Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 2, April, 2026

Rain and Snow Phase Classification with Wet Bulb Profile Encoding and Elevation Adaptive CatBoost

Rigos Balaskas[1,*], Giannis Koris[1], Dimos Chatzis[1]

[1] *School of Information and Communication Technologies, University of Piraeus, Piraeus, 18534, Greece*

Received 12 February 2026; Revised 13 June 2026; Accepted 19 June 2026; Published 26 June 2026

Abstract

Reliable precipitation phase classification requires information about the atmospheric column and underlying terrain. This paper proposes an elevation-adaptive CatBoost framework that combines compact wet-bulb profile encoding, terrain-relative descriptors, and calibrated class probabilities. The encoder represents warm-layer depth, integrated positive thermal exposure, near-surface cooling, zero-crossing height, and profile availability through terrain-relative thermal descriptors. Continuous elevation interactions and density-controlled training weights address uneven sampling support between lowlands and uplands. The evaluation examines classification performance, terrain transfer, and probability reliability under different input conditions. There are 24,000 precipitation cases and five elevation bands in the benchmark. In the performance comparison, the proposed configuration has a macro F1 score of 0.910 and a balanced accuracy of 0.914, while surface-only CatBoost achieved 0.867 and 0.871, respectively. After removing profile encoding, the macro F1 score is 0.880; after removing elevation adaptation, it is 0.891. The Brier score and expected calibration error decrease from 0.168 and 0.061 to 0.122 and 0.024, respectively. These contrasts describe the reported aggregate comparison rather than a test of statistical significance. The reliability-bin definition and the relationship between pooled and fold-averaged diagnostics require reconciliation before the probability scores support operational calibration claims. The contribution is a compact representation and evaluation framework whose transferability remains conditional on independent geographic testing, documented observation labels, and consistent prediction records.

Keywords: Precipitation Phase, Wet Bulb Profile Encoding, Elevation Adaptation, CatBoost Classification, Probability Calibration

1. Introduction

A comparison of air-temperature and wet-bulb-temperature partitioning schemes in the Colorado River Basin shows that phase partitioning affects snowpack simulation and projected streamflow sensitivity [1]. Recent work on road-surface snow detection connects timely snow-cover identification with winter road monitoring and maintenance [2]. Dynamical downscaling over High Mountain Asia reveals spatially varying orographic precipitation gradients that depend on terrain orientation and elevation [3]. A dynamic wet-bulb-temperature threshold method identifies relative humidity and elevation as influential controls on precipitation phase partitioning [4]. Operational phase classification requires atmospheric information that represents the pathway traversed by falling precipitation. The classifier should distinguish rain, snow, and mixed precipitation, accommodate incomplete profiles, and quantify uncertainty near phase boundaries. Vertical profile information is particularly relevant when a warm layer aloft and a cold surface layer exert different influences on descending hydrometeors. Compact descriptors of this structure connect physical reasoning with statistical classification. The proposed framework combines terrain-aligned thermal features, continuous elevation context, and calibrated tree-ensemble predictions. The framework supports station-scale diagnostics with documented

* Corresponding author: bala.rig@cs.unipi.gr

http://dx.doi.org/10.68406/jaci.2026.vol4iss2no13:180-195

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

input timing and terrain alignment. The evaluation design links profile availability, terrain support and phase-dependent errors to the architecture of the classifier.

Atmospheric profiles and surface measurements describe complementary parts of the precipitation pathway. Near-surface wet-bulb temperature characterizes evaporative cooling potential, while a warm layer aloft indicates an opportunity for melting before particles reach the ground. Sounding-based phase classification research represents melting and refreezing conditions using thermal-profile areas relative to the freezing isotherm, with particular relevance to inversion profiles [5]. Terrain further complicates profile interpretation because pressure levels may intersect the local surface at different elevations. Physically consistent mesoscale model evaluation in complex terrain accounts for differences in model terrain, measurement height, and time-varying valley stratification [6]. A classifier trained on an imbalanced station network can associate elevation with sampling practices and regional climate. A recent review identifies spatial autocorrelation, imbalanced sampling, and model generalization as central challenges in environmental geospatial learning [7]. These dependencies are particularly relevant when a model is transferred to an unmonitored catchment. Elevation therefore enters the framework as a continuous contextual feature accompanied by information about atmospheric conditions and sampling support. Station identifiers are reserved for grouping evaluation records, while transferable atmospheric and terrain descriptors form the predictor set. Missing profile layers are represented explicitly, and the support available for each thermal descriptor is retained.

Tree ensembles can learn interactions among thermal and terrain descriptors. A drought-monitoring comparison evaluates CatBoost, XGBoost, LightGBM, random forests, and extra trees within an ensemble learning framework [8]. A recent precipitation-forecasting study combines CatBoost regression with a separate LightGBM extreme-event classifier, illustrating the use of tree ensembles with meteorological predictors [9]. The sampling and evaluation design must control geographic leakage, account for uneven network support, and assess physically plausible extrapolation. A recent evaluation framework distinguishes discrimination, calibration, overall probability performance, and classification as separate dimensions of predictive assessment [10]. Accurate classification of common rain events can coexist with poor probability calibration for mixed-phase cases near the transition. The architecture combines profile summaries, elevation-relative inputs, bounded density weighting, and a separate calibration stage. The first hypothesis is that physically organized profile information provides a more transferable signal than an unrestricted set of pressure-level temperatures. A second hypothesis is that continuous terrain adaptation will reduce the dependence on the elevation distribution of the training stations. Both hypotheses require identical evaluation folds, predictors available at the forecast issue time, and thresholds selected exclusively from development data. The evaluation compares classwise performance, elevation-dependent transfer, and sensitivity to missing atmospheric inputs.

Assessing a precipitation phase classifier across contrasting elevations requires a close connection between its atmospheric representation, the composition of the observing network, and the conditions under which predictions will be used. Work on environmental benchmark design emphasizes transparent definitions of the prediction target and the evaluation population. Comparisons against crowdsourced rain and snow observations reveal substantial shortcomings in widely used precipitation-phase products and emphasize the value of independent phase labels [11]. Spatial block cross-validation research examines how the separation of training and test locations and the choice of block geometry affect estimated predictive accuracy [12]. The contribution developed here is therefore a complete experimental framework, including a terrain aligned profile encoder, a bounded elevation adaptation mechanism, and an evaluation plan that separates spatial transfer from within network discrimination. The analysis examines how thermal descriptors and terrain context affect discrimination and probability reliability under distinct evaluation conditions. Section 2 reviews threshold, profile based, and ensemble approaches, with attention to the assumptions that motivate this design. Section 3 defines the profile encoding, elevation adaptation, and probabilistic training pipeline, including the mathematical objects needed to reproduce the proposed method. Section 4 describes the benchmark population and the analyses of discrimination, transfer, calibration, and perturbation sensitivity. Section 5 states the methodological contribution, its present limitations, and the research directions relevant to operational deployment. The numbered bibliography identifies the published studies supporting the corresponding methodological and contextual claims. The methodological contribution lies in the integration and controlled evaluation of these components; priority claims for the individual algorithms remain outside the scope of this study.

2. Related Work

2.1. Surface Thresholds and Atmospheric Profiles

Surface threshold methods remain attractive because their inputs are widely available and their decision rules are transparent. Hu et al. [13] compared six empirical precipitation-phase schemes across meteorological, topographic, and geographic conditions. A Northwest United States hydrological study finds that adding relative humidity to temperature-based rain-snow partitioning reduces bias in several snow and streamflow metrics [14]. These methods can be expressed as deterministic boundaries or as smooth probabilities, with the latter better suited to ambiguous events. Their simplicity makes their assumptions visible: a surface diagnostic approximates the cumulative effect of an atmospheric pathway

using measurements taken at its lower boundary. A regional fit can absorb that pathway into its coefficients, but the coefficients may change when circulation, terrain, or observing practice changes. An engineering comparison should therefore include a surface method tuned under the same development protocol as the proposed classifier. It should also report how the transition range was estimated and whether the tuning sample contains the stations later used for testing.

Profile based approaches introduce information about warm layers, cold layers, and changes in vertical temperature gradient. Jennings et al. [15] compared machine-learning classifiers with phase-partitioning benchmarks and identified persistent limits to discrimination using near-surface meteorology. Radar-based research over the Tibetan Plateau uses differences in vertical reflectivity and related precipitation structure to improve rain-type classification [16]. Compact descriptors of vertical structure are useful because a raw profile may contain redundant levels, irregular spacing, and sections below local ground. A thermal integral provides a compact proxy for the thermal conditions encountered by falling particles. Fall speed, particle size, precipitation intensity, and exchange with the surrounding air alter the relationship between an integral and the resulting phase. The proposed encoder consequently preserves complementary thermal descriptors and their validity masks, allowing exposure magnitude, layer geometry, and information availability to enter the model separately. This choice also permits an ablation that compares physically organized summaries with raw terrain aligned temperatures while keeping model capacity approximately comparable.

