Print ISSN: 3104-932X
Online ISSN: 3104-9338

# Interpretable Severity Stratification of Weather-Associated County Power-Outage Events

Huachen Zhang

School of Energy, Belarusian National Technical University, Minsk, Belarus

Received: Nov 3,2026

Revised: Nov 9, 2026

Accepted: Nov 11, 2026

Published online:November 17, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 3, No. 1 (January 2027)

∗ Corresponding Author: Huachen Zhang (zhh86764@gmail.com)

**Abstract.** This study examines whether information available at the first observed outage hour can help distinguish severe weather-associated county power-outage events from less severe events. The analysis uses the public EAGLE-I and National Weather Service Valid Time Event Code (VTEC) harmonized dataset. The downloaded primary CSV contains 3,572,394 county-hour records from 2 January 2015 to 3 January 2023. Because the final year is incomplete, the main analysis is restricted to 1,373,715 reconstructed events in complete calendar years 2015–2022 across 3,010 counties and 49 state-level jurisdictions. Event severity is defined as the event peak customer-outage count divided by a county-specific 75th-percentile outage baseline. A severe event is one at or above the 90th percentile of this ratio in the 2015–2020 training period. Regularized logistic regression and random forest models use only first-record information, including the matched VTEC phenomena, county location, calendar timing, number of simultaneous phenomena, and source-record availability. Models are trained on a reproducible 200,000-event sample and evaluated on a 100,000-event sample drawn from the 2021–2022 holdout period. The random forest achieves a ROC-AUC of 0.693 and average precision of 0.186, compared with a held-out severe-event prevalence of 0.100. When the highest-risk 10% of events are selected, precision is 0.221 and recall is 0.220. The results indicate modest ranking value, but also show that first-record warning and contextual features alone miss most severe events. The analysis is conditional on a positive outage already being recorded and does not estimate outage incidence or establish weather causation.

***Keywords:*** *power system resilience; extreme weather; power outages; event severity; EAGLE-I; VTEC; machine learning*

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

# 1. Introduction

Electricity distribution systems are exposed to storms, heat, cold, floods, and other hazards that can produce concentrated customer interruptions. Resilience studies therefore examine not only whether an interruption occurs, but also how strongly service is degraded and how quickly it is restored. Quantitative resilience metrics differ in their system boundaries, time horizons, and treatment of disruption and recovery, making explicit outcome definitions essential for comparing analyses [1]. County-level outage archives have expanded the evidence available for this work, while also retaining uneven data coverage and reporting artifacts [2].

Several research strands motivate the present analysis. Weather-related line failure studies estimate component damage from wind or hurricane conditions, including hybrid fragility and machine-learning approaches for typhoon-related transmission-line damage [3]. Other work has used Bayesian networks to predict distribution-line outages under hurricanes and to support proactive resilience planning [4]. At a broader spatial scale, outage and weather mapping has been used to identify locations and event types associated with severe customer impacts [5]. These studies address different system components and questions; county-level outage records cannot directly reveal which transmission line or distribution asset failed.

Recent data releases make it possible to study the association between customer-outage records and matched weather warnings at hourly resolution. The EAGLE-I record provides county-level customer counts at sub-hourly intervals [2], and a harmonized archive joins processed outage records with National Weather Service (NWS) watches, warnings, and advisories represented by VTEC phenomenon codes [6]. The joined file is useful for examining warning-associated outages, but it contains positive outage observations that overlap a matched warning. It does not supply a complete set of county-hours with no outage and no warning. Consequently, it cannot support an unbiased estimate of outage occurrence probability without adding a suitable control sample.

This paper studies a narrower problem: among reconstructed county events for which a positive outage record is present, how much can the warning phenomenon and contextual information at the first observed outage hour distinguish events with unusually high impact relative to the county's own baseline? We use a chronological holdout, define severity using a training-period threshold, and prevent event peak magnitude and event duration from entering the feature set. The analysis is a retrospective severity-stratification exercise, not a pre-event forecast or a causal estimate of weather damage.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

The study makes three limited contributions. First, it documents a reproducible event-level construction and explicitly audits the timestamps in the downloaded release before selecting the study period. Second, it compares a regularized linear classifier with a tree ensemble under a year-based holdout rather than a random row split. Third, it evaluates how the results change when the first event record has higher source-record availability. The findings are intended as a baseline for future work that can incorporate measured meteorology, customer denominators, and asset-level grid data.

