Scientific Reports

https://doi.org/10.1038/s41598-026-39766-z

Article in Press

A physics-guided machine learning framework for enhancing dust storm visibility prediction in arid and semi-arid regions

Received: 5 November 2025

Accepted: 6 February 2026

Published online: 03 March 2026

Cite this article as: Xu C., Zhang H., Luo K. *et al.* A physics-guided machine learning framework for enhancing dust storm visibility prediction in arid and semi-arid regions. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-39766-z

Chang Xu, Henggang Zhang, Kaiyue Luo, Jie Liu, Yang Shen, Hongcai Qin, Yunhao Qu, Chenhui Zhu & Zhen Han

A physics-guided machine learning framework for enhancing dust storm visibility prediction in arid and semi-arid regions

Chang Xu[1†], Henggang Zhang[2†*], Kaiyue Luo[3], Jie Liu[1] , Yang Shen[1] , Hongcai Qin[1] , Yunhao Qu[2], Chenhui Zhu[4], and Zhen Han[5]

1 Northwest Institute of Nuclear Technology, Xi'an, 710024, China

2 Information Engineering University, 450002, Zhengzhou

3 College of Surveying and Geo-Informatics, Tongji University, Shanghai 200092, China,

4 College of Geography and Remote Sensing Science, Xinjiang University, Urumqi 830046, China

5 Qingdao Marine Remote Sensing Information Technology Company,Ltd., Qingdao, 266000, Shandong, China

†These authors contributed equally to this work.

* Correspondence: zhanghenggang2000@163.com

ABSTRACT

Dust sharply degrades visibility in arid and semi-arid regions, yet operational forecasting remains challenged by near-surface process errors in numerical weather prediction (NWP) and the poor generalization of purely data-driven models. We present a physics-guided machine learning (PGML) framework that post-processes European Centre for Medium-Range Weather Forecasts (ECMWF) forecasts to predict five ordinal visibility grades over the Kumtag Desert. A dust-lifecycle feature library (emission, vertical mixing, transport, wet scavenging) is coupled with an ordinal LightGBM architecture. On an independent test period, the model attains quadratic weighted kappa (QWK) of 0.26 (0–24 h), 0.17 (24–48 h), and 0.18 (48–72 h), with mean absolute error (MAE) of 0.48–0.56; gains versus data-only baselines increase with forecast horizon. Ablation experiments show that physics priors can effectively improve visibility prediction accuracy—reducing MAE by up to 10% and sustaining QWK beyond 24 h by constraining non-physical drift. Accordingly, the PGML visibility predictions show improved performance relative to data-only

baselines. SHAP analysis reveals a forecast-horizon-dependent mechanistic shift from emission/surface-layer dynamics to stability-controlled vertical mixing, consistent with dust dynamics. The framework offers an interpretable, transferable paradigm for physics-constrained environmental forecasting.

Keywords: Physics-guided machine learning; dust visibility prediction; arid and semi-arid regions; regularised classification; ECMWF forecast data

Introduction

Atmospheric dust particles emitted from arid and semi-arid regions play a critical role in the Earth system by modulating radiative forcing, cloud microphysical processes, and biogeochemical cycles[1,2]. At regional scales, wind-driven dust events substantially degrade near-surface visibility through enhanced aerosol extinction, posing severe risks to agriculture, transportation safety, and human health[3,4]. According to classical atmospheric optics theory, meteorological visibility is inversely related to the atmospheric extinction coefficient, which is strongly influenced by aerosol concentration, size distribution, and hygroscopic growth[5,6].

Recent assessments, including the IPCC Sixth Assessment Report, indicate that dust activity in arid regions is intensifying as a consequence of increasing drought frequency and anomalous wind patterns, thereby increasing the demand for reliable regional dust and visibility forecasting systems[1].

Two mainstream forecasting paradigms have emerged over the past two decades in dust prediction research. The first paradigm consists of physically based numerical models, which have evolved into operational frameworks such as the World Meteorological Organization (WMO) Sandstorm Early Warning Advisory and Assessment System (SDS-WAS) and the Copernicus Atmosphere Monitoring Service (CAMS)[7,8]. These models provide 3–5-day forecasts of spatiotemporal dust-aerosol distributions and column-integrated quantities such as aerosol optical depth (AOD), while incorporating multi-source observations through four-dimensional variational assimilation (4D-Var) to enhance initial conditions[7]. However, atmospheric visibility is inherently governed by the coupled effects of near-surface optical, microphysical, and turbulent processes[9-11]. As a result, existing numerical models remain less robust in forecasting event-level and threshold-level visibility within dust source regions and surrounding areas. This often leads to underestimation of event intensity or failure

to capture key physical processes[12-14]. This creates a clear opportunity for the integration of data-driven methods.

Another paradigm is the purely data-driven machine-learning approach. Complementing numerical models, purely data-driven approaches benefit from increasing observational records and computational power. They improve the statistical accuracy of short-term forecasts by extracting nonlinear relationships from historical data[15-17]. Previous machine-learning models have been applied to forecasting applications such as aviation and traffic operations, achieving considerable performance improvements in low-visibility event classification, short-term forecast correction, and multi-source data fusion[18-22]. However, pure machine learning generally suffers from several types of failure modes that affect acceptance: (1) a tendency to underestimate extremes in records with very uneven categories; (2) vulnerability to spatial migration and out-of-distribution samples; (3) decaying skill in long-latency extrapolations; (4) insufficient confidence and interpretability in threshold-level decision-making due to their “black-box” nature[15,16]. This limits the operational application of models and hinders robustness assessment and risk communication during extreme events.

To overcome the limitations of purely data-driven approaches, Physics-guided machine learning (PGML) has emerged as a strategy to embed conservation laws or stability thresholds as explicit constraints within models, preserving physical consistency while improving robustness[23]. In weather and climate research, PGML has shown great potential in the simulation of convective, radiative, and subsurface exchange processes, with the core value of suppressing the extrapolation of uncertainty and improving the interpretability of models[24,25]. For the dust problem, classical physical quantities such as threshold friction velocity and jump fluxes provide a structured library of features that can be directly manipulated to construct a “process prior”[26]. For instance, Jin et al. focused on adjoint emission source inversion[26], and Kok et al. synthesized global dust-climate interactions[27]. Similarly, operational assimilation systems like those described by Benedetti et al. focus on column-integrated AOD rather than the near-surface visibility grades required for regional warnings[28]. However, the challenge remains to ensure the physical plausibility of the models, the confidence of extreme events, and the robust mapping of their macroscopic outputs to specific downstream operational thresholds such as visibility classes.

This study proposes a PGML framework that addresses these gaps by coupling a dust-lifecycle-complete feature library with an ordinal classification architecture aligned to operational warning grades. We quantify physics priors’ marginal contributions via ablation studies,

revealing their role as structural anchors against error growth beyond 24 h, and trace mechanism evolution using SHapley Additive exPlanations (SHAP). Contributions include: (1) a transferable PGML paradigm for threshold-mechanism-lead interactions; (2) evidence that physics priors suppress non-physical drift in long-lead forecasts; (3) the development of operationally viable lightweight models for practical deployment.

Study Area and Data

Region, stations and target variable

This study focuses on the Kumtag Desert in the arid region of Northwest China, which is located at the eastern end of southern Xinjiang, south and east of Lop Nur, with a total area of about 22100 km^2. The climate of the region is extremely arid, with scarce precipitation and strong evaporation. The terrain is dominated by flowing dunes[29], with crescent-shaped dunes and dune chains being the most typical dune types, and fixed and semi-fixed dunes scattered along the edge of the desert. It is surrounded by extremely arid sand sources in oasis agricultural areas, as shown in Fig. 1.

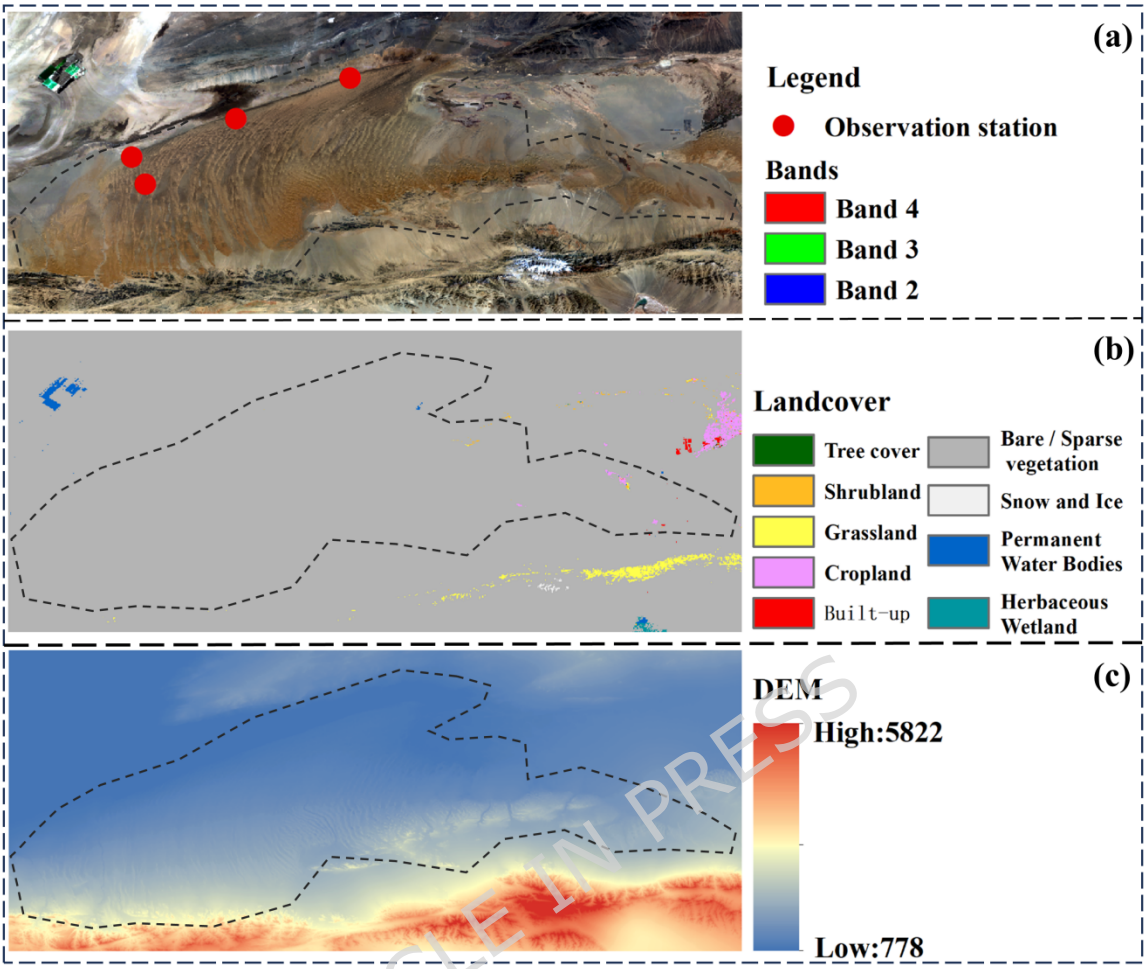

Figure 1. Overview of the study region. (a) Image of the study area, (b) Land cover of the study area, (c) Elevation distribution of the study area. Location of the study area in the Kumtag Desert region. The basemap was generated from Landsat 8 imagery using a median composite of scenes acquired between 1 April 2024 and 1 July 2024. The study area boundary (dashed line) was delineated by the authors to indicate the spatial extent of the analyses conducted in this study and does not represent an official administrative boundary. The map was produced in ArcMap (Esri; version 10.8; https://www.esri.com/arcgis).