2.2. Terrain Transfer and Ensemble Learning

Elevation enters precipitation classification through both physical and statistical pathways. An analysis across climatic, physiographic, and vegetation domains in Mexico finds that the predictive value of elevation for temperature depends on environmental context [17]. A station-based comparison in the Chinese Tianshan Mountains finds that precipitation-type separation accuracy varies among gridded datasets and responds to humidity and temperature biases [18]. A model can inadvertently combine these effects into a site-specific shortcut, especially if the same locations appear in training and testing. Continuous elevation, elevation mismatch, and height above ground are therefore conceptually different predictors. The first locates the station within a regional terrain distribution, the second describes a sampling discrepancy, and the third aligns the atmospheric column to the actual surface. Their distinction is central to transfer experiments. A held-out elevation band probes extrapolation in terrain context, whereas held out stations within familiar bands probe interpolation across locations. Each test addresses a distinct transfer condition and requires reporting the support available in the remaining training sample.

Ensemble methods can model nonlinear interactions without requiring a manually specified global response surface. A seasonal precipitation correction study compares LightGBM with linear regression, XGBoost, and CatBoost using model output and station observations [19]. An atmospheric prediction study compares CatBoost with other tree ensembles using satellite, meteorological reanalysis, and ground-observation predictors, demonstrating its applicability to heterogeneous atmospheric inputs [20]. These applications motivate estimator selection; precipitation-phase accuracy requires direct evaluation on phase labels. Its physical relevance depends on controlled comparisons with alternative learners. A stronger comparison holds the atmospheric descriptors constant and changes only the estimator, then holds the estimator constant and changes only the descriptors. For a broader model comparison, random forests and another gradient boosting implementation would provide complementary controls. In the present manuscript, the evaluation focuses on the latter comparison because the proposed contribution concerns representation and terrain adaptation. The profile encoder is deterministic, so it can be inspected independently of the learner. This separation makes it possible to diagnose whether a gain arises from atmospheric information, the elevation distribution of the sample, or additional tuning freedom.

2.3. Reliability and Evidence Gaps

Classification accuracy alone provides an incomplete description of a precipitation phase system. Research on multivariate probabilistic verification recommends complementary proper scoring rules to characterize different application-relevant aspects of predictive performance [21]. Recent calibration software supports reliability diagrams, class-conditional assessment, alternative binning schemes, and statistical tests for probabilistic classifiers [22]. Mixed precipitation creates an additional challenge because its label can represent a short interval containing several hydrometeor types and changing phase composition. A recent survey reviews sample selection, label correction, robust loss functions, and related approaches for learning from noisy class labels [23]. These issues affect the interpretation of every apparent improvement. An estimator with higher macro F1 may still produce probabilities unsuitable for a warning threshold, and a well calibrated model may remain weak for rare elevation regimes. The proposed evaluation therefore pairs class balanced measures with probability scores, reliability diagrams, and explicit support counts.

The unresolved methodological need is a compact, auditable connection between profile information, terrain context, and evaluation under geographic transfer. Raw profiles retain detail but can carry irrelevant pressure level geometry; threshold models are interpretable but may omit a warm layer aloft; unrestricted site features can exploit local associations that disappear outside the network. The present design addresses these limitations through terrain relative

interpolation, complementary thermal summaries, restrained weighting of underrepresented elevation ranges, and separate calibration. The central research gap concerns the joint evaluation of physical representation, terrain support, and microphysical variability under geographic transfer. Station, storm, and season separation provide complementary conditions for examining whether a feature representation retains its value beyond familiar environments. This requirement shapes the methods and analyses below, including the joint assessment of discrimination, probability quality, and uncertainty.

3. Wet Bulb Profile Encoding and Elevation Adaptive Learning

3.1. Terrain Relative Profile Representation

Each input record consists of an atmospheric column, surface meteorology, station elevation, grid elevation, and quality indicators. The prediction target contains three mutually exclusive classes: rain, snow, and mixed precipitation. Freezing rain and ice pellets fall outside this classification scope and receive separate flags when the available labels distinguish them. The profile is sampled at eight heights above the ground: 0, 100, 250, 500, 750, 1000, 1500, and 2000 meters. Before interpolation, the original levels below the site surface were removed. Linear interpolation operates between bracketing valid levels, with the supported profile top defining the upper processing limit. Wet bulb temperature enters the encoder as an atmospheric field calculated upstream from temperature, humidity, and pressure using a documented psychrometric solver. Mixed precipitation denotes the supplied observation category. Its observation interval and the treatment of transitions require explicit label metadata when the framework is transferred to another observing network.

Terrain alignment establishes a consistent geometric reference through the following equation. Here, e denotes station elevation above mean sea level, h_j is the height of sampling level j above local ground, and T_w represents wet bulb temperature. The resulting q_j values form the terrain relative profile vector. Sampling is restricted to heights supported by the source column. Heights are expressed in metres and temperatures in degrees Celsius throughout preprocessing. Each sample is accompanied by a validity mask, while the fraction of valid levels enters the feature set as a separate predictor. A completely unavailable column produces missing descriptors and an availability flag, preserving its distinction from a measured isothermal profile. Interpolation support must record the actual bracketing levels and the extent of internal gaps, since a valid level fraction alone cannot establish continuity of the column.

$$q_j = T_w(e + h_j), j = 1, \ldots, 8 \quad (1)$$

The processing sequence in Figure 1 begins with input timing and quality checks, proceeds through terrain alignment and profile encoding, and ends with probability calibration and phase assignment.

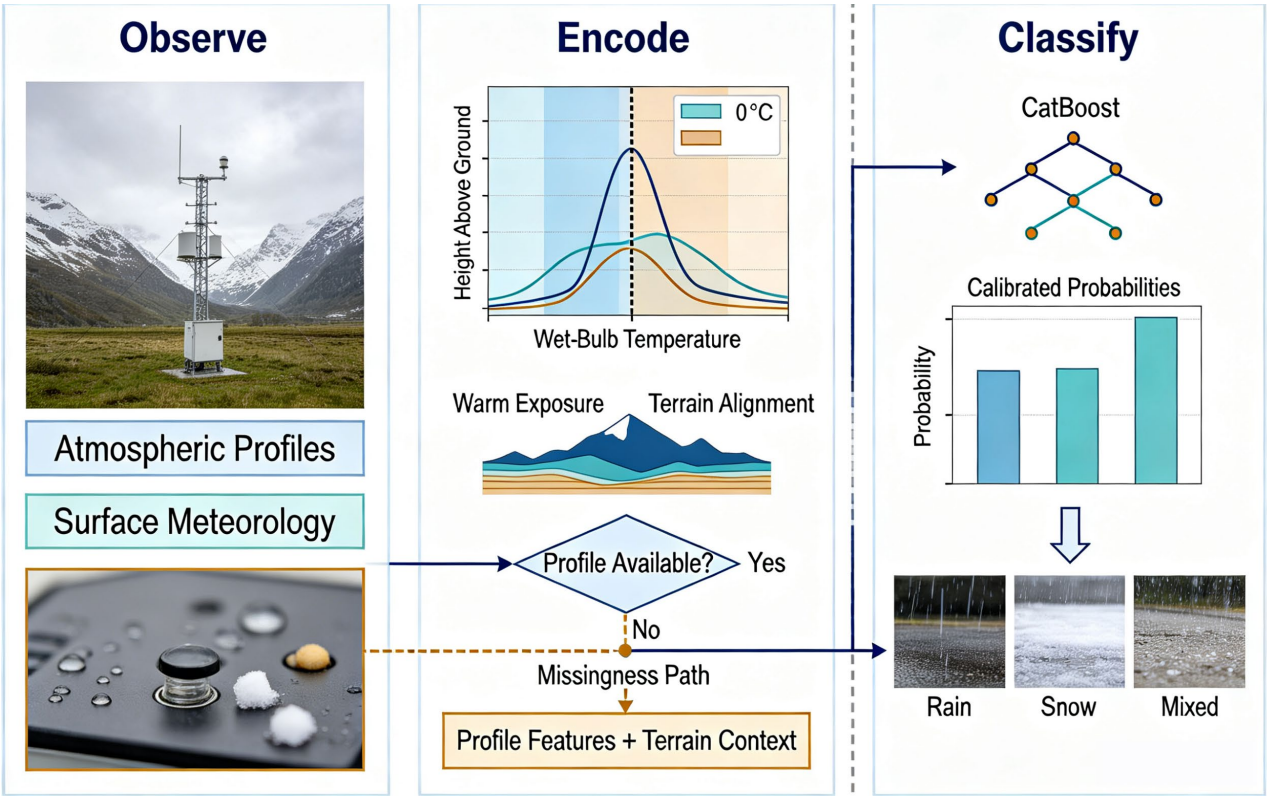


Figure 1. Proposed data preparation and prediction workflow

Its quality branch routes unavailable profiles through an explicit missingness path before feature assembly. This arrangement distinguishes unavailable column information from a genuinely near-freezing atmospheric column. A missingness path identifies the available predictors; it establishes a separately trained surface fallback only when that alternative model and its switching rule are specified. The encoder retains surface wet bulb temperature, the warmest valid level, warm layer depth, the lowest zero crossing height, near surface gradient, positive thermal exposure, and valid level fraction. When several zero crossings occur, their count is retained together with the lowest crossing and a multiple crossing flag. Warm layer depth is calculated by summing the vertical thicknesses of all positive temperature segments.

