Analysis of Correlation between Cold Weather Meteorological Variables and Electricity Outages

Melissa R. Allen-Dumas, Sangkeun Lee, Supriya Chinthavali
Computational Sciences and Engineering Division
Computational Science and Mathematics Division
Geospatial Science and Human Security Division
Oak Ridge National Laboratory
Oak Ridge, TN USA
allenmr@ornl.gov, lees4@ornl.gov, chinthavalis@ornl.gov

***Abstract*—The significance of the impact of weather on the electric grid has grown as climate change continues to increase the frequency and intensity of extreme weather events. In recent years (2021-2022) in particular, extreme winter weather has affected the grid in locations in the US rarely exposed to extreme low temperatures, snow and icing conditions. Here we analyze the correlation between cold weather meteorological variables and electricity outages during two large winter storm events, Uri (February 2021) and Landon (February 2022) using Random Forest machine learning and Pearson's correlation coefficient. Our geographical focus across the two storms is the state of Texas. Extrapolation of the method to winter weather impacts over other years and additional locations is proposed.**

***Index Terms*—machine learning, electric grid, extreme weather, winter storms, outages**

I. Introduction

Technologies such as internet, transportation charging and building heating and cooling are driving changes in customer expectations of power system reliability [1]. Yet continuing increases in frequency and intensity of extreme weather events challenge energy reliability across the United States power grid. Annual reports such as the State of Reliability reports issued by the North American Electric Reliability Corporation (NERC) routinely find that the leading causes of large electricity outages are weather related [2]. Furthermore, nearly half of all major outage events for the years 2015-2019 were caused by extreme winter weather associated with low temperatures, high winds, heavy snow, hail, and blizzards [3]. In more recent years, Winter Storm Uri (February 2021) caused electricity power outages for 4.5 million customers at its peak, and left many customers without power for several days [4]. Winter Storm Landon [5], in February 2022, manifested as a 2,000-mile-long expanse of snow and ice from the Southern Rockies and Plains into the Midwest and northern New England causing massive outages across its track. In the current age of "big data," many researchers are applying machine learning (ML) techniques to predict power outages based on winter weather, land use, grid asset status, vegetation management and other conditions [6]–[9]. Here, we use a Random Forest machine learning method and Pearson's correlation coefficient to understand the relationship between 1 km gridded daily weather variables and county-level daily customer outages so that utilities can employ these data-driven approaches to aid storm response planning, long-term asset management and optimal crew mobilization ahead of extreme winter storm events.

This manuscript has been authored by UT-Battelle LLC under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes.DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

II. Data and Methods

We perform two analyses using components of the Advanced data SCiENce toolkit for Non-Data Scientists (ASCENDS) tool (Section II-C). The first analysis uses the ASCENDS implementation of the Python Scikit-learn Random Forest (RF) Regressor [10], which is a non-parametric model that fits a user-chosen number of classifying decision trees on various samples drawn from the dataset, and is evaluated using the root mean squared error (RMSE). The second analysis employs ASCENDS to perform a Pearson correlation among each weather and outage variable with an outage prediction by solving the equation:

$$r = \frac{\Sigma[(X[:,i] - mean(X[:,i])) * (y - mean(y))]}{(std(X[:,i]) * std(y))} \quad (1)$$

with the X matrix including all weather and outage variables, and y representing the outage predictions. Here, std indicates the standard deviation of the data distribution.

A. Data

Data analyzed for the study were gridded weather observations and Texas county-level electric outage reports during the five days each of two major winter storms, Uri (February 13-17, 2021), and Landon (February 2-6, 2021). The 1km x 1km gridded weather observations were acquired from Oak Ridge National Laboratory's (ORNL) reanalysis dataset, Daymet (Version 3) [11] for which grid cell parameters

are calculated based on the coordinates of each grid cell's centroid. Daymet weather variables included in this study were maximum and minimum daily temperature, daylight average incident short wave radiation, cumulative precipitation and snow water equivalent and average vapor pressure. County level customer outage counts were obtained from the archives of the Department of Energy (DOE) Environment for Analysis of Geo-Located Energy Information (EAGLE-I) situational awareness platform for near real-time energy status.

B. Data Preparation

For ingestion of the data into the ASCENDS tool for the correlation analysis, the maximum daily weather data and the daily averaged outage data were converted to csv and aggregated to the county level so that variables in the two data sets could be mapped one-to-one. This aggregation was facilitatied by the ArcGIS tool using a spatial join.