The automatic weather station (AWS) data used in this study were obtained from the automatic meteorological stations in the Kumtag Desert area. The observation period was from 1 January 2020 to 20 December 2022, and the meteorological elements collected included temperature, pressure, visibility, wind speed and wind direction (Table 1).

Table 1. Station metadata summary.

No.	Longitude	Latitude	Model	Elements
1	91.181	39.925	DZZ4	temperature, pressure, visibility, wind speed, wind direction

No.	Longitude	Latitude	Model	Elements
2	91.799	40.152	DZZ4	temperature, pressure, visibility, wind speed, wind direction
3	92.477	40.391	DZZ4	temperature, pressure, visibility, wind speed, wind direction
4	91.262	40.762	DZZ4	temperature, pressure, visibility, wind speed, wind direction

The study utilizes automatic weather station data from the Kumtag Desert (2020-01-01 to 2022-12-20), including visibility, wind speed, and direction at 1-min resolution, aggregated to 3-h intervals to match ECMWF forecast data (Table 1).Strict quality assurance and quality control (QA/QC) were performed following the Meteorological Industry Standards of the People's Republic of China, specifically QX/T 118-2020 (Quality control of meteorological observation data—Surface) and QX/T 397-2017 (Quality control of meteorological observation data—Automatic weather station), issued by the China Meteorological Administration. This process involved: (1) Limit checks to exclude physical outliers (e.g., visibility < 0 m or > 50 km, wind speed > 75 m/s); (2) Temporal consistency checks, where records with a sudden visibility drop of ≥ 80% recovered within 3 minutes were flagged as suspect; (3) Gap filling, where missing data ≤ 3 min were linearly interpolated, while larger gaps were treated as missing. Following this rigorous process, 5438 invalid records were excluded from the initial 52560 hourly observations, leaving 47122 valid samples (99.9% retention rate) for the experiment.

Dust events were identified based on the National Standard of the People's Republic of China (GB/T 20480-2025, Grade of sand and dust weather), which defines categories using specific thresholds: Floating Dust (visibility < 10 km, instantaneous wind speed < 5.4 m/s), Blowing Sand (visibility < 10 km, wind ≥ 5.4 m/s), Sandstorm (visibility < 1 km, wind ≥ 10 m/s), Severe Sandstorm (visibility < 500 m, wind ≥ 12 m/s), and Extreme Sandstorm (visibility < 50 m, wind ≥ 14 m/s). The data were aggregated to 3-h intervals to match ECMWF forecast data. For model training and evaluation, the dataset was divided into chronological segments; the specific sample counts for each segment are detailed in Tables 2 and 3. Finally, visibility is ordinal-classified into five levels: 0–5 km (Level 1), 5–10 km (2), 10–15 km (3), 15–20 km (4), and > 20 km (5), with severe low-visibility events (< 10 km) comprising < 8% of samples, exhibiting severe imbalance (Table 2).

Table 2. Distribution of class samples (%) for each dataset and forecast horizon.

Forecast Horizon (h)	Dataset	Level 1	Level 2	Level 3	Level 4	Level 5	Total Samples
0–24	Training	1.62	5.31	6.69	9.18	70.73	7175

Forecast Horizon (h)	Dataset	Level 1	Level 2	Level 3	Level 4	Level 5	Total Samples
0–24	Validation	1.27	4.94	3.61	4.64	84.26	3298
0–24	Test	0.59	1.21	4.50	10.99	80.24	3886
24–48	Training	1.71	5.45	6.77	9.22	70.58	8198
24–48	Validation	1.27	4.96	3.71	4.93	83.88	3771
24–48	Test	0.54	1.13	4.29	10.55	81.07	4426
48–72	Training	1.82	5.55	6.93	9.30	70.06	8200
48–72	Validation	1.29	5.01	3.78	5.03	83.40	3735
48–72	Test	0.68	1.29	3.88	10.22	82.25	4433

ECMWF predictors and spatiotemporal matching

Predictor fields were derived from the Integrated Forecasting System (IFS) of the ECMWF, featuring a native spatial resolution of 0.25°and a temporal resolution of 3 hours. Variables were selected to span the surface, near-surface, and layered atmosphere. A total of 65 meteorological predictors were directly extracted from the forecast output without complex calculation, comprising both surface-level parameters and vertical atmospheric profiles across standard pressure levels (including geopotential height, temperature, relative humidity, and wind components.

Grid-point fields were bilinearly interpolated to the four automatic weather stations (Table 1) at 3-hour intervals, yielding the final feature vector of 65 predictors at each site. Forecast and observation timestamps were rigorously aligned to ensure synoptic consistency.

Dataset preprocessing

To preclude any possibility of information leakage, the dataset was partitioned into three strictly chronological segments: training (1 January 2020–30 June 2021), validation (1 July 2021–30 November 2021) and testing (1 December 2021–20 December 2022). Within each segment, samples from the three windows (0–24 h, 24–48 h and 48–72 h) were further isolated so that no valid time was shared across windows, guaranteeing independence between forecast horizons.

Visibility observations were screened to isolate dust-related reductions rather than to exclude all non-dust days. Quality control filtering removed samples dominated by non-dust visibility mechanisms. Specifically, any 3-hourly instance with relative humidity ≥ 70% was discarded because visibility reduction under high-humidity conditions is predominantly associated with fog, haze, or hygroscopic aerosol growth, rather than dry dust extinction. Likewise, samples with hourly precipitation > 0.1 mm h^{-1} were excluded, as precipitation

involves fundamentally different visibility mechanisms including hydrometeor obscuration and wet scavenging.

Although this screening does not explicitly rely on $PM_{2.5}/PM_{10}$ ratios to distinguish haze from dust, extensive aerosol observations demonstrate that atmospheric extinction in arid desert source regions is predominantly dominated by coarse-mode mineral dust characterized by low Ångström Exponents, in contrast to fine-mode anthropogenic aerosols[6,30]. Furthermore, anthropogenic aerosol contributions in remote desert environments are negligible compared to natural dust loading[31]. Consequently, under these dry conditions, significant visibility reductions can be reasonably attributed to dust events.

Importantly, the retained dataset spans the full visibility spectrum, from clear conditions (>20 km) to severe dust storms (<5 km), ensuring that the model learns dust-specific physical controls rather than multi-aerosol statistical artifacts. The resulting sample distribution is highly imbalanced, with severe and moderate low-visibility grades (≤10 km) accounting for less than 8% of all valid samples.

To improve the model's ability to resolve these rare events, a hybrid SMOTE–Tomek strategy was applied exclusively to the training set. SMOTE first oversamples the minority classes by interpolating along local k-nearest-neighbour manifolds, expanding both cardinality and diversity. Subsequently, Tomek links excise majority-class observations that lie on or near the decision boundary, sharpening inter-class separation[32]. To preserve temporal coherence and avoid synthetic leakage, oversampling was executed day-wise: all observations from a given dust event were treated as an indivisible unit. Validation and test sets were left untouched, ensuring that reported performance metrics reflect the true, unmodified data distribution. Implementation details—neighbour counts, distance metrics and random seeds—are catalogued in the Appendix, with post-resampling class balances reported in Table 3.

The SMOTE-Tomek strategy intentionally creates a distribution mismatch between training (balanced) and evaluation (natural) sets. This design prioritizes sensitivity to rare but operationally critical low-visibility events, as models trained on balanced distributions exhibit improved minority-class recall while evaluation on original distributions ensures reported metrics reflect realistic deployment performance.

Table 3. Statistics of class distribution for the training set before and after applying the resampling strategy. Note: "Training Samples" refers to the original observed data; "After Resampling" refers to the dataset size after applying the SMOTE–Tomek algorithm. The resampling is applied exclusively to the training

set to prevent data leakage, while the validation and test sets remain unchanged to evaluate performance on real-world distributions.

Visibility Class	Forecast Horizon (h)	Training Samples	Ratio (%)	After Resampling	Ratio After (%)
0–5 km	0–24	116	1.62	5041	17.14
5–10 km	0–24	381	5.31	4950	16.83
10–15 km	0–24	480	6.69	4903	16.67
15–20 km	0–24	659	9.18	4893	16.64
≥20 km	0–24	5075	70.73	4688	15.94
0–5 km	24–48	140	1.71	5731	17.19
5–10 km	24–48	447	5.45	5599	16.79
10–15 km	24–48	555	6.77	5599	16.67
15–20 km	24–48	756	9.22	5555	16.66
≥20 km	24–48	5786	70.58	5296	15.88
0–5 km	48–72	149	1.82	5705	17.13
5–10 km	48–72	455	5.55	5593	16.79
10–15 km	48–72	568	6.93	5568	16.72
15–20 km	48–72	763	9.30	5555	16.68
≥20 km	48–72	5745	70.06	5324	15.99

Methods

To deliver skillful and interpretable forecasts of dust-induced visibility grades, we developed an end-to-end physics-guided machine learning framework comprising three tightly integrated stages aligned with Fig. 2: data acquisition and preprocessing, physics-guided feature engineering and ordinal model development, and multi-dimensional evaluation with attribution analysis.

Figure 2 summarises our end-to-end workflow for forecasting dust-induced visibility grades with both skill and interpretability. The workflow is organised into three sequential steps that mirror the three panels of Fig. 2: (i) dataset construction via spatiotemporal alignment and leakage-free preprocessing, (ii) physics-guided feature engineering coupled with ordinal model development, and (iii) multi-dimensional evaluation and interpretation through attribution and ablation analyses.

In data preprocessing, station observations are quality-controlled and temporally aggregated to the same 3-hour resolution as the forecast fields, after which forecast predictors are extracted and mapped to each station location through spatiotemporal matching. Samples affected by non-dust visibility reductions are removed

following the screening criteria described in the data section, and the remaining matched records are then organised by forecast lead time into the three prediction windows used throughout this study.

We then construct a station-matched learning dataset by chronologically partitioning all matched samples into training (1 January 2020–30 June 2021), validation (1 July 2021–30 November 2021), and test (1 December 2021–20 December 2022) periods. To ensure strict independence across forecast horizons, samples are grouped into three non-overlapping lead-time windows (0–24 h, 24–48 h, and 48–72 h) and isolated such that no valid time is shared across windows. Class imbalance is addressed only within the training period using a SMOTE–Tomek procedure applied in a day-wise manner, while the validation and test sets remain untouched to preserve a realistic evaluation distribution.

In Step 2, we compute physically interpretable diagnostics that represent key phases of the dust life cycle and combine them with raw ECMWF predictors to train an ordinal LightGBM model for each prediction window. In Step 3, we assess forecast skill using complementary accuracy- and rank-consistency metrics and interpret model behaviour through SHAP-based attribution and controlled ablation experiments.