# 2. Data and Methods

## 2.1. Data source and coverage audit

The primary input is the public county-hour outage and VTEC archive deposited with the dataset descriptor [6]. The descriptor identifies a 2015–2024 coverage period. We downloaded the released primary CSV and independently inspected its timestamp field: it contains 3,572,394 rows, with observed timestamps from 2 January 2015 03:00 UTC through 3 January 2023 01:00 UTC. Only 1,644 source rows fall in 2023, and there are no 2024 rows in the downloaded file. We therefore treat 2023 as incomplete and exclude it from the main analysis. This release-level discrepancy is reported because it materially affects reproducibility and the valid temporal split.

The primary file represents a county-hour only when a positive processed outage metric overlaps at least one matched NWS VTEC warning. The variable max rolling avg is the maximum four-hour trailing-average customer-outage value associated with the hour. The dataset also supplies an event identifier, event-level peak outage count, a county-specific baseline used in event reconstruction, matched warning phenomenon codes, and diagnostics for the availability of the underlying 15-minute source observations [6]. A missing source observation is not equivalent to a confirmed zero outage; the data descriptor explicitly notes that absent records can reflect either zero-outage periods or source-reporting gaps [2,6].

Events were grouped by the five-digit county FIPS code and the within-county new event no identifier. A single event record was created from each group. The first chronological county-hour row defined the information point for the classifier; the event peak count was retained only to construct the outcome. The complete-year study sample contains 1,373,715 county events from 2015 through 2022, representing 3,010 county FIPS codes and 49 state-level jurisdiction labels. The analysis is restricted to the data's observed, warning-matched, positive-outage population.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

Table 1. Core fields and their role in the analysis.

| Field | Meaning in the released file | Use in this study |
|---|---|---|
| fips code; state; county | County identifier and names | Event key and location feature |
| Hour | UTC start time of the county-hour record | First-observed event time |
| Matched Events; Matched Count | Matched NWS VTEC phenomenon codes and count | First-record warning features |
| source completeness pct | Share of 16 expected 15-minute source records present in the trailing four-hour window | First-record availability feature and sensitivity filter |
| new event no | Reconstructed event identifier within a county | Event aggregation key |
| Max outage | Peak processed customer-outage count within an event | Outcome numerator; excluded from predictors |
| median sum | County-specific 75th-percentile positive outage baseline | Outcome denominator; excluded from predictors |

## 2.2. Event severity and predictors

For event $e$, relative severity is defined as $R_e$ = Max outage$_e$ / median sum$_e$. The denominator is the county-specific 75th-percentile baseline used by the released event-segmentation workflow. It provides a within-county scale for comparing the peak of an event with a typical upper-range positive outage value. It is not a percentage of customers served, an estimate of unserved energy, or a directly comparable measure of absolute customer impact between counties. Events with non-positive or non-finite denominators were excluded; no imputation was applied to the target.

The binary outcome marks an event severe when $R_e$ is at least the 90th percentile of $R_e$ in the complete 2015–2020 training period. The threshold is computed before drawing the model sample and then held fixed for the 2021–2022 evaluation period. For the released file used here, this threshold is 12.67. This relative, training-derived threshold avoids selecting a cutoff using the holdout outcomes, but it should be interpreted as a high-severity tail within this dataset rather than an engineering damage criterion.

Predictors are taken from the first positive county-hour record in each event: state, month and hour represented using sine and cosine terms, the matched phenomenon codes, the number of matched codes, and source completeness pct. Phenomenon codes are represented as binary indicators for the 30 most frequent codes in the training sample; unseen or less frequent codes do not receive separate indicators. NWS VTEC assigns codes such as SV to severe thunderstorms, TO to tornadoes, HT to heat, WI to wind, and WW to winter weather [7]. No event peak, later warning code, event duration, event mean availability, or other event-wide

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

quantity is used as a predictor. This design limits look-ahead leakage, although the first positive outage record occurs after the onset of customer impact and is not a pre-disaster warning time.

### 2.3. Temporal evaluation and models

All complete events from 2015–2020 form the training pool (968,968 events), and complete events from 2021–2022 form the held-out pool (404,747 events). To keep model fitting and repeated feature evaluation tractable, a fixed random sample of 200,000 training events and 100,000 holdout events is used for model estimation and reporting. Seeds are 17 and 29, respectively. The sample is drawn without replacement from each period; the observed severe-event prevalence in the evaluation sample is 10.03%. The remaining eligible events contribute to descriptive statistics but not to model fitting or model-score calculations.

The first model is a regularized logistic regression with class-balanced weights and standardized numeric inputs. The second is a random forest with 80 trees, maximum depth 10, a minimum leaf size of 20, and class-balanced subsampling. A training-prevalence baseline assigns every evaluation event the severe-event prevalence in the training sample. Both fitted models use the same predictors, training sample, evaluation sample, and fixed severity threshold. The RF permutation importance is calculated by the reduction in average precision after permuting one feature at a time, using a fixed 15,000-event subset of the evaluation sample and two repeats.

### 2.4. Evaluation metrics and sensitivity analysis

Discrimination is summarized by the area under the receiver operating characteristic curve (ROC-AUC) and average precision (AP), which summarizes precision across recall levels for the imbalanced severe-event class. We also report precision and recall among the 10% of evaluation events assigned the highest model scores. This is a fixed-capacity triage measure: precision is the severe-event share of the flagged group, and recall is the share of all severe evaluation events captured by that group. We do not interpret class-weighted model probabilities as calibrated event probabilities. The sensitivity analysis evaluates the fitted random forest only on held-out sampled events whose first-record source availability is at least 75% of the expected 15-minute observations.

## 3. Results

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

## 3.1. Sample characteristics

The complete-year event table includes 1,373,715 reconstructed county events. Relative severity has a median of 0.338 and an interquartile range of 0.106–1.227. Thus, for at least half of events, the event peak does not exceed the county-specific upper-quartile baseline used in reconstruction. The 90th-percentile training cutoff is much higher, at 12.67. The median source-record availability at the first observed event hour is 43.75%, indicating that less than half of the 16 expected 15-minute source records are present for the median event start. Figure 1 shows the annual event counts and median relative severity for complete years. The increase in recorded event counts over time should not be interpreted as a change in physical outage incidence because EAGLE-I coverage and reporting availability vary across years [2].

Severe-thunderstorm phenomenon code SV is present at the first observed outage hour in 750,631 events, followed by heat (HT; 186,056), wind (WI; 124,616), winter weather (WW; 109,400), and tornado (TO; 102,406). These categories are not mutually exclusive: one event can carry more than one matched phenomenon code. Figure 2 therefore reports code-specific event counts, not a partition of the event population. The distribution describes the matched records and does not measure the frequency of weather hazards across all county-hours.