Next, the data was reorganized for correlation between meteorological variables for a given day (t) to the outage data for the next day (t+1). Customer outages were tracked using both the number of customers outaged per county and the percentage of customers outaged per county. The total number of customers per county are those reported by all utilities serving a given county. In some cases, not all customers were reported, which contributed to a measure of uncertainty for total customer count. In those cases, the calculated percentage of customers outaged (pcout) was sometimes over 100, in which case the values were replaced with 100. Each row (county) of the data included the following column attributes:

- GeoID: combination state and county FIPS code (Federal Information Processing Series–unique numerical identities for each state and county)
- all_custom: total electricity customer count ($\frac{countypopulation}{no.firms+no.residences}$) [12] for the GeoID
- tmin(t) and tmax(t): today's minimum and maximum temperature
- tmin(t-1) and tmax(t-1): yesterday's minimum and maximum temperature
- prcp(t): today's cumulative precipitation
- prcp(t-1): yesterday's cumulative precipitation
- swe(t): today's cumulative snow water equivalent
- swe(t-1): yesterday's cumulative snow water equivalent
- srad(t): today's daylight average incident short wave radiation
- srad(t-1): yesterday's daylight average incident short wave radiation
- vp(t): today's average vapor pressure
- vp(t-1): yesterday's average vapor pressure
- out(t): today's raw outage count for GeoID
- out(t-1): yesterday's raw outage count for GeoID
- pcout(t):today's percentage of customers outaged for GeoID
- pcout(t-1): yesterday's percentage of customers outaged for GeoID
- pcout(t+1): this will be the percent of customers outaged on the next date

C. Machine Learning with ASCENDS

The Advanced data SCiENce toolkit for Non-Data Scientists (ASCENDS) is a set of command-line and web-based GUI tools for performing various advanced data analysis and machine learning techniques [13], [14]. The toolkit focuses on two different machine learning tasks: classification and regression. The classification part of the toolkit predicts a category (Y) from input variables (X). Regression in the toolkit is used to train a predictive model that approximates a continuous output variable (y) from input variables (X). The toolkit supports linear, logistic and other types of regression, random forests, support vector machines and neural networks. Additionally, capabilities for feature selection based on various criteria and automatic hyperparameter tuning are provided.

III. Results and Discussion

Using the daily weather and outage data for five days each from the two winter storms, we used the RF regressor to answer the question, "If we know today and yesterday's weather information, can we predict tomorrow's outages?" For this analysis, which was run using information focused on Texas outages for both Uri and Landon, we randomly shuffled the data, then used 85% of the data for training the RF model to predict pcout(t+1). We then used 15% of the data to validate the result.

Figures 1, 2 and 3 show the predicted and actual pcout(t+1) values. The x-axis in each figure is the weather-predicted percentage of customers outaged and the y-axis is the actual percentage of customers outaged. If the model perfectly predicted pcout(t+1), we would see the red dots are perfectly lined up at a 45° angle. However, none of the results show this type of predictability.

Figure 1 can be interpreted such that there is a small positive correlation for Winter Storm Uri of yesterday's weather to today's outages, as it shows a somewhat significant RMSE of 11.091.

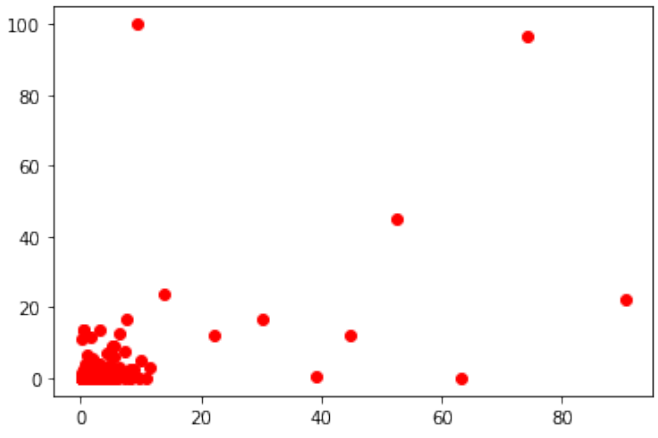

Fig. 1. Prediction result from the random forest model trained with the Winter Storm Uri (Texas counties). The x-axis shows the predicted percentage of customers outaged and the y-axis shows the actual value. RMSE = 11.091.

Using the same geographical data but for the 2022 Winter Storm Landon, the RF model yielded a lower RMSE of 7.787 (Figure 2), which can be interpreted such that there is a greater positive correlation among the weather variables and the percentage of customers outaged during that storm.

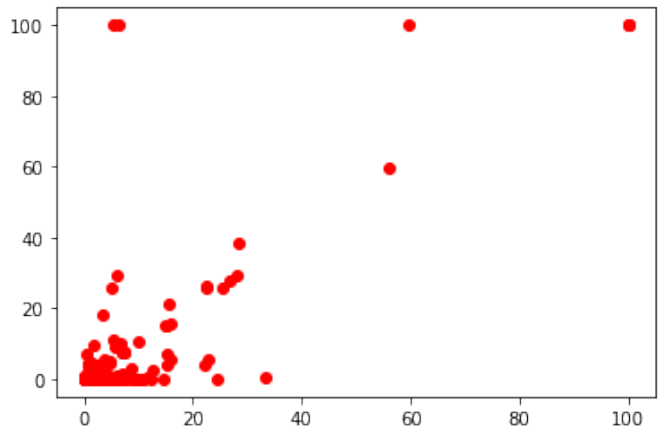

Fig. 2. Prediction result from the random forest model trained with the Winter Storm Landon (Texas counties). The x-axis shows the predicted percentage of customers outaged and the y-axis shows the actual value. RMSE = 7.787.

Combining the Texas county data from both winter storms provided the best correlation among weather variables and outages, and produced the lowest RMSE at 5.769. The scatterplot for this correlation is shown in Figure 3.

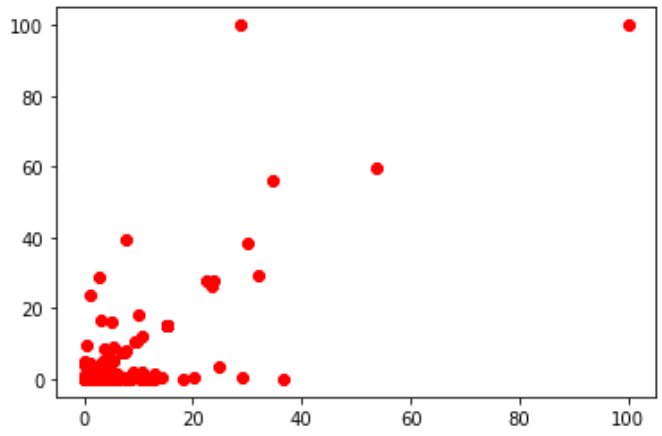

Fig. 3. Prediction result from the random forest model trained with combined data collected from Winter Storm Uri and Landon (Texas counties). The x-axis shows the predicted percentage of customers outaged and the y-axis shows the actual value. RMSE = 5.769.

The initial conclusions from this first analysis show that that prediction from the combined data from both storms was better than that of either Uri or Landon alone. Additionally, the Landon data and model showed better prediction results than that of Uri. Possible explanations for this result are 1) because the size of dataset was larger (more counties with outages reported) in the case of the Landon storm and in the case of the combined data, more data led to better results, and 2) the way the dataset is shuffled may impact the result, so repetition of the analysis using a variety of shuffling methods is needed to make sure results are consistent. It is encouraging that combined data led to better result, since it opens up the possibility of using additional historical winter storm data and making the prediction model better over time.

The second analysis evaluated the strength of the relationship between the relative movements of two variables using Pearson's correlation coefficient. Figure 4 shows that in the case of the Uri Texas county dataset, pcout(t) and pcout(t-1) have a strong positive correlation. That is, the previous days' outage percentages were highly positively correlated to the percentage of customers out of power on the next day, pcout(t+1). It can be interpreted from this result, and was certainly observed during the storms, that outages persisted in the same regions for several days. Additionally, maximum daily temperature (tmax) and daily average incident short wave radiation (srad) were positively correlated with outages, which is a bit counterintuitive, but interesting to note. There were no strong negative correlations observed.