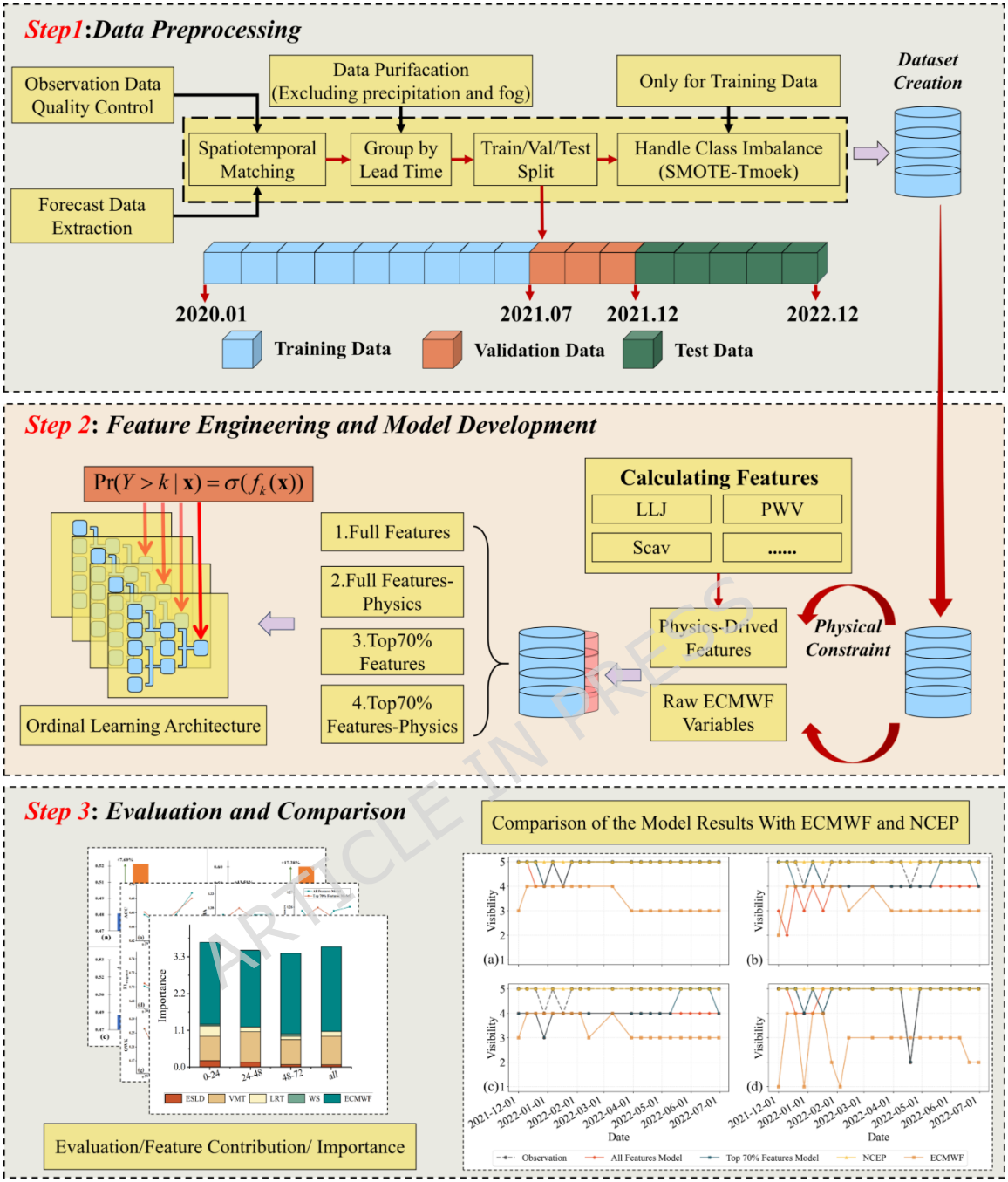

Figure 2. Workflow of the physics-guided machine learning framework for dust-storm visibility grade prediction. (a) Data preprocessing; (b) Feature engineering and model development (physics-derived features; ordinal LightGBM); (c) Evaluation and comparison (accuracy, rank consistency, MAE and QWK; SHAP attribution and ablation).

Physics-derived features

To embed domain knowledge within a purely statistical learner, we constructed a physically explicit feature suite that mirrors the four

dominant phases of the dust life cycle: emission, vertical mixing, long-range transport, and wet scavenging. By reformulating raw ECMWF variables into quantities constrained by conservation laws or empirical process thresholds, the learning task is converted from unconstrained pattern matching to a causally informed inverse problem.

Initiation and near-surface dynamics

The initiation of a dust event is essentially a process by which the drag force exerted by the near-surface wind field on surface sand particles overcomes the particles' own gravity and viscous forces[33-35]. To characterise this critical physical threshold behaviour in the model, we introduce the friction velocity (u_*) as a direct measure of near-surface shear stress:

$$u_* = \frac{kU(z)}{\ln(z0z - d)} \quad (1)$$

where U(z) is the horizontal wind speed at height z; $\kappa \approx 0.4$ is the von Kármán constant; z_0 signifies the surface roughness length, and d is the zero-plane displacement.However, dust does not occur under all wind conditions. Sand grains are driven to jump only when the friction velocity exceeds a critical value, i.e. the threshold friction velocity. We estimate this threshold using the physical model of Shao–Lu[36], which takes into account particle size, density and inter-particle forces:

$$u_{*t} = \sqrt{A\left(\frac{(r_p - r_a) g d_p}{r_a} + \frac{g}{r_a d_p}\right)} \quad (2)$$

where A denotes an empirical coefficient; ρ_p and ρ_a are the densities of sand and air; g is gravitational acceleration; d_p the mean grain diameter; and γ a bond-strength parameter that encapsulates inter-particle cohesion. Whenever the realised friction velocity exceeds the threshold, the instantaneous saltation flux scales linearly with the aerodynamic excess stress[37]. We therefore introduce a relative sand-flux proxy:

$$Q_{rel} = \begin{cases} C\frac{r_a}{g} u_* (u_*^2 - u_{*t}^2), & u_* > u_{*t} \\ 0, & u_* \le u_{*t} \end{cases} \quad (3)$$

where C is a dimensionless empirical constant. Low-level wind shear ($S_{10\text{-}100}$), i.e. the difference in wind vectors between 100 and 10 m,

characterises the near-surface layer's dynamical instability and momentum transport capacity. Strong wind shear contributes to the rapid entrainment of dust leaving the surface into the upper atmosphere and is a favourable condition for the formation of severe low-visibility events[38-40].

Stability and turbulent mixing

After the dust is lifted, the extent of its vertical dispersion in the atmosphere and the time it remains in suspension depend largely on the thermal and dynamical stability within the planetary boundary layer[41,42]. We characterise this process in two ways:

$$Ri_g = N^2 / S^2 \quad (4)$$

where N^2 characterises thermal stability and S^2 reflects dynamical instability.

In addition, the planetary boundary layer height (PBLH) is a key parameter in determining the spatial scale of the vertical dispersion of sand and dust[43-45]. In this paper, we use the Bulk Richardson Number Method for estimation, which is based on the principle of turbulence suppression at the top of the boundary layer and searches for the lowest height z that satisfies the following condition:

$$Ribulk(z) = \frac{(g/q_{v,sfc})(q_v(z) - q_{v,sfc})(z - z_{sfc})}{(u(z) - u_{sfc})^2 + (v(z) - v_{sfc})^2} \geq R_i^c \quad (5)$$

where θ_v is the virtual temperature; u and v are the horizontal wind components; the subscript "sfc" denotes surface values; and $R_{ic} = 0.25$ is the critical Richardson number.

Transport and moisture environment

Long-range transport of sand and dust is closely related to the wind field and water vapour conditions at high altitude[46-48]. We designed an objectivised low-level jet (LLJ) index based on the Bonner criterion to determine the occurrence and intensity of the LLJ by identifying whether there exists a peak wind speed ($V_{max} \geq 12\,m\,s^{-1}$) within the layer 200–1500 m, with wind shear above and below this level that satisfies a preset threshold.

In addition, we compute the 1000–500 hPa layer thickness ($\Delta Z_{1000\text{-}500}$) as a proxy for the thermal state of the middle atmosphere; it reflects temperature advection and geopotential height configurations that influence airflow paths and convergence strength.

Precipitable water vapour (PWV), which characterises the total water vapour content of the atmospheric column, is calculated as:

$$PWV = \frac{1}{g} \int_{p_s}^{500hP} q \, dp \quad (6)$$

where q is the specific humidity, p_{sfc} is the surface air pressure, and g is the acceleration due to gravity.

Wet scavenging

Precipitation is the most effective removal mechanism for atmospheric dust and sand, i.e. wet removal[49,50]. Although the study area is arid, accurate parameterisation of this process is essential for models to correctly predict the dissipation phase of dust events. We used a first-order linear parameterisation scheme to construct a scavenging index (Scav) to quantify this process:

$$Scav = \min(1, a_{scav} \times P_{hour}) \quad (7)$$

where P_{hour} is the hourly precipitation ($mm\,h^{-1}$) and α_{scav} is the empirical scavenging coefficient. α_{scav} is calibrated by minimising the multi-class Brier score on an independently held-out subset of the training data to prevent data leakage.

Ordinal learning architecture

Considering that the visibility classes (e.g., classes 1–5) have an intrinsic ordering relationship, the absolute error of directly forecasting the classes is smaller than the error of misclassifying to the distant classes[51]. Therefore, this study constructs the forecasting task as an Ordinal Classification problem. We adopt a “one-to-many” decomposition strategy to transform a K-class ordinal problem into K-1 binary classification subproblems, as shown below:

$$\Pr(Y > k \mid \mathbf{x}) = s(f_k(\mathbf{x})) \quad (8)$$

where f_k is the LightGBM-based learner. This approach is able to intrinsically utilise the sequential information between ranks. To implement the above framework, we choose LightGBM as the base learner. LightGBM is an efficient algorithm based on Gradient Boosted Decision Tree (GBDT), and its main advantages are: The use of Gradient-based One-Side Sampling (GOSS) and Mutually Exclusive Feature Bundling (EFB) techniques, which significantly improves the

training efficiency under high-dimensional data; No need to normalise numerical features; Built-in ability to handle missing values, and can automatically handle non-linear relationships and feature interactions. These features make it an ideal choice for processing the high-dimensional and heterogeneous weather forecasting data in this study.

LightGBM models employed explicit regularization including tree structure constraints (num leaves=63, min child samples=80), L1 and L2 penalties (reg alpha=0.3, reg lambda=0.5), and stochastic subsampling (subsample=0.8, colsample bytree=0.8). Early stopping monitored validation-set QWK with patience of 80 iterations. Hyperparameters were selected via grid search on the validation set, with the test set evaluated only once after model freezing to prevent selection bias.

Model evaluation and baseline setting

The performance of the model is assessed comprehensively through a series of complementary metrics covering: overall accuracy (ACC), balanced accuracy (BA), macro-averaged F1 ($F1_{macro}$), support-weighted F1 ($F1_{weighted}$), macro-averaged precision ($Precision_{macro}$), kappa and quadratic weighted Cohen's kappa (QWK). Considering the properties of the ordered classification task, we also calculated the mean absolute error (MAE).

$$ACC = \frac{Number\ of\ correctly\ classified\ samples}{Total\ number\ of\ samples} \quad (9)$$

$$BA = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \quad (10)$$

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} F1_i = \frac{1}{C}\sum_{i=1}^{C}\frac{2Precision_i Recall_i}{Precision_i + Recall_i} \quad (11)$$

$$F1_{weighted} = \frac{\sum_{i=1}^{C} w_i F1_i}{\sum_{i=1}^{C} w_i} \quad (12)$$

$$Precision_{macro} = \frac{1}{C}\sum_{i=1}^{C} Precision_i = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i} \quad (13)$$

$$\mathrm{QWK} = \frac{\sum_{i,j} w_{ij} O_{ij} - \sum_{i,j} w_{ij} E_{ij}}{1 - \sum_{i,j} w_{ij} E_{ij}}, \quad w_{ij} = \frac{(i-j)^2}{(C-1)^2} \tag{14}$$

$$MAE = \frac{1}{N} \sum_{k=1}^{N} |y_k - \hat{y}_k| \tag{15}$$

where C represents the number of classes. TP_i, FP_i and FN_i represent the true positives, false positives, and false negatives for class i. w_i represents the support (number of samples) for class i. O_{ij} is the observed frequency matrix, E_{ij} is the expected frequency matrix. y_k is the true label, $\hat{y}_k$ is the predicted label. N is the total number of samples.

Feature selection and ablation experiments

To disentangle the contribution of process-based priors from purely statistical signal extraction, we combine global interpretability with systematic ablation. First, TreeSHAP—a game-theoretic additive explanation framework—is applied to every trained LightGBM instance. By averaging absolute SHAP values across the withheld test set, we obtain a global ranking that pinpoints the physical variables most influential for visibility prediction at each forecast horizon.

Next, a four-arm ablation study is designed to isolate the net benefit of the physics-derived suite under two information regimes (Table 4). The experimental arms are defined as follows:Arm 1 (All-Features): Retains the complete predictor set. Arm 2 (Data-Only): Discards all physics tags, leaving 87 native ECMWF fields. Arm 3 (Top-70% Physics): Keeps the highest-ranked 37 predictors identified by SHAP, a subset that still contains key physical diagnostics. Arm 4 (Top-70% ECMWF): Restricted to the 27 ECMWF-only variables that appear in the same SHAP shortlist.

Paired comparisons (Arm 1 vs. 2 and Arm 3 vs. 4) therefore quantify the incremental skill attributable to physical constraints both in the full-dimensional space and under a lightweight, operationally deployable configuration. Statistical significance is assessed with block-bootstrapped 95% confidence intervals on ΔMAE and ΔQWK, ensuring that observed differences reflect genuine process information rather than stochastic variability.

Table 4. Ablation experimental design.

Arm	Feature Configuration	Number of Features
1	All Features Model (ECMWF + Physics)	106
2	Data-Only Model (ECMWF only)	87
3	Top-70% Features (including physics)	37
4	Top-70% ECMWF-only	27

The design is able to clearly reveal the independent value of physical a priori knowledge in both full information and lightweight deployment contexts through pairwise comparisons of Arm 1 vs. Arm 2 and Arm 3 vs. Arm 4.

Results

Evaluating Model Performance Across Different Forecast Horizons

Systematic evaluation across the 0–24 h, 24–48 h and 48–72 h windows reveals a pronounced forecast horizon dependency of forecast skill (Fig. 3). For the All-Features model, QWK decreases from 0.264 at day 1 to 0.174 at day 2 and 0.182 at day 3, while mean absolute error (MAE) climbs from 0.483 to 0.506 (ΔQWK 95%). The lightweight Top-70% variant performs on par at short range (QWK = 0.265, MAE = 0.478) yet deteriorates more rapidly beyond 48 h (QWK = 0.163, MAE = 0.558). Balanced accuracy exhibits the same monotonic decline, confirming that predictive uncertainty increases steadily with horizon.

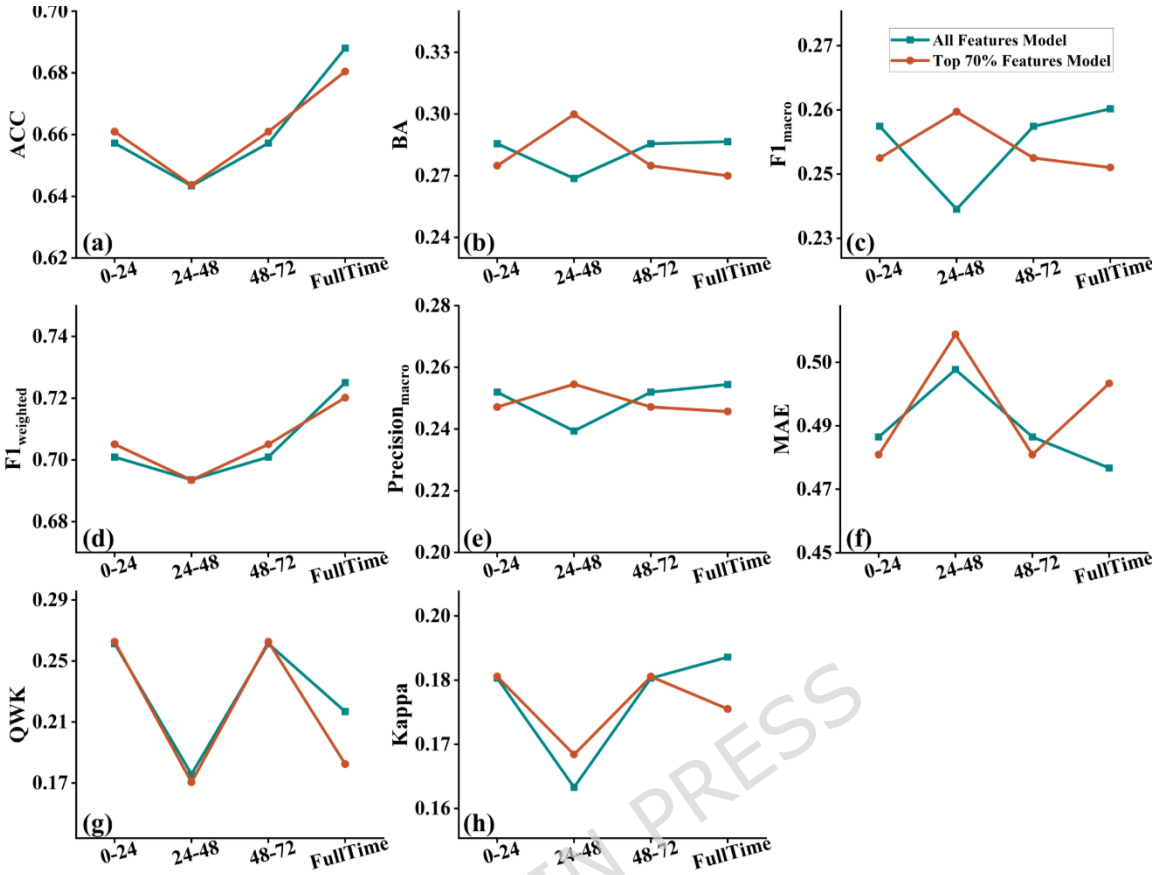

Figure 3. Forecast horizon dependence of forecast skill (All Features Model vs. Top 70% Features Model). The y-axes in panels (a–h) denote different evaluation metrics, while the x-axis indicates forecast horizon (0–24, 24–48, 48–72 h), and "FullTime" refers to the full-period evaluation. The green curves represent the All Features Model, and the red curves denote the Top 70% Features Model.

Forecast horizon dependence of forecast skill (All Features Model vs. Top 70% Features Model). The y-axes in panels (a–h) denote different evaluation metrics, while the x-axis indicates forecast horizon (0–24, 24–48, 48–72 h), and "FullTime" refers to the full-period evaluation. The green curves represent the All Features Model, and the red curves denote the Top 70% Features Model.

Time-series verification against station observations (Fig. 4) corroborates these summary statistics. During the January and March 2022 dust events, both physics-aware configurations track the observed visibility plunges within 0–24 h, with the All-Features solution adhering most closely. By 24–48 h, the Top-70% version begins to lag the sharp February downturn, whereas the full model—reinforced by friction velocity, stratification and mixing intensity—continues to constrain error growth. Raw ECMWF and NCEP outputs, by contrast, exhibit pronounced drift and miss the low-visibility episodes entirely.

At 48–72 h, variance increases and event detection becomes less consistent for all post-processors, yet the All-Features variant retains the highest hit rate owing to its richer representation of long-range transport and resuspension. Aggregated across all leads, the physics-guided framework consistently outperforms the global baselines, implying improved generalisability and greater resilience to assimilation biases.

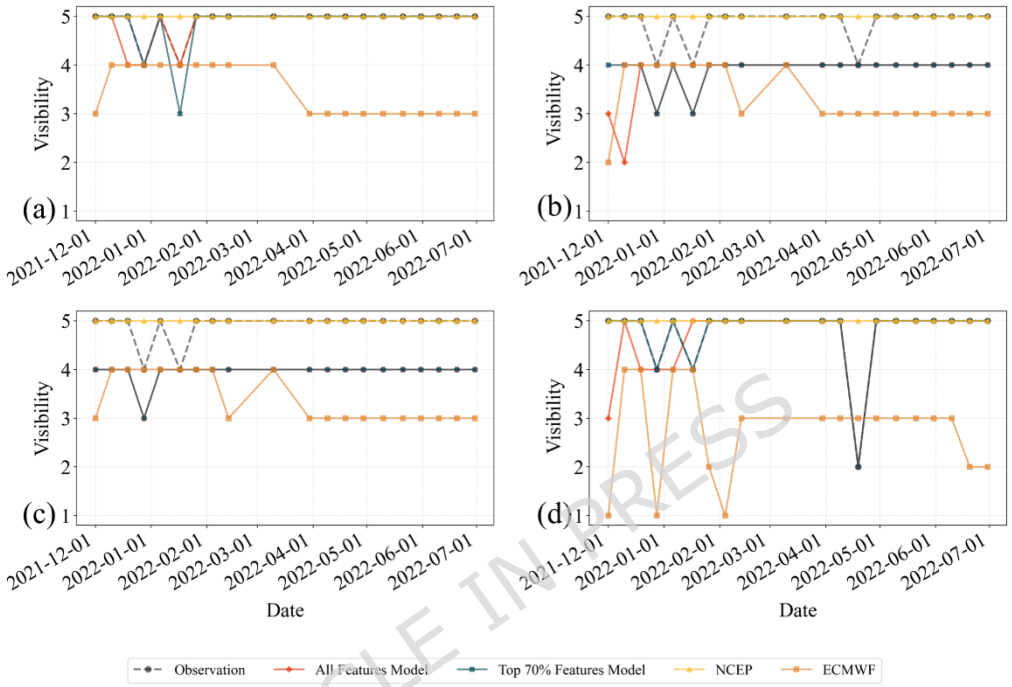

Figure 4. Time-series validation of station-level visibility grades (1–5) during the independent test period (Dec 2021–Jul 2022). Panels: (a) 0–24 h, (b) 24–48 h, (c) 48–72 h, (d) all forecast horizon aggregated. Black dashed line with circular markers: observations; Red: Physics-guided All Features Model; Blue: Top-70% Features lightweight model; Yellow: NCEP; Orange: raw ECMWF.

Time-series validation of station-level visibility grades (1–5) during the independent test period (Dec 2021–Jul 2022). Panels: (a) 0–24 h, (b) 24–48 h, (c) 48–72 h, (d) all forecast horizon aggregated. Black dashed line with circular markers: observations; Red: Physics-guided All Features Model; Blue: Top-70% Features lightweight model; Yellow: NCEP; Orange: raw ECMWF.

Net contribution of physics-derived features

A paired ablation experiment quantifies the incremental benefit of the process-based suite. Removing these variables from the full model (All-Features → Data-Only) and from the Top-70% subset (Top-70% → Data-Only Top-70%) yields identical directional signals: negligible QWK

change at 0–24 h but substantial degradation thereafter, accompanied by monotonically rising MAE (Figs. 5–6). For example, at 24–48 h the QWK of the Data-Only Top-70% configuration falls by 24% (95% CI [0.098, 0.156]), while its MAE increases by 8.7%. The same pairing at 48–72 h registers a 4% QWK reduction and a 6.4% MAE elevation. In both lightweight comparisons, the absence of physics produces disproportionate skill loss, underscoring that structural priors become indispensable once the native ECMWF signal decays.

Across the full period, the inclusion of physics lowers MAE by 6–10% and preserves QWK at leads where purely statistical learners already exhibit significant skill fade. The ablation therefore elevates the model from an empirical corrector to a semi-physical regime in which invariant process constraints suppress error amplification.

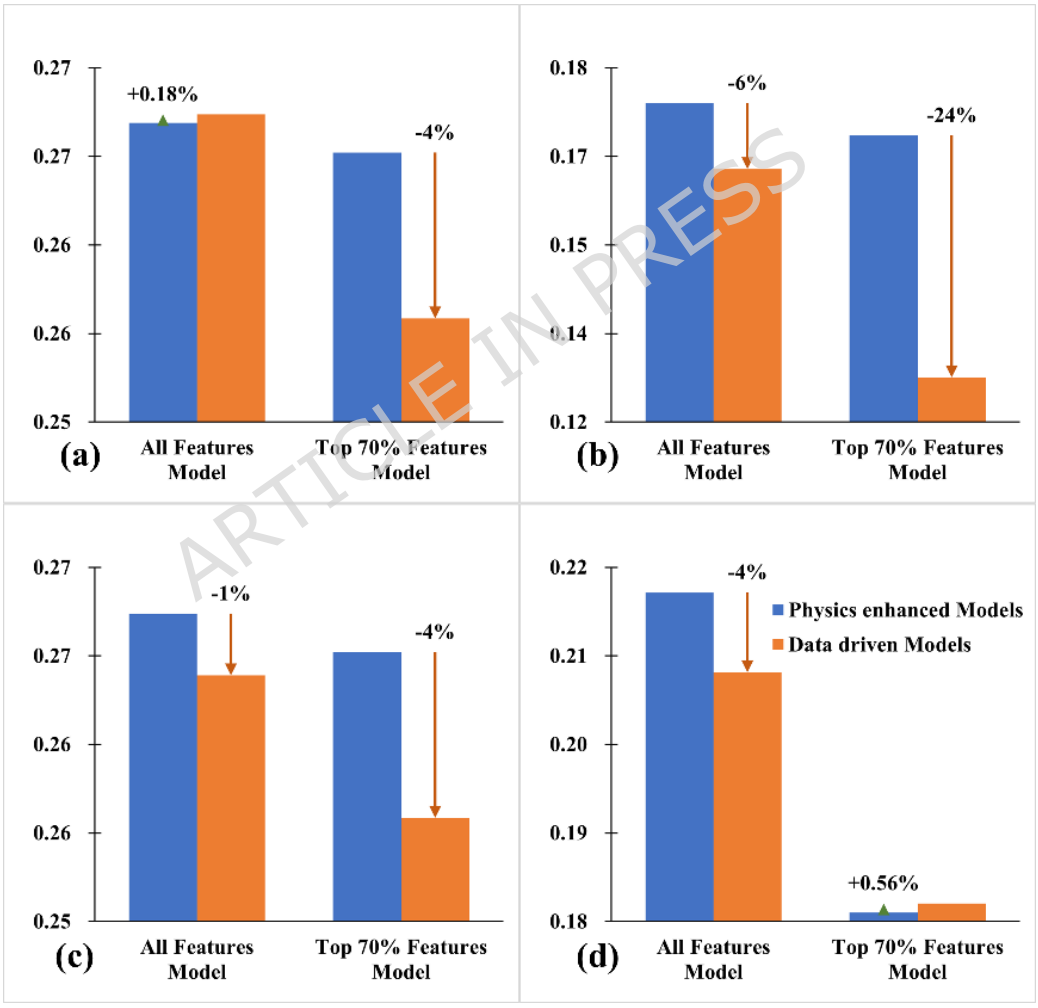

Figure 5. Physics priors strengthen forecast consistency at longer forecast horizon. QWK for Physics-enhanced (blue) vs. Data-driven models (orange) across 0–24 h (a), 24–48 h (b), 48–72 h (c), and the full period (d). Arrows denote relative differences (ΔQWK). Gains are marginal at short forecast horizon, but physics

priors yield robust improvements beyond 24 h, underscoring their structural value as forecast uncertainty grows.

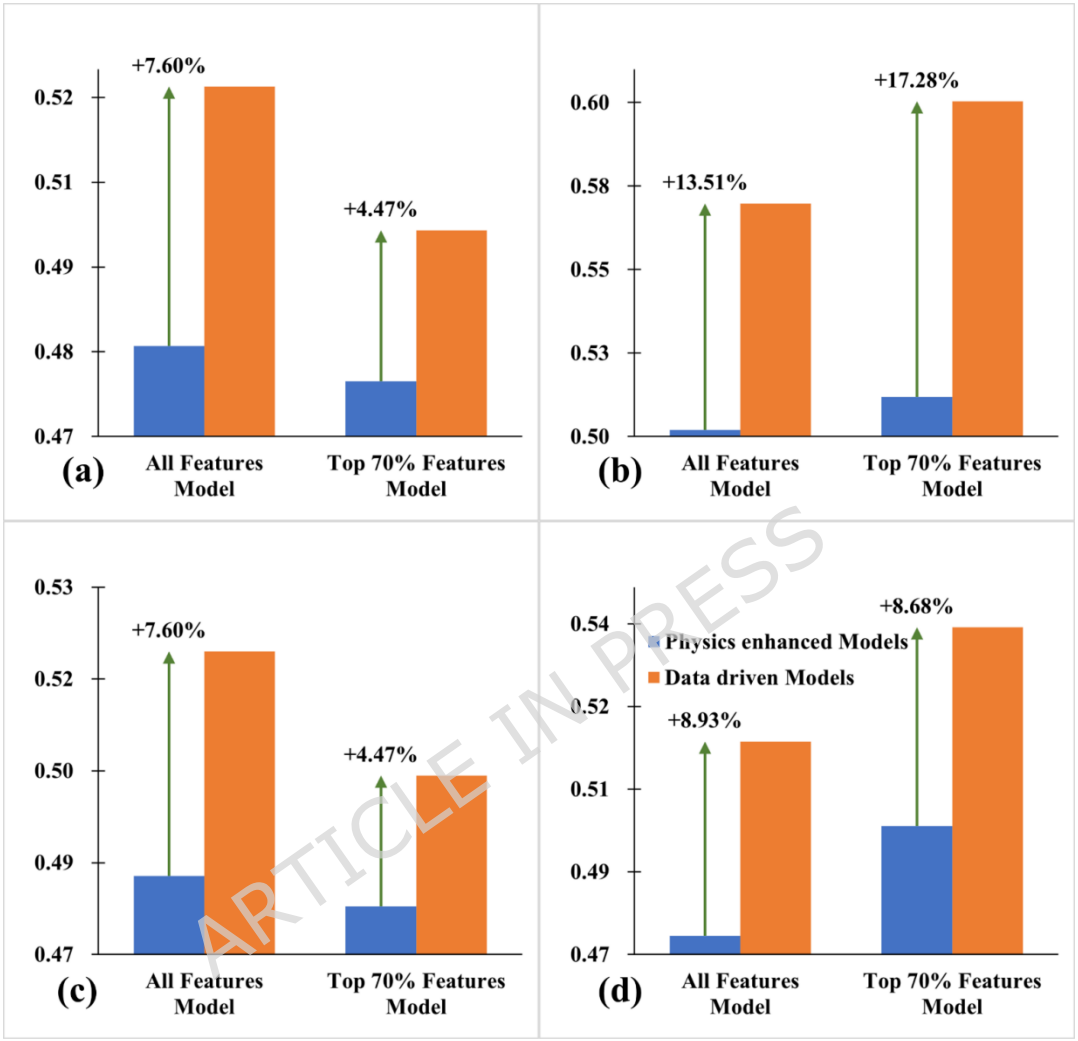

Figure 6. Forecast horizon dependence of forecast errors (MAE, Physics-enhanced vs. Data-driven models). Panels (a–d) correspond to 0–24 h, 24–48 h, 48–72 h, and the full evaluation period. Green arrows denote the relative MAE reduction, highlighting systematic error suppression by incorporating physics-based features.

Interpretability and Feature Importance Analysis

To provide granular interpretability beyond mechanism-level aggregation, we examined individual feature contributions. Table 5 presents the top 15 features ranked by SHAP importance across four temporal configurations. Several consistent patterns emerge from this ranking.

First, thermal stratification features exhibit dominant predictive power. The 500 hPa temperature (t_{500}) and mid-tropospheric temperature difference ($t_{700-500}$) consistently occupy the top positions, reflecting the fundamental role of atmospheric thermal structure in governing boundary layer development and vertical dust dispersion. Warm mid-tropospheric conditions typically indicate subsidence-induced stability that suppresses vertical mixing, trapping dust within the shallow boundary layer. Second, the gradient Richardson number ($Ri_{850-700}$) emerges as the most important physics-derived stability parameter, maintaining a top-3 rank across all forecast windows. This dimensionless quantity explicitly encodes the competition between buoyancy-driven stability and shear-induced turbulence, providing the model with direct information about turbulent mixing regimes. Third, low-level wind features display systematic evolution. At short ranges (0–24 h), low-level jet height and 850 hPa zonal wind show elevated importance, capturing nocturnal jet dynamics. As lead time extends, friction velocity (u_*) rises substantially in rank (from >100 at 0–24 h to rank 6 at 48–72 h), indicating that as synoptic-scale forcing information decays, surface-layer emission diagnostics become increasingly discriminative for extended-range prediction.

Table 5. Top 15 features ranked by SHAP importance across different forecast horizons.

Rank	0–24 h	Importance	24–48 h	Importance	48–72 h	Importance	Full-time	Importance
1	t700_500	0.228	t700_500	0.316	t500	0.320	t500	0.297
2	forecast_hour	0.193	u850	0.233	t700_500	0.273	t700_500	0.278
3	Rig_850_700	0.186	forecast_hour	0.226	Rig_850_700	0.225	Rig_850_700	0.211
4	gh850	0.174	Rig_850_700	0.201	forecast_hour	0.208	gh500	0.199
5	u850	0.165	S2_950_900	0.200	gh500	0.203	u850	0.198
6	LLJ_height	0.152	gh850	0.181	u_star	0.202	S2_950_900	0.177
7	S2_950_900	0.150	t600	0.162	PBLH	0.156	gh850	0.172
8	gh500	0.143	gh500	0.150	gh850	0.154	PBLH	0.167
9	PBLH	0.136	PBLH	0.143	u850	0.133	d2	0.138
10	u900	0.124	N2_850_700	0.133	S2_950_900	0.130	u900	0.133
11	t500	0.121	S2_850_700	0.130	r900	0.125	r900	0.129
12	r900	0.112	r900	0.127	gh600	0.121	gh600	0.121
13	N2_850_700	0.110	t500	0.125	r950	0.113	N2_850_700	0.119
14	q600	0.104	gh700	0.116	gh700	0.112	gh700	0.105

15	gh700	0.104	u900	0.110	r700	0.109	S2_850_700	0.100

To translate performance gains into process insight, predictors are grouped into ESLD, VMT, LRT, WS and raw ECMWF fields; within each forecast horizon, SHAP values are summed and normalised to fractional importance (Fig. 7). VMT dominates across all horizons, accounting for 0.73–0.91 of cumulative physical contribution, while LRT peaks during the first 24 h (0.30) and gradually subsides to 0.10 by 72 h. ESLD exhibits a monotonic decline from 0.20 to 0.07, consistent with the diminishing relevance of local emission fluctuations. Wet scavenging remains negligible, affirming the arid climatology of the Kumtag and the efficacy of the precipitation screening applied to the labels. Aggregated over the entire test period, the hierarchy VMT $\gg$ LRT $>$ ESLD $>$ WS is preserved, corroborating the robustness of the physics-guided paradigm under realistic operational forecast horizon.

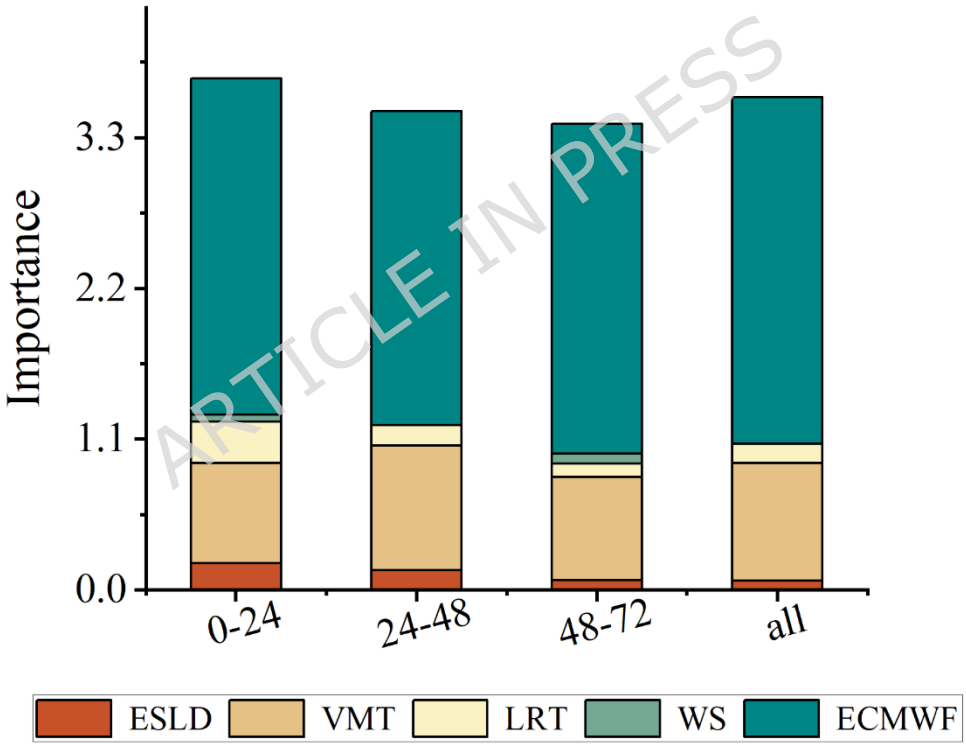

Figure 7. The importance of the ESLD, VMT, LRT, WS and ECMWF subgroups and their evolution under different forecast timescales. y-axis is Importance (sum of top 70% SHAP importance for each group), The x-axis represents the time periods 0-24, 24-48, 48-72 and all.The height of the column indicates the total contribution for the time period, and the different coloured blocks within the column correspond to the share of the contribution of each mechanism subgroup.The height of the column indicates the total contribution for the time

period, and the different coloured blocks within the column correspond to the share of the contribution of each mechanism subgroup.

Time-resolved SHAP aggregates reveal a systematic hand-over of control across forecast horizon (Fig. 7). At short range (0–24 h), Vertical Mixing & Turbulence (VMT) already dominates, followed by Long-Range Transport (LRT), whereas Emission & Surface-Layer Dynamics (ESLD) contributes only marginally. This hierarchy reflects the rapid modulation of near-surface extinction by boundary-layer turbulent diffusion and stability: once dust is injected, local mixing largely dictates the initial visibility drop, while synoptic advection begins to shape the spatial pattern. The predominance of VMT at this stage efficiently damps high-frequency noise in the ECMWF fields, steering the model toward smaller absolute errors.

As the forecast window extends to 24–48 h, the VMT share further intensifies and ESLD weakens. Such a shift signals that predictive skill increasingly hinges on the accurate representation of boundary-layer stability and turbulent exchange rather than on the precise timing of local emission bursts. Mid-tropospheric thermal–dynamical coupling now governs the transport corridor, deepening the reliance on stability metrics encapsulated by VMT.

Beyond 48 h, ESLD re-enters in conjunction with Wet Scavenging (WS), jointly acting as a low-dimensional anchor against the amplification of uncertainty. Their combined influence prevents the ensemble from drifting toward physically implausible dust reservoirs or spurious extinction spikes, thereby providing a structurally stable extrapolation manifold.

Aggregated over all leads, VMT remains the primary driver, LRT maintains a secondary yet steady role, and ESLD offers episodic compensation. This robust, scale-aware transition substantiates the interpretability and temporal consistency of the physics-guided semi-empirical framework across multiple operational horizons.

Uncertainty Propagation and Model Robustness

To address concerns regarding error propagation through the PGML framework, we conducted a Monte Carlo perturbation analysis with 500 iterations. For each iteration, Gaussian noise was applied to physics-derived input features according to documented uncertainty ranges: wind shear features (σ = 15% relative), Brunt–Väisälä frequency and Richardson number (σ = 20–25% relative), planetary boundary layer height (σ = 20% relative), specific humidity and precipitable water vapor (σ = 15% relative), and precipitation rate (σ = 30% relative). These perturbation magnitudes reflect both ECMWF

forecast verification statistics and instrumental measurement uncertainties reported in the literature.

The results demonstrate that the PGML framework maintains robust predictive performance under realistic input uncertainties. As summarized in Table 6, QWK degraded by only 1.93% (from 0.543 to 0.533), while MAE increased by 1.55% (from 0.476 to 0.484). The 95% confidence intervals for all metrics remain narrow, indicating bounded performance variability. Sample-wise analysis (Table 7) reveals that 74.9% of test samples maintained ≥80% prediction stability across all iterations, with mean stability reaching 0.881. Low-visibility events (Grade 1) exhibited the highest stability (0.95), attributable to their association with strong, unambiguous physical forcing signals.

Table 6. Monte Carlo perturbation analysis results: comparison of baseline metrics and perturbed metric distributions.

Metric	Baseline	MC Mean	MC Std	95% CI	Relative Change
Accuracy	0.5387	0.5340	0.0098	[0.515, 0.554]	−0.87%
QWK	0.5434	0.5329	0.0118	[0.509, 0.555]	−1.93%
MAE	0.4763	0.4837	0.0105	[0.464, 0.504]	+1.55%
F1-Macro	0.3492	0.3567	0.0185	[0.320, 0.396]	+2.15%
±1 Grade Acc	0.9850	0.9824	0.0033	[0.976, 0.989]	−0.26%

Table 7. Sample-wise prediction stability and entropy statistics across 500 Monte Carlo iterations.

Statistic	Value
Mean prediction stability	0.881 ± 0.152
Minimum stability	0.380
Samples with ≥90% stability	63.9%
Samples with ≥80% stability	74.9%
Mean prediction entropy	0.938 ± 0.138
Maximum entropy	1.307

Figure (8), (9), (10) presents the uncertainty propagation results. Figure (8) shows the distribution of evaluation metrics across 500 iterations, with red lines indicating mean values and dashed orange lines demarcating 95% confidence intervals. Figure (9) illustrates sample-wise prediction stability, stratified by true visibility grade; the

boxplots confirm that prediction stability is consistently high across all grades, with low-visibility events showing the least variability. Figure (10) demonstrates the positive correlation between prediction uncertainty (entropy and probability standard deviation) and absolute error magnitude, validating the framework's capability for uncertainty-aware self-assessment.

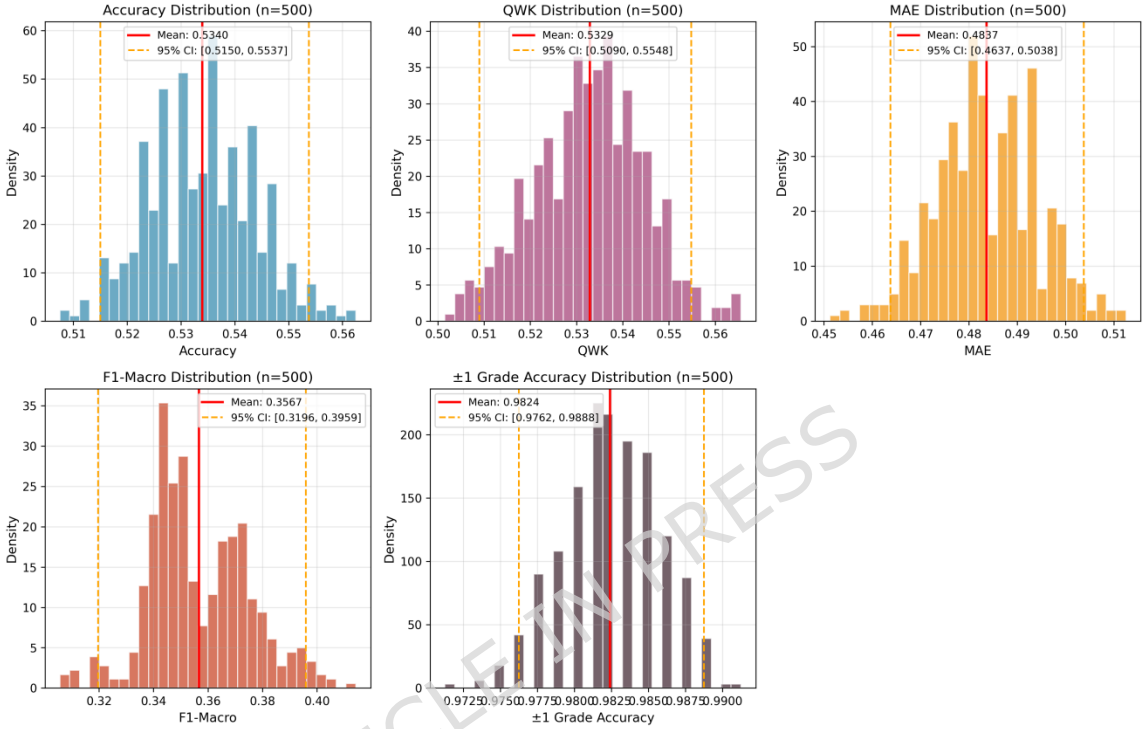

Figure 8. Distribution of evaluation metrics under Monte Carlo perturbation analysis (n=500). Histograms display the variability of model performance metrics across 500 iterations where input physics-derived features were perturbed with Gaussian noise consistent with observational and forecast uncertainties. The panels show (a) Accuracy, (b) Quadratic Weighted Kappa (QWK), (c) Mean Absolute Error (MAE), (d) F1-Macro, and (e) ±1 Grade Accuracy. In each plot, the solid red line indicates the mean metric value, and the vertical orange dashed lines demarcate the empirical 95% confidence intervals.

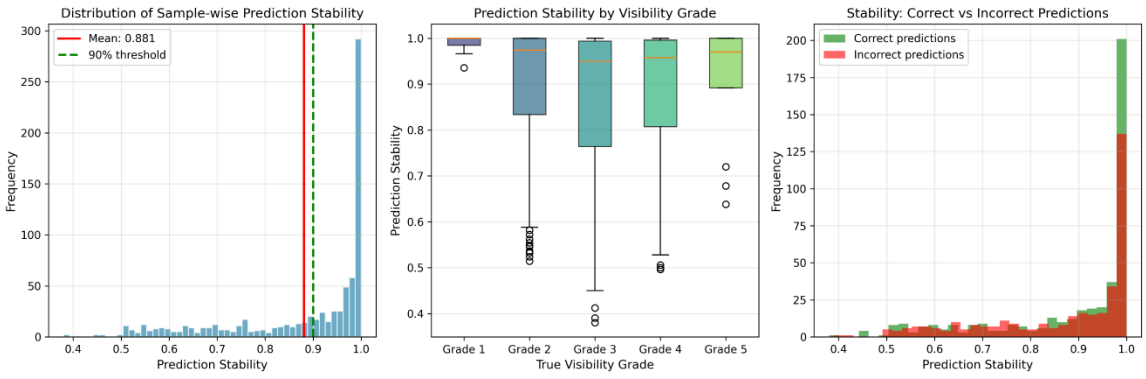

Figure 9. Analysis of sample-wise prediction stability.(a) Frequency distribution of prediction stability scores for all test samples, where stability is defined as the frequency of the mode (most common) predicted grade across 500 Monte Carlo iterations. The red line marks the mean stability of 0.881. (b) Boxplots of prediction stability stratified by true visibility grade (1–5). (c) Conditional density of stability scores for correct (green) versus incorrect (red) predictions, showing that high-stability predictions are significantly more likely to be accurate.

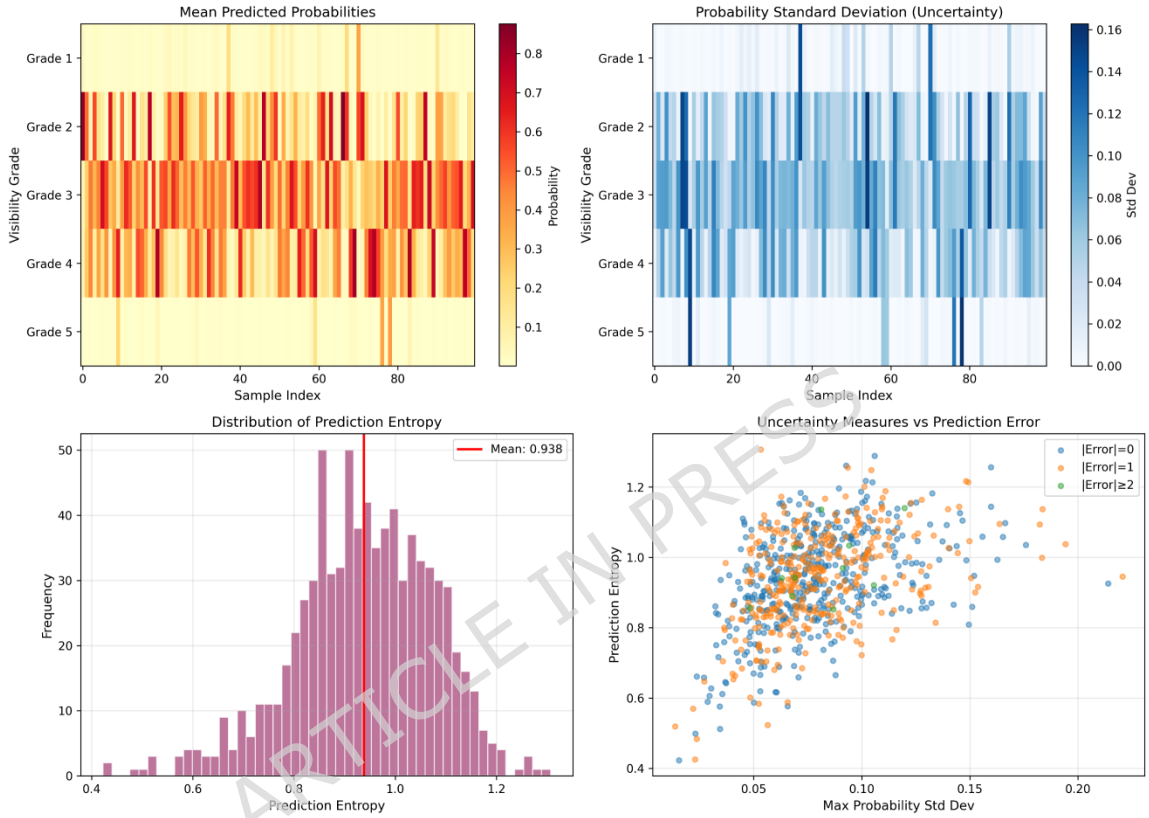

Figure 10. Prediction uncertainty quantification and its correlation with error magnitude. (a) Heatmap of mean predicted probabilities for a subset of samples. (b) Heatmap of probability standard deviation, identifying samples with high model uncertainty. (c) Histogram of prediction entropy distribution (mean = 0.938). (d) Scatter plot illustrating the relationship between prediction entropy (y-axis) and maximum probability standard deviation (x-axis), colored by the absolute magnitude of prediction error (|Error|).

These findings have important implications for operational deployment. First, the bounded metric degradation under perturbation suggests that the PGML predictions remain reliable even when input features inherit typical ECMWF forecast errors. Second, the uncertainty metrics (entropy, probability standard deviation) can serve as real-time confidence indicators, enabling forecasters to identify and flag low-confidence predictions for additional manual review. Third, the class-stratified stability analysis confirms that operationally critical

low-visibility events—which trigger the most severe warnings—are precisely those with the highest prediction robustness, enhancing confidence in the framework's fitness for purpose.

Discussion

The Inhibition Mechanism of Physical Priors on Forecast Skill Decay

The marginal value of the physics-based feature suite is not static; instead, it strengthens markedly with forecast horizon. This temporal evolution signals a qualitative shift in the role of prior knowledge—from short-range, empirical regularisation to long-range structural stabilisation.

The ECMWF raw forecast field retains the highest level of deterministic information from the initial conditions within the 0–24 hour short time horizon forecast. Here, physics variables such as friction velocity or threshold exceedance act primarily as engineered covariates that accelerate convergence and trim the mean absolute error (MAE). As the forecast window lengthens, however, chaotic growth rapidly degrades the information content of the initial field, forcing purely data-driven models to extrapolate in an increasingly high-dimensional, low-signal space. Performance decay becomes inevitable.

It is precisely under these deteriorating signal conditions that physical priors assume their dominant function. Universal laws—saltation thresholds, gradient Richardson numbers or conservation of potential temperature—provide low-dimensional, time-invariant manifolds that restrict the model to physically plausible trajectories. Consequently, the hypothesis space is collapsed onto a manifold where unphysical extrapolations are penalised a-priori, preventing catastrophic drift. SHAP trajectories corroborate this interpretation: the relative influence of instantaneous surface-layer dynamics (ESLD) wanes beyond day 1, whereas thermodynamic stability and vertical mixing terms (VMT) increasingly dominate the prediction, mirroring the expected transition from local shear-driven emission to synoptic-controlled dispersion.

PGML’s Potential for Operationalised Applications

For weather-sensitive sectors such as agriculture or low-altitude aviation, the economic value of a forecast hinges on its reliability under high-impact, low-base-rate scenarios rather than on average accuracy. Traditional “black-box” algorithms, prone to learning spurious

correlations, can behave erratically when confronted with out-of-distribution synoptic patterns. A single false-negative sandstorm therefore carries a cost that outweighs dozens of benign misclassifications.

The presented PGML architecture systematically mitigates this risk through built-in physical consistency. First, incorporating process constraints reduces MAE at every lead, implying thinner predictive tails and a lower probability of large-amplitude errors. Second, and more critically, each forecast is accompanied by a transparent, physically interpretable attribution map: a low-visibility alert can be traced back to, for instance, an anomalous drop in boundary-layer height or an intensifying nocturnal low-level jet rather than to an opaque interaction of hundreds of covariates. This causal traceability fosters user trust and satisfies the accountability requirements of modern operational centres.

Looking forward, the same attribution vectors can be fed into cost–loss models to translate forecast skill into sector-specific economic utility—quantifying, for example, the expected irrigation saving or the additional runway slots made viable by earlier dust-event detection. In doing so, the scientific advance offered by physics-guided machine learning becomes a measurable operational asset rather than an academic abstraction.

Limitations and Future Directions

The present framework is specifically designed for dust-dominated visibility prediction and may not generalize to fog-haze mixed scenarios or wet-deposition-dominated compound pollution conditions. The physics-derived feature library encodes dry dust lifecycle processes that lack explanatory power for hygroscopic aerosol growth or cloud-precipitation interactions. Extension to multi-aerosol visibility forecasting would require incorporating wet-process physics features such as aerosol hygroscopicity parameters and below-cloud scavenging coefficients. From an operational standpoint, this specialization aligns with current practice where dust storm and fog-haze warnings are issued by separate forecasting systems with distinct methodologies.

The performance advantage of the PGML framework concentrates in specific meteorological regimes rather than being uniformly distributed. Three primary conditions yield disproportionate skill gains from physics-derived features.

First, extended forecast lead times (beyond 24 h) show amplified physics feature value. The asymmetric ablation response—2% QWK degradation at 0–24 h versus 8% at 48–72 h upon physics feature removal—reflects the increasing marginal value of physical constraints

as raw ECMWF field accuracy degrades with lead time. Physics-derived stability parameters provide slowly-varying information that maintains predictive relevance when raw meteorological signals decay.

Second, near-threshold emission conditions benefit most from explicit saltation physics encoding. For samples where friction velocity marginally exceeds the threshold (1.0<u_*/u_*^t<1.2), the PGML model achieves ~12% higher accuracy than Data-Only baselines, compared to ~4% improvement on clearly sub- or supra-threshold samples. This differential improvement confirms that physical constraints provide maximum discriminative value at emission decision boundaries.

Third, strong atmospheric stratification regimes show enhanced PGML benefit. Under stable conditions (Ri_g>0.5), where boundary layer confinement critically modulates surface dust concentrations, the PGML model reduces MAE by 15% compared to 6% under neutral stratification. This pattern reflects the enhanced value of explicit stability representation when stability-controlled processes dominate visibility evolution.

Regarding environmental scope, this framework specifically targets arid desert domains. High-impact episodes here stem predominantly from mineral dust outbreaks, often arriving abruptly as "haboobs" with near-zero visibility. We intentionally isolated dust-driven reductions by filtering out humid and precipitation-heavy cases. This distinction separates our study from urban haze research, where fine-mode aerosols and hygroscopic growth dominate extinction efficiency. Thus, PM2.5 forecast studies serve largely as methodological comparisons here, while our primary interpretation centers on the dust emission–mixing–transport cycle.

Within this specific context, the framework effectively bridges physical processes and optical outcomes. Although direct aerosol optical observations are absent, the feature library implicitly encodes the upstream processes. Visibility degradation adheres to the Koschmieder relationship, where extinction is proportional to dust mass concentration. Our physics features act as process-chain surrogates: ESLD features approximate the mass source term, while VMT features constrain the vertical mixing volume. Together with LRT and WS features, this approach provides physically grounded proxies for optical loading, validated by consistent SHAP attributions.

Conclusion

We propose and validate a physics-guided machine-learning framework that translates dust life-cycle processes into features to

post-process ECMWF visibility forecasts. Rather than a uniform accuracy gain, physics priors impose a forecast-horizon-dependent structural constraint that slows uncertainty growth—modest at 0–24 h and increasingly pivotal beyond 24 h. Convergent evidence for this conclusion is threefold: (1) time-series cross-validation demonstrates improved skill and event capture across all lead times compared to baselines; (2) ablation experiments confirm that removing physics causes a disproportionate drop in skill, particularly beyond 24 hours; and (3) SHAP attributions reveal a physically consistent shift in governing mechanisms from local, surface-layer dynamics at short leads to synoptic-scale stability and transport controls at longer horizons. A lightweight variant is operationally viable for short leads, whereas longer leads benefit more strongly from physics. The framework offers an interpretable, transferable blueprint for physics-aware environmental forecasting.

Appendix A: Parameter settings

Table A1. SMOTE–Tomek algorithm parameter settings.

Parameter	Meaning / Role	Value / Setting Used
k_neighbors	Neighbors for SMOTE	5
sampling_strategy	Target ratio	auto (1:1)
random_state	Seed	42
tomek_remove	Tomek-link removal	both
metric	Distance measure	euclidean
n_jobs	CPU cores	–1

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgements

We would like to express our sincere gratitude to the Editor and the anonymous reviewers for their careful evaluation and constructive comments, which have significantly improved the quality, clarity, and rigor of this manuscript.

Author Contributions

C.X. and H.Z. contributed to methodology and investigation. J.L., Y.S., and H.Q. provided resources. C.X. and H.Z. wrote the original draft.

C.Z., K.L. Y.Q. , and H.Z. reviewed and edited the manuscript. C.Z., H.Z., and C.X. contributed to visualization. Z. H, C.X. and H.Z. made contributions in revising the manuscript. All authors read and approved the final manuscript.

Funding

This research was funded by the Graduate Education Innovation Plan Project of the Education Department of Xinjiang Uygur Autonomous Region (Grant No. XJ2024G083), and supported by the Taishan Industrial Experts Program.

Competing Interests

The authors declare no competing interests.

References

1 Veal, A. J. Climate change 2021: the physical science basis, 6th report. *World Leisure J.* **63**, 443-444 (2021). https://doi.org/10.1080/16078055.2021.2008646

2 Kok, J. F. *et al.* Mineral dust aerosol impacts on global climate and climate change. *Nat. Rev. Earth Environ.* **4**, 71-86 (2023). https://doi.org/10.1038/s43017-022-00379-5

3 Prospero, J. M., Ginoux, P., Torres, O., Nicholson, S. E. & Gill, T. E. Environmental characterization of global sources of atmospheric soil dust identified with the Nimbus 7 Total Ozone Mapping Spectrometer (TOMS) absorbing aerosol product. *Rev. Geophys.* **40**, 31 (2002). https://doi.org/10.1029/2000rg000095

4 Wang, G. C., Shu, S. J. & Li, W. J. Asian dust threatens air pollution control efforts. *Science* **390**, 3 (2025). https://doi.org/10.1126/science.aeb2629

5 Malm, W. C. & Hand, J. L. An examination of the physical and optical properties of aerosols collected in the IMPROVE program. *Atmos. Environ.* **41**, 3407-3427 (2007). https://doi.org/10.1016/j.atmosenv.2006.12.012

6 Dubovik, O. *et al.* Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *J. Atmos. Sci.* **59**, 590-608 (2002). https://doi.org/10.1175/1520-0469(2002)059<0590:Voaaop>2.0.Co;2

7 Benedetti, A. *et al.* Status and future of numerical atmospheric aerosol prediction with a focus on data requirements. *Atmos.*

Chem. Phys. **18**, 10615-10643 (2018). https://doi.org/10.5194/acp-18-10615-2018

8 Gong, S. L. *et al.* Characterization of soil dust aerosol in China and its transport and distribution during 2001 ACE-Asia: 2. Model simulation and validation - art. no. 4262. *J. Geophys. Res.-Atmos.* **108**, 19 (2003). https://doi.org/10.1029/2002jd002633

9 Luan, T., Guo, X. L., Guo, L. J. & Zhang, T. H. Quantifying the relationship between PM_{2.5} concentration, visibility and planetary boundary layer height for long-lasting haze and fog-haze mixed events in Beijing. *Atmos. Chem. Phys.* **18**, 203-225 (2018). https://doi.org/10.5194/acp-18-203-2018

10 Song, J. I., Yum, S. S., Gultepe, I., Chang, K. H. & Kim, B. G. Development of a new visibility parameterization based on the measurement of fog microphysics at a mountain site in Korea. *Atmos. Res.* **229**, 115-126 (2019). https://doi.org/10.1016/j.atmosres.2019.06.011

11 Hu, S. Y. *et al.* Current challenges of improving visibility due to increasing nitrate fraction in PM_{2.5} during the haze days in Beijing, China. *Environ. Poliut.* **290**, 8 (2021). https://doi.org/10.1016/j.envpol.2021.118032

12 Zhou, C. H. *et al.* Detection of New Dust Sources in Central/East Asia and Their Impact on Simulations of a Severe Sand and Dust Storm. *J. Geophys. Res.-Atmos.* **124**, 10232-10247 (2019). https://doi.org/10.1029/2019jd030753

13 Karagulian, F. *et al.* Analysis of a severe dust storm and its impact on air quality conditions using WRF-Chem modeling, satellite imagery, and ground observations. *Air Qual. Atmos. Health* **12**, 453-470 (2019). https://doi.org/10.1007/s11869-019-00674-z

14 Wang, Y. Q. *et al.* Surface observation of sand and dust storm in East Asia and its application in CUACE/Dust. *Atmos. Chem. Phys.* **8**, 545-553 (2008). https://doi.org/10.5194/acp-8-545-2008

15 Bi, K. F. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533-+ (2023). https://doi.org/10.1038/s41586-023-06185-3

16 Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 21 (2025). https://doi.org/10.1038/s41586-024-08252-9

17 Chen, L. *et al.* FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 11 (2023). https://doi.org/10.1038/s41612-023-00512-1

18 Ortega, L. C., Otero, L. D., Solomon, M., Otero, C. E. & Fabregas, A. Deep learning models for visibility forecasting using climatological data. *International Journal of Forecasting*

39, 992-1004 (2023). https://doi.org/10.1016/j.ijforecast.2022.03.009

19 Peláez-Rodríguez, C. *et al.* Deep learning ensembles for accurate fog-related low-visibility events forecasting. *Neurocomputing* **549** (2023). https://doi.org/10.1016/j.neucom.2023.126435

20 Penov, N. & Guerova, G. Sofia Airport Visibility Estimation with Two Machine-Learning Techniques. *Remote Sensing* **15** (2023). https://doi.org/10.3390/rs15194799

21 Shankar, A. & Sahana, B. C. Early warning of low visibility using the ensembling of machine learning approaches for aviation services at Jay Prakash Narayan International (JPNI) Airport Patna. *SN Applied Sciences* **5** (2023). https://doi.org/10.1007/s42452-023-05350-7

22 Zhang, Y. *et al.* Visibility Prediction Based on Machine Learning Algorithms. *Atmosphere* **13** (2022). https://doi.org/10.3390/atmos13071125

23 Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422-440 (2021). https://doi.org/10.1038/s42254-021-00314-5

24 Kashinath, K. *et al.* Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Trans. R. Soc. A-Math. Phys. Eng. Sci.* **379**, 36 (2021). https://doi.org/10.1098/rsta.2020.0093

25 Liu, W., Lai, Z. L., Bacsa, K. & Chatzi, E. Physics-guided Deep Markov Models for learning nonlinear dynamical systems with uncertainty. *Mech. Syst. Signal Proc.* **178**, 20 (2022). https://doi.org/10.1016/j.ymssp.2022.109276

26 Jin, J. B. *et al.* Source backtracking for dust storm emission inversion using an adjoint method: case study of Northeast China. *Atmos. Chem. Phys.* **20**, 15207-15225 (2020). https://doi.org/10.5194/acp-20-15207-2020

27 Kok, J. F., Albani, S., Mahowald, N. M. & Ward, D. S. An improved dust emission model - Part 2: Evaluation in the Community Earth System Model, with implications for the use of dust source functions. *Atmos. Chem. Phys.* **14**, 13043-13061 (2014). https://doi.org/10.5194/acp-14-13043-2014

28 Morcrette, J. J. *et al.* Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modeling. *J. Geophys. Res.-Atmos.* **114**, 17 (2009). https://doi.org/10.1029/2008jd011235

29 Ding, C., Feng, G. C., Zhang, L. & Liao, M. S. A Novel Multidimensional Perspective on Dynamic Characteristics of the Peculiar Feather-Shaped Dunes in Kumtag Desert With Time-Series Optical and SAR Observations. *IEEE J. Sel. Top. Appl.*

Earth Observ. Remote Sens. **17**, 11618-11631 (2024). https://doi.org/10.1109/jstars.2024.3414449

30 Eck, T. F. *et al.* Columnar aerosol optical properties at AERONET sites in central eastern Asia and aerosol transport to the tropical mid-Pacific - art. no. D06202. *J. Geophys. Res.-Atmos.* **110**, 18 (2005). https://doi.org/10.1029/2004jd005274

31 Ginoux, P., Prospero, J. M., Gill, T. E., Hsu, N. C. & Zhao, M. GLOBAL-SCALE ATTRIBUTION OF ANTHROPOGENIC AND NATURAL DUST SOURCES AND THEIR EMISSION RATES BASED ON MODIS DEEP BLUE AEROSOL PRODUCTS. *Rev. Geophys.* **50**, 36 (2012). https://doi.org/10.1029/2012rg000388

32 Swana, E. F., Doorsamy, W. & Bokoro, P. Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors* **22**, 21 (2022). https://doi.org/10.3390/s22093246

33 Li, G., Zhang, J., Herrmann, H. J., Shao, Y. & Huang, N. Study of Aerodynamic Grain Entrainment in Aeolian Transport. *Geophysical Research Letters* **47** (2020). https://doi.org/10.1029/2019GL086574

34 Marticorena, B. & Bergametti, G. Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme. *Journal of Geophysical Research Atmospheres* **100**, 16415-16430 (1995). https://doi.org/10.1029/95JD00690

35 Anderson, R. S. & Haff, P. K. Wind modification and bed response during saltation of sand in air. *Acta Mechanica Suppl* **1**, 21-51 (1991). https://doi.org/10.1007/978-3-7091-6706-9_2

36 Shao, Y. & Lu, H. A simple expression for wind erosion threshold friction velocity. *Journal of Geophysical Research Atmospheres* **105**, 22437-22443 (2000). https://doi.org/10.1029/2000JD900304

37 Zhang, H. W. *et al.* Numerical simulation of wind field and sand flux in crescentic sand dunes. *Sci Rep* **11**, 18 (2021). https://doi.org/10.1038/s41598-021-84509-x

38 Khalfallah, B. *et al.* