Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Probabilistic Framework for Reconstructing Weather-Related Power Outage Frequency

Mohammad Siddiqur Rahman

MohammadSiddiqur.Rahman@ucf.edu

University of Central Florida

Thomas Wahl

University of Central Florida

Meghana Nagaraj

University of Central Florida

Research Article

Keywords: Power outages, Probabilistic modeling, Climate variability, Grid resilience, Tropical cyclones, Frequency reconstruction

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11381609/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Probabilistic Framework for Reconstructing Weather-Related Power Outage Frequency

Mohammad Siddiqur Rahman[1,2*], Thomas Wahl[1,2] and Meghana Nagaraj[1,2]

[1*]Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, 32816, FL, USA.
[2]Center for Integrated Coastal Research, University of Central Florida, Orlando, 32816, FL, USA.

*Corresponding author(s). E-mail(s): MohammadSiddiqur.Rahman@ucf.edu;
Contributing authors: t.wahl@ucf.edu; meghana.nagaraj@ucf.edu;

Abstract

Weather-related power outage models are usually evaluated on how well they predict a select number of individual events, whereas long-term resilience planning also requires understanding of how outage frequencies are affected by climate variability and change. We develop a probabilistic framework that reconstructs weather-related outage frequency by summing calibrated daily class probabilities. It combines EAGLE-I outage observations with hydrometeorological data across all 67 Florida counties from January 2015 to September 2024. A multiclass gradient-boosted tree model separates days with no weather-related outage from those with minor (1–2 h), moderate (2–8 h), and major (at least 8 h) outages, and is evaluated by leave-one-year-out cross-validation. The predicted probabilities are well calibrated, with expected calibration errors of 0.005 or less, and summing them recovers every class total within 3%. Including static attributes in the model raises the spatial Kling–Gupta efficiency from 0.69 to 0.97 while leaving temporal skill unchanged, showing that weather controls when outages occur and local attributes control where. For nine water years, predicted and observed annual totals correlate at 0.82 across counties. Across 26 tropical cyclones, totals correlate at 0.95 for all outages and 0.85 for major outages, far exceeding a county-by-month baseline. Skill is lowest for major and summer outages, reflecting convective and restoration processes that daily county-scale predictors do not

resolve. The framework offers an open and transferable route to reconstructing outage climatology and to isolating climate-driven changes in outage frequency under a stationary-grid assumption.

Keywords: Power outages, Probabilistic modeling, Climate variability, Grid resilience, Tropical cyclones, Frequency reconstruction

1 Introduction

Extreme weather events are among the most disruptive hazards affecting human health, critical infrastructure, and environmental systems, with power outages representing one of the most frequent and consequential impacts (Watson et al. 2022; Yang et al. 2020; Liu and Zhong 2017; Do et al. 2023). In the United States, weather-related outages have increased both in frequency and intensity over recent decades, creating significant challenges for emergency response, infrastructure management, and long-term resilience planning (NOAA NCEI 2025; Carvallo and Casey 2024). Florida is particularly vulnerable due to its exposure to hurricanes, severe thunderstorms, and other extreme weather conditions where compounding effects between different weather variables cause outages (Rahman et al. 2025). For example, Hurricane Irma in 2017 resulted in widespread electrical disruptions, affecting approximately two-thirds of Florida customers (U.S. Energy Information Administration 2017; Chakalian et al. 2019). However, weather-related outages are not limited to catastrophic events. While major hurricanes produce large-scale and highly visible disruptions, more frequent moderate weather events generate localized outages that cumulatively represent a substantial burden on communities and utilities. Despite their high frequency of occurrence, these smaller-scale events remain comparatively understudied and are often not considered when training and testing outage models, especially among the growing number of studies using statistical or machine learning approaches (Han et al. 2009; Shashaani et al. 2018; Quiring et al. 2011; McRoberts et al. 2018).

Using data driven approaches for power outage modeling is inherently challenging because outage observations are sparse and characterized by strong class imbalance, zero inflation, and spatial heterogeneity. At the same time there are complex nonlinear relationships between weather conditions, infrastructure characteristics, and outage outcomes. Existing outage models often rely on data from a limited number of major storms and/or on proprietary utility datasets (Shashaani et al. 2018; Quiring et al. 2014; McRoberts et al. 2018). The latter typically have high spatial resolution and models perform well when trained with them, but the data remains inaccessible to many emergency management agencies at county, state, and federal levels (McRoberts et al. 2018). In contrast, open-access datasets such as EAGLE-I provide data in a generalized form and are easy to access but aggregated to coarse county-level resolution. The increasing frequency of weather extremes combined with aging electrical infrastructure highlights the need for publicly accessible frameworks capable of characterizing outage risk across large spatial domains and over extended periods of time.

Power-outage modeling has progressed from statistical approaches, including linear regression (Reed 2008), negative binomial regression (Liu et al. 2005), generalized linear models (Liu et al. 2008), and generalized additive models (Han et al. 2009), to tree-based machine-learning methods that capture nonlinear interactions among meteorological and infrastructure-related predictors (McRoberts et al. 2018; Quiring et al. 2011; Yang et al. 2020; Zhu et al. 2026). These models nevertheless face two related challenges: strong class imbalance and incomplete representation of spatial vulnerability. The predominance of no-outage days can cause models to favor the majority class and poorly detect rare but operationally important events. Previous studies have addressed this problem using multi-stage frameworks that separate outage occurrence from magnitude (McRoberts et al. 2018; Shashaani et al. 2018) and synthetic minority oversampling to increase the representation of rare outages (Mohammadian et al. 2021; Ma et al. 2025). Model performance also depends on combining dynamic weather with

static attributes that represent local differences in land cover, terrain, and grid infrastructure, making certain areas more susceptible to outages than others (McRoberts et al. 2018; Quiring et al. 2011). However, evidence for the value of static attributes is mixed. McRoberts et al. (2018) reported a 17% improvement after adding static predictors within a two-stage framework, whereas Quiring et al. (2011) found that proxy variables such as land cover and soil type provided little improvement when detailed infrastructure information was unavailable. This contrast motivates evaluating the benefits of adding static attributes separately for capturing temporal and spatial variability rather than judging their value using a single performance measure.

Existing probabilistic outage models have generally focused on event- or component-level prediction (Guikema et al. 2014; Arora and Ceferino 2023), although recent studies have extended probabilistic modeling across several scales. Hughes et al. (2024) combined physics-based structural fragility with machine-learning probabilities for nonstructural transmission outages and summed line–storm probabilities to estimate expected outage totals. Spaulding and Lin (2026) converted cell-level tropical-cyclone fragility probabilities into spatially aggregated outage impacts using infrastructure exposure, while Emmanouil (2025) combined conditional outage distributions with projected changes in weather hazards to estimate changes in outage recurrence intervals. These studies demonstrate the value of probabilistic representations, but their primary targets are component failure, event-scale impact, or climate-conditioned risk.

Despite recent progress, important challenges remain in developing transferable models of weather-related power outages. First, models trained predominantly on major events may not adequately represent the response of the power grid to moderate events and can consequently misrepresent outage risk for a given area (Watson et al. 2022). Second, the limited public availability of detailed infrastructure characteristics necessitates the use of static geographic and socioeconomic attributes as proxies for

grid characteristics, but their value for improving model transferability across locations remains unclear. Third, outage observations are strongly imbalanced and the relationships between weather, local characteristics, and outage severity vary spatially. How such imbalance affects outage severity classification in a spatially heterogeneous setting therefore warrants further consideration.

To fill some of the identified gaps, the study addresses three questions: (i) Can calibrated daily probabilities recover observed outage frequencies when aggregated over time? (ii) How much do static local attributes improve the spatial reconstruction of outage frequency? (iii) For which spatial and temporal scales is the reconstruction reliable?

2 Data and methods

This study follows the workflow outlined in Figure 1. Power outage observations are obtained from the Environment for the Analysis of Geo-Located Energy Information (EAGLE-I), and dynamic weather and static local features are derived from a publicly available regional reanalysis (Table 1 and Table 2). All datasets are preprocessed, harmonized, and aggregated to a common county-level spatial resolution and daily temporal scale (Sections 2.1 and 2.2). Weather-related outage days and their severity classes are defined through an explicit sensitivity analysis (Section 2.3). The integrated dataset is then used to train and evaluate a probabilistic tree-based framework that estimates how many county-days of each severity class occur in a given time window (Section 2.4).

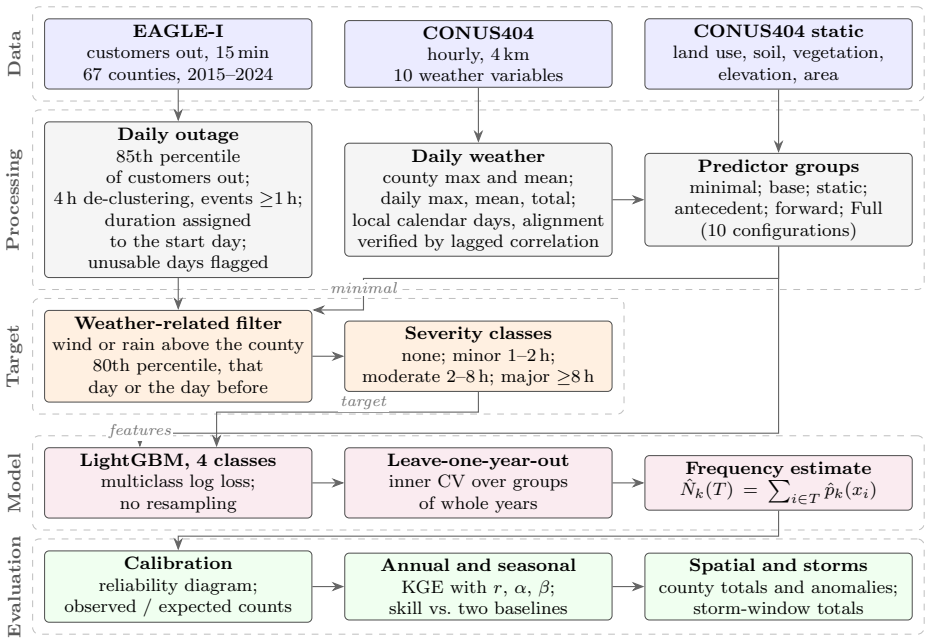


Fig. 1 Workflow for data processing and weather-related outage frequency modeling.

2.1 Power outage data and processing

EAGLE-I power outage data is managed by Oak Ridge National Laboratory and provides one of the most comprehensive records of power outages in the United States, covering more than 90% of customers (Brelsford et al. 2024). The dataset reports the number of customers without power at 15-minute intervals for each county. We use data from all 67 Florida counties from 1 January 2015 to 30 September 2024, giving 238,587 county-days.

Outage events are identified following the framework described in Rahman et al. (2025). An 85th percentile (ranges from 16 to 474 customers) threshold is first applied to each county to remove minor disturbances. Records reporting zero customers out are excluded when the percentile is computed, because these records appear only in the later years of the archive as earlier years do not report zero outage, and therefore reflect a change in reporting convention rather than in outage behavior. Consecutive

records above the threshold are grouped into independent events using a 4-hour declustering window, selected as the window producing the highest correlation between outage duration and weather variables. Each record contributes 0.25 h to the event duration, events shorter than one hour are discarded, and events lasting more than a day are assigned in full to the day on which they started. Daily outage duration for a county is the sum of the durations of all events starting on that day (note that based on this processing the duration assigned to a day can be longer than 24 hours).

The record contains 77,660 outage days, a rate of about 334 per 1,000 county-days, and their durations are heavily right-skewed (Fig. 2a). The median daily outage duration is 2.75 h, the 90th and 99th percentiles are 8.25 h and 26.75 h; 11.0% of outage days last at least eight hours, and the longest events extend beyond 100 h. Annual totals show year to year variability especially in 2024, which only goes until September (Fig. 2b). Outage days occur throughout the year with a pronounced warm-season maximum in June through August (Fig. 2c),

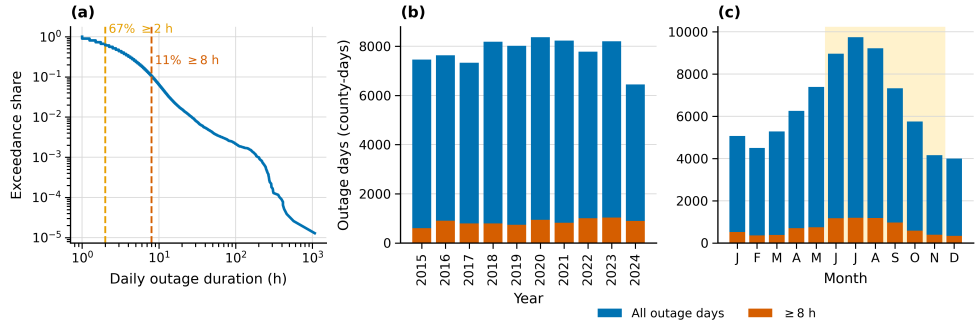


Fig. 2 Daily county-level outage record for Florida, 2015–2024. (a) Exceedance distribution of daily outage duration on outage days, with the class boundaries at 2 and 8 h indicated. (b) Outage days per year, with the subset lasting at least 8 h shown in orange. (c) Outage days per month, aggregated over all years, yellow shading represents Hurricane season.

2.1.1 Temporal alignment and data coverage

EAGLE-I timestamps and the reanalysis are both recorded in UTC. Because Florida's convective maximum spans local evening hours, which straddle UTC midnight, all records are converted to local calendar days (America/New_York) before daily aggregation, so that an evening storm and the outages it causes fall on the same date. Cross-correlating hourly customers out with hourly precipitation and wind confirms this alignment: the correlation peaks at a lag of zero hours for precipitation and one hour for wind (Fig. S1), indicating a shared time base and a short outage response time.

Every county is retained on a complete daily calendar. Days on which the outage record cannot be trusted are instead flagged and excluded from model training and evaluation while still receiving predictions. These are days before a county's first or after its last EAGLE-I record, days inside gaps longer than seven days with no record for that county, and days on which fewer than 20% of counties report anything, which indicates a failure of the reporting system rather than an absence of outages. This affects 6,019 of 238,587 county-days (2.5%), leaving 232,568 county-days for training and evaluation.

2.2 Weather and environmental data

2.2.1 Dynamic feature variables

Dynamic features are obtained from CONUS404, a convection-permitting regional reanalysis produced at 4 km spatial and hourly temporal resolution (Rasmussen et al. 2023) (Table 1). Using a single meteorological source for all dynamic variables avoids the inhomogeneities that arise when observational networks with different histories are combined.

All gridded variables are spatially aggregated over each county using the maximum over the intersecting grid cells (`cellmax`) and the mean over those cells (`cellmean`),

and are then aggregated to daily values as the daily maximum (`dmax`), mean (`dmean`), and, for accumulated quantities, total (`dsum`). The cell-maximum statistic represents the most severe conditions experienced anywhere in the county, whereas the cell-mean statistic represents how widespread those conditions are. Features are named by variable, then spatial statistic, then temporal statistic, so that `WSPD10_cellmax_dmax` (see Table 1 for abbreviations of weather variables) denotes the daily maximum of the county-maximum 10 m wind speed. This yields 54 daily weather features.

Table 1 Dynamic features, all from the CONUS404 reanalysis (Rasmussen et al. 2023) at 4 km spatial and hourly temporal resolution. All variables are aggregated to county-level daily values using the maximum and mean over intersecting grid cells and the daily maximum, mean, and, for accumulated quantities, total.

Variable	Description
Precipitation (PREC_ACC_NC)	Hourly accumulated precipitation
Graupel (GRAUPEL_ACC_NC)	Hourly accumulated graupel (an indicator of deep convection and of the storms)
10 m wind speed (WSPD10)	Wind speed at 10 m above ground
2 m temperature (T2)	Air temperature at 2 m
2 m dew point (TD2)	Dew point temperature at 2 m
Top-layer soil moisture (SMOIS_TOP)	Volumetric water content of the uppermost soil layer
Mixed-layer CAPE (MLCAPE)	Convective available potential energy for a mixed low-level parcel
Planetary boundary layer height (PBLH)	Depth of the turbulently mixed layer
Precipitable water (PWAT)	Total column water vapor
Canopy water (CANWAT)	Water intercepted and held by vegetation

Each feature is screened individually against the outage record using the area under the receiver operating characteristic curve (AUC), which is the probability that a day drawn at random from the more affected group has a higher value of the variable than a day drawn at random from the less affected group. A value of 0.5 means the variable carries no information, and 1.0 means the two groups are perfectly separated. Because the AUC depends only on the ranks of the values, it is comparable across

variables with different units and distributions. The daily maximum of the county-maximum 10 m wind speed is the most informative single variable, both for separating outage days from days without outages (AUC 0.651) and for separating outage days of at least eight hours from shorter ones (AUC 0.670), followed by county-mean daily precipitation (0.641 and 0.658). No individual variable separates the groups strongly, which motivates the multivariate treatment that follows.

2.2.2 Static feature variables

Static features represent local characteristics that may influence differences in outage response among counties exposed to similar weather conditions (Table 2). Land cover and soil texture are summarized as the fraction of each county's area in each CONUS404 land-use and soil category, together with the annual maximum and minimum green vegetation fraction and their range, mean terrain elevation, and county area. Because land cover fractions are expressed relative to county area, they are comparable across counties of different sizes.

Table 2 Static county attributes with their source, resolution, and description.

Variable	Source	Resolution	Description
Land-use class fractions (IVGTYP)	CONUS404 land-use classification	County	Fraction of county area in each land-use class; fractions sum to one
Soil-texture class fractions (ISLTYP)	CONUS404 soil classification	County	Fraction of county area in each soil-texture class
Green vegetation fraction (SHDMAX, SHDMIN, range)	CONUS404	County	Annual maximum, minimum, and range of green vegetation fraction
Terrain elevation (HGT)	CONUS404	County	Mean elevation
County area	TIGER shapefiles	County	Total land area

2.3 Defining weather-related outages and severity classes

Not all outages are caused by weather. Equipment failure, planned maintenance, animal contact, and vehicle strikes produce outages that no weather-driven model can be expected to predict, and including them would attribute unexplained variance to the weather predictors. Outage days are therefore classified as weather-related only when the weather on or immediately before the outage day was locally extreme. Two choices define the resulting labels: which weather variables and threshold define "locally extreme", and which durations separate severity classes. Because no ground truth exists for whether an individual outage was weather-caused, both are selected through a systematic comparison rather than by convention.

2.3.1 Tropical cyclone windows

Two of the choices below, and part of the evaluation, are scored against the periods when a tropical cyclone affected Florida. These windows are derived from the HURDAT2 Atlantic best-track database (Landsea and Franklin 2013). A six-hourly best-track fix counts when the system is at least at tropical-storm strength, defined as a maximum sustained wind of at least 34 kn or a tropical-storm or hurricane status code, and lies within 200 km of the Florida state boundary. Each storm's window runs from its first qualifying fix to its last, padded by one day at each end and rounded to whole local days. Twenty-six systems satisfy these criteria between 2015 and 2024; their windows cover 11,374 county-days, 4.9% of the record, and contain 4,722 outage days (Table S1).

2.3.2 Filter selection

We evaluate 220 candidate definitions formed by crossing 11 variable sets, five percentile thresholds (75th to 95th), two threshold bases (the percentile computed over a county's outage days or over all of its days), and two timing rules (the outage day

only, or the outage day or the day before). Candidates are scored on two criteria computed among outage days. The first, J_{8h}, is the share of outage days lasting at least eight hours that a filter retains minus the share of shorter outage days it retains. The second, J_{storm}, is the share of outage days falling within tropical cyclone windows that it retains minus the share of all other outage days. Both are Youden indices, equal to zero for a filter that retains outage days at random and equal to one for perfect separation; their mean is the selectivity score. Candidates must also retain at least 15% of outage days and show a coefficient of variation of the retained share across years no greater than 0.12, which excludes definitions that are too restrictive to support estimation or that introduce artificial inter-annual variability.

Selectivity increases sharply with the addition of wind and precipitation and then saturates (Fig. S2). Definitions based on rainfall intensity alone are clearly inferior, and adding graupel or convective available potential energy to wind and precipitation lowers the storm selectivity, because both variables are elevated on many days without damaging weather. The selected definition classifies an outage day as weather-related when either the daily maximum 10 m wind speed or the county-mean daily precipitation total exceeds the county's 80th percentile of outage days, evaluated on the outage day or the day before. It retains 43.0% of outage days, 63.4% of outage days lasting at least eight hours, and 65.2% of outage days within tropical cyclone windows, with a coefficient of variation of the retained share across years of 0.050.

This definition (wind + rain total, p80, outage days, day or day before) ranks third of 220 candidates by selectivity score (0.232, Table S3), and the two candidates above it differ by less than 0.002. The leading candidate differs only in using the cell-maximum rather than the county-mean daily precipitation total, so the ranking does not separate the leading candidates. We retain the county-mean formulation because it measures how widespread the rainfall is across a county rather than conditions at a single grid cell, and because its retained share is more stable between years

(coefficient of variation 0.050 against 0.067). Wind-only definitions rank close behind but retain only 30.3% of outage days, and they exclude outages caused by flooding and saturated-ground tree failure, which are physically part of the process being modeled.

Selectivity is insensitive to the percentile threshold between the 75th and 90th percentiles (Fig. S3, Table S2), so the threshold primarily governs sample size. The 80th percentile maximizes the selectivity score (0.232, against 0.230 at the 85th and 0.216 at the 90th) while retaining the largest number of outage days. Moving to the 90th percentile would discard 41% of the labeled outage days and reduce the retained share of storm-period outage days from 65.2% to 44.9% without becoming more selective. The 85th percentile is reported as a sensitivity case.

2.3.3 Severity classes

Retained outage days are divided into three severity classes by duration, and all remaining county-days form a fourth class representing the absence of a weather-related outage. Thirty-eight combinations of the minor upper bound A and the major lower bound B are compared (Fig. S4). Each combination is scored by the mean AUC with which the best weather variable separates adjacent classes, subject to the requirement that every class contains at least 10% of retained outage days and at least 30 days in every year. Separability increases monotonically with B, but the number of major days per year falls rapidly, so the most separable definitions are also the least estimable.

The selected classes are minor ($1 \leq d < 2$ h), moderate ($2 \leq d < 8$ h), and major ($d \geq 8$ h), which ranks second of 38 combinations with a mean adjacent-class AUC of 0.595 (Table S4). The individual values are 0.561 for moderate against minor and 0.630 for major against moderate, so weather separates major outages from moderate ones appreciably better than moderate from minor, but neither separation is sharp. Duration is therefore only partly determined by the meteorological forcing, which is expected, because the same storm produces different durations depending

on network condition, vegetation, and available restoration resources. The classes are best understood as ordered severity bands whose boundaries are supported by the data rather than as meteorologically distinct regimes. The highest-ranked combination differs only in the minor bound ($A = 1.5$ h) and scores 0.598; because durations are recorded in 15-minute increments and events shorter than one hour are excluded, that definition restricts the minor class to durations of 1.00 and 1.25 h, so the 2-hour bound is preferred. The eight-hour major threshold is consistent with definitions of major outage events used in reliability reporting (Do et al. 2023).

The resulting classes are physically distinct (Fig. S5). Outage days that the filter does not classify as weather-related occur under markedly lower wind and rainfall than any of the three weather-related classes (median daily maximum wind 7.08 $\mathrm{m\,s^{-1}}$ and median county-mean rainfall 0.26 mm), and both quantities increase monotonically across the severity classes, reaching 11.13 $\mathrm{m\,s^{-1}}$ and 12.50 mm for major outages. The distributions overlap, as expected when identical forcing produces different outcomes depending on local network condition, vegetation, and the precise location of a storm within a county. The filter is therefore interpreted as identifying weather-related rather than definitively weather-caused outages. The final dataset contains 199,143 county-days with no weather-related outages, 8,830 with minor, 19,178 with moderate, and 5,417 county-days with major outages (Fig. S6, Table 3).

Table 3 Severity classes and their frequency over 2015–2024, counted over the 232,568 county-days with usable outage data.

Class	Definition	County-days	Share (%)
No outages	No or not weather-related	199,143	85.6
Minor	$1 \leq d < 2$ h	8,830	3.8
Moderate	$2 \leq d < 8$ h	19,178	8.2
Major	$d \geq 8$ h	5,417	2.3
Total		232,568	100.0

2.4 Weather-related outage frequency modeling

2.4.1 Modeling objective

The quantity of interest is the number of county-days of each severity class in a period, rather than the classification of any individual day. Frequency is therefore estimated by summing predicted class probabilities over the days of interest, an approach known as probabilistic classify-and-count in the quantification literature (Bella et al. 2010; González et al. 2017) and equivalent to calibration-in-the-large in the validation of risk models (Van Calster et al. 2016). The expected number of days of class k in period T is

$$\hat{N}_k(T) = \sum_{i \in T} \hat{p}_k(x_i), \tag{1}$$

where $\hat{p}_k(x_i)$ is the predicted probability of class k for county-day i.

Summing event-level probabilities into an expected count is used in related settings. Preisler et al. (2004) summed daily ignition probabilities to expected numbers of wildfires, Hughes et al. (2024) summed transmission-line outage probabilities across line–storm observations, and Spaulding and Lin (2026) multiplied a local fragility probability by the number of possible outage locations and summed the resulting expected failures spatially. The present formulation differs in that the observational unit is a county-day, the outcomes are four mutually exclusive duration classes, and aggregation is primarily over time rather than over network components.

2.4.2 Model framework

We train a gradient-boosted decision tree ensemble (LightGBM) with a multiclass objective over the four severity classes. Gradient boosting builds trees in sequence, each fitted to the errors of its predecessors. It suits this application because outage response depends on interactions among weather variables and local characteristics rather than on any single predictor, and because the feature set combines precipitation totals, wind speeds, land cover fractions, and terrain attributes on different scales

without requiring transformation. The model is trained by minimizing multiclass log loss, a proper scoring rule that rewards calibrated probabilities.

2.4.3 Feature sets

Features are grouped so that the contribution of each group can be isolated. The *base* set contains all same-day weather features and season, expressed as month and the sine and cosine of the day of year. Three sets each add one group to this base. *Static* adds the static county attributes. *Antecedent* adds backward-looking weather: previous-day values, three-day backward windows, precipitation accumulated over the preceding 3 to 30 days, an antecedent precipitation index with a daily decay of 0.9, and antecedent soil moisture with its anomaly relative to the county's own mean for that calendar month. *Forward* adds forward-looking weather: the following one and two days, three- and five-day forward windows, five- and seven-day centered windows, and counts of days exceeding the county's 95th percentile of wind within those windows. A *relative* set adds county-normalized features, which express each day relative to that county's own distribution and therefore carry information about timing but not about the county's absolute level. Three combined sets test whether groups act together: *forward_plus* adds interactions that use only same-day or forward inputs, while *forward_static* and *antecedent_static* add the static attributes to the forward and antecedent sets. A *minimal* set containing only the two variables used to define the labels provides a lower reference, and a *full* set contains every group. Because each single-group set differs from *base* by exactly one group, the difference in skill between them measures that group's contribution. All window operations are performed on a complete daily calendar for each county, so a window never spans a gap in the record or crosses a county boundary; multi-day accumulations require at least 70% of their constituent days and are rescaled to the full window length.

2.4.4 Training and validation

Models are validated by leave-one-year-out cross-validation. For each complete calendar year, a model is trained on all remaining years and used to predict every day of the held-out year, so each annual estimate is entirely out of sample. Annual frequencies are reported by water year, from 1 October to 30 September, following the CONUS404 reporting convention; the record contains nine complete water years, from October 2015 to September 2024. Spatial results use every evaluated county-day and therefore cover the full record. Hyper-parameters and the number of boosting iterations are selected within the training years alone, using an inner cross-validation in which the training years are partitioned into three groups of whole years, so that validation is never performed on days from the same year as the training data. Three configurations of tree complexity and regularization are compared, and the one with the lowest mean validation log loss is refit on all training years.

2.4.5 Probability calibration

Because frequency estimates are sums of predicted probabilities, calibration is evaluated directly. For each class, predicted probabilities are divided into bins, and the mean predicted probability in each bin is compared with the observed frequency in that bin; a well-calibrated model lies close to the 1:1 line in the reliability diagram. Calibration is summarized by the expected calibration error,

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \left| \mathrm{obs}_b - \mathrm{pred}_b \right|, \tag{2}$$

where n_b is the number of county-days in bin b, obs_b the observed class frequency in that bin, and pred_b the mean predicted probability. We also report the ratio of the total observed count to the total expected count obtained by summing predicted probabilities, which should be close to one if the frequency estimates are unbiased in aggregate.

2.4.6 Model evaluation

For each outage class, observed and predicted annual frequencies are compared using the Kling–Gupta efficiency,

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}, \tag{3}$$

where r is the Pearson correlation between observed and predicted annual frequencies, $\alpha = \sigma_{\mathrm{pred}}/\sigma_{\mathrm{obs}}$ is the ratio of their standard deviations, and $\beta = \mu_{\mathrm{pred}}/\mu_{\mathrm{obs}}$ is the ratio of their means. The three components describe different aspects of the reconstruction: r measures whether active and quiet years are correctly ranked, α whether the amplitude of inter-annual variability is reproduced, with $\alpha < 1$ indicating a predicted series smoother than the observations, and β whether the long-term mean is reproduced.

Skill is also measured against two reference models. The climatology baseline uses the overall class proportions estimated from the training years, and the seasonal baseline uses county-by-month class proportions, which represents the frequency that can be inferred from location and time of year without the daily meteorological state. Skill relative to a baseline is

$$S = 1 - \frac{\mathrm{MSE}_{\mathrm{model}}}{\mathrm{MSE}_{\mathrm{baseline}}}, \tag{4}$$

so that $S = 1$ is perfect prediction, $S = 0$ equals the baseline, and $S < 0$ is worse than the baseline. We use mean squared error because it gives greater weight to large errors during unusually active periods, and report mean absolute error in physical units where the typical magnitude of the error is informative.

2.4.7 Uncertainty and sensitivity to individual years

Uncertainty in the statewide annual metrics is quantified by resampling counties with replacement 2,000 times, reconstructing the statewide water-year series for each sample, and recomputing the metrics; the 2.5th and 97.5th percentiles define the 95% confidence interval. This measures how far the results depend on the particular set of counties analysed. A water-year jackknife measures the influence of individual years: each complete water year is removed in turn and the temporal metrics are recomputed from the remaining years.

2.4.8 Seasonal, storm-scale, and daily performance

Performance is evaluated at both aggregated and daily scales. The mean monthly frequency is used to assess the seasonal cycle, and complete meteorological seasons (DJF, MAM, JJA, and SON) are evaluated with the same KGE decomposition and baseline-relative skill applied to seasonal totals, with December assigned to the following year's winter and incomplete seasons at the ends of the record excluded. Performance is also compared among tropical (June–November) and non-tropical (December–May) cyclone seasons, to test whether skill extends beyond named storms. Within each tropical cyclone window, observed and predicted outage totals are compared using correlation and mean absolute error for both the model and the seasonal baseline.

Daily discrimination is evaluated using the AUC defined in Section 2.3, applied to the predicted probabilities rather than to a weather variable. The first AUC measures how well the summed probability of the three outage classes separates outage days from days without a weather-related outage, and the second measures how well the conditional probability of a major outage separates major outages from shorter ones. Both are computed by weather regime and compared with the seasonal baseline. Because the AUC evaluates the ranking of days rather than the absolute level of the probabilities, it complements the calibration and frequency diagnostics.

3 Results

3.1 Calibration of the predicted probabilities

The frequency estimates are sums of daily class probabilities, so those probabilities must be reliable before proceeding to reconstruction. Observed frequencies follow the predicted probabilities closely across the whole range for all four classes in the held-out data (Fig. 3). The expected calibration error is 0.005 or smaller for every class, and the ratio of observed to expected counts over all 232,568 evaluated county-days is 0.999 for the no-outage class, 1.004 for minor outages, 0.999 for moderate outages, and 1.028 for major outages (Table S5). Summing the daily probabilities therefore recovers the observed totals to within 3% for every class. This result supports the use of summed daily probabilities to estimate outage frequency rather than counting only the most likely class.

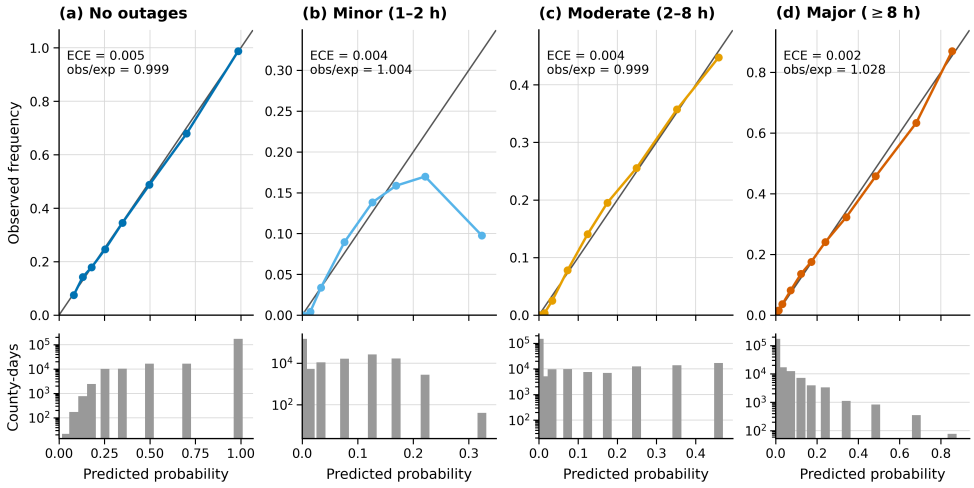


Fig. 3 Reliability of the predicted daily class probabilities. The upper panels compare the mean predicted probability in each bin with the observed frequency in that bin, with the 1:1 line shown in black. The lower panels give the number of county-days in each bin on a logarithmic scale. ECE is the expected calibration error and obs/exp is the ratio of observed to expected counts over all evaluated county-days.

Two departures are visible. Minor outages are slightly underpredicted in the sparsely populated bins above a predicted probability of 0.25, which contain fewer than 100 county-days in total. Major outages are underpredicted by 3% overall, and the underprediction is concentrated in the bins between 0.2 and 0.8. Both patterns indicate that the model is conservative on the days when severe outcomes are most likely, which also appears in the annual results below.

3.2 Predictor groups

The feature-set comparison shows a clear trade-off between temporal and spatial reconstruction (Fig. S7, Table 4). Temporal KGE ranges from 0.68 to 0.72 for all outage classes combined, and from 0.46 to 0.57 for major outages. The exception is the minimal configuration, which uses only the two variables that define the labels and reaches 0.54 and 0.24. Adding the full same-day weather state to those two variables therefore produces the single largest gain in this study, and every further group of predictors adds little beyond it.

Table 4 Feature-set comparison for all weather-related outages (Any) and major outages. Temporal skill is MSE skill relative to the seasonal baseline. The full dataset is retained for the remaining analyses because it combines strong temporal reconstruction, the highest baseline-relative skill, and high spatial reconstruction.

	Temporal KGE		Skill vs. seasonal		Spatial KGE	
Feature set	Any	Major	Any	Major	Any	Major
Minimal	0.543	0.240	0.432	0.299	0.359	0.297
Base	0.681	0.490	0.343	0.436	0.685	0.668
Static	0.721	0.547	0.342	0.493	0.969	0.925
Antecedent	0.688	0.465	0.532	0.368	0.702	0.684
Forward	0.702	0.517	0.535	0.449	0.689	0.660
Forward+	0.698	0.519	0.474	0.447	0.685	0.651
Relative	0.682	0.535	0.583	0.516	0.675	0.665
Forward+static	0.710	0.571	0.564	0.488	0.963	0.911
Antecedent+static	0.692	0.506	0.626	0.467	0.969	0.910
Full	0.703	0.547	0.654	0.547	0.969	0.922

The configurations separate more clearly in space than in time. Sets that include the static county attributes reproduce the county totals with a spatial KGE between 0.91 and 0.97, while sets built from weather alone reach 0.65 to 0.70. Adding static attributes to the forward-looking set raises its spatial KGE from 0.69 to 0.96 while leaving its temporal KGE almost unchanged.

We therefore use the full data set for the remaining analyses. This choice does not mean that the full model maximizes every individual metric. Instead, it gives the most balanced reconstruction across temporal variability, spatial structure, and improvement over the seasonal baseline.

3.3 Annual frequency and inter-annual variability

Predicted annual frequencies track the observed record across the nine complete water years from October 2015 to September 2024 (Fig. 4, Table S6). The model reproduces minor outages best, with a KGE of 0.81 and a correlation of 0.96, and predicts 8,040 county-days against 8,055 observed. Moderate outages reach a KGE of 0.60 and a correlation of 0.86. Combining the three outage classes gives a KGE of 0.68 and a correlation of 0.82, with 30,427 predicted county-days against 30,699 observed. Annual deviations stay within 10% for minor outages, moderate outages, and the combined total in every year, and the bias ratio lies between 0.99 and 1.00 for those three targets.

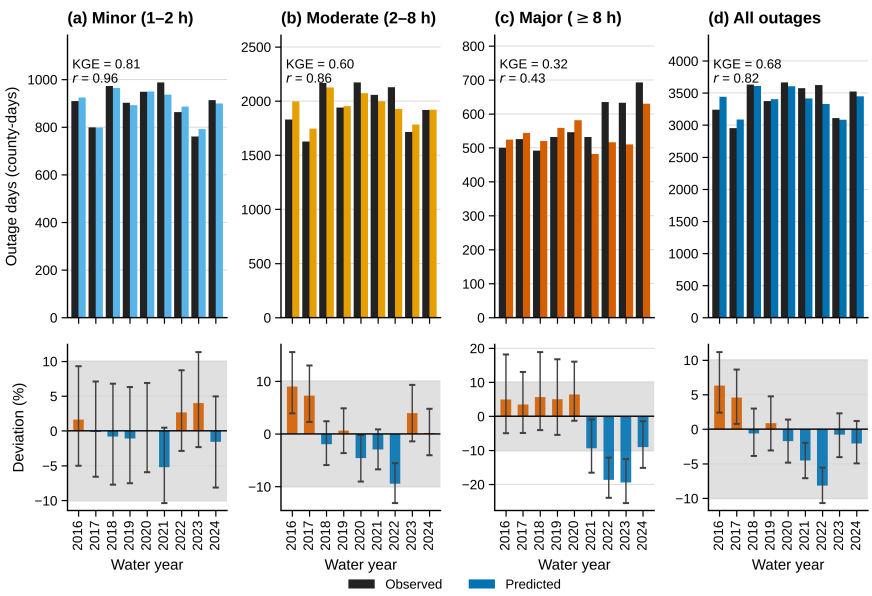


Fig. 4 Observed and predicted annual frequency of weather-related outage days for the nine complete water years from October 2015 to September 2024. The upper panels give total county-days per water year. The lower panels give the deviation of the prediction from the observation, with 95% intervals obtained by resampling counties and a grey band marking ±10%. Orange bars indicate overprediction and blue bars underprediction.

The main limitation is the amplitude of year-to-year variability. All four series have α between 0.62 and 0.81, so the predicted annual series is smoother than the observations. Major outages are the hardest target. The correlation falls to 0.43 and the KGE to 0.32, although the total over the nine years remains close to the observed value at 4,867 county-days against 5,089. The errors follow a clear pattern in time. The model overpredicts major outages by 3.5 to 6.4% in the five water years from 2016 to 2020, then underpredicts them by 9.4, 18.7, 19.4 and 9.1% in 2021 through 2024. The observed number of major outage days rises from about 500 per year in the first part of the record to 693 in 2024, and the predictions stay between 482 and 630. The variability ratio of 0.63 reflects the same behavior. The county-bootstrap intervals indicate that these deficits are systematic rather than a result of only a few counties. The model therefore captures the broad frequency level better than the full amplitude of severe-outage years.

A leave-one-year-out jackknife of the evaluation shows that the bias ratio remains between 0.94 and 1.01 and the variability ratio below 1 for every class in every fold, so neither the absence of bias nor the damping of inter-annual variability depends on a particular year. Skill is more sensitive: omitting the 2017 water year, which contains Hurricane Irma, lowers the KGE for all outage classes from 0.68 to 0.38, and omitting 2024 removes the major class's advantage over the seasonal baseline (skill 0.45 to -0.01). With nine evaluation years, single active years therefore carry a substantial share of the measured skill.

3.4 Seasonality

The model reproduces the seasonal cycle of all three classes (Fig. 5). Predicted and observed rates peak together in June through August at about 265 outage days per 1,000 county-days, fall to about 75 in November and December, and agree within a few days per 1,000 county-days in every month. The county-by-month baseline follows the same cycle by construction, so this comparison shows that the model learns the seasonal signal from the weather rather than from the calendar. It does not by itself demonstrate skill.

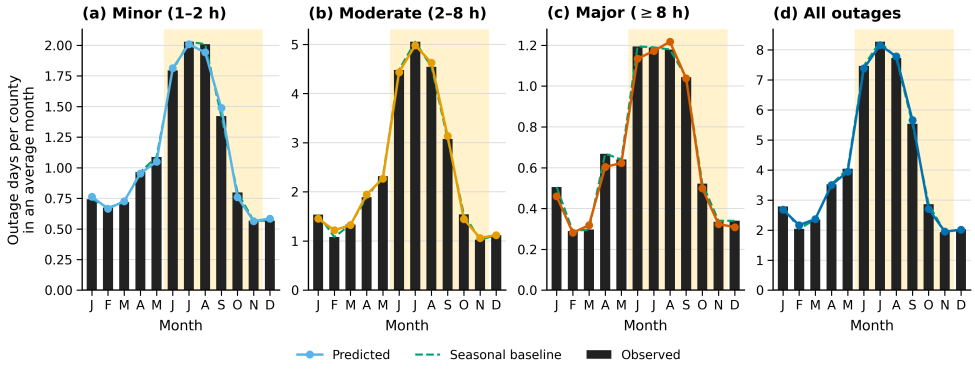


Fig. 5 Mean seasonal cycle of weather-related outage days by class, averaged over all evaluated years. The shaded band marks the Atlantic hurricane season from June to November. The seasonal baseline reproduces this cycle by construction and is shown for reference; skill relative to it is assessed on inter-annual variation in Figs. 6

Season-by-season results show that the model captures much of the observed inter-annual variability, with most observed seasonal totals falling within the 95% bootstrap confidence intervals of the predictions (Fig. 6, Table S7). For all outage classes combined (Fig. 6a–d), KGE is highest in winter (0.98), followed by spring and autumn (0.86 in both seasons), and lowest in summer (0.67). The model outperforms the seasonal baseline in every season, with MSE skill ranging from 0.56 in summer to 0.97 in winter and autumn. For major outages ($\geq$ 8 h; Fig. 6e–h), performance follows the same seasonal ordering but is generally lower. KGE decreases from 0.87 in winter to 0.59 in spring and autumn and 0.35 in summer, while MSE skill ranges from 0.52 in summer to 0.86 in winter.

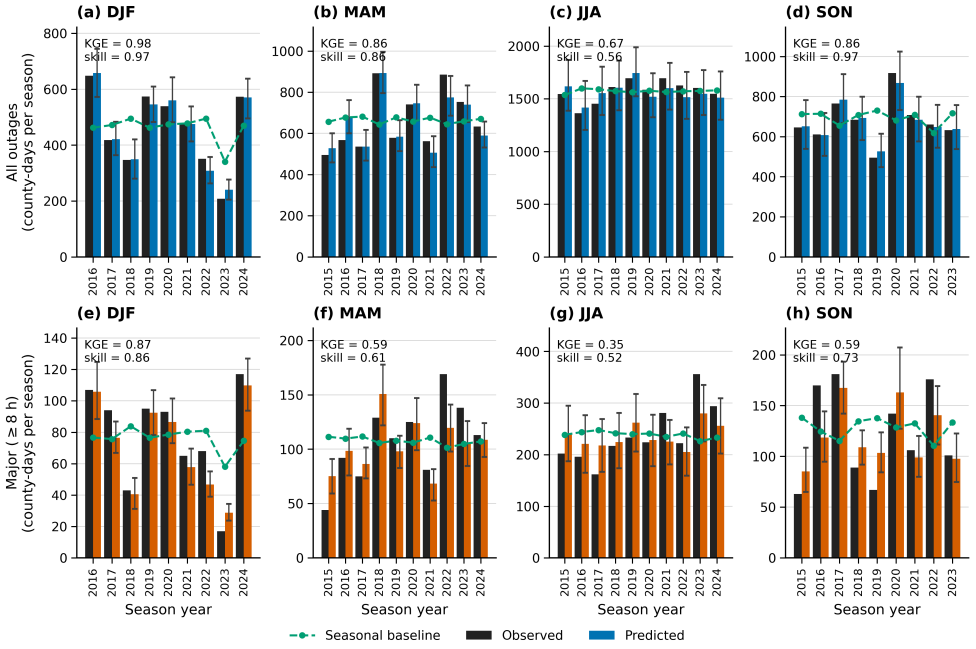


Fig. 6 Observed and predicted seasonal totals of weather-related outage days by standard season: (a)–(d) all outage classes, (e)–(h) major outages ($\geq$8 h). Error bars give 95% intervals from resampling counties, and the dashed line is the county-by-month baseline.

The model therefore reproduces cool-season and transition-season variability more closely than summer variability. Summer contributes the largest number of outage

county-days but remains the most difficult season to reproduce. Observed summer totals for all outage classes range from 1,363 to 1,686 county-days across the ten summers, a range of approximately 20%, whereas the predicted totals show less inter-annual variation. This under-representation is more pronounced for major outages, for which the summer variability ratio is 0.41, the lowest among the four seasons. Winter shows the closest agreement in variability. Observed winter totals range from 208 to 646 county-days, nearly a factor of three, and the corresponding variability ratio is 0.99. These results indicate that the model captures winter inter-annual variability particularly well, whereas the daily county-scale predictors explain less of the variability associated with summer thunderstorm activity.

3.5 Tropical cyclones and daily skill by regime

Twenty-six tropical cyclones meet the proximity and intensity rule for Florida during the study period (Section 2.3). Summed over all storm windows, the model predicts 3,106 weather-related outage days against 3,079 observed, and 820 major outage days against 846 observed (Fig. 7). Predicted and observed totals correlate at 0.95 across storms for all outage classes and at 0.85 for major outages. The county-by-month baseline predicts 2,052 and 343 county-days for the same windows, with correlations of 0.64 and 0.13. Weather information therefore supplies nearly all of the storm response, and the baseline, which knows only the county and the month, cannot distinguish an active storm window from an ordinary day in the same month.

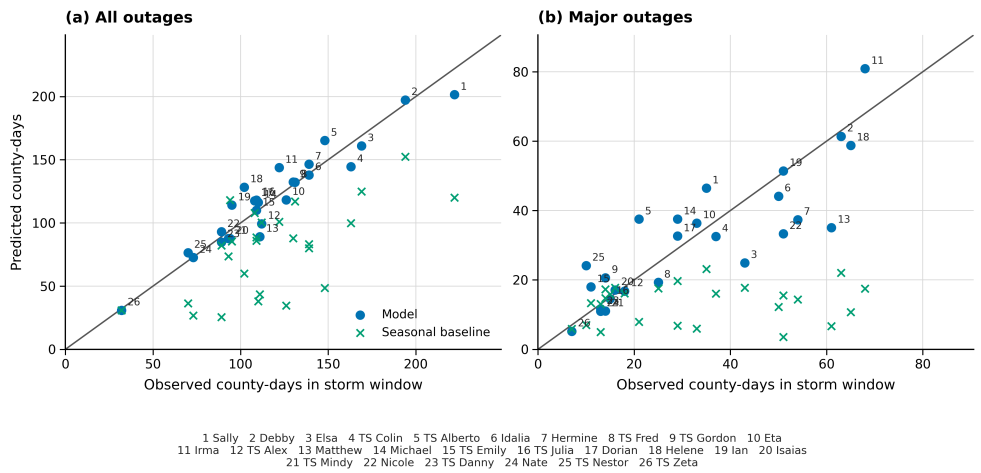


Fig. 7 Model performance during tropical cyclones and by weather regime. (a) and (b) compare observed and predicted county-days within each storm window for all outage classes and for major outages, with the county-by-month baseline shown for reference.

Individual storms show where the model departs from the observations. It reproduces the largest events closely: Hurricane Irma, the most impactful storm in terms of outages in the record, and Hurricanes Debby and Helene all fall near the 1:1 line in both panels. The clearest shortfalls are for storms whose observed impact is large but whose weather footprint over Florida is comparatively modest, including Matthew and Nicole, which passed offshore along the east coast, and Elsa. The model overpredicts for a small number of weaker systems, including Tropical Storms Nestor and Alberto. The pattern is consistent with the damping of extremes seen in the annual results: predictions for the most severe windows are pulled towards the mean of comparable weather days.

Daily skill differs by weather regime (Table S8). The model separates outage days from days without outages in every regime, with an AUC of 0.908 on tropical-cyclone days, 0.945 on other hurricane-season days and 0.963 in the off-season, against 0.679, 0.771 and 0.663 for the seasonal baseline. Separating major outages from shorter ones is harder and depends strongly on the regime. The model reaches 0.804 on tropical-cyclone days, where severity follows the storm itself, but only 0.684 on other

hurricane-season days, and 0.748 in the off-season. The baseline reaches 0.497 on tropical-cyclone days, which is no better than chance. Warm-season convective storms therefore produce outages whose duration depends on factors that daily county-scale predictors do not resolve, which is the same limitation that gives summer the lowest seasonal skill (Section 3.4).

Tropical-cyclone days are rare but disproportionate. They account for 11,374 of 232,568 evaluated county-days, or 4.9%, yet carry an outage rate of 271 days per 1,000 county-days against 184 in the rest of the hurricane season and 94 in the off-season, and 27.5% of their outage days are major, against 16.2% overall (Table S8).

3.6 Spatial pattern

Over the whole period the model reproduces the geography of outage exposure (Fig. 8, Table S9). County totals correlate with the observations at 0.988 for minor outages, 0.997 for moderate outages and 0.991 for major outages. The median absolute deviation per county is 5.3%, 3.1% and 9.5% respectively. For minor and moderate outages, 96% and 93% of counties fall within 20% of their observed totals. For major outages the share drops to 72%.

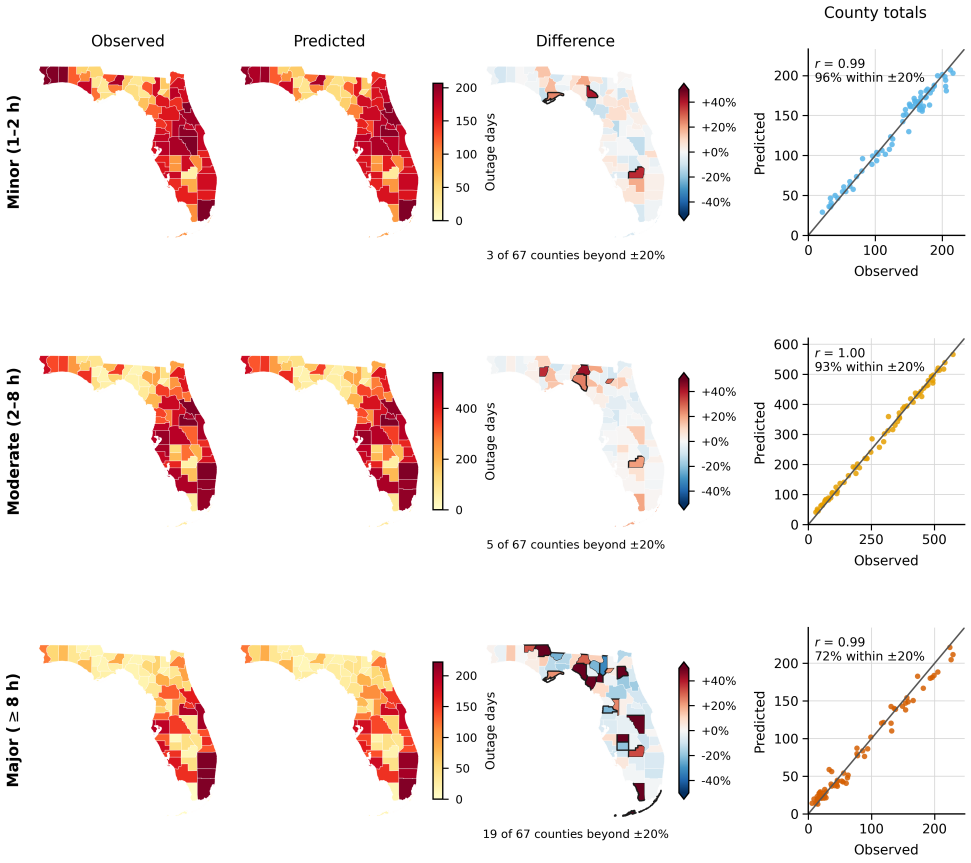


Fig. 8 Observed and predicted county totals of weather-related outage days by class, summed over all evaluated county-days. The third column gives the difference as a percentage of the observed total and the fourth column compares the totals county by county.

The difference maps show that errors are largest in the counties with the fewest outage days, where a small absolute error produces a large percentage error. The remaining differences for major outages are not randomly distributed. The model predicts too many major outage days in several inland counties of the northern peninsula and too few along parts of the southeast coast.

County-level inter-annual anomalies are harder to predict than either the statewide totals or the long-term county pattern. The correlation between observed and predicted county-year anomalies is 0.54 for all outage classes combined and 0.27 to 0.35 for individual classes, and it changes little between predictor configurations. The model

therefore identifies which years are active statewide and which counties are exposed on average, but it captures county-level departures from each county's own average only moderately well.

4 Discussion

4.1 Frequency estimation

The calibration results show why the distinction drawn in Section 1, between predicting an event and estimating a frequency, matters in practice. Counting the most probable class would have recovered only a small share of the observed minor and major outage days, because neither class is ever the most likely outcome on a typical day. Summing the predicted probabilities recovers the observed totals to within 3% for every class over the full record (Section 3.1). Our use of probability summation is most directly related to Hughes et al. (2024), who summed binary line–storm outage probabilities to estimate total outages, and to Spaulding and Lin (2026), who aggregated exposure-weighted cell fragilities into spatial outage impacts. The present study changes both the unit and the target by summing calibrated multiclass probabilities over county-days to reconstruct duration-specific frequencies. It also differs from the climate-risk framework of Emmanouil (2025), which estimates return-period changes by combining a stationary conditional outage relationship with projected shifts in hazard distributions. Thus, the main contribution here is not a new definition of expected count, but its calibrated multiclass implementation for long-term, county-scale outage-frequency reconstruction.

4.2 Roles of dynamic and static attributes

The predictor comparison separates two questions that are often merged. Weather variables determine when outages occur, and static county attributes determine where

they occur. Adding the static attributes to a weather-only configuration raises the spatial KGE from about 0.69 to 0.97 while leaving the temporal KGE almost unchanged (Section 3.2). Static attributes are constant in time, so they cannot by themselves explain year-to-year variation, and the result confirms that they do not.

This result offers an explanation for the mixed picture in earlier work (McRoberts et al. 2018; Quiring et al. 2011). Both findings can hold at once if the two studies were testing different dimensions of performance. Static attributes help where the target has a spatial dimension, and they add little where the target is a time series. A study that evaluates only one dimension will therefore reach only one of these conclusions.

The county-normalized predictors behave in a way that follows from the same argument. Expressing wind and precipitation relative to each county's own distribution improves the ranking of days within a county, and it raised skill against the seasonal baseline in our comparison, but it carries no information about a county's absolute level of exposure. Configurations built from these variables alone reproduced the timing of outages while distorting the spatial pattern. The practical implication is that a model intended to serve both purposes needs predictors of both kinds, which is why we retained the full configuration.

4.3 Major outages as target

Three results point to the same limitation. The predicted annual series is smoother than the observations for every class, and most strongly for major outages ($\alpha = 0.63$). The model under-predicts the most damaging storm windows. Summer, which contributes the largest number of outage days, has the lowest seasonal skill.

Part of this is intrinsic to the method. A probabilistic model assigns a moderate probability to many similar days rather than a high probability to the few days that turn out to be extreme, so summing probabilities reproduces the mean of a group of comparable days and not the tail. The quantification literature describes the same

behavior as a bias towards the class distribution of the training data, and proposes adjustments that rescale the predicted prevalence (Bella et al. 2010; Saerens et al. 2002). Applying such an adjustment to a climate application would require assuming how the prevalence changes, which is the quantity being estimated, so we did not apply one here. The damping should instead be reported as a known property: the framework estimates the level of severe-outage activity well and its amplitude conservatively.

Part of it is physical. Duration depends on how long restoration takes, and restoration depends on crew availability, access, and the number of simultaneous faults, none of which is contained in daily county-scale weather (Stevens and Belligoni 2024; Hallenstein et al. 2019; Rahman et al. 2025). The weather variables separate major outages from shorter ones only moderately even in the descriptive analysis, with an adjacent-class AUC of 0.630 (Section 2.3), and the daily discrimination results show the same pattern: severity is easiest to predict during tropical cyclones, where the storm itself dominates the outcome, and hardest during ordinary convective weather (Table S8).

Summer is the clearest case. Florida's warm-season outages are driven by thunderstorms whose damaging winds occur on scales of a few kilometers and a few minutes (Tervo et al. 2019). CONUS404 resolves convection explicitly at 4 km, but aggregating to county-daily values removes much of that detail. A county-mean daily rainfall total cannot distinguish a slow-moving storm that affects one substation from widespread light rain. Sub-daily or sub-county predictors would be the natural next step, at the cost of a target that EAGLE-I cannot resolve, since it reports at the county level.

4.4 The rise in major outages after 2020

The observed number of major outage days increases from about 500 per year early in the record to 693 in 2024, while the predictions stay between 482 and 630 (Section 3.3). A weather-driven model reproduces only the variability that the weather carries, so this gap has two possible explanations, and the data available here cannot separate

them. The first is the damping described above, which would produce exactly this pattern if the later years contained more extreme weather. The second is a change in the outage record or in the system that the weather does not explain. Candidates include changes in EAGLE-I reporting coverage (Brelsford et al. 2024), growth in the number of customers served, and changes in vegetation management or restoration practice. Florida utilities have also carried out extensive storm-hardening programmes over this period (Weiss and O'Quinn 2025), which would be expected to reduce rather than increase outage duration, and which therefore does not explain the direction of the change. Distinguishing these explanations would require infrastructure and restoration records that are not publicly available, and we note the trend as an open question rather than attributing it.

4.5 Limitations

Nine complete water years is a short sample for inter-annual evaluation. The jackknife shows that no single year controls the absence of bias or the damping of variability, but that skill does depend on individual years: removing the 2017 water year, which contains Hurricane Irma, lowers the KGE for all outage classes from 0.68 to 0.38 (Section 3.3). Differences between predictor configurations are much smaller than this, which is why we treat the configuration comparison as a sensitivity analysis rather than a selection procedure.

No record states whether an individual outage was caused by weather, so weather-related outage days are defined by a filter, and the filter is part of the result. We addressed this by selecting the filter through an explicit comparison of 220 candidates and by reporting the sensitivity of the outcome to that choice, but a different reasonable filter would produce a different labeled dataset. The filter also introduces a mild circularity, since days are labeled using two of the variables the model later uses as predictors. The predictor comparison bounds the size of this effect: a configuration

using only those two variables reaches a KGE of 0.54 for all outage classes, against 0.70 for the full configuration (see Table 4), so most of the skill comes from information beyond the definition itself.

EAGLE-I reports at the county level, so the analysis cannot resolve which feeders or substations failed, and county attributes cannot represent the configuration or condition of the distribution network. Studies with utility records report that network characteristics matter for outage prediction (McRoberts et al. 2018). The county scale is the price of using a public dataset with consistent coverage across a whole state for a decade.

Because a multi-day event is assigned in full to the day it starts, weather on the following days is informative about the label by construction. Such predictors are admissible for hindcasting and for driving the model with climate model output, where the whole weather sequence is available, but they are not available in real time, and the skill reported here should not be read as forecast skill.

4.6 Applications

The calibrated daily probabilities can be aggregated over any period for which the required meteorological fields are available, supporting historical reconstruction, climate projection, and long-term resilience planning.

Because CONUS404 extends back to 1979 (Rasmussen et al. 2023), the model can reconstruct how frequently each outage-severity class would occur if the present-day power system were exposed to earlier weather. These estimates represent a counterfactual reconstruction rather than the outages that actually occurred.

The framework could likewise be driven with downscaled historical and future climate simulations while maintaining a stationary power grid and weather–outage

relationship, following Emmanouil (2025). Differences in probability-summed frequencies between baseline and future periods would then represent the isolated climate signal rather than a forecast of an evolving grid.

Finally, expected annual county-days by severity class can support comparisons among counties, climate periods, and resilience-planning scenarios (Paul et al. 2024; Hughes et al. 2024). The model identifies long-term spatial exposure and years with high statewide activity, but captures county-specific annual anomalies only moderately well. It is therefore better suited to long-horizon planning and broad resource-capacity assessment than to year-ahead allocation for individual counties with an option to scale to entire U.S..

5 Conclusions

We developed a county-level probabilistic framework for reconstructing the frequency of weather-related power outages from publicly available outage, meteorological, and geographic data. Rather than assigning each county-day to its most probable class, the framework estimates frequency by summing calibrated daily class probabilities. Across 232,568 evaluated county-days, expected calibration error was at most 0.005, observed-to-expected count ratios ranged from 0.999 to 1.028, and the total frequency of every severity class was recovered within 3%. Counting only the most probable class, in contrast, substantially underestimated rare outages. The results also distinguish the factors controlling temporal and spatial reconstruction. Weather variables captured when outages occurred, whereas static county attributes increased spatial KGE from approximately 0.69 to 0.97 with little change in temporal skill. Annual frequencies were nearly unbiased, with bias ratios of 0.96–1.00. The model reproduced the ranking of years for minor, moderate, and all outage classes combined, with correlations of 0.82–0.96, and outperformed the county-by-month climatology in every season. Major outages were more difficult, with an annual correlation of 0.43.

The principal limitation was the damping of extremes. Variability ratios of 0.62–0.81 indicate that the model reproduced average frequency more reliably than the magnitude of unusually active years, with the strongest damping for major and summer outages. Nevertheless, across 26 tropical cyclones, predicted and observed storm totals correlated at 0.95 for all outage classes and 0.85 for major outages, substantially exceeding the climatological baseline. The framework is therefore best suited to reconstructing long-term outage frequency and comparing climate periods rather than predicting the precise impact of individual county-years. Because it relies on public data and a reproducible target definition, the approach is transferable to other regions and to climate-model applications that isolate changes in meteorological forcing under a stationary-grid assumption. Projected changes in severe outages should, however, be interpreted conservatively.