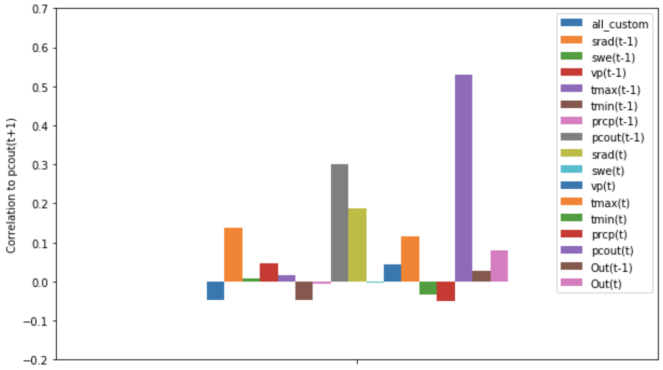

Fig. 4. Pearson's correlation of the percentage of customers outaged tomorrow (t+1, x-axis) in Winter Storm Uri to each of the weather and outage variables. The y-axis shows the strength and sign of the correlation.

The same type of correlation performed using the Landon Texas county data did not show any strong positive or negative correlations (Figure 5). However, minimum temperature (tmin) and snow water equivalent (swe) were weakly positively correlated to the percentage of customers outaged on the next day, pcout(t+1).

Finally, a Pearson's correlation using Texas county data from both storms (Figure 6) showed that while daylight average incident shortwave radiation and cumulative snow water equivalent were weakly positively correlated to the percentage of customers outaged tomorrow, and average vapor pressure, maximum and minimum temperature and cumulative precipitation were more highly and negatively correlated with the percentage of customers outaged tomorrow, the highest positive correlation occurred in the combined dataset again with the previous day's outages.

IV. Conclusions

In this study, we employed the Random Forest machine learning method and Pearson's correlation coefficient to ana-

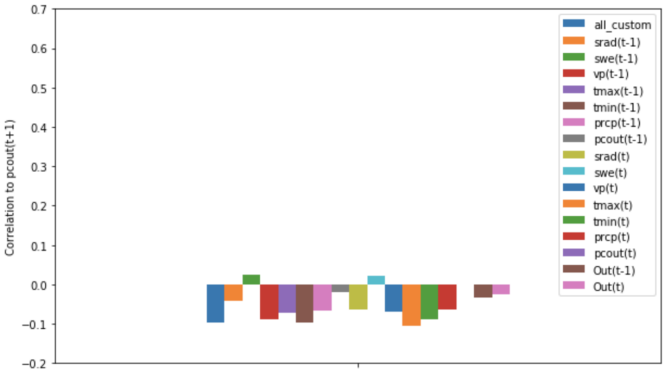

Fig. 5. Pearson's correlation of the percentage of customers outaged tomorrow (t+1, x-axis) in Winter Storm Landon to each of the weather and outage variables. The y-axis shows the strength and sign of the correlation.

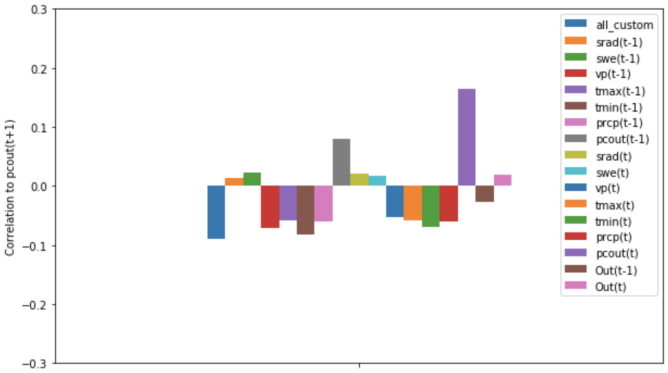

Fig. 6. Pearson's correlation of the percentage of customers outaged tomorrow (t+1, x-axis) in both winter storms to each of the weather and outage variables. The y-axis shows the strength and sign of the correlation.