Positive thermal exposure is quantified by integrating the warm portion of the supported profile, as expressed in the equation below.

$$A_{+} = \int_{0}^{H} \max[T_w(e + h),0]dh \tag{2}$$

Adjacent valid samples are connected by linear segments, and H denotes the supported top height above local ground. The resulting descriptor has units of degree Celsius metres. Integration divides a segment at its zero crossing and accumulates its positive portion, preserving shallow warm regions within segments whose endpoints have different signs. The magnitude of negative thermal exposure is accumulated from the absolute value of the subzero portion and retained as a nonnegative cooling descriptor. Its interpretation concerns thermal conditions along the precipitation pathway; a complete refreezing energy budget would additionally require particle and exchange properties. The relationship between surface temperature and integrated warm exposure is particularly informative during inversions, when conditions aloft differ substantially from those near the ground.

3.2. Continuous Elevation Adaptation

Elevation adaptation combines continuous terrain descriptors with bounded training weights. Station elevation enters the predictor set directly, while the difference between station and grid elevation represents the geometric mismatch between the prediction location and its atmospheric source. Interaction features combine elevation with surface wet bulb temperature, positive thermal exposure, and lowest zero crossing height. These interactions are constructed after centering and scaling the numerical inputs using training fold medians and interquartile ranges. The stored transformations are subsequently applied to validation and test records. This representation allows the learner to express changes in thermal response along the terrain gradient. Five elevation bands support evaluation and training density estimation, while the prediction features retain continuous elevation values across band boundaries. Each interaction is the product of the corresponding standardized variables. A zero interquartile range requires a fixed unit scale, recorded with the fitted transformation, to preserve a defined numerical mapping.

Uneven terrain support is addressed using a density weighting equation that limits the influence of sparsely represented bands. For training record i, $b(i)$ identifies its elevation band and $n_{b(i)}$ is the number of training records in that band. The quantities N and B denote the total number of training records and the number of elevation bands, respectively. Clipping restricts the initial weights to the interval from 0.5 to 2.0. These weights are then rescaled to have a mean of one within the training fold. Terrain density weighting adjusts the geographic distribution represented by the training objective, while class prevalence retains its original contribution. Macro F1, balanced accuracy, and classwise recall provide complementary checks on phase imbalance. Any additional class weighting requires a separate ablation because it changes the optimization target. Bands without training records provide no density estimate; extrapolation into those bands retains an out-of-support flag. The clipping limits apply before mean normalization, so the final normalized weights may extend beyond those initial limits.

$$w_i = \mathrm{clip}\left(\frac{N}{Bn_{b(i)}}, 0.5, 2.0\right) \tag{3}$$

Figure 2 organizes the architecture into a profile branch, a terrain branch, a controlled feature merge, an ordered boosting learner, and a disjoint calibration module. The terrain branch supplies station elevation, grid mismatch, and interaction features. Elevation density information supports training weights and diagnostic reporting. Dashed connections distinguish labels and grouping information used during development from predictors available at inference. This separation makes the information flow auditable and helps identify leakage through station codes or label dependent transformations. The calibration module receives scores for records reserved specifically for calibration. Final testing follows model selection, transformation fitting, and calibration parameter estimation. Deployment uses the approved input fields together with the stored preprocessing parameters, fitted classifier, and calibration parameter.

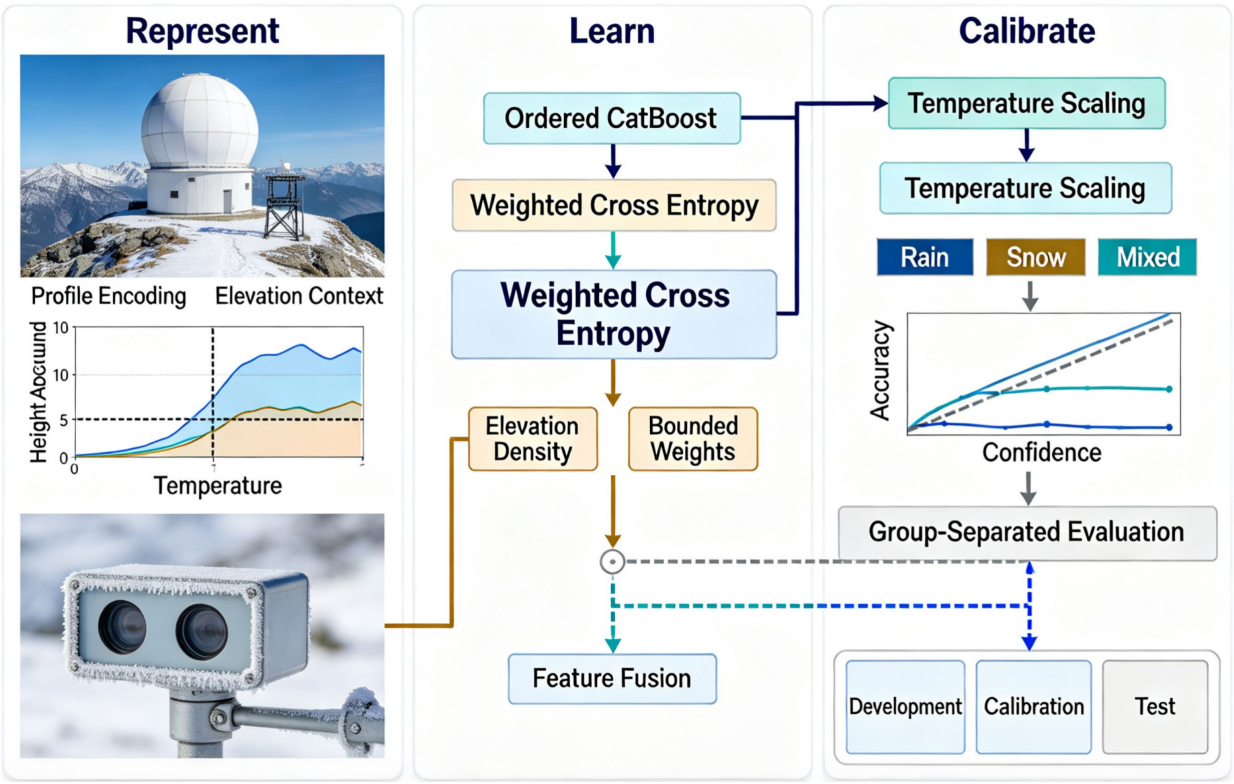


Figure 2. Architecture of the profile encoder, terrain adaptation module, ordered CatBoost learner, and disjoint probability calibration path

Elevation response curves provide an additional diagnostic under fixed atmospheric conditions. Their shape can reveal discontinuities associated with weighting or sparse training support, while interpretation remains conditional on the regional climate represented during development. The system stores the elevation range covered by training and flags predictions outside that range. High mountain stations retain their actual elevations when an extrapolation flag is issued. Missing numerical profile features remain compatible with CatBoost missing value handling, and explicit masks identify the information available for each prediction. Station identifiers, record identifiers, storm labels, and future observations are reserved for grouping, evaluation, and provenance tracking. Restricting these fields to their intended roles helps preserve the distinction between transferable atmospheric predictors and persistent associations specific to the development sample.

3.3. Training Calibration and Reproducibility

Training minimizes weighted multiclass cross entropy through the equation presented below. For record i, y_i denotes the assigned phase and p_{ik} is the probability of class k, obtained by applying a softmax transformation to the model scores. The number of classes is $K = 3$, and the indicator function selects the probability associated with the assigned class. The weights w_i are the normalized terrain density weights. This objective connects model fitting to probability estimation while allowing operating thresholds to be selected separately. The initial CatBoost configuration uses ordered boosting, depth six, a learning rate of 0.04, a maximum of 1200 trees, and early stopping after 100 validation iterations without improvement. Reproducibility requires recording the software version, processing mode, random seeds, stopping iteration, and complete parameter dictionary.

$$\mathcal{L} = -\frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \sum_{k=1}^{K} \mathbf{1}(y_i = k) \log p_{ik} \tag{4}$$

Temperature scaling provides the probability calibration stage. A single positive scalar is selected on the calibration partition by minimizing unweighted cross entropy. Scaling preserves the ordering of class scores within each record while adjusting the concentration of the probability vector. The unweighted calibration target reflects the deployment population, giving calibration a population specific role distinct from terrain balanced fitting. Changes in the deployment elevation distribution therefore require renewed reliability assessment. The primary phase decision selects the class with the largest calibrated probability. An optional review mechanism flags records whose maximum probability falls below a threshold selected using validation data, with 0.60 serving as the initial candidate. Coverage and conditional error describe this mechanism alongside primary scores calculated over the complete test population. Because a single positive temperature rescales every class score equally, it preserves the primary argmax decision. Its direct effects concern

probability concentration and the coverage of any confidence-based review rule; the candidate review threshold requires a separate coverage–error assessment.

Data separation precedes every transformation that depends on the sample distribution. The outer evaluation groups stations, while a stricter storm grouped variant separates records associated with the same evolving weather system. Inner folds support model complexity and threshold selection, and each development partition includes a disjoint calibration subset. A season separated test introduces an additional temporal challenge by reserving a complete period for independent evaluation. All baseline configurations use identical partition membership and issue time information. The ablation protocol removes profile summaries, elevation adaptation, and calibration individually to examine their respective contributions. A common prediction schema records class probability, group membership, and profile availability for every configuration, enabling paired comparisons across the same evaluation cases. Station and storm separation impose different constraints. A joint split must prevent either group from crossing the development–test boundary; reporting one grouping scheme establishes only that particular transfer condition.

The experimental record preserves individual predictions together with aggregated diagnostics. Each prediction row includes a stable case key, station group, storm group, elevation band, target label, probability vector, profile validity state, and model configuration identifier. Figures and tables are derived from these records through a shared analysis pipeline, maintaining consistency between graphical displays and numerical comparisons. Computational assessment records hardware, batch size, preprocessing time, inference time, and measurement repetitions. Separating preprocessing from inference identifies the contribution of terrain alignment and profile encoding to total latency. The subsequent analysis examines classification errors, transfer across elevation bands, probability calibration, and sensitivity to incomplete or perturbed atmospheric inputs. Multiclass Brier reporting requires a declared class-sum or class-mean convention and the same convention across configurations. The current aggregate ledger does not specify that normalization, which limits comparison with scores reported by other studies.

4. Experimental Evaluation and Analysis

4.1. Benchmark Population and Profile Structure

The benchmark population comprises 24,000 precipitation cases at 120 stations, including 12,000 rains, 8400 snow, and 3600 mixed cases. The five elevation bands have 6400, 5600, 4800, 4000 and 3200 records, respectively, and their centers are 250, 750, 1250, 1750 and 2250 m. The corresponding station counts decrease from 32 to 16 across increasing elevation, producing uneven sampling support between lowlands and uplands. The analysis examines population characteristics and classification diagnostics. The DISDRODB framework standardizes disdrometer metadata and processing workflows, including quality control, temporal resampling, and precipitation-phase classification [24]. Radar-disdrometer observations in eastern Canada show overlapping probabilities of solid, liquid, and mixed precipitation within a near-surface temperature interval, supporting an explicitly probabilistic treatment of phase ambiguity [25]. Environmental regionalization research shows that explicit spatial covariates can affect machine-learning predictions, reinforcing the need to distinguish spatial information used as predictors from location identifiers used for evaluation [26]. A recent natural-disaster modelling study groups validation records by disaster event to address dependence among observations belonging to the same event [27]. Grouping variables are reserved for evaluation design and partitioning. The following analysis distinguishes sampling structure, physical descriptors, and model performance. The population totals specify sample composition. Observation dates, geographic coverage, profile products, collocation tolerances, and label instruments remain necessary metadata for an independently reproducible benchmark; the totals alone establish none of these properties.

Figure 3(a) displays 240 surface wet bulb temperature and elevation pairs, with 80 cases for each phase. The overlap near the zero-degree reference is the main diagnostic feature: similar surface thermal conditions accompany different assigned phases, creating a setting in which column information could provide additional discrimination. The elevation spread also prevents a single near surface threshold from describing the entire population. Figure 3(b) shows warm exposure distributions whose five medians decline from 280 through 240, 200, and 170 to 140-degree Celsius metres. The endpoint contrast of 140-degree Celsius metres represents a halving of median exposure across the terrain sequence. Distribution overlap remains substantial, however, so elevation and positive thermal exposure encode complementary information. Within each band, the upper tail identifies profiles with comparatively large melting opportunity; their frequency and extent require examination within each elevation band. These distributions support a feature design that preserves exposure magnitude alongside terrain context, allowing the learner to distinguish a typical high elevation column from an unusually warm column at the same station height. The equal phase counts in the scatter display differ from the benchmark prevalence; visual overlap in this selected subset describes feature ambiguity rather than population class frequencies.

Figure 3(c) presents empirical cumulative distributions from 180 supported zero crossings per phase. The snow curve rises at lower crossing heights, the rain curve shifts toward higher heights, and the mixed curve occupies an intermediate

range. At a common cumulative probability, separation along the height axis describes differences in the vertical distance available beneath the lowest thermal crossing. The transition region between curves is especially informative because it contains profiles for which crossing height alone offers limited discrimination. Combining that height with warm exposure distinguishes the geometry of a crossing from the magnitude of the warm layer above it. Profiles containing several crossings require the additional crossing count defined in the encoder, since the lowest crossing represents only one part of their thermal history. The current cumulative distributions are conditional on a supported crossing and therefore describe a specific subset of the population. In a population comparison, the fraction of profiles with a crossing should accompany the conditional distribution so that regional differences in crossing availability can be separated from differences in crossing height. Profiles with no supported crossing and profiles lacking sufficient vertical support require separate counts before this conditional distribution can represent the complete benchmark.

Figure 3(d) compares valid profile fractions for 60 cases in each elevation band. Displayed medians are approximately 0.93, 0.90, 0.90, 0.84, and 0.78, placing the highest band close to the 0.75 completeness reference. The contrast between the two upper bands is relevant to interpretation because declining information availability can coincide with the terrain regime in which transfer is most difficult. A model that receives fewer valid layers may lose warm layer evidence even when the surface measurement remains reliable. The wider lower tails identify cases requiring explicit missingness indicators and a documented fallback path. The reference line is a diagnostic aid; an acceptance threshold would require checking which specific layers remain available. Losing the upper warm layer and losing a redundant near surface sample can produce the same completeness fraction while affecting the encoded exposure differently. Missingness mechanisms should consequently be retained at the level of individual profile samples, with complete upper column outages evaluated separately from scattered gaps.

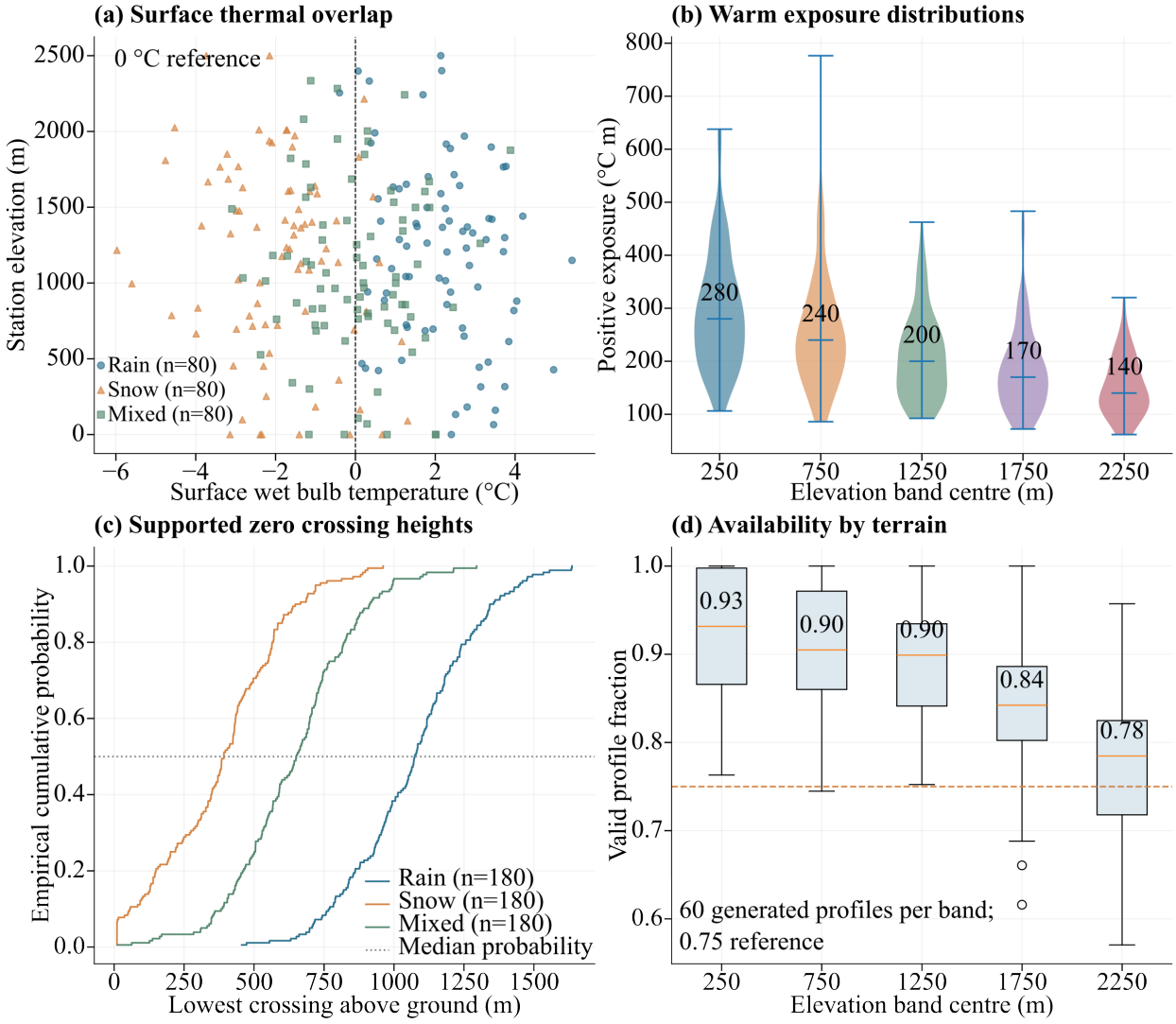


Figure 3. Benchmark structure. (a) Surface wet bulb temperature and elevation scatter. (b) Warm exposure violin distributions by elevation band. (c) Zero crossing height empirical cumulative distributions by phase. (d) Profile completeness box plots by elevation band

4.2. Discrimination Ablations and Terrain Transfer

Table 1 presents the fold average performance comparison for six configurations. Surface thresholding has macro F1 0.812, balanced accuracy 0.820, Brier score 0.238, and expected calibration error 0.094. Surface CatBoost reaches 0.867, 0.871, 0.168, and 0.061, respectively. The full configuration is assigned 0.910, 0.914, 0.122, and 0.024, producing a 0.043 macro F1 contrast and a 0.046 Brier score contrast against the surface ensemble. For multiclass evaluation, the arithmetic mean of classwise F1 scores gives equal importance to each class irrespective of its sample count [28]. A recent meteorological-topographic modelling study compares gradient-boosted learners on a common independent validation period and applies paired tests to their prediction errors [29]. Figure 4(a) places the six macro F1 values alongside sensitivity spreads of 0.012, 0.010, 0.009, 0.009, 0.008, and 0.007. The raw profile and encoded profile configurations differ by 0.008, while adding terrain adaptation to encoded profiles changes the score by 0.019. These contrasts separate the question of atmospheric access from the question of how that information is represented. Uncertainty in these contrasts requires evaluation with predictions grouped by station or storm. The displayed sensitivity spreads have no stated sampling construction and therefore provide descriptive ranges rather than confidence intervals or evidence of statistical significance.

Table 1. Fold average performance comparison

Configuration	Macro F1	Balanced accuracy	Brier score	Calibration error
Surface threshold	0.812	0.820	0.238	0.094
Surface CatBoost	0.867	0.871	0.168	0.061
Raw profile CatBoost	0.883	0.887	0.151	0.052
Encoded no adaptation	0.891	0.895	0.142	0.046
Terrain no encoding	0.880	0.884	0.156	0.055
Full configuration	0.910	0.914	0.122	0.024

Figure 4(b) displays the pooled test allocation for 3000 rains, 2100 snow, and 900 mixed cases. Correct predictions number 2871, 1995, and 688, yielding phase recalls of 0.957, 0.950, and approximately 0.764. Of the mixed cases, 106 enter the rain column and 106 enter the snow column, indicating symmetric confusion with the two pure phases in this allocation. Direct rain to snow and snow to rain errors total 68, whereas mixed to pure phase errors total 212. The latter pattern places the main diagnostic burden on transition discrimination. Pooled accuracy is approximately 0.926, pooled macro F1 is approximately 0.910, and pooled balanced accuracy is approximately 0.890. The table reports a fold average balanced accuracy of 0.914, whereas the displayed counts give a pooled value of approximately 0.890. Different aggregation rules can produce different scores, but that possibility alone does not reconcile these particular values. Case identities, fold membership, and class supports are needed to establish whether both summaries refer to the same prediction population; the discrepancy remains an unresolved reporting limitation.

Figure 4(c) compares precision recall curves with area labels of 0.970 for rain, 0.960 for snow, and 0.860 for mixed precipitation. The mixed curve declines more rapidly as recall approaches its upper range, expressing a larger precision cost when the operating point attempts to recover nearly all mixed events. The mixed phase curve places greater emphasis on the tradeoff between recovering ambiguous cases and retaining precision at demanding recall levels. Figure 4(d) compares six normalized dimensions for the surface, encoded, and full configurations. Its outward movement for the full configuration is most useful as a visual check of whether gains are concentrated in discrimination or shared with reliability and stability. The unnormalized metrics in the table retain the quantitative meaning needed for comparison. The 0.030 macro F1 contrast between the full and no encoding configurations includes both extra atmospheric information and its representation. An individual descriptor contribution would require removing that descriptor while retaining the remaining features and development budget. The radar panel requires the direction, reference range, and normalization rule for each axis to support quantitative interpretation. Until those mappings are supplied, the raw metrics in the table provide the basis for comparisons.

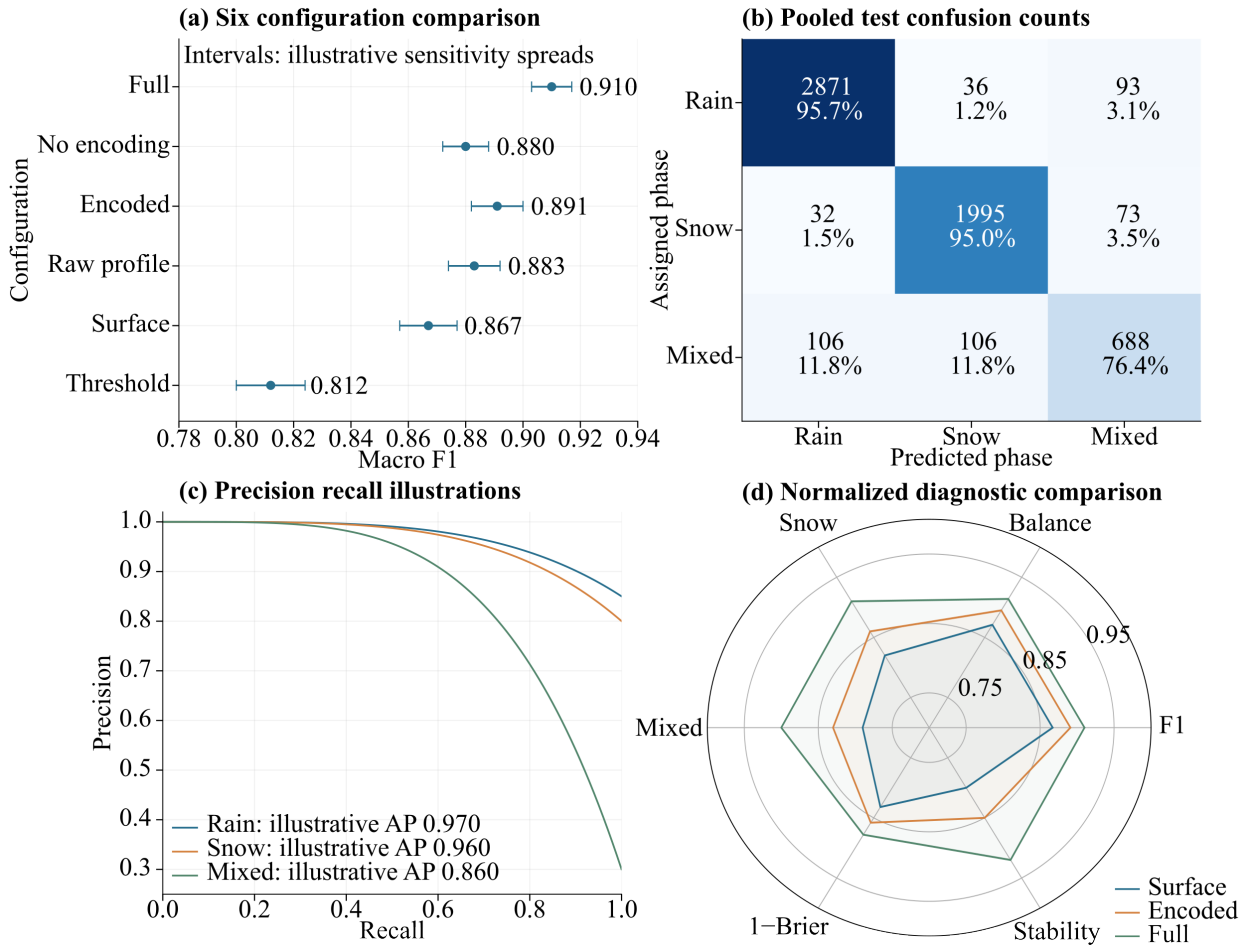


Figure 4. Discrimination diagnostics. (a) Macro F1 comparison with sensitivity spreads. (b) Pooled class confusion counts. (c) Precision recall curves for rain, snow, and mixed precipitation. (d) Normalized diagnostic radar comparison

Figure 5(a) compares surface only scores of 0.880, 0.870, 0.850, 0.820, and 0.780 with full configuration scores of 0.920, 0.920, 0.910, 0.900, and 0.880 across increasing elevation. Their differences grow from 0.040 to 0.100. The surface model loses 0.100 between the endpoints, while the full configuration loses 0.040, making sensitivity to terrain distribution a distinct target of the adaptation design. A subsurface modelling comparison finds that random and spatial validation can produce different performance estimates, with spatial validation providing a more realistic assessment of extrapolation in that application [30]. A held band experiment must therefore reserve the relevant stations before fitting the encoder transformations, density weights, model settings, and calibration parameter. The five terrain centres provide a common reference for the bandwise comparison. Their interpretation depends on the support available in each band and the degree of storm overlap between development and testing. Similar bandwise scores can arise from different storm populations, so class prevalence and independent group counts should accompany a geographically independent evaluation.

Figure 5(b) adds the raw profile, no adaptation, and no encoding configurations to the terrain comparison. The no adaptation sequence declines from 0.910 to 0.810, increasing its separation from the full configuration from 0.010 to 0.070. This pattern locates the intended adaptation effect in the sparsely represented upper terrain range. The raw profile sequence reaches 0.800 in the highest band, while the no encoding sequence reaches 0.840. Their different endpoints illustrate how atmospheric representation and terrain context can produce different transfer responses even when a pooled ranking appears similar. Continuous elevation interactions should be assessed for smoothness between the plotted centres, especially around density band boundaries. The weighting bands serve a sampling function, whereas the elevation predictor remains continuous. A sharp jump at a boundary would indicate sensitivity to the implementation or training support and should be investigated through neighbouring elevation perturbations with the atmospheric column held fixed.

Figure 5(c) displays 150 absolute probability errors against station minus grid elevation mismatch, with reference limits at minus and plus 300 m. Errors extend upward toward both mismatch extremes, illustrating how a geographically displaced column can alter probability quality even if its most likely class remains unchanged. The broad scatter near a common mismatch also shows that terrain discrepancy alone provides an incomplete account of error magnitude. Profile structure and label ambiguity remain relevant conditioning variables. Figure 5(d) displays five-fold loss distributions

with 50 sensitivity values each and approximate medians of 0.124, 0.120, 0.117, 0.134, and 0.127. The median range of about 0.017 identifies fold composition as a plausible source of variation worth examining alongside aggregate loss. The fourth fold has the highest median, making its terrain support, mixed phase prevalence, and profile completeness natural inspection targets. These sensitivity ranges describe variation within the diagnostic comparison; sampling intervals require resampling independent station or storm groups.

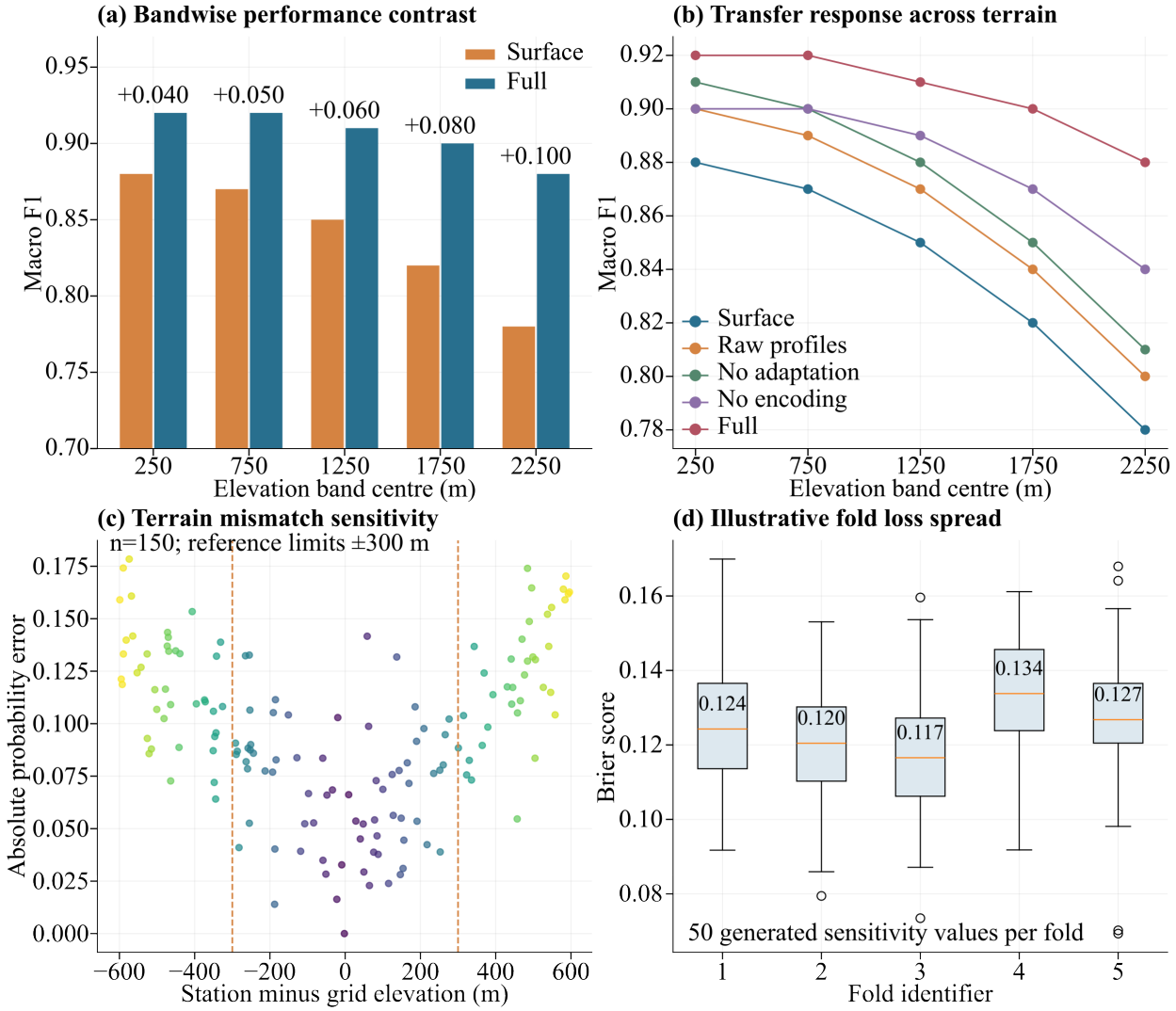


Figure 5. Terrain transfer and stability. (a) Two model grouped bar comparison across five bands. (b) Five configuration transfer curves. (c) Absolute probability error against elevation mismatch. (d) Fivefold-level loss box plots

4.3. Probability Calibration and Input Sensitivity

Probability reliability describes the agreement between predicted confidence and classification outcomes in engineering applications. The analysis uses nine confidence bins centred from 0.10 to 0.90 in increments of 0.10, with support counts totalling 6000 cases. Calibration assessment research shows that estimated calibration error depends on binning choices, motivating explicit reporting of the binning procedure when interpreting reliability diagrams [31]. For three normalized class probabilities, the largest probability is at least one third. The nonempty bins centred at 0.10 and 0.20 are therefore incompatible with the stated top-label definition. Interpreting the display as class-conditional reliability would require class identities and binary outcomes that are absent from the reported aggregates. The current display cannot establish top-label calibration until its probability definition and case-level assignments are reconciled. Temperature scaling is fitted on a disjoint calibration partition, and final test probabilities provide the verification sample. The terrain-balanced training objective and the deployment population may differ in their class and elevation distributions, making calibration dependent on the population used for fitting. Retaining unscaled and scaled probability vectors allows score changes to be traced to improved separation, reduced confidence, or redistribution of probability among phase classes.

Figure 6(a) shows accuracies before and after calibration against the nine confidence bin centres. The calibrated accuracies are 0.12, 0.22, 0.32, 0.42, 0.52, 0.62, 0.72, 0.82, and 0.93, giving a support weighted absolute gap of 0.023. The corresponding pre calibration gap is approximately 0.059. The arithmetic reduction of about 0.036 describes

agreement between the displayed coordinates and the identity reference. Because the lower bins conflict with the top-label probability definition, this reduction is a diagnostic calculation rather than a verified top-label calibration improvement. The uppermost confidence bin remains 0.03 above its nominal confidence, while the preceding bins each have a gap of 0.02. Its larger population therefore gives this residual an appreciable contribution to the aggregate score. The table's fold average calibration error of 0.024 belongs to a distinct aggregation scheme. A valid reliability analysis requires an explicit probability target, admissible bin boundaries, handling of empty bins, and class-conditional outcomes. The fold average calibration error also requires reconciliation with the underlying prediction records.

Figure 6(b) presents the nine support counts of 50, 90, 160, 300, 450, 650, 1000, 1500, and 1800. The first two bins fall below the displayed 100 case reference, whereas the last three contain 4300 cases, approximately 72 percent of the total. The displayed support is concentrated in the upper coordinate ranges. The assignment of all counts to admissible probability bins requires verification before these counts describe the confidence distribution of a three-class classifier. A change in the lowest bin can produce a visually conspicuous displacement while contributing little to the weighted gap. Conversely, a small displacement of the highest bin affects a much larger fraction of the population. This support imbalance motivates examining both local reliability and the distribution of issued probabilities. The reference count serves an inspection function; interval precision also depends on observed accuracy and the independence of contributing cases. A station or storm grouped interval would be appropriate when the same weather system supplies many predictions to a common confidence bin.

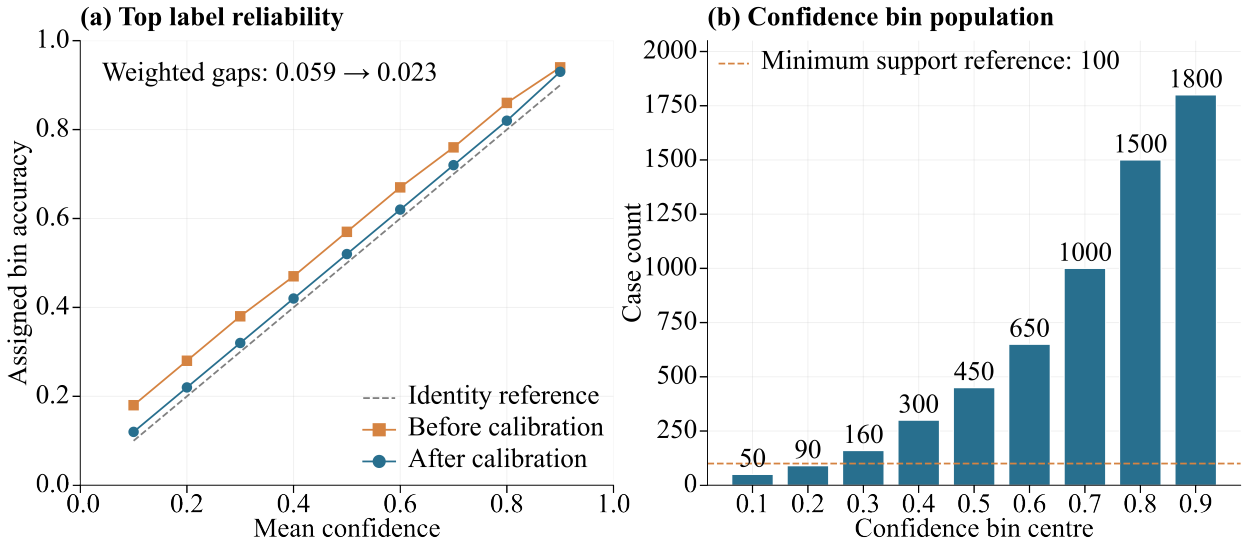


Figure 6. Probability reliability and bin support. (a) Reliability curves before and after calibration with identity reference. (b) Nine confidence bin support counts with a 100-case reference

Input sensitivity is examined at profile missingness levels of 0, 10, 20, 30, and 40 percent. The perturbation mechanism determines which parts of the encoded atmospheric history are lost. A meteorological imputation study evaluates missing-data scenarios informed by power outages, observer absences, and sensor failures, supporting stress tests that reflect distinct mechanisms of input loss [32]. Removing the highest supported layers can suppress warm exposure, while removing alternating interior levels primarily changes interpolation support. A structured stress test should retain the original case population and compare predictions before and after each controlled outage. Figure 7(a) assigns macro F1 values of 0.910, 0.905, 0.897, 0.884, and 0.866, with sensitivity spreads of 0.007, 0.008, 0.010, 0.013, and 0.017. The endpoint decline is 0.044, and the score crosses the displayed 0.880 reference between 30 and 40 percent missingness. The growing spread identifies an additional sensitivity dimension: degraded inputs can increase variation across conditions as well as shift the central score. Both effects require layer specific validity records for diagnosis. The missingness percentages specify input loss, while the layer-selection rule and repetition scheme remain necessary to reproduce each point. The displayed sequence alone does not identify whether the outages are random, contiguous, or concentrated in upper layers.

The shape of the missingness response suggests where a fitted experiment should concentrate its inspection effort. The decline from 0 to 10 percent is 0.005, compared with 0.018 from 30 to 40 percent, indicating increasing sensitivity as profile support diminishes. A plausible diagnostic question is whether critical warm layers disappear more frequently in the latter regime or whether several weaker descriptors jointly lose information. Comparing exposure, crossing height, and near surface gradient before and after deletion would distinguish those possibilities. The width of the spread rises by 0.010 between the endpoints, so a single central score would conceal part of the stress response. A fallback analysis should additionally stratify by surface temperature and phase because columns near the transition may depend more

strongly on vertical information. Examining cases on either side of the reference score can identify the missing layers associated with the largest probability changes and the phase categories most affected by those changes.

Figure 7(b) presents a joint response surface on five temperature perturbation levels between minus 2 and plus 2 degrees Celsius and five elevation mismatch levels between minus 600 and plus 600 m. The central macro F1 value is 0.910. At zero elevation mismatch, the two extreme temperature perturbations give 0.862; at zero temperature perturbation, the extreme elevation mismatches give 0.875. Their combined corner values are 0.827, producing a 0.083 centre to corner contrast. The symmetric contours describe equal score changes under perturbations of opposite sign, while their spacing reflects the relative sensitivity to the two input dimensions. An operational forecast-verification study identifies uncertain observations, limitations of simplified metrics, and differences between verification information and users' experiences as practical evaluation challenges [33]. A paired sensitivity analysis should retain case identity, rerun terrain alignment and feature encoding after each perturbation, and retain the resulting probability vectors. Temperature changes can create or remove zero crossings, while elevation changes can alter the supported portion of a column. Those changes in feature availability should be inspected alongside the probability shift, particularly where a perturbed ground reference removes a shallow near surface layer. Symmetry of the displayed response surface alone cannot establish equal physical sensitivity to warming and cooling or to opposite terrain offsets. That interpretation requires paired case-level changes and the corresponding feature-support diagnostics.

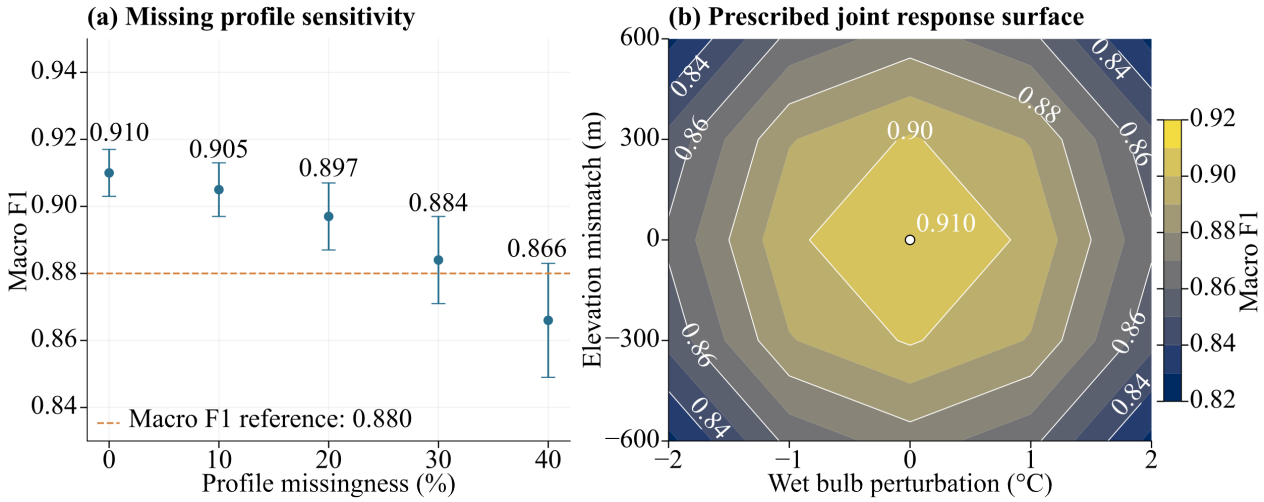


Figure 7. Robustness diagnostics. (a) Macro F1 under five profile missingness levels. (b) Filled contour response over temperature perturbation and elevation mismatch

5. Conclusion

This paper introduces a rain and snow phase classification framework that integrates terrain-relative wet-bulb profile encoding with elevation-adaptive CatBoost learning. The profile representation preserves warm-layer exposure, near-surface cooling, vertical crossings, and information availability, while continuous terrain descriptors and bounded density weights address uneven geographic support. A separate calibration path is employed for class prediction and probabilistic decision-making. The evaluation establishes a reporting structure for discrimination, terrain transfer, probability reliability, and perturbation sensitivity. The framework provides a methodological design for comparison with operational precipitation diagnostics. Its present evidence concerns the reported benchmark; geographic independence and calibration consistency remain conditions for stronger deployment claims.

The evaluation covers a restricted range of phase categories, terrain conditions, and input perturbations. Generalization across climates, uncertainty associated with correlated storm sampling, and operational value require further assessment. The compact profile summaries approximate thermal conditions along the particle pathway, while the representation of particle histories, microphysical processes, and terrain effects remains incomplete. Mixed precipitation labels may lack time-of-occurrence or instrument-specific information; flexible calibration is needed for class- and terrain-dependent reliability. These limitations constrain interpretation and motivate more detailed assessment of the intermediate descriptors.

Future research should combine independently quality-controlled precipitation labels with collocated atmospheric profiles and documented terrain metadata. Validation should separate stations, storms, and seasons, compare alternative learners under matched tuning budgets, and assess the value of profile encoding under realistic observation failures. Further work should examine class-conditional calibration, out-of-support detection, and decision thresholds informed by explicit engineering costs. Releasing the encoder, partition definitions, prediction records, and provenance checks would support regional transfer assessments and identify physical descriptors that remain useful as observing systems change. An operational assessment should additionally document the available forecast lead time, source-profile latency,

preprocessing cost, and end-to-end prediction time. These measurements would establish whether the representation can meet a specific warning or monitoring deadline.

6. Declarations

6.1 Author Contributions

Rigos Balaskas contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Giannis Koris contributes to validation, analysis, investigation, data collection, draft preparation. Dimos Chatzis contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 Declaration of Competing 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.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Wang, Z., Vivoni, E. R., Whitney, K. M., Xiao, M., & Mascaro, G. (2024). On the sensitivity of future hydrology in the Colorado River to the selection of the precipitation partitioning method. Water Resources Research, 60(6), Article e2023WR035801. https://doi.org/10.1029/2023WR035801

[2] Pan, G., Zhou, X., Zheng, W., Liu, Z., & Zhang, A. (2025). RSSD: A regional-level resource-saving snow detection model for winter road surface maintenance. Expert Systems with Applications, 277, Article 127155. https://doi.org/10.1016/j.eswa.2025.127155

[3] Wolvin, S., Strong, C., Rupper, S., & Steenburgh, W. J. (2024). Climatology of orographic precipitation gradients over High Mountain Asia derived from dynamical downscaling. Journal of Geophysical Research: Atmospheres, 129(20), Article e2024JD041010. https://doi.org/10.1029/2024JD041010

[4] Liu, L., Zhang, L., Zhang, Q., Wang, G., Zhou, Z., Li, X., & Tang, Z. (2025). A non-sigmoidal-curve-dependent dynamic threshold method improves precipitation phase partitioning in the Northern Hemisphere. Water Resources Research, 61(4), Article e2024WR038636. https://doi.org/10.1029/2024WR038636

[5] Shi, S., & Liu, G. (2024). Improvements on phase classification using atmospheric melting and refreezing energy based on soundings. Journal of Geophysical Research: Atmospheres, 129(10), Article e2023JD040030. https://doi.org/10.1029/2023JD040030

[6] Simonet, G., Rotach, M. W., & Lehner, M. (2026). Physically consistent mesoscale model evaluation in complex terrain. Quarterly Journal of the Royal Meteorological Society, 152(775), Article e70063. https://doi.org/10.1002/qj.70063

[7] Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2024). Challenges in data-driven

geospatial modeling for environmental research and practice. Nature Communications, 15(1), Article 10700. https://doi.org/10.1038/s41467-024-55240-8

[8] Xu, X., Chen, F., Wang, B., Harrison, M. T., Chen, Y., Liu, K., Zhang, C., Zhang, M., Zhang, X., Feng, P., & Hu, K. (2024). Unleashing the power of machine learning and remote sensing for robust seasonal drought monitoring: A stacking ensemble approach. Journal of Hydrology, 634, Article 131102. https://doi.org/10.1016/j.jhydrol.2024.131102

[9] Muftuoglu, T. D. (2026). High-accuracy day-ahead precipitation forecasting with an interpretable CatBoost regressor and a LightGBM-based extreme-event alerting pipeline. Water Science, 40(1), Article 32. https://doi.org/10.1007/s44533-026-00033-9

[10] Van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L., Kerr, K. F., Barreñada, L., Varoquaux, G., Singh, K., Moons, K. G. M., Hernandez-Boussard, T., Timmerman, D., McLernon, D. J., van Smeden, M., & Steyerberg, E. W. (2025). Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance. The Lancet Digital Health, 7(12), Article 100916. https://doi.org/10.1016/j.landig.2025.100916

[11] Yu, G., Jennings, K. S., Hatchett, B. J., Nolin, A. W., Hur, N., Collins, M., Heggli, A., Tonino, S., & Arienzo, M. M. (2024). Crowdsourced data reveal shortcomings in precipitation phase products for rain and snow partitioning. Geophysical Research Letters, 51(24), Article e2024GL112853. https://doi.org/10.1029/2024GL112853

[12] Stock, A. (2025). Choosing blocks for spatial cross-validation: Lessons from a marine remote sensing case study. Frontiers in Remote Sensing, 6, Article 1531097. https://doi.org/10.3389/frsen.2025.1531097

[13] Hu, J., Che, T., He, Y., Huang, W., & Yang, X. (2024). An intercomparison of empirical schemes for partitioning precipitation phase. Journal of Hydrology: Regional Studies, 53, Article 101757. https://doi.org/10.1016/j.ejrh.2024.101757

[14] Singh, B., Liu, M., Abatzoglou, J. T., Adam, J. C., & Rajagopalan, K. (2025). Incorporating relative humidity in precipitation phase partitioning reduces model bias for some snow and streamflow metrics across the Northwest US. Journal of Hydrology, 662, Article 133833. https://doi.org/10.1016/j.jhydrol.2025.133833

[15] Jennings, K. S., Collins, M., Hatchett, B. J., Heggli, A., Hur, N., Tonino, S., Nolin, A. W., Yu, G., Zhang, W., & Arienzo, M. M. (2025). Machine learning shows a limit to rain-snow partitioning accuracy when using near-surface meteorology. Nature Communications, 16(1), Article 2929. https://doi.org/10.1038/s41467-025-58234-2

[16] Fu, Y., Yang, L., Wu, Z., Zhang, P., Gu, S., Chen, L., & Nan, S. (2024). A new algorithm of rain type classification for GPM dual-frequency precipitation radar in summer Tibetan Plateau. Advances in Atmospheric Sciences, 41(11), 2093–2111. https://doi.org/10.1007/s00376-024-3384-7

[17] Antúnez, P. (2026). When and why elevation fails as a temperature proxy: Hierarchical controls on lapse rates across climate, physiography, and vegetation in Mexico. Agricultural and Forest Meteorology, 389, Article 111376. https://doi.org/10.1016/j.agrformet.2026.111376

[18] Yang, C., Li, X., Zhang, X., Wu, J., & Li, L. (2025). Evaluation and comparison of separated precipitation types from multi-sources data in the Chinese Tianshan mountainous region. Journal of Mountain Science, 22(2), 489–504. https://doi.org/10.1007/s11629-024-9258-8

[19] Zhang, B., Yu, H., Hu, Z., Yue, P., Tang, Z., Luo, H., Wang, G., & Cheng, S. (2025). A machine learning-based observational constraint correction method for seasonal precipitation prediction. Advances in Atmospheric Sciences, 42(1), 36–52. https://doi.org/10.1007/s00376-024-4191-x

[20] Haseeb, M., Tahir, Z., Mehmood, S. A., Arif, H., Kousar, S., Ghafoor, S., & Mehmood, K. (2026). Ensemble artificial intelligence fusing satellite, reanalysis, and ground observations for improved PM2.5 prediction. Atmosphere, 17(4), Article 411. https://doi.org/10.3390/atmos17040411

[21] Pic, R., Dombry, C., Naveau, P., & Taillardat, M. (2025). Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. Advances in Statistical Climatology, Meteorology and Oceanography, 11(1), 23–58. https://doi.org/10.5194/ascmo-11-23-2025

[22] Fan, K. L., Pennello, G., Liu, Q., Petrick, N., Samala, R. K., Samuelson, F. W., Thompson, Y. L. E., & Cao, Q. (2025). Calzone: A Python package for measuring calibration of probabilistic models for classification. Journal of Open Source Software, 10(114), Article 8026. https://doi.org/10.21105/joss.08026

[23] Song, B., Zhao, S., Dang, L., Wang, H., & Xu, L. (2025). A survey on learning from data with label noise via deep neural networks. Systems Science & Control Engineering, 13(1), Article 2488120. https://doi.org/10.1080/21642583.2025.2488120

[24] Ghiggi, G., Candolfi, K., Billault-Roux, A.-C., Longchamp, R., Pham-Ba, S., Weil, C., Uijlenhoet, R., & Berne, A. (2026). disdrodb: An open-source Python package for standardized processing, sharing, and analysis of disdrometer data. Atmospheric Measurement Techniques, 19(14), 4943–4987. https://doi.org/10.5194/amt-19-4943-2026

[25] Bédard-Therrien, A., Anctil, F., Thériault, J. M., Chalifour, O., Payette, F., Vidal, A., & Nadeau, D. F. (2025). Leveraging a radar-based disdrometer network to develop a probabilistic precipitation phase model in eastern Canada. Hydrology and Earth System Sciences, 29(4), 1135–1158. https://doi.org/10.5194/hess-29-1135-2025

[26] Ohmer, M., Doll, F., & Liesch, T. (2025). Incorporating spatial information for regionalization of environmental parameters in machine learning models. Mathematical Geosciences, 57(2), 251–273. https://doi.org/10.1007/s11004-024-10163-4

[27] Talan, T., Giray, T., Korkmaz, A., Kosunalp, S., Tumen, V., & Akin, E. (2026). Predicting and explaining economic losses from natural disasters: A post-hoc machine learning and explainable AI approach. Applied Sciences, 16(19), Article 9497. https://doi.org/10.3390/app16199497

[28] Pivin-Bachler, J. R., & van den Broek, E. L. (2025). SIMBA: A robust and generalizable measure of data imbalance. Patterns, 6(12), Article 101395. https://doi.org/10.1016/j.patter.2025.101395

[29] Sun, P., Wang, M., Meng, D., Xu, Y., Cheng, C., & Ju, W. (2026). Multi-source meteorological–topographic modeling of monthly power generation for mountain photovoltaic stations using gradient-boosted trees. Energies, 19(12), Article 2936. https://doi.org/10.3390/en19122936

[30] Zhan, C., Jiao, J. J., Ma, Y., Wang, H., Soltanian, M. R., Carroll, K. C., & Dai, Z. (2026). Impacts of validation strategy and geological complexity on machine learning-based subsurface structure modeling. Journal of Hydrology, 666, Article 134797. https://doi.org/10.1016/j.jhydrol.2025.134797

[31] Jiang, Z., Liu, A., & Van Durme, B. (2024). Addressing the binning problem in calibration assessment through scalar annotations. Transactions of the Association for Computational Linguistics, 12, 120–136. https://doi.org/10.1162/tacl_a_00636

[32] Toure, M., Klutse, N. A. B., Sarr, M. A., Bhuiyan, M. A. E., Kenne, A. D., Thiaw, W. M., Badiane, D., Gaye, A. T., Ndiaye, O., & Mbow, C. (2025). Machine learning approaches for imputing missing meteorological data in Senegal. Applied Computing and Geosciences, 27, Article 100281. https://doi.org/10.1016/j.acags.2025.100281

[33] Pagano, T. C., Casati, B., Landman, S., Loveday, N., Taggart, R., Ebert, E. E., Khanarmuei, M., Jensen, T. L., Mittermaier, M., Roberts, H., Willington, S., Roberts, N., Sowko, M., Strassberg, G., Kluepfel, C., Bullock, T. A., Turner, D. D., Pappenberger, F., Osborne, N., & Noble, C. (2024). Challenges of operational weather forecast verification and evaluation. Bulletin of the American Meteorological Society, 105(4), E789–E802. https://doi.org/10.1175/BAMS-D-22-0257.1