Acknowledgements. We acknowledge DOE-CESER, the agency responsible for funding the creation of EAGLE-I data. Support for DOI 10.13139/ORNL-NCCS/1975202 dataset is provided by the U.S. Department of Energy, project EAGLE-I under Contract DE-AC05-00OR22725. Project EAGLE-I used resources of the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Cybersecurity, Energy Security, and Emergency Response of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Statements and Declarations

Funding. Mohammad Siddiqur Rahman and Thomas Wahl were supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under the Solar Energy Technologies Office Award Number DE-EE0010418. This work is also supported by the National Science Foundation as part of the Megalopolitan Coastal Transformation Hub (MACH) under NSF award ICER-2103754.

This is MACH contribution number 117. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability. EAGLE-I power outage data are available from Oak Ridge National Laboratory at https://doi.ccs.ornl.gov/dataset/d26451af-7ea4-577c-8285-78b72cd8a2dc. CONUS404 reanalysis data are available through the U.S. Geological Survey at https://www.sciencebase.gov/catalog/item/6372cd09d34ed907bf6c6ab1, and HURDAT2 best-track data are available from NOAA at https://www.aoml.noaa.gov/hrd/hurdat/hurdat2.html. The processed dataset and code used to reproduce the analyses and figures are available at https://github.com/CoRE-Lab-UCF/ProbabilisticPowerOutage.

Author contributions. This study was conceived by M S R and T W. M S R developed the methodology, conducted the analysis, and wrote the first draft of the article under the guidance of T W. M N contributed by generating ideas, providing valuable insights during technical discussions and editing the manuscript.

References

Arora P, Ceferino L (2023) Probabilistic and machine learning methods for uncertainty quantification in power outage prediction due to extreme events. Natural Hazards and Earth System Sciences 23(5):1665–1683

Bella A, Ferri C, Hernández-Orallo J, et al (2010) Quantification via probability estimators. In: IEEE International Conference on Data Mining (ICDM), pp 737–742

Brelsford C, Tennille S, Myers A, 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

Carvallo JP, Casey J (2024) Weather-related power outages rising. Climate Matters, URL https://www.climatecentral.org/climate-matters/weather-related-power-outages-rising, accessed September 20, 2026

Chakalian PM, Kurtz LC, Hondula DM (2019) After the lights go out: Household resilience to electrical grid failure following Hurricane Irma. Natural Hazards Review 20(4):05019001. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000335

Do V, McBrien H, Flores NM, et al (2023) Spatiotemporal distribution of power outages with climate events and social vulnerability in the USA. Nature Communications 14(1):2470. https://doi.org/10.1038/s41467-023-38084-6

Emmanouil S (2025) A parametric framework for evaluating power outage risk trends under changing climatic conditions. Stochastic Environmental Research and Risk Assessment 39:1851–1876. https://doi.org/10.1007/s00477-025-02937-x

González P, Castaño A, Chawla NV, et al (2017) A review on quantification learning. ACM Computing Surveys 50(5):1–40

Guikema SD, Nateghi R, Quiring SM, et al (2014) Predicting hurricane power outages to support storm response planning. IEEE Access 2:1364–1373. https://doi.org/10.1109/ACCESS.2014.2365716

Hallenstein J, Rich D, Cordiano V (2019) Review of processes for establishing and communicating estimated time of restoration by electric IOUs. Tech. Rep. PA-18-08-005, Florida Public Service Commission, Office of Auditing and Performance Analysis

Han SR, Guikema SD, Quiring SM (2009) Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models. Risk Analysis 29:1443–1453. https://doi.org/10.1111/j.1539-6924.2009.01280.x

Hughes W, Nyame S, Taylor W, et al (2024) A probabilistic method for integrating physics-based and data-driven storm outage prediction models for power systems. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 10(2):04024021. https://doi.org/10.1061/AJRUA6.RUENG-1171

Landsea CW, Franklin JL (2013) Atlantic hurricane database uncertainty and presentation of a new database format. Monthly Weather Review 141(10):3576–3592. https://doi.org/10.1175/MWR-D-12-00254.1

Liu H, Davidson RA, Rosowsky DV, et al (2005) Negative binomial regression of electric power outages in hurricanes. Journal of Infrastructure Systems 11(4):258–267

Liu H, Davidson RA, Apanasovich TV (2008) Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. Reliability Engineering & System Safety 93(6):897–912

Liu Y, Zhong J (2017) Risk assessment of power systems under extreme weather conditions—a review. In: 2017 IEEE Manchester PowerTech. IEEE, Manchester, UK, pp 1–6, https://doi.org/10.1109/PTC.2017.7981141

Ma J, Li B, Omitaomu OA, et al (2025) Establishing nationwide power system vulnerability index across US counties using interpretable machine learning. Applied Energy 397:126360. https://doi.org/10.1016/j.apenergy.2025.126360

McRoberts DB, Quiring SM, Guikema SD (2018) Improving hurricane power outage prediction models through the inclusion of local environmental factors. Risk Analysis

38:2722–2737. https://doi.org/10.1111/risa.12728

Mohammadian M, Aminifar F, Amjady N, et al (2021) Data-driven classifier for extreme outage prediction based on Bayes decision theory. IEEE Transactions on Power Systems 36(6):4906–4914. https://doi.org/10.1109/TPWRS.2021.3086031

NOAA NCEI (2025) U.S. billion-dollar weather and climate disasters. National Centers for Environmental Information, https://doi.org/10.25921/stkw-7w73, URL https://www.ncei.noaa.gov/access/billions/, accessed September 20, 2026

Paul S, Poudyal A, Poudel S, et al (2024) Resilience assessment and planning in power distribution systems: Past and future considerations. Renewable and Sustainable Energy Reviews 189:113991. https://doi.org/10.1016/j.rser.2023.113991

Preisler HK, Brillinger DR, Burgan RE, et al (2004) Probability based models for estimation of wildfire risk. International Journal of Wildland Fire 13(2):133–142

Quiring SM, Zhu L, Guikema SD (2011) Importance of soil and elevation characteristics for modeling hurricane-induced power outages. Natural Hazards 58:365–390. https://doi.org/10.1007/s11069-010-9672-9

Quiring SM, Schumacher AB, Guikema SD (2014) Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling. Bulletin of the American Meteorological Society 95:47–58. https://doi.org/10.1175/BAMS-D-12-00012.1

Rahman MS, Wahl T, Nagaraj M, et al (2025) Characteristics of power outages from compound weather extremes in Florida. Environmental Research: Climate 4(3):035008. https://doi.org/10.1088/2752-5295/adf252

Rasmussen RM, Chen F, Liu C, et al (2023) CONUS404: Four-kilometer long-term regional hydroclimate reanalysis over the conterminous United States. U.S. Geological Survey data release, https://doi.org/10.5066/P9PHPK4F, version 3.0, June 2026

Reed DA (2008) Electric utility distribution analysis for extreme winds. Journal of Wind Engineering and Industrial Aerodynamics 96(1):123–140. https://doi.org/10.1016/j.jweia.2007.04.002

Saerens M, Latinne P, Decaestecker C (2002) Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation 14(1):21–41. https://doi.org/10.1162/089976602753284446

Shashaani S, Guikema SD, Zhai C, et al (2018) Multi-stage prediction for zero-inflated hurricane induced power outages. IEEE Access 6:62432–62449. https://doi.org/10.1109/ACCESS.2018.2877078

Spaulding A, Lin N (2026) High-resolution simulation and modeling of tropical-cyclone induced power outages in the United States. Applied Energy 426:128511. https://doi.org/10.1016/j.apenergy.2026.128511

Stevens KA, Belligoni S (2024) Policymaking in the dark: The impact of power outage information asymmetry on local government resilience efforts in Florida. International Journal of Disaster Risk Reduction 105:104381. https://doi.org/10.1016/j.ijdrr.2024.104381

Tervo R, Karjalainen J, Jung A (2019) Short-term prediction of electricity outages caused by convective storms. IEEE Transactions on Geoscience and Remote Sensing 57(11):8618–8626. https://doi.org/10.1109/TGRS.2019.2921809

U.S. Energy Information Administration (2017) Hurricane Irma cut power to nearly two-thirds of Florida's electricity customers. Today in Energy, URL https://www.eia.gov/todayinenergy/detail.php?id=32992

Van Calster B, Nieboer D, Vergouwe Y, et al (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology 74:167–176

Watson PL, Spaulding A, Koukoula M, et al (2022) Improved quantitative prediction of power outages caused by extreme weather events. Weather and Climate Extremes 37. https://doi.org/10.1016/j.wace.2022.100487

Weiss A, O'Quinn H (2025) State spotlight on resilience: The Florida Public Service Commission and storm protection plans. Tech. rep., National Association of Regulatory Utility Commissioners, supported by the U.S. Department of Energy, Grid Deployment Office

Yang F, Watson P, Koukoula M, et al (2020) Enhancing weather-related power outage prediction by event severity classification. IEEE Access 8:60029–60042. https://doi.org/10.1109/ACCESS.2020.2983159

Zhu N, Sun N, Chen B, et al (2026) Grid-based damage prediction for distribution network under typhoon disasters based on two-stage LightGBM. In: 2026 IEEE 2nd International Conference on Power and Integrated Energy Systems (ICPIES), pp 998–1002, https://doi.org/10.1109/ICPIES69888.2026.11650849

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Probabilisticreconstructionofpoweroutagefrequencysupply.pdf