree	0	3

d.)

SOU	**O**	**M**
Trans	**4**	**4**
Pole	8	4
Service	14	8
Second	**4**	**3**
Primary	14	5
Tree	9	4

TABLE 6. Same as Table 5, except for 11 September 2009.

the strong correlation between maximum temperature and damage. Another possible source of error is the omission of lightning data in the damage model, as this particular thunderstorm event was observed to have frequent lightning at Newark Liberty International Airport (KEWR) for several hours. Other surrounding stations such as Teterboro Airport (KTEB), Trenton Mercer County Airport (KTTN) and Philadelphia International Airport (KPHL) also observed lightning for several hours.

5.2 11 September 2009: “WARM”

A weak surface cyclone formed late on 10 September and tracked across New Jersey from south to north through the day of 11 September while dissipating. This system was responsible for wind gusts of 31 to 37 mph (14−17 m s^{-1}), rainfall of 0.5 to 1.5″ (13−38 mm), and maximum temperatures were 64 to 68 ºF (18−20ºC). The damage model’s “Warm Cyclone” mode was applied for each division (Fig. 5; Table 6).

The model performed quite well in Central where the total damage estimate error was only six elements with small errors for each element except poles. The damage is underestimated again in Metropolitan with the only observed damage assigned to trees. The damage estimate for Palisades is overestimated mainly attributed to an overestimation of service wire and tree damage. The Southern division’s damage is underestimated by the model due to substantial errors in pole, service wire, and tree forecasts. The large underestimation in Southern may be caused by the selection for the input station as Trenton Mercer County Airport (KTTN) is in the northernmost part of the territory and may not fully represent the surface weather conditions experienced during this storm.

6. DISCUSSION

6.1 CONCLUSIONS

The second attempt at creating a statistical damage model for PSE&G yielded encouraging results as measured by relatively high correlation coefficients for most storm modes and divisions with the best correlations for ‘Thunderstorm’ and ‘Mix’ mode (Table 3). The verification of the model yielded reasonable error scores, which encourage future work (Table 4). It is important to note that a bias towards higher damage exists in the verification statistical analysis due to an auto-correction applied to the model output, which set negative damage estimates to zero.

6.2 FUTURE WORK

The immediate task at hand is to perform backwards elimination for every regression equation to create the operational model. The implementation phase of the damage model will involve the use of a webpage interface where a forecaster may enter a forecast of the necessary variables and obtain a deterministic damage forecast for each division.

The next phase of damage modeling may include more stations. Specifically, Philadelphia International Airport (KPHL), South Jersey Regional Airport in Mount Holly, NJ (KVAY), McGuire Air Force Base (KWRI), Linden Airport (KLDJ), Morristown Municipal Airport (KMMU), and Essex County Airport in Caldwell, NJ (KCDW) are all first-order stations in and around the PSE&G service territories. Perhaps GIS may be utilized to sort the damage data by location to assign the damage to the closet weather station, but such information is unavailable at this time. New variables may be considered for any new model attempt such as wind duration, thunderstorm coverage and duration, lightning information, and wind direction. The correlation of the Storm Factors with the damage data may suggest that transformations on the data are necessary to improve the model.

REFERENCES

Cerruti, B. J., S. G. Decker, L. A. Bowers, W. K. Wittman, and J. Carlson, 2009: Undergraduate forecasting and nowcasting for a major urban public utility. Preprints, *8th Symp. on the Urban Environment,* Phoenix, AZ, Amer. Meteor. Soc., J22.6.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

NCDC, cited 2009: U.S. Storm Events Database. [Available online at http://www.ncdc.noaa.gov.]

SPC, cited 2009: U.S. Local Storm Database. [Available online at http://www.spc.noaa.gov.]

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 648 pp.

Wittman, W., L. A. Bowers, J. Carlson, R. Dunk, and S. Glenn, 2006: Weather forecast benchmarking and plant damage modeling for public utilities. *Proc. 3rd Annual Meeting,* Baltimore, MD, MACOORA.