

Explainable Machine Learning for Extreme Weather Event Prediction Using Long-Term Environmental Measurements

Jose Tupayachi

Oak Ridge National Laboratory

Ziwei Liu

University of Tennessee at Knoxville

Kevin Birdwell

Oak Ridge National Laboratory

Xueping Li

University of Tennessee at Knoxville

Xiao-Ying Yu

yuxiaoying@ornl.gov

Oak Ridge National Laboratory

Research Article

Keywords: Extreme Event Detection, Data Fusion, SHAP, Machine Learning Modeling, Environmental Measurements, Big Data Analytics

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9944156/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable Machine Learning for Extreme Weather Event Prediction Using Long-Term Environmental Measurements

Jose Tupayachi^{1,2}, Ziwei Liu², Kevin Birdwell³, Xueping Li²,
Xiao-Ying Yu^{1*}

¹Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA.

²Department of Industrial and Systems Engineering, University of Tennessee, Knoxville, TN, 37996, USA.

³Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA.

*Corresponding author(s). E-mail(s): yuxiaoying@ornl.gov;
Contributing authors: tupayachisja@ornl.gov; zliu72@vols.utk.edu;
birdwellkr@ornl.gov; xli27@utk.edu;

Abstract

Prediction of extreme weather events is essential for operational preparedness at research facilities and critical infrastructure. This study presents an explainable multi-tower ML framework for forecasting extreme events using 15-min meteorological observations from six instrumented towers at the Oak Ridge Reservation. Using exploratory data analysis, five operationally relevant events are determined of high occurrence, including temperature-moisture hazard (above 20°C and 15 g/m³ absolute humidity), wind chill (at or below 4.8°C with winds of at least 3 mph), low temperature (below 3°C), high wind (above 19 mph), and atmospheric stagnation (winds below 4 mph). A data-fusion strategy pivots observations from all six towers into a single wide-format feature space, enabling each model to learn cross-tower spatial patterns. We evaluate six architectures spanning three paradigms: gradient boosting (LightGBM, XGBoost), deep learning (LSTM-FCN, TST), and convolutional neural networks (ResNet+, XceptionTime). The 300 experimental configurations span event-specific forecast horizons from 30 min to 6 hrs. An ablation study quantifies the contribution of seven background meteorological variables, including mixing height, dew point, solar radiation, atmospheric stability, and surface pressure gradients. Gradient

boosting achieves the highest overall performance based on MCC, compared against DL and CNN. Event-specific accuracy is excellent for temperature moisture, wind chill, and low temperature, while high wind remains challenging due to rapid onset dynamics. SHAP-based feature importance analysis reveals that tower features, such as rolling temperature minima, peak wind speed, and absolute humidity, dominate extreme weather prediction, while physics-based gradient features contribute at extended horizons up to six hours. This work provides the first operational benchmark using ML models based on ORR-HiResMet for extreme event prediction at critical facilities.

Keywords: Extreme Event Detection, Data Fusion, SHAP, Machine Learning Modeling, Environmental Measurements, Big Data Analytics

1 Introduction

Extreme weather events present persistent and growing challenges for the operation of research facilities, industrial sites, and critical infrastructure. Accurate detection of such events with sufficient lead time enables proactive risk mitigation, optimizes resource allocation, and improves safety planning. The Oak Ridge Reservation (ORR) is located in East Tennessee within the Ridge and Valley Province of the Appalachian Region and encompasses major research and national security facilities, including Oak Ridge National Laboratory and the Y-12 National Security Complex. These facilities maintain a multi tower meteorological monitoring network that provides high-resolution temporal environmental data. Monitoring towers continuously acquire high-resolution meteorological observations; however, to date, no explainable benchmark study has established a near real-time detection framework for extreme weather events in this environment. Long-term monitoring across the complex terrain of the ORR remains important for ensuring safe and reliable site operations. The availability of rich quality-assured environmental data provides an opportunity to develop predictive models using rapidly advancing machine learning (ML) techniques capable of resolving local-scale meteorological variability (Steckler et al. 2025). The application of ML methods is particularly valuable for improving operational resilience, resource allocation, and safety planning under rapidly evolving atmospheric conditions. Beyond the original ORR-HiResMet variables, we incorporate expert-derived gradient variables and meteorologist-informed atmospheric descriptors specifically designed to capture terrain-induced variability. These physics-informed gradient variables are introduced to enhance predictive performance at longer forecasting horizons, where models base solely on baseline observational variables demonstrated reduced event detection capabilities (see Section 2.2).

In this work, we establish a framework that can handle multi-year environmental data that is designed for event threshold calibration and descriptive statistics. Although the long-term dataset can be used to inform extreme event discovery and quality assurance, extreme event detection based on ML is lacking. First, an Exploratory Data Analysis (EDA) is used to determine five distinct extreme weather events using methods such as Sen’s slope, Mann-Kendall test, and cluster analysis (Liu

et al. 2026; Zhou et al. 2022). Literature on data-driven weather event prediction spans several research efforts. In addition, we used data from 2022 with gradient features and meteorologist insights to provide more constraints in ML applications (see Section 2.2). Several popular ML, deep learning (DL), and convolutional neural network (CNN) methods are employed and compared for their effectiveness in extreme event modeling. These models are selected according to their recent applications in temporal environmental and atmospheric measurements.

State-of-the-art DL approaches such as GraphCast (Lam et al. 2023) and convolutional neural network (CNN) based models (Weyn et al. 2019; Fang et al. 2021; Bochenek and Ustrnul 2022) have demonstrated strong capability in synoptic-scale weather prediction. However, these methods operate on coarse gridded atmospheric fields and they are not directly suitable for binary extreme-event detection at the level of individual instrumented towers. Station-level sequence models, including Temporal Convolutional Networks (TCNs)

and Long Short-Term Memory (LSTM) architectures (Hewage et al. 2020, 2021), have been primarily applied to single-site wind-speed regression tasks. While effective for local temporal dynamics, these approaches typically do not incorporate multi-tower data fusion or explicitly account for terrain-stratified event classification. In parallel, classical ML baselines, such as gradient boosting methods, are frequently used for meteorological classification; however, they are rarely evaluated in a unified setting alongside DL and CNN-based models under consistent terrain and atmospheric conditions. Large-scale environmental and time-series prediction tasks paradigms use Xiong et al. (2024) work to demonstrate that hybrid gradient-boosting and recurrent architectures (GSK-BiGRU-XGBoost) deliver strong short-term meteorology-driven forecasts, motivating our side-by-side evaluation of XGBoost, LightGBM, and recurrent/transformer baselines on multi-tower observations. Hancock and Khoshgoftaar (2020) provide an interdisciplinary review of gradient boosting on big-data workloads, supporting our choice of tree-based models as competitive baselines for high-dimensional fused tabular inputs. Beyond environmental applications, ML for time-indexed operational data is an active topic within the journal, as illustrated by Wu et al. (2026), who benchmark static, ensemble, and LSTM models for industrial predictive maintenance on chronologically ordered sensor records,



Figure 1 ORR meteorological tower locations. Tower is abbreviated as TOW, i.e., Tower A is TOWA and so on.

139 highlighting the breadth of temporal-sequence modeling under the journal’s scope.
140 More broadly, prior studies on extreme-event classification (Amnuaylojaroen 2025; V
141 et al. 2025; Cai et al. 2025) often operate at hourly to daily temporal resolutions and
142 focus on single-event prediction. Many of these works address class imbalance using
143 Synthetic Minority Over-Sampling Technique (SMOTE), which generates synthetic
144 samples to balance training distributions. In contrast, this study does not rely on
145 synthetic oversampling. Instead, it models observed extreme-event distributions and
146 employs Shapley Additive exPlanations (SHAP) alongside gradient-based attribution
147 methods to enable a comparative interpretability analysis across both traditional ML
148 and deep/ convolutional learning frameworks. This explainability-first design is con-
149 sistent with the recent studies toward interpretable deep models on high-dimensional
150 data, exemplified by the Polynomial-SHAP framework of Ejiyi et al. (2026).

151 We evaluate three families of models ranging from gradient boosting, DL, and
152 convolutional architectures to provide a comprehensive comparison among modeling
153 paradigms (see Table 2). Gradient boosting decision trees represent state-of-the-art
154 methods for tabular weather prediction tasks (Friedman 2001; Natekin and Knoll
155 2013). LightGBM (Ke et al. 2017) employs leaf-wise tree growth with histogram-
156 based binning for efficient large-scale training, while XGBoost (Chen and Guestrin
157 2016) utilizes level-wise growth with regularization. Their inherent compatibility with
158 SHAP explainability makes them particularly suitable for operational deployment.
159 We evaluate two DL architectures designed for time series classification. LSTM-
160 FCN (Long Short-Term Memory with Fully Connected Network) (Karim et al. 2019)
161 combines LSTM layers for sequential pattern capture with a FCN head for clas-
162 sification. TST (Temporal Sequence Transformer) (Zerveas et al. 2020) employs
163 multi-head self-attention mechanisms to capture long-range temporal dependencies
164 without recurrence.

165 In addition, two convolutional architectures are evaluated for temporal pattern
166 extraction. ResNet+ (Ismail Fawaz et al. 2019, 2020) employs residual convolutional
167 blocks with skip connections, enabling deep feature hierarchies while mitigating van-
168 ishing gradients. XceptionTime (Rahimian et al. 2020) uses depthwise separable
169 convolutions for parameter-efficient multi-scale feature extraction. Recent benchmarks
170 on tabular data consistently demonstrate that well-tuned gradient boosting methods
171 achieve competitive or superior performance compared to DL (Rasp et al. 2020). By
172 evaluating all three model families, we provide first empirical evidence for the optimal
173 modeling strategy for multi-tower extreme event prediction. A central methodological
174 contribution of this work is the multi-tower data fusion strategy. Instead of training
175 separate models for each tower, observations from all six towers are integrated into a
176 unified wide-format feature space indexed by timestamp. Modeling efforts based on
177 this strategy aim to close the knowledge gaps through five joint contributions: (i) pro-
178 viding the first ML and DL benchmark on the ORR-HiResMet dataset using six years
179 of quality assured data from six instrumented towers between 2017 and 2022 to obtain
180 overall statistical trend; (ii) offering controlled, six-architecture, three-paradigm com-
181 parisons among Gradient Boosting (GB) (i.e., LightGBM, XGBoost), DL (i.e., LSTM,
182 Gated Recurrent Unit (GRU)), and CNN (i.e., 1D-CNN, TCN), under strictly identi-
183 cal data splits, feature engineering, and K-fold testing; (iii) validating a spatial fusion
184

strategy and assembling concurrent observations into a single wide-format feature matrix that exploits cross-tower correlations along valley to ridge elevation gradients; (iv) handling high class imbalance (event rates 1.6-24%) via inverse-frequency weights and Matthews Correlation Coefficient (MCC), Bayesian hyperparameter optimization (Optuna Tree-structured Parzen Estimators (TPE) (Akiba et al. 2019)) without data augmentation; and (v) establishing an explainability framework using SHAP TreeExplainer (Lundberg and Lee 2017), SHAP DeepExplainer and integrated gradients for sequential DL, which provides first-difference attribution for convolutional models and yields terrain-stratified feature-importance maps that expose the physical mechanisms underpinning each identified extreme weather event type.

2 Monitoring Stations and Data Description

This section describes the data sources and processing pipeline for extreme weather event prediction using multi-tower meteorological observations at ORR. Six instrumented towers cover complex terrain from valley floors (258 m) to ridge tops (354 m), where each tower records atmospheric variables (i.e., temperature, wind speed and direction, humidity, and pressure), forming the basis for ML models. We further construct features across heights and time-aligned tower observations, and augment them with physics-informed gradient variables (e.g., mixing height, atmospheric stability, surface pressure gradients) derived by site meteorologists to provide boundary-layer context. Annual event occurrence rates are shown in Supplementary Materials Figure S4.

2.1 Study Area and Data Collection

The study area of ORR is known for its complex terrain with an observational network that enables high-resolution meteorological monitoring and subsequent modeling. Table 1 summarizes diverse terrain types from valley floors to ridge slopes, with elevations ranging from 258 m to 354 m above sea level. The tower network covers a compact geographic footprint (~ 7 km north-south, ~ 8 km east-west). It captures substantial topographic diversity, with an elevation range of 96 m and terrain types from flat valley floors (TOWA, TOWB; slope $< 1^\circ$) to steeply-inclined ridge slopes (TOWF; 15.1°). TOWF is the only station situated in a 55% canopy cover, while the valley-floor stations are in open fields. TOWD features the tallest instrumentation (60 m), providing unique vertical profile observations. This topographic diversity the multi-tower fusion model to allow learning of terrain-dependent atmospheric signatures that a single-tower approach cannot capture. The tower network samples a systematic terrain gradient: TOWA and TOWB are flat valley-floor open fields (~ 260 m MSL, $\leq 0.6^\circ$ slope, > 1100 m from ridge); TOWD is a north-facing grassland slope (261 m, 10.9° , 736 m to ridge, 60 m mast); TOWF sits at the ridge crest (354 m, 15.1° , 55% deciduous canopy); and TOWS, TOWW, and TOWY are Y-12 shrubland slopes (285-352 m, 3 5.5° , 256-552 m to ridge). This layout encodes the physical gradients cold-air pooling, drainage flows, canopy turbulence that single-station networks cannot resolve. The full six-year record is used for event-threshold calibration and quality assurance while the predictive modeling uses 2022 data, compiled by the site meteorologist (Section 2.2).

Measured variables include temperature (TempC), relative humidity (RelHum), absolute humidity (AbsHum), sustained wind speed (WSpdMph), peak wind speed (PkWSpdMph), vertical wind speed (VSSpdMph), barometric pressure (BarPresMb), wind direction (WDir), precipitation (PrecipIn), and solar radiation (SolarRadWm2) at multiple heights per tower. TOWY lacks AbsHumw while TOWF has sparse 2017 observations.

Table 1 Meteorological tower network characteristics at the ORR.

Tower	Lat.	Lon.	Elev. (m)	hr. (m)	Terrain Type	Slope (°)	Canopy (%)	Land Use
TOWA	35.920	-84.305	260	33	Valley Floor	0.0	5	Open Field
TOWB	35.933	-84.302	258	33	Valley Floor	0.6	5	Open Field
TOWD	35.925	-84.325	262	60	Valley Slope	10.9	20	Grassland
TOWF	35.953	-84.303	354	10	Ridge Slope	15.1	55	Deciduous Forest
TOWS	35.980	-84.254	322	25	Valley Slope	3.4	33	Shrubland
TOWY	35.987	-84.254	285	33	Valley Slope	5.5	25	Shrubland

Notes: Lat. = Latitude; Lon. = Longitude; Elev. = Elevation; hr. = Height.

2.2 Site Meteorological Features

In order to augment local tower observations with physics-based atmospheric context, we integrate seven gradient variables compiled by the site meteorologist. These variables capture the synoptic and mesoscale drivers of extreme weather onset at timescales from hr to sub-daily, organized into three categories. The first category characterizes boundary-layer state through mixing height (measured in meters above ground level) and atmospheric stability class, which together describe the depth and turbulent structure of the boundary layer. The second category encompasses thermodynamic forcing via dew point temperature (in °C) and solar radiation (W/m^2), both of which drive surface energy exchange and near-surface moisture availability. The third category captures large-scale and near-surface pressure forcing through three variables: 500-mbar geopotential height (in m), surface pressure gradient direction °, and surface pressure gradient magnitude (in mbar/km), which together set the large-scale flow and near-surface forcing relevant to wind events. Each variable is available at hourly resolution and aligned to match the 15-min tower observations prior to feature engineering. Each base variable then undergoes similar feature engineering pipeline as tower observations, including lag windows, rolling statistics, and first-order derivatives.

2.3 Extreme Event Definitions & Research Hypotheses

Five extreme weather events are defined based on operational thresholds (Liu et al. 2026; Zhou et al. 2022) relevant to the site (Supplementary Materials Table S2). Events are labeled when the threshold condition is met continuously for the specified minimum duration across any of the six towers or union logic. This work explores three central hypothesis. **Feature importance and terrain influence** hypothesis (**H1**), valley-floor and valley-slope towers are expected to dominate feature importance for thermodynamic events (E1-E3) due to cold-air pooling and stronger temperature signal-to-noise, while ridge-top towers are expected to dominate dynamic wind events (E4-E5), with cross-tower aggregations ranking among the top predictors for all event types. **Temporal detection horizon and event predictability** (**H2**),

temperature-driven events (E1-E3) are expected to sustain $MCC > 0.80$ at 4-6 hr horizons due to thermodynamic inertia, while wind-driven events (E4-E5) are expected to degrade steeply beyond 1 hr, with gradient boosting models outperforming DL/CNN approaches by 0.15-0.20 MCC across all horizons. **Multi-event feature synergy and measurement variables enhancement (H3)**, gradient variables are expected to benefit dynamically-driven wind events (E4) most at extended horizons (≥ 1.5 hr) where pressure-gradient dominates, while contributing marginally or negatively for thermodynamic events (E1-E3) and for sequence-based DL/CNN models susceptible to dimensionality noise.

3 Methodology

3.1 Machine Learning Approaches

Table 2 Modeling approaches for extreme weather event prediction.

Model	Approach	Applications & References
LightGBM	GB	Solar irradiance, temperature anomaly (Ke et al. 2017; Voyant et al. 2017)
XGBoost	GB	Precipitation nowcasting, environmental monitoring (Chen and Guestrin 2016; Shi et al. 2015)
LSTM-FCN	DL	Sequence classification, time series forecasting (Karim et al. 2019)
TST	DL	Long-range temporal dependencies (Zerveas et al. 2020)
ResNet+	CNN	Time series classification (Ismail Fawaz et al. 2019, 2020)
XceptionTime	CNN	Multivariate time series (Rahimian et al. 2020)

Notes: GB = Gradient Boosting; DL = Deep Learning; CNN = Convolutional Neural Network.

GB models operate on the wide-format tabular feature matrix produced by multi-tower fusion, augmented with lag features ($k \in \{1, \dots, 24\}$ steps) and rolling statistics (mean, std, min, max) over windows $\{2, 6, 12, 24\}$ steps (Supplementary Section S1.5). Both use 1,000 estimators, learning rate 0.05, column and row subsampling at 0.9 and 0.8, respectively, and early stopping on validation AUC. LightGBM uses leaf-wise growth (31 leaves) with `is_unbalance=True`, while XGBoost uses level-wise growth (max depth 6) with a computed `scale_pos_weight` (Supplementary Table S4). DL and CNN models consume fixed-length sliding-window sequences of $L \in \{12, 24\}$ timesteps (3–6 hr of 30-min observations; Supplementary Section S5.1), trained with AdamW ($lr = 10^{-3}$, weight decay 10^{-4}), batch size 64, up to 100 epochs, early stopping patience 15, and weighted binary cross-entropy to handle class imbalance (Supplementary Table S7). All six architectures use a purged venetian-blind train–test split followed by stratified 3-fold cross-validation, with final performance reported on the test set (Supplementary Section S2.1).

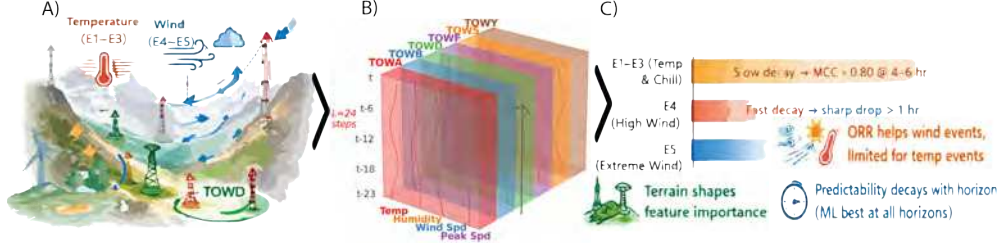


Figure 2 Multi-tower fusion framework. (A) The meteorological tower deployed for continued measurements. (B) Observations are pivoted into a fused tensor and used to predict event target $\hat{y}_{t+h}$ for horizons $h \in \{30 \text{ min}, \dots, 6 \text{ hr}\}$. (C) Event detection and feature-importance outputs. Additional details are seen in Supplementary Figure S3.

3.2 Feature Engineering

To predict event onset, we engineer features that capture temporal dynamics and atmospheric conditions across multiple scales and locations. Our approach comprises three main components: temporal feature extraction, spatial aggregation, and event-specific resampling. First, for each variable x_t measured at time t , we extract features at multiple temporal scales to obtain temporal feature extraction. Lagged values at offsets $k \in \{1, 2, \dots, 24\}$ time steps encode the recent atmospheric trajectory over intervals from 15 min to 6 hrs. Rolling-window statistics (mean, minimum, maximum) over windows $w \in \{2, 6, 12, 24\}$ time steps capture short-term trends, dips, and spikes. First-order temporal differences indicate the rate of change, which is particularly important for rapid-onset events. Distance to threshold features $d_t^{(\theta)} = |x_t - \theta|$ provide direct information on nearness to event onset thresholds. Second, cross-tower aggregations (mean, standard deviation across towers) capture network-wide conditions relevant to event onset to achieve spatial aggregation. Combined with the temporal features above, the full feature space comprises approximately 4,500 engineered features when site variables are included. Implementation details are provided in Supplementary Materials Section S1.5. Third, to align the temporal resolution of the training data with the characteristic timescale of each event, we apply event-specific resampling prior to model training. Events with short minimum durations (E1, E4, E5: 0.5–1.0 hr) are resampled to 30-min intervals, while the longer-duration low temperature event (E3: 2.0 hr) resolution is retained. Numeric features are aggregated using the mean values, and event labels are aggregated using the maximum. If any constituent timestep has an active event, the resampled interval is labeled positive.