lyze and understand the correlation between weather variables and electricity outages during two winter storm events. We observed that the larger data set (from Winter Storm Landon) and the combination of data from two different winter storm events can lead to better outage prediction results, which is encouraging, because it shows that we can improve the quality of outage predictions by collecting more data. Thus, we can enhance the model over time using the data collected from multiple events. However, from both analyses conducted we conclude that we have not yet found a clear, consistent, strong correlation across the two winter storms affecting Texas for the years 2021 and 2022 among weather variables and percentage of customers out of power the next day. One explanation may be the limitation of the Pearson's correlation coefficient. It only captures linear correlation between variables, so there may still be more complex correlation hidden in the data. In the next study, we will explore more correlation coefficients such as, among others, the Maximal Information Coefficient [15]. Also, There may have been many external factors other than the meteorological variables (e.g., infrastructure difference, readiness for events, etc.) that contributed to the outages in the Texas counties that are related to the winter storms that can be explored in future studies. Additionally, similar analyses of the effects of winter storms on different parts of the United States, such as the effect of Winter Storm Landon on both the Midwest and the Northeast may suggest clues to the proportional influence of weather and infrastructure on grid robustness and resiliency to extreme weather and overall reliability of service during winter storm events. The inclusion of grid asset location and health as part of a future empirical model workflow could add robustness to model predictability (as in [6]). What we learned from this study is that the quality and consistency of the data is the crucial. How to deal with missing data, data with error (e.g., percentage value over 100), inconsistent data availability across geographical regions, etc. must be further explored.

Acknowledgements

The authors wish to thank the North American Energy Resilience Model (NAERM) project for supporting this study.

References

[1] C. Watts, C. McCarthy, and B. Levite, "Consumer-centric reliability metrics," *IEEE Power and Energy Magazine*, vol. April, pp. 117–124, 2022.

[2] NERC, "2021 State of Reliability: An assessment of 2020 bulk power system performance," NERC State of Reliability, 2021, 2021. [Online]. Available: https://www.nerc.com/pa/RAPA/PA/Performance Analysis DL/NERC_SOR_2021.pdf

[3] S. Ekisheva, R. Rieder, J. Norris, M. Lauby, and I. Dobson, "Impact of extreme weather on north american transmission system outages," in *2021 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2021, pp. 01–05.

[4] J. W. Busby, K. Baker, M. D. Bazilian, A. Q. Gilbert, E. Grubert, V. Rai, J. D. Rhodes, S. Shidore, C. A. Smith, and M. E. Webber, "Cascading risks: Understanding the 2021 winter blackout in Texas," *Energy Research & Social Science*, vol. 77, p. 102106, 2021.

[5] Newsweek, "Winter Storm Landon Update," 2022.

[6] M. Angalakudati, J. Calzada, V. Farias, J. Gonynor, M. Monsch, A. Papush, G. Perakis, N. Raad, J. Schein, C. Warren *et al.*, "Improving emergency storm planning using machine learning," in *2014 IEEE PES T&D Conference and Exposition*. IEEE, 2014, pp. 1–6.

[7] S. Mukherjee, R. Nateghi, and M. Hastak, "A multi-hazard approach to assess severe weather-induced major power outage risks in the us," *Reliability Engineering & System Safety*, vol. 175, pp. 283–305, 2018.

[8] D. Cerrai, M. Koukoula, P. Watson, and E. N. Anagnostou, "Outage prediction models for snow and ice storms," *Sustainable Energy, Grids and Networks*, vol. 21, p. 100294, 2020.

[9] W. O. Taylor, P. L. Watson, D. Cerrai, and E. N. Anagnostou, "Dynamic modeling of the effects of vegetation management on weather-related power outages," *Electric Power Systems Research*, vol. 207, p. 107840, 2022.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] P. Thornton, M. Thornton, B. Mayer, Y. Wei, R. Devarakonda, R. Vose, and R. Cook, "Daymet: Daily surface weather data on a 1-km grid for North America, Version 3, ORNL DAAC, Oak Ridge Tennessee," 2017.

[12] M. R. Allen, S. J. Fernandez, J. S. Fu, and M. M. Olama, "Impacts of climate change on sub-regional electricity demand and distribution in the southern united states," *Nature Energy*, vol. 1, no. 8, pp. 1–9, 2016.

[13] S. Lee, J. Peng, A. Williams, and D. Shin, "Ascends: advanced data science toolkit for non-data scientists," *Journal of Open Source Software*, vol. 5, no. 46, 2020.

[14] J. Peng, S. Lee, A. Williams, J. A. Haynes, and D. Shin, "Advanced data science toolkit for non-data scientists–a user guide," *Calphad*, vol. 68, p. 101733, 2020.

[15] J. B. Kinney and G. S. Atwal, "Equitability, mutual information, and the maximal information coefficient," *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.