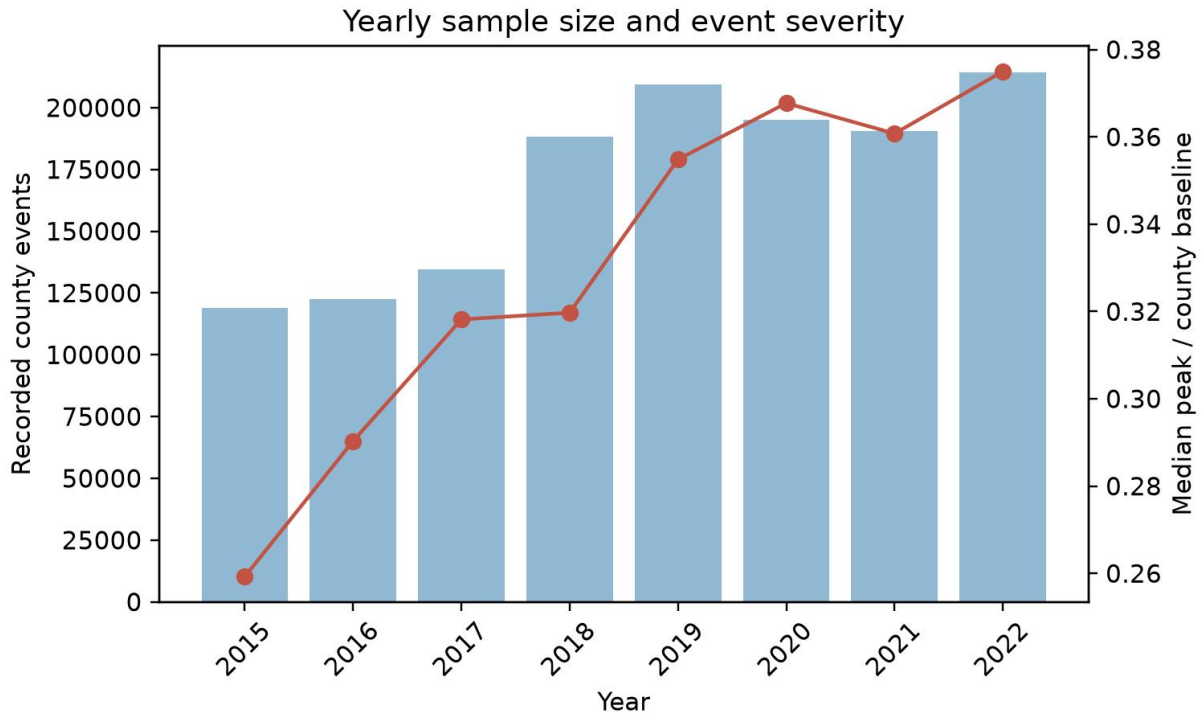



Figure 1. Annual number of reconstructed county events and median relative severity in complete years (2015–2022).

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

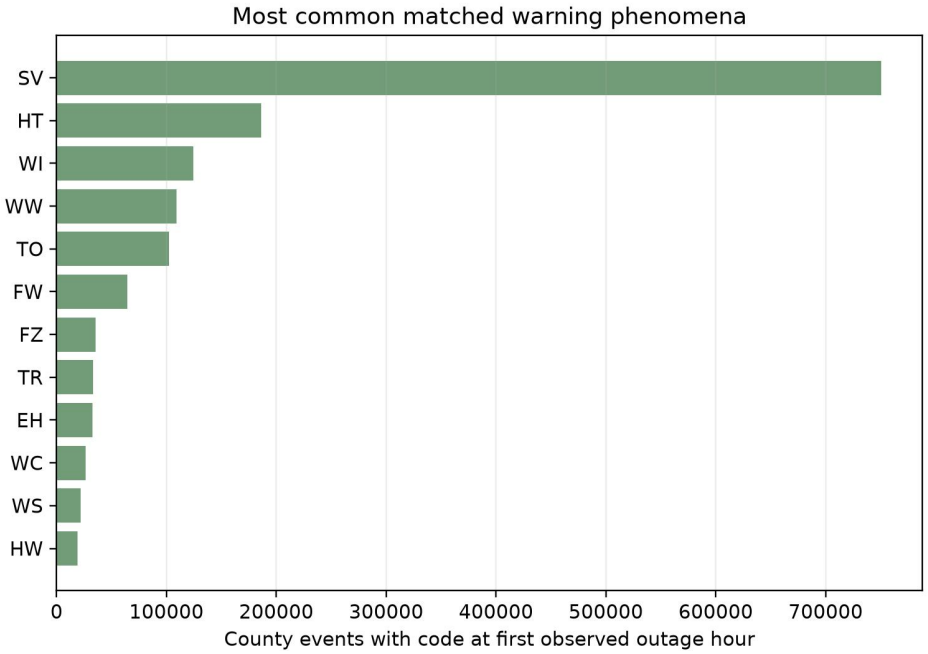



Figure 2. Most common matched NWS VTEC phenomenon codes at the first observed outage hour; codes may overlap within events.

Table 2. Study periods and event samples.

| Period or subset | Events | Role |
|---|---|---|
| 2015–2020 complete years | 968,968 | Training pool |
| 2021–2022 complete years | 404,747 | Temporal holdout pool |
| Training sample | 200,000 | Model fitting; seed 17 |
| Holdout sample | 100,000 | Model evaluation; seed 29 |
| Holdout events with source availability ≥75% | 32,113 | Sensitivity subset |

## 3.2. Held-out severity discrimination

Both models rank severe events above the constant-prevalence baseline, but their discrimination remains moderate. On the 100,000-event temporal holdout sample, logistic regression reaches ROC-AUC 0.681 and AP 0.179. The random forest reaches ROC-AUC 0.693 and AP 0.186, compared with baseline AP 0.100, approximately the held-out prevalence. Figure 3 shows the ROC curves. The difference between the two fitted models is small, and the absolute AP indicates that most high-severity events remain difficult to identify from the restricted first-record feature set.

At a fixed alert budget of 10% of the holdout sample, the random forest has precision 0.221 and recall 0.220: about 22% of its flagged events meet the severe-event definition, and it captures about 22% of all severe events. Logistic regression yields precision 0.212 and recall 0.211 at the same alert budget. A random selection of 10% of the holdout events would have

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

expected precision close to the 10.03% prevalence and recall close to 10%. This comparison shows useful but limited enrichment rather than reliable detection of all severe events.

Table 3. Event-severity discrimination on the 2021–2022 holdout sample.

| Model | ROC-AUC | Average precision | Precision in top 10% | Recall in top 10% |
|---|---|---|---|---|
| Training-prevalence baseline | 0.500 | 0.100 | 0.100 | 0.100 |
| Regularized logistic regression | 0.681 | 0.179 | 0.212 | 0.211 |
| Random forest | 0.693 | 0.186 | 0.221 | 0.220 |

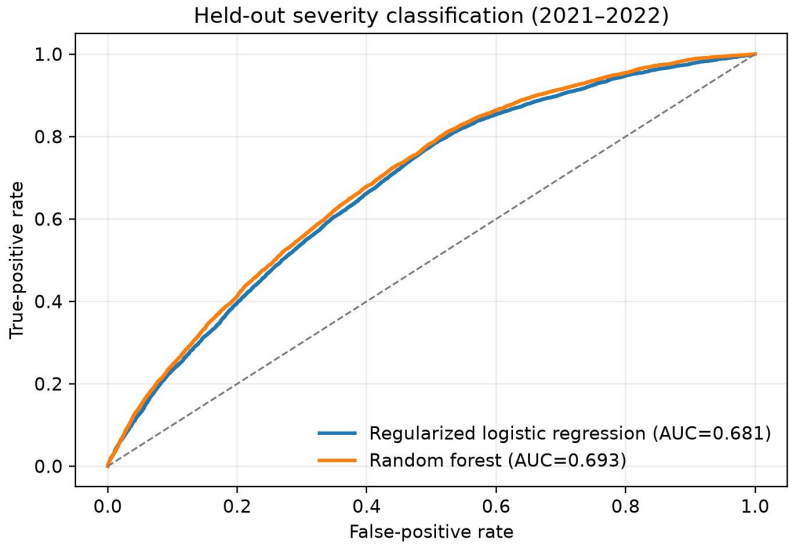



Figure 3. ROC curves for the two fitted models on the sampled 2021–2022 temporal holdout.

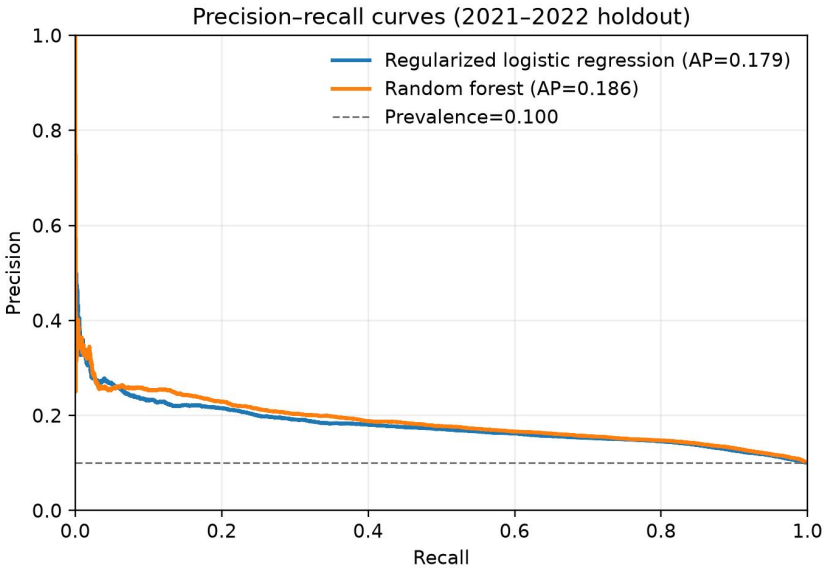



Figure 4. Precision–recall curves for the severe-event class; the dashed line marks the holdout prevalence.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

## 3.3. Feature contributions and source-availability sensitivity

Permutation importance places the severe-thunderstorm indicator SV first, followed by source-record availability at the event's first observed hour, the tornado indicator TO, and time-of-day terms. These values quantify the predictive information available to this fitted model in this holdout sample; they do not establish that a phenomenon code causes a larger outage. State indicators and other warning codes add smaller changes to AP. Figure 5 reports the leading features with their repeat-to-repeat variability.

The higher-availability sensitivity subset contains 32,113 events, or 32.1% of the sampled holdout. Its severe-event prevalence is 11.25%. The random forest has ROC-AUC 0.693 and AP 0.201 in this subset. The AP is slightly higher than the all-event holdout value, while ROC-AUC is essentially unchanged. Because the subset changes the outcome prevalence and event composition, its AP should not be compared without that context. The similar AUC suggests that the model's rank ordering does not materially improve simply by restricting the evaluation to higher-availability starts.

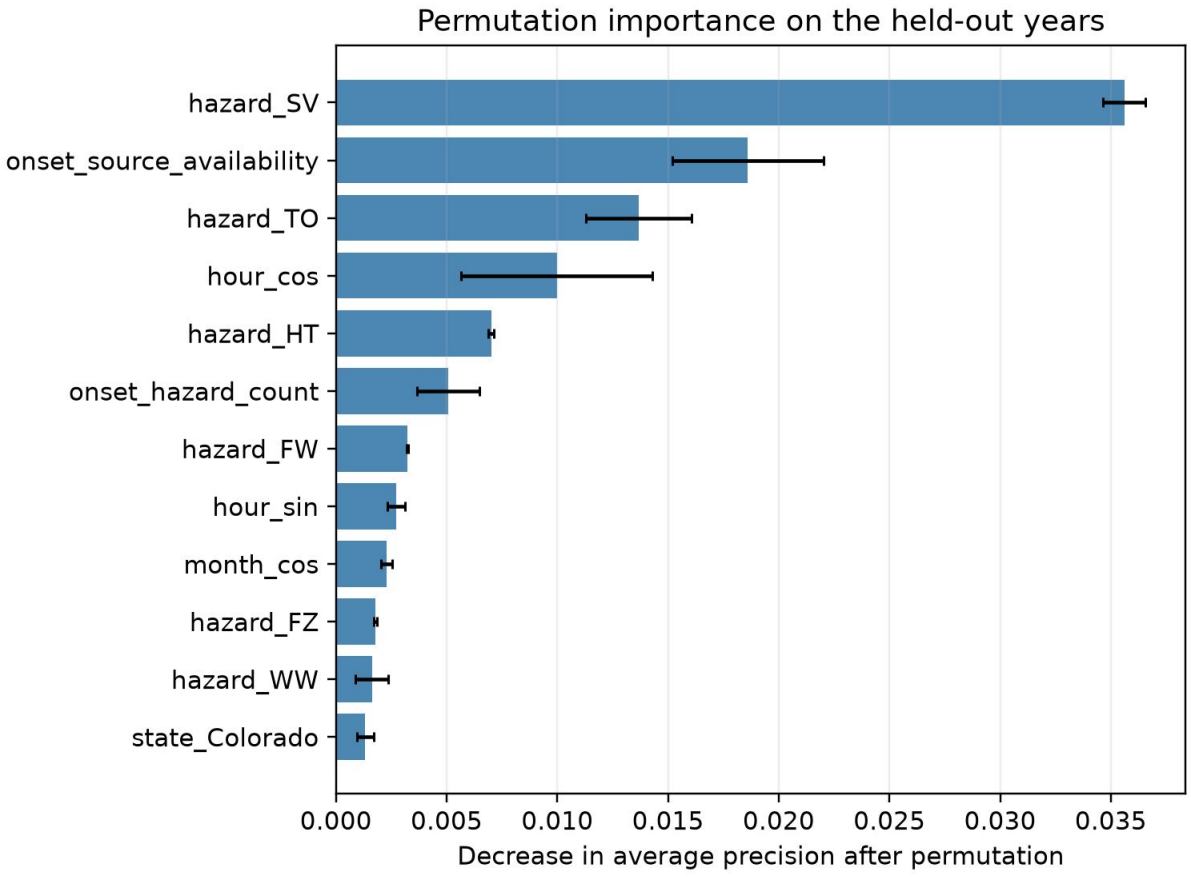



Figure 5. Leading random-forest permutation-importance values on the held-out sample; bars show mean AP decrease and error bars show repeat-to-repeat standard deviation.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