3.3 Gradient Variables and Ablation Study Design

To quantify the contribution of physics-based gradient variables, we conduct an ablation study comparing the full feature set with a tower-only configuration that excludes the human-extracted site features (~ 300 features). Both configurations use identical model hyperparameters, temporal train-test splits, class weighting strategies, and evaluation protocols to attribute the physics-informed variables. The seven selected site

1. Data Collection & Filtering

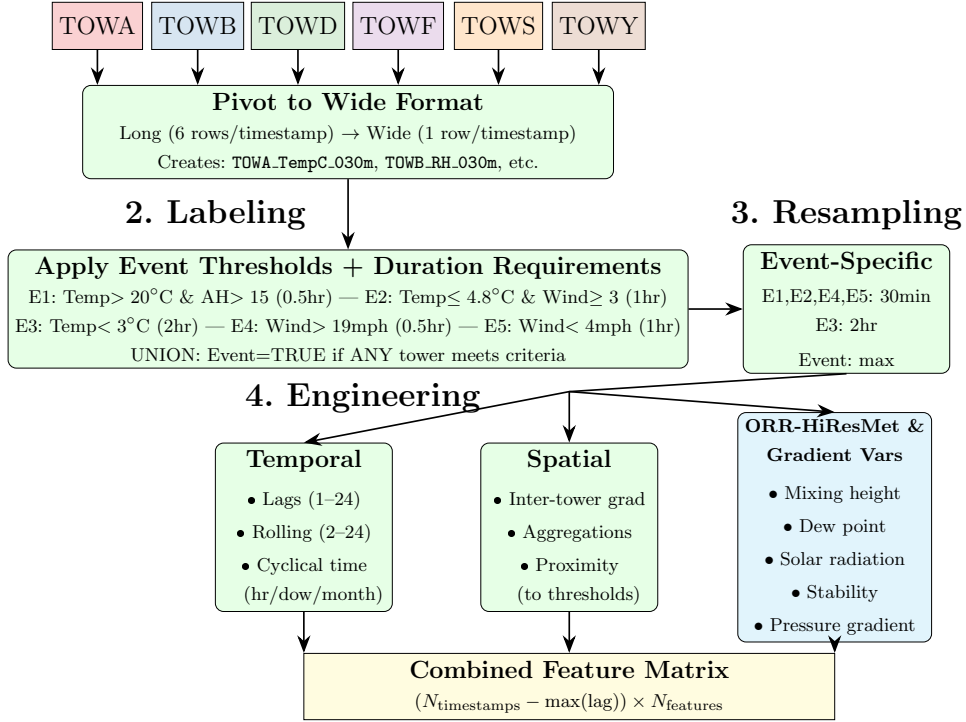


Figure 3 The data processing pipeline including raw observations, pivoting, event labeling, resampling, and feature engineering as key steps.

gradient variables and their meteorological significance are described in Supplementary Materials Section S1.4.

3.4 Class Imbalance in Extreme Event Prediction

While these statistical considerations address the learning challenge, the underlying predictability of extreme events is also strongly governed by physical geography. Complex terrain modulates local meteorology through processes such as cold-air drainage and nocturnal katabatic flows in valleys, flow channeling that concentrates or accelerates winds, thermally driven circulations induced by differential slope heating, and mechanical effects including flow separation and turbulence around ridges (Stull 2011; Wallace and Hobbs 2006). In this context, the ORR exhibits all of these mechanisms, motivating a multi-tower modeling framework designed to explicitly capture localized terrain–atmosphere interactions. Complete definitions of evaluation metrics, including Balanced Accuracy, Cohen’s Kappa, and Average Precision, as well as the

class-weighting formulation, are provided in Supplementary Materials Section S2.4. Complex terrain plays a critical role in shaping local meteorology through cold-air drainage and nocturnal katabatic flows in valleys, flow channeling that concentrates or accelerates winds, thermally driven circulations induced by differential slope heating, and mechanical effects such as flow separation and turbulence around ridges (Stull 2011; Wallace and Hobbs 2006).

3.5 Implementation Details

This section describes the modeling simulation setup and evaluation protocol used for model selection. The simulations employ LightGBM and XGBoost (CPU-based), LSTM-FCN and TST, and ResNet+ and XceptionTime (GPU-based). For the DL and CNN models, two sequence lengths ($L \in 12, 24$) are additionally evaluated. In total, we conduct 300 modeling simulations. An 80%/20% chronological train-test split prevents data leakage by partitioning 2022 into a training set and a held-out test set. Seasonal variability is preserved within the training partition through the Venetian blind cross-validation scheme described below, which interleaves 14-day blocks across folds so that winter, spring, summer, and early-autumn conditions are represented in every fold. The October-December hold-out set further provides a strictly out-of-period evaluation covering late-autumn and early-winter conditions. Training partition undergoes 3-fold stratified time-block cross-validation (Kohavi 1995), where data are divided into 14-day blocks and distributed across folds to balance event frequencies while preserving temporal order. Prediction labels are generated by shifting the target event indicator forward by the detection horizon (e.g., features at time t predict event occurrence at $t + h$, where h denotes the horizon-specific offset). To address class imbalance, class weights are applied inversely proportional to event frequency. The ORR-HiResMet paired with gradient variables ablation study compares full-feature and tower-only conditions across all approaches. A schematic overview of the complete processing pipeline is shown in Figure 3. Systematic hyperparameter using bayesian optimization methods, including TPE implemented in Optuna, a Python hyperparameter optimization library (Akiba et al. 2019), efficiently explore high-dimensional hyperparameter spaces by modeling the objective function and balancing exploration versus exploitation (Bergstra and Bengio 2012). Recent work demonstrates the importance of hyperparameter optimization for event detection models, with performance gains over default configurations (Rasp et al. 2020). The hyperparameters reported in Supplementary Materials Table S4 represent informed defaults derived from preliminary Optuna simulations. This providing a baseline for operational deployment.

4 Results

Evaluated models span gradient boosting (ML: XGBoost, LightGBM), deep learning (DL: LSTM-FCN, TST), and convolutional neural networks (CNN: ResNet+, XceptionTime). These methods are applied within a unified simulation pipeline to ensure fair comparison across event types and forecasting horizons. Key findings are summarized across the three modeling approaches and evaluated over multiple forecasting

horizons and ORR variable conditions. We first report overall predictive performance and then analyze event-specific behavior, detection sensitivity, SHAP-based feature importance, and ablation study. 300 simulations across 6 models $\times$ 5 events $\times$ multiple horizons $\times$ and ablation indicator. Time series cross-validation and training configurations are detailed in Supplementary Materials Section S2.1. Supplementary Materials Table S5 summarizes the best-performing model per event type across all approaches at the shortest detection horizons. Gradient boosting methods consistently achieve the highest performance, followed by CNN and DL models. Results indicate strong predictive capability for temperature-moisture (E1) and wind chill (E2) events (MCC > 0.96), excellent performance for low temperature (E3) prediction at 2-hr horizons (MCC 0.937), followed by high wind events (E4, MCC 0.764), and low wind conditions (E5, MCC 0.864).