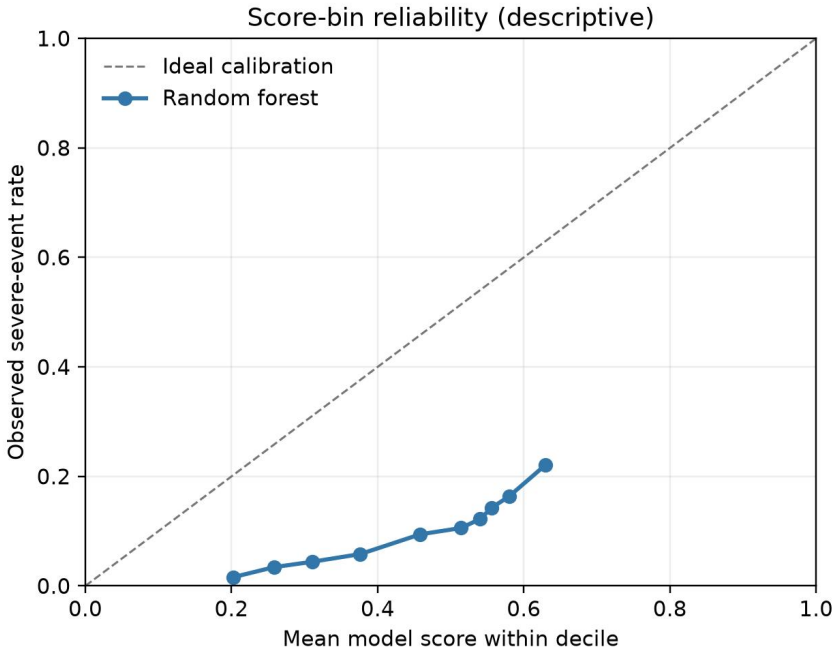



Figure 6. Mean random-forest score and observed event rate by score decile. Scores are not calibrated probabilities.

## 3.4. Capacity, year-wise, and score-reliability checks

At 5%, 10%, and 20% alert capacities, the forest's precision is 0.253, 0.221, and 0.192, while recall rises from 0.126 to 0.220 and 0.384. Logistic regression shows the same capacity trade-off, with slightly lower precision and recall at each operating point. This expands the single 10% triage result into a capacity curve and makes the decision trade-off explicit.

Table 4. Holdout performance at alternative alert capacities.

| Model | Flagged share | Precision | Recall |
|---|---|---|---|
| Random forest | 5% | 0.253 | 0.126 |
| Random forest | 10% | 0.221 | 0.220 |
| Random forest | 20% | 0.192 | 0.384 |
| Logistic regression | 5% | 0.228 | 0.114 |
| Logistic regression | 10% | 0.212 | 0.211 |
| Logistic regression | 20% | 0.183 | 0.365 |

Performance is directionally consistent across the two held-out calendar years: forest ROC-AUC/AP are 0.687/0.189 in 2021 and 0.699/0.184 in 2022. The modest year-to-year movement supports temporal stability over this short two-year window, but does not substitute for longer rolling validation. The score-decile plot shows that average scores exceed observed event rates in every bin, consistent with the class-weighted fitting scheme. Therefore, scores are useful for ranking in this experiment and should not be read as calibrated probabilities.

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

Table 5. Random-forest performance by held-out year.

| Year | n | Severe prevalence | ROC-AUC | AP |
|---|---|---|---|---|
| 2021 | 47,245 | 0.102 | 0.687 | 0.189 |
| 2022 | 52,755 | 0.098 | 0.699 | 0.184 |

## 4. Discussion

The main result is that a small set of first-record contextual variables contains some information about which already-recorded, warning-associated county events will reach the upper tail of relative severity. A random forest modestly outperforms the regularized linear model and a constant-prevalence baseline, but at a 10% triage budget it still misses roughly four out of five severe events. This level of performance may justify further evaluation as one input to event monitoring, but it does not support autonomous resource allocation or public-facing risk claims.

The leading warning indicators are consistent with the value of retaining hazard type in an event-level analysis. Earlier resilience research has combined outage and weather records to map impacts at county or state scale [5], and line-level studies have modeled weather-related component failures using physical and probabilistic inputs [3,4]. The present results are not directly comparable: the outcome here is customer-outage severity within reconstructed county events, while line-damage studies model individual grid assets and their failure mechanisms. County-level VTEC matches can help characterize event context, but they do not identify the damaged asset, local wind speed, vegetation contact, or restoration constraints.

The modest discrimination also reflects information limits in the dataset. The event's first positive outage record is observed after customer impact begins; it cannot serve as a true pre-event forecast timestamp. The model uses phenomenon codes and coarse location and time terms rather than measured wind, precipitation, temperature, flood depth, vegetation, feeder topology, customer counts served, or repair-crew availability. Moreover, the data descriptor's source-record availability field is not verified completeness. Its median value of 43.75% in this study means that sparse source records are common at event starts. The sensitivity analysis shows that filtering to a higher-availability subset does not substantially change ROC-AUC, but it cannot correct the possibility that unobserved records reflect either zero outages or collection gaps.

The file audit is another practical contribution. Although the associated data descriptor describes a 2015–2024 archive [6], the downloaded primary CSV used in this analysis ends on 3 January 2023. We exclude the incomplete 2023 segment and report the discrepancy rather

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

than silently treating the stated descriptor period as observed coverage. Researchers attempting to reproduce the work should confirm the exact file version and timestamp range before applying the reported split. A corrected or more complete release could support later-year external validation and would likely change both sample composition and model estimates.

Several extensions are needed before operational use. First, a complete county-time panel with explicit source availability and verified zero-outage periods is required to model event occurrence rather than severity conditional on a positive record. Second, customer-denominator data should convert customer counts to county-level proportions and should be aligned to the same year as each outage record. Third, measured meteorology and asset-level topology would permit a clearer link between hazard exposure, line or feeder failure, and service impact. Finally, future validation should use multiple rolling time splits, calibration analysis, and geographically blocked tests. Resilience decision support should connect severity estimates to explicit mitigation or restoration decisions rather than interpret model rankings as resilience scores [1,8].

## 5. Conclusion

This study evaluates severity stratification among recorded county-level power-outage events matched to NWS weather phenomena. From the complete 2015–2022 period in the downloaded file, 1.373 million reconstructed events were identified. A random forest trained on first-record warning and context features achieved ROC-AUC 0.693 and AP 0.186 on a sampled 2021–2022 holdout, with approximately 22% precision and 22% recall when the highest-risk 10% were flagged. The gain over the 10% severe-event prevalence is measurable but modest. The findings support further investigation of warning-aware impact triage, while making clear that the model neither predicts outages before they begin nor attributes outages causally to weather. The incomplete temporal coverage in the downloaded release, sparse source-record availability, lack of no-outage controls, and county-level rather than asset-level resolution limit the conclusions. A complete, quality-audited outage and meteorology panel is needed to extend this analysis to operational failure prediction and resilience planning.

## References

[1] Huang, Y., Rong, X., Bie, Z., Li, J., Huang, B., Zhao, T., & Li, G. (2025). Artificial intelligence for resilient power system: Motivations, advances, and challenges. Smart Power & Energy Security, 1(2), 65–75. https://doi.org/10.1016/j.spes.2025.06.001

[2] Brelsford, C. M., Tennille, S. A., Myers, A. T., et al. (2024). A dataset of recorded electricity outages by United States county 2014–2022. Scientific Data, 11, 271. https://doi.org/10.1038/s41597-024-03095-5

[3] Hou, H., Yu, S., Wang, H., Xu, Y., Xiao, X., Huang, Y., & Wu, X. (2020). A hybrid

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.

prediction model for damage warning of power transmission line under typhoon disaster. IEEE Access, 8, 85038–85050. https://doi.org/10.1109/ACCESS.2020.2992528

[4] Omogoye, O. S., Folly, K. A., & Awodele, K. O. (2021). Enhancing the distribution power system resilience against hurricane events using a Bayesian network line outage prediction model. The Journal of Engineering, 2021(11), 731–744. https://doi.org/10.1049/tje2.12091

[5] Lee, S. M., Chinthavali, S., Bhusal, N., Stenvig, N., Tabassum, A., & Kuruganti, T. (2024). Quantifying the power system resilience of the US power grid through weather and power outage data mapping. IEEE Access, 12, 5237–5255. https://doi.org/10.1109/ACCESS.2023.3347129

[6] Saki, S., Sofia, G., Kar, B., & Anagnostou, E. (2026). A spatiotemporally harmonized dataset of utility-reported power outages and NWS VTEC weather warnings for the contiguous United States (2015–2024). Scientific Data, 13, 1387. https://doi.org/10.1038/s41597-026-08356-z

[7] National Weather Service. (n.d.). Decoding the Valid Time Event Coding (VTEC). https://www.weather.gov/bmx/vtec (accessed 4 November 2026).

[8] Shuai, H., Li, F., She, B., Wang, X., & Zhao, J. (2023). Post-storm repair crew dispatch for distribution grid restoration using stochastic Monte Carlo tree search and deep neural networks. International Journal of Electrical Power & Energy Systems, 144, 108477. https://doi.org/10.1016/j.ijepes.2022.108477

© 2027 The Author(s). Published by Hong Kong Dawn Clarity Press Limited. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Licence (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/.