4.1 Event-Specific Performance Analysis

E1 models achieve the strongest overall performance, with XGBoost test F1 0.973 / MCC 0.964 at 30 min, dominated by absolute humidity rolling-window maxima from TOWD and TOWA at 15 m. E2 is the second-strongest event, with LightGBM at AUC 0.999 / F1 0.971 / MCC 0.963. Performance degrades with horizon (F1 0.971 to 0.955 from 30 min to 1.5 hr.), and temperature rolling-window minima from TOWA/TOWY dominate. E3 ML models at the 2-hr horizon also perform excellently (F1 0.949 / MCC 0.937). The custom distance-to-threshold feature is the single most important SHAP one (value 3.47), and ML retains strong skill at 4 hr. (F1 0.877) and 6 hr. (F1 0.825), supporting multi-hr early warning. Surface pressure gradient direction (gradient variables) enters the top-10 features at the 6-hr horizon, indicating that physics-based gradient variables becomes increasingly valuable at extended detection times. E4 modeling proves the most challenging event among all approaches. LightGBM achieves best 30-min performance (F1 0.785 / MCC 0.764), with CNN (MCC 0.573) and DL (MCC 0.509) underperforming significantly, reflecting the difficulty in capturing turbulent wind dynamics from sequence data. Performance drops to F1 0.652 at 1.5 hr. Peak wind speed rolling maxima from TOWS (ridge-slope, 334 m) dominate SHAP importance, and ORR surface pressure gradient magnitude joins the top-10 at 1.5 hr. E5 obtains strong ML performance (AUC 0.988 / F1 0.915 / MCC 0.864) at 30 min, benefiting from the higher event rate and gradual onset relative to E4, with CNN (MCC 0.719) and DL (MCC 0.708) yielding moderate results. The tower-mean peak wind speed feature is ranked first, and solar radiation derivative (gradient variables) appears among top features at 1.5 hr. This links atmospheric stability transitions to stagnation onset. These event-specific patterns are consistent with the broader temporal behavior observed across all event types, where F1 and MCC scores degrade across detection horizons for all modeling approaches. E2, E3 exhibit the most gradual degradation, with ML models maintaining MCC > 0.80 even at 4-6 hr horizons, whereas E4, E5 show a steeper decline beyond 1 hr. The performance gap between ML and DL/CNN approaches also widens at longer horizons.

Table 3 Performance among all event detection horizons (best ML model per event-horizon using enhanced ORR Variables). Arrows (↓) indicate degradation from the shortest horizon.

Event	Horizon	Best Model	Test AUC	Test F1	Test MCC	Test Precision	Test Recall	Trend
E1 TempMoisture	30min	XGBoost	0.999	0.971	0.962	0.966	0.977	
E1 TempMoisture	1hr	LightGBM	0.998	0.958 ↓	0.944 ↓	0.941	0.976	↓
E1 TempMoisture	1.5hr	XGBoost	0.998	0.952 ↓	0.936 ↓	0.930	0.976	↓
E2 WindChill	30min	LightGBM	0.999	0.969	0.961	0.947	0.992	
E2 WindChill	1hr	LightGBM	0.999	0.966 ↓	0.956 ↓	0.943	0.990	↓
E2 WindChill	1.5hr	XGBoost	0.998	0.955 ↓	0.942 ↓	0.928	0.983	↓
E3 LowTemp	2hr	XGBoost	0.998	0.945	0.933	0.920	0.972	
E3 LowTemp	4hr	LightGBM	0.989	0.877 ↓	0.849 ↓	0.824	0.938	↓
E3 LowTemp	6hr	XGBoost	0.980	0.825 ↓	0.782 ↓	0.761	0.901	↓
E4 HighWind	30min	LightGBM	0.984	0.785	0.764	0.697	0.898	
E4 HighWind	1hr	LightGBM	0.956	0.694 ↓	0.659 ↓	0.624	0.781	↓
E4 HighWind	1.5hr	LightGBM	0.944	0.652 ↓	0.614 ↓	0.572	0.759	↓
E5 LowWind	30min	LightGBM	0.988	0.915	0.864	0.854	0.986	
E5 LowWind	1hr	XGBoost	0.963	0.845 ↓	0.747 ↓	0.748	0.971	↓
E5 LowWind	1.5hr	XGBoost	0.937	0.798 ↓	0.668 ↓	0.685	0.956	↓

Notes: AUC = Area Under the Curve; F1 = F1 score (harmonic mean of precision and recall); MCC = Matthews Correlation Coefficient; Arrows (↓) indicate degradation from the shortest horizon.

4.2 Event Detection Analysis & Physics Linkage

Horizon sensitivity between temperature and wind driven events reflects fundamental differences in atmospheric predictability arising from the governing physics of thermodynamic and dynamic processes. Temperature fields evolve relatively slowly because they are governed by the thermodynamic energy equation, where radiative and advective forcing is moderated by the thermal inertia of the land surface and the atmospheric boundary layer (Stull 2011). In the Ridge and Valley Province, the soil-vegetation system dampens rapid fluctuations, producing smooth and strongly auto-correlated temperature time series. This persistence means that the current temperature state remains informative several hr ahead, explaining why E2 (wind chill) and E3 (low temperature) preserve MCC values above 0.80 even at 4–6 hr horizons. In practice, the thermodynamic boundary layer retains substantial atmospheric “memory,” enabling longer predictability windows for temperature-driven events. In contrast, wind fields, particularly peak wind gusts, are dominated by turbulent momentum transport operating on much shorter timescales. Gust intensity depends on the interaction between the mean flow, terrain-induced perturbations such as flow separation and lee-side recirculation, and stochastic turbulent fluctuations governed by the Turbulence Kinetic Energy budget (Supplementary Materials Eq. 20) (Stull 2011; Wallace and Hobbs 2006). The characteristic eddy turnover time in the surface layer is typically on the order of tens of seconds to several min, far shorter than the prediction horizons evaluated in this study. As a result, atmospheric momentum exhibits far weaker temporal persistence than temperature, causing predictive skill to decay rapidly with increasing horizon. This behavior is reflected in E4 (high wind), where MCC decreases from 0.764 at 30 min to 0.614 at 1.5 hr, corresponding to a relative decline of approximately 20%. The complex terrain surrounding the Oak Ridge Reservation further amplifies this predictability contrast. Valley regions promote cold-air pooling and persistent temperature inversions that evolve gradually over time, enhancing predictability for low-temperature events such as E3. Conversely, ridge-top wind gusts are influenced by transient rotor circulations, localized turbulence, and flow separation processes that respond sensitively to small upstream perturbations (Wallace and Hobbs 2006).

This terrain-dependent turbulence likely explains the strong SHAP importance of the ridge-slope tower TOWS (334 m elevation) for E4 (high wind), while valley towers such as TOWA and TOWY contribute more strongly to E2 and E3 temperature-related events. The role of ORR Variables also reflects this separation of atmospheric timescales. Features such as surface pressure gradients, 500 mb geopotential height, mixing height, and atmospheric stability describe larger-scale physical forcing mechanisms that evolve over periods of roughly 6–12 hr. Consequently, these variables contain limited information about short-lived sub-hourly turbulent gusts, but become increasingly valuable at extended horizons where local turbulence memory has diminished. This behavior is consistent with the ORR Variables ablation results, in which the largest MCC improvement occurs for E4 at 1.5 hr (+0.062). At this horizon, large-scale pressure-gradient and stability forcing becomes the dominant remaining source of predictive information after the influence of local turbulent fluctuations has decayed.

4.3 Feature Importance Analysis

SHAP analysis (methodology in Supplementary Materials Section S4) reveals consistent patterns across event types, with top features per event summarized in Supplementary Materials Table S6. Rolling-window statistics dominate feature importance across all events, with 2-timestep rolling maxima and minima providing the strongest predictive signals. Cross-tower aggregations (tower-mean features) rank among the top features for wind events, confirming the value of multi-tower fusion. Distance-to-threshold features are highly predictive. The temperature distance-to-threshold feature has the highest single-feature SHAP value (3.47) for E3. ORR Variables features appear in top-10 rankings primarily at extended horizons (≥ 1.5 hr.), with surface pressure gradient magnitude and direction being the most informative ORR variables. Temporal derivatives contribute meaningfully to wind events, capturing rapid changes in atmospheric conditions preceding event onset.

Figure 4 summarises the aggregated feature importance for XGBoost (SHAP importance) and LightGBM (gain importance) under both the full and ablated feature sets. Temperature and humidity base variables at multiple measurement heights (e.g. TempC_015m, AbsHum_015m) accumulate the largest aggregate importance of all five events using XGBoost. LightGBM gain rankings are consistent in direction but differ in absolute scale. Detection horizon and ORR Variables ablation analyses are provided in Supplementary Materials Figure S7.

4.4 Ablation Study Results

Table 4 shows quantitative results of the gradient variables inclusion performance gains. Per-family explainability results for DL models are provided in Supplementary Materials Figure S6. The ablation study reveals event-dependent contributions from observed variables. High wind events (E4) show highest benefits from this effort, with MCC gains of +0.015 at 30 min and +0.062 at 1.5 hrs. Surface pressure gradient magnitude emerges among the top-10 SHAP features at extended horizons, indicating that physics-based pressure gradient forcing improves wind event prediction. Low wind events (E5) show smaller improvements (+0.007 MCC at 30 min,

+0.006 at 1.5 hr.), with the solar radiation derivative capturing atmospheric stability transitions relevant to stagnation onset. Temperature events (e.g., E1, E2, E3) have marginal or slightly negative impact, as local temperature observations already provide strong predictive signals for thermodynamic events. The slight negative Δ MCC at short horizons (-0.002 for E1 and E2) suggests that selected site observation features may introduce noise when the local signal is already highly informative. These results confirm that using physics-based site variables gradient features primarily benefits dynamically-driven events (i.e., wind) for event detection horizons with longer time, where local tower measurements alone are insufficient.

Table 4 Ablation Study: MCC change when physics-based features are added (best model per event-horizon). Bold Δ values $> \pm 0.02$; rows grouped by meaningful ($\Delta\% \geq 2\%$) vs. negligible ($|\Delta\%| < 2\%$) change.

Event	Horizon	AUC (with)	AUC (w/o)	MCC (with)	MCC (w/o)	Δ MCC	$\Delta\%$	Direction
<i>Meaningful improvement ($\Delta\% \geq +2\%$) ORR Variables provide actionable gains</i>								
E4 HighWind	1.5hr	0.944	0.936	0.614	0.552	+0.062	+11.2%	↑
E4 HighWind	30min	0.984	0.983	0.764	0.749	+0.015	+2.0%	↑
<i>Negligible change ($\Delta\% < 2\%$) within estimation noise, no consistent directional benefit</i>								
E5 LowWind	1.5hr	0.937	0.936	0.668	0.662	+0.006	+0.9%	↑
E5 LowWind	30min	0.988	0.988	0.864	0.857	+0.007	+0.8%	↑
E3 LowTemp	2hr	0.998	0.998	0.933	0.937	-0.004	-0.4%	↓
E2 WindChill	1.5hr	0.998	0.998	0.942	0.939	+0.003	+0.3%	↑
E1 TempMoisture	1.5hr	0.998	0.998	0.936	0.943	-0.007	-0.7%	↓
E2 WindChill	30min	0.999	0.999	0.961	0.963	-0.002	-0.2%	↓
E1 TempMoisture	30min	0.999	0.999	0.962	0.964	-0.002	-0.2%	↓

Notes: AUC = Area Under the Curve; MCC = Matthews Correlation Coefficient; with = with ORR Variables; w/o = without ORR Variables; ORR = Oak Ridge Reservation; Δ = absolute MCC change; E1 = Temperature-Moisture; E2 = Wind Chill; E3 = Low Temperature; E4 = High Wind; E5 = Low Wind (Stagnation). ↑/↑ = improvement; ↓ = degradation.

4.5 Cross-Tower Generalization

The multi-tower fusion approach enables the model to leverage spatial correlations within the tower network. Performance varies depending on tower characteristics. Ridge-top towers (TOWF at 354m, TOWS at 334m) contribute the most informative wind features, while valley towers (TOWA, TOWB) dominate temperature-related predictions (tower characteristics in Supplementary Materials Figures S2). The 60 m instrumentation height at TOWD provides unique vertical profile information, with peak wind speed features at multiple heights ranking highly for E4. Network-wide features, specifically tower-mean aggregations, consistently appear in top-10 SHAP rankings, validating that the multi-tower fusion strategy is more useful for interpreting extreme events in complex terrain over relying on data from single-towers.

4.6 Computational Efficiency

The computational efficiency of gradient boosting, combined with its superior performance on this tabular multi-tower dataset, makes ML the recommended approach for applications in extreme weather event modeling related to site operations. DL and CNN approaches do not offer better performance than ML for any event type, consistent with recent benchmarks showing gradient boosting dominance on tabular data (Rasp et al. 2020).

5 Implications

The ML modeling approach of extreme weather events demonstrated in this work offers concrete benefits for site operation applications. Specifically, 30-min to 6-hr. detection can support proactive resource allocation and safety measures, with gradient boosting consistently achieving the strongest performance. The multi-tower fusion approach enables the use of a single model to learn spatial patterns associated with each extreme weather event. Tower-mean aggregation features are consistently the top SHAP predictors, supporting the effectiveness of the fusion strategy. This finding is further reinforced by the strong performance of the ML models, which surpasses the DL models, requiring less computational resources. Our results reveal that rapid-onset wind events decay sharply. In contrast, under low wind stagnation conditions, 30-min windows dominate at 1 hr detections and 120-min windows at 4 hr. SHAP analysis shows that tower-local features dominate at all horizons, contrary to the expectation that physics-based gradient features would govern. Findings suggest dense surface networks can partially substitute for physics-based gradient products when latency or availability is a concern.

In addition, ML modeling results can be translated into actionable alert windows. Specifically, low temperature (E3) offers reliably event detection 4-6 hrs ahead ($MCC > 0.78$) for warming warning and freeze protection. Wind chill (E2) is highly predictable at 1-1.5 hrs ($MCC > 0.94$) for stop-work in outdoor. High wind (E4) provides reliable 30-min alerts, sufficient for crane shutdowns and equipment-securing protocols. Low wind / stagnation (E5) yields 30-min alerts for ventilation adjustments at facilities that handle hazardous materials. Finally, Temperature-moisture can potentially support both 30-min and 1.5-hr staged heat-stress mitigation. These representative field applications, essential for field personnel at critical facilities.

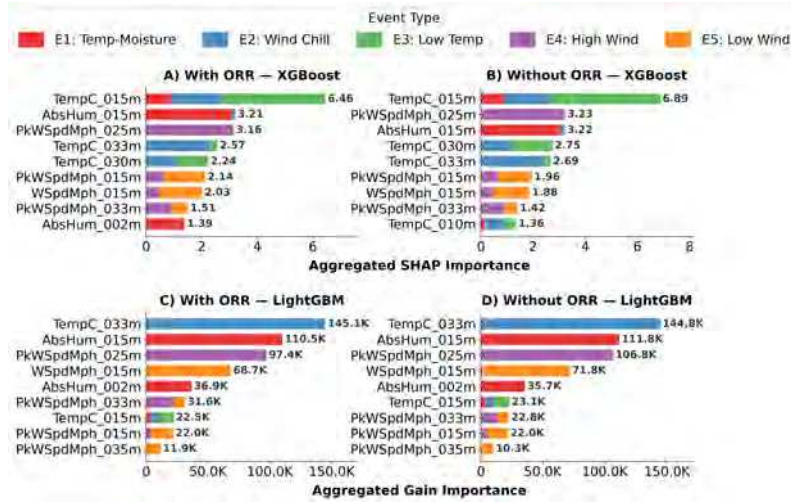


Figure 4 Feature importance for gradient boosting models. (AB) SHAP importance for XGBoost with and without ORR Variables; (CD) gain importance for LightGBM with and without digested site variables. Bar colours denote event type: E1 red, E2 blue, E3 green, E4 purple, and E5 orange.

6 Conclusion

We present an explainable multi-tower ML framework for investigating five extreme weather events discovered using EDA using long-term meteorological records at the ORR. 300 modeling runs are performed covering six architectures and three popular ML modeling paradigms. The results show that gradient boosting achieves mean MCC of 0.829, outperforming DL (0.663) and CNN (0.625), with the largest gaps for high wind (ML 0.764 vs. CNN 0.573) and low temperature events (0.937 vs. 0.670). Hypothesis 1 is confirmed, namely valley towers (TOWA, TOWY) dominate SHAP importance for thermodynamic events. For instance, TOWY temperature minimum at higher elevation is highly related to wind chill event (E2) at 22.4%. The variable indicating distance to target controls low temperature event (E3) at 41.8% dominate high wind prediction of high wind event (E4) with peak wind of 16.0%. Tower-mean aggregations rank among top predictors for all event types, suggesting the usefulness of the multi-tower fusion strategy. Hypothesis 2 is also confirmed. Thermal inertia sustains $MCC > 0.80$ at 4-6 hr for E2 and E3, while turbulent momentum memory decays rapidly in high wind conditions (i.e., E4 MCC: 0.764 at 30 min to 0.614 at 1.5 hr). Gradient boosting outperforms DL and CNN by 0.17-0.20 MCC across different time horizons. Hypothesis 3 is validated for dynamic events. Physics-based gradient features selectively improve high wind prediction and contribute negligibly to thermodynamic events, where local tower observations supply near-complete predictive signal. Therefore, gradient variables are not necessary for modeling temperature extreme events. These parameters become important for long detection horizon concerning wind extreme events. SHAP attribution links each alert to a physical mechanism located at ridge-slope regions by using peak wind for high wind events E4. Furthermore, valley temperature minima for low temperature events (e.g., E2, E3) can support a tiered alert architecture. For instance, a system can consider a watch period of 2-6 hr for thermal events. For wind events, a warning of 30 min-1 hr can be employed as a CPU-efficient method, transferable to monitoring networks used in complex terrain.

Limitations and future work

DL and CNN models may benefit from further architecture optimization and extended sequence lengths beyond $L \in \{12, 24\}$. For example, XceptionTime fails to learn low temperature event (E3) patterns in several configurations, suggesting sensitivity may be limited to the 2-hr resampling interval. Since only gradient variables are only available in 2022, evaluation of multiyear is not feasible yet. Fixed event thresholds may not optimally reflect seasonal operational risk. Equal tower weighting in data fusion could be improved by terrain-aware spatial weighting. Furthermore, the 30-min alert window for wind events, although physically justified (Section 4.2), may be insufficient for procedures that require longer lead times. Future directions may include extension to multi-year analyzed data.

Acknowledgements. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US DOE. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish

or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<https://www.energy.gov/doe-public-access-plan>).

Declarations

Funding. Research efforts were supported by the Nuclear Safety Research and Development (NSR&D) program sponsored by the National Nuclear Security Administration (NNSA) Office of Environment, Health, Safety and Security (EHSS) of the U.S. Department of Energy (DOE).

Conflict of interest. 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 availability. The datasets analyzed during the current study are available as follows. **ORR-HiResMet:** Five years meteorological data from the Oak Ridge Reservation (Steckler et al. 2025). **Model outputs:** Trained models, hyperparameter configurations, SHAP importance rankings, and prediction results are available at the project repository. **Code:** Python implementation of the multi-tower preprocessing pipeline, feature engineering, gradient boosting training, and ORR Variables ablation study scripts are accessible via the project repository. **Terrain elevation:** Data were obtained from the Open-Elevation open-source service (<https://open-elevation.com>). All code, trained models, and processed datasets are available from the project repository to enable reproducibility and extension to other meteorological monitoring networks.

Author contributions. Jose Tupayachi: Conceptualization, methodology, data curation, software development, visualization, formal analysis, writing original draft, review & editing. Ziwei Liu: Data analysis. Kevin Birdwell: Data acquisition, meteorological instrumentation, domain expertise, review & editing. Xueping Li: Validation, review & editing. Xiao-Ying Yu: Conceptualization, supervision, project administration, funding acquisition.

References

- Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *arXiv preprint arXiv:1907.10902*. <https://doi.org/10.1145/3292500.3330701>.
- Amnuaylojaroen, T. 2025, 12. Enhancing extreme weather event classification ensemble machine learning modelling: an application to mainland southeast asia region. *Theoretical and Applied Climatology* 157. <https://doi.org/10.1007/s00704-025-05925-x>.

783 Bergstra, J. and Y. Bengio. 2012. Random search for hyper-parameter optimization.
784 *The Journal of Machine Learning Research* 13(1): 281–305. [https://doi.org/10.](https://doi.org/10.5555/2188385.2188395)
785 [5555/2188385.2188395](https://doi.org/10.5555/2188385.2188395) .

786

787 Bochenek, B. and Z. Ustrnul. 2022. Machine learning in weather prediction and climate
788 analyses—applications and perspectives. *Atmosphere* 13(2). [https://doi.org/10.](https://doi.org/10.3390/atmos13020180)
789 [3390/atmos13020180](https://doi.org/10.3390/atmos13020180) .

790

791 Cai, Y., C. Huang, F. Zheng, G. Li, S. Lai, L. Zhu, and Q. Zhu. 2025. Pre-
792 dicting tropical cyclone extreme rainfall in guangxi, china: An interpretable
793 machine learning framework addressing class imbalance and feature optimization.
794 *Meteorological Applications* 32(3): e70052. <https://doi.org/10.1002/met.70052>.
795 <https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.70052> .

796

797 Chen, T. and C. Guestrin 2016. Xgboost: A scalable tree boosting system. In *Pro-*
798 *ceedings of the 22nd acm sigkdd international conference on knowledge discovery*
799 *and data mining*, pp. 785–794.

800

801 Ejayi, C.J., F.O. Eze, M.A. Hossin, E.S.E.B. Aggrey, M.B. Ejayi, A. Ifeanyi, et al.
802 2026. A robust framework for clinically interpretable hypothyroidism diagnosis with
803 Polynomial-SHAP and DeepSeqNet. *Journal of Big Data* 13(1). [https://doi.org/](https://doi.org/10.1186/s40537-025-01332-1)
804 [10.1186/s40537-025-01332-1](https://doi.org/10.1186/s40537-025-01332-1) .

805

806 Fang, W., Q. Xue, L. Shen, and V.S. Sheng. 2021. Survey on the application of deep
807 learning in extreme weather prediction. *Atmosphere* 12(6). [https://doi.org/10.](https://doi.org/10.3390/atmos12060661)
808 [3390/atmos12060661](https://doi.org/10.3390/atmos12060661) .

809

810 Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine.
811 *Annals of statistics*: 1189–1232. <https://doi.org/10.1214/aos/1013203451> .

812

813 Hancock, J.T. and T.M. Khoshgoftaar. 2020. CatBoost for big data: an inter-
814 disciplinary review. *Journal of Big Data* 7(1): 94. [https://doi.org/10.1186/](https://doi.org/10.1186/s40537-020-00369-8)
815 [s40537-020-00369-8](https://doi.org/10.1186/s40537-020-00369-8) .

816

817 Hewage, P., A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and
818 Y. Liu. 2020. Temporal convolutional neural (tcn) network for an effective weather
819 forecasting using time-series data from the local weather station: P. hewage et al.
820 *Soft Computing* 24(21): 16453–16482. <https://doi.org/10.1007/s00500-020-04954-0>
821 .

822

823 Hewage, P., M. Trovati, E. Pereira, and A. Behera. 2021, February. Deep learning-
824 based effective fine-grained weather forecasting model. *Pattern Anal. Appl.* 24(1):
825 343–366. <https://doi.org/10.1007/s10044-020-00898-1> .

826

827 Ismail Fawaz, H., G. Forestier, J. Weber, L. Idoumghar, and P.A. Muller. 2019.
828 Deep learning for time series classification: a review. *Data Mining and Knowledge*
Discovery 33(4): 917–963. <https://doi.org/10.1007/s10618-019-00619-1> .

Ismail Fawaz, H., B. Lucas, G. Forestier, C. Pelletier, D.F. Schmidt, J. Weber, G.I. Webb, L. Idoumghar, P.A. Muller, and F. Petitjean. 2020, Sept. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery* 34(6): 1936–1962. <https://doi.org/10.1007/s10618-020-00710-y> .

Karim, F., S. Majumdar, H. Darabi, and S. Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116: 237–245. <https://doi.org/10.1016/j.neunet.2019.04.014> .

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.Y. Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30. <https://doi.org/10.5555/3294996.3295074> .

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. pp. 1137–1143. <https://doi.org/10.5555/1643031.1643047> .

Lam, R., A. Sanchez-Gonzalez, M. Willson, P. Wirsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, et al. 2023. Learning skillful medium-range global weather forecasting. *Science* 382(6677): 1416–1421. <https://doi.org/10.1126/science.adi2336> .

Liu, Z., M. Steckler, K. Birdwell, and X.Y. Yu. 2026. Exploratory data analysis of long-term oak ridge reserve meteorological data for extreme weather event discovery. *Scientific Reports*. <https://doi.org/10.21203/rs.3.rs-8855721/v1> .

Lundberg, S.M. and S.I. Lee 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, Red Hook, NY, USA, pp. 4768–4777. Curran Associates Inc.

Natekin, A. and A. Knoll. 2013. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics* 7: 21. <https://doi.org/10.3389/fnbot.2013.00021> .

Rahimian, E., S. Zabihi, S.F. Atashzar, A. Asif, and A. Mohammadi 2020. Xceptiontime: Independent time-window xceptiontime architecture for hand gesture classification. pp. 1304–1308.

Rasp, S., P.D. Dueben, S. Scher, J.A. Weyn, S. Mouatadid, and N. Thuerey. 2020, Nov. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems* 12(11). <https://doi.org/10.1029/2020ms002203> .

Shi, X., Z. Gao, H. Wang, D.Y. Yeung, W.K. Wong, and W.c. Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214*. <https://doi.org/10.48550/arXiv.1506.04214> .

875 Steckler, M.R., K.R. Birdwell, H. Xu, and X.Y. YU. 2025. Descriptor: High temporal
876 resolution meteorological data at oak ridge reservation (orr-hiresmet). *IEEE Data*
877 *Descriptions* 2: 118–124. <https://doi.org/10.1109/IEEEDATA.2025.3571018> .
878
879 Stull, R.B. 2011. *Meteorology for Scientists and Engineers* (3rd ed.). University of
880 British Columbia.
881
882 V, R.C., A.C. Johnvictor, and P.S. N. 2025. Comparative analysis of machine learning
883 approaches for heatwave event prediction in india. *Scientific Reports* 15: 22431.
884 <https://doi.org/10.1038/s41598-025-04634-9> .
885
886 Voyant, C., G. Notton, C. Darras, A. Fouilloy, S. Kalogirou, and M.L. Nivet. 2017.
887 Machine learning methods for solar radiation forecasting: A review. *Renewable*
888 *Energy* 105: 569–582. <https://doi.org/10.1016/j.renene.2016.12.095> .
889
890 Wallace, J.M. and P.V. Hobbs. 2006. *Atmospheric science: an introductory survey*
891 (2nd ed.). Academic Press.
892
893 Weyn, J.A., D.R. Durran, and R. Caruana. 2019. Can machines learn to pre-
894 dict weather? using deep learning to predict gridded 500-hpa geopotential height
895 from historical weather data. *Journal of Advances in Modeling Earth Sys-*
896 *tems* 11(8): 2680–2693. <https://doi.org/https://doi.org/10.1029/2019MS001705>.
897 <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001705> .
898
899 Wu, X., J. Chang, C. Long, and H. Ghorbani. 2026. Machine learning insight in quality
900 control and maintenance management of drilling mud pumps. *Journal of Big Data*.
901 <https://doi.org/10.1186/s40537-026-01475-9> .
902
903 Xiong, G., J. Zhang, X. Fu, J. Chen, and A.W. Mohamed. 2024. Seasonal short-term
904 photovoltaic power prediction based on gsk-bigrx-xgboost considering correlation
905 of meteorological factors. *Journal of Big Data* 11(1): 164. <https://doi.org/10.1186/s40537-024-01037-x> .
906
907 Zerveas, G., S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff 2020. A
908 transformer-based framework for multivariate time series representation learning.
909
910 Zhou, H., H. Ren, P. Royer, H. Hou, and X.Y. Yu. 2022. Big data analytics for
911 long-term meteorological observations at hanford site. *Atmosphere* 13(1). <https://doi.org/10.3390/atmos13010136> .
912
913
914
915
916
917
918
919
920

Supplementary Files

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

- [supplements.pdf](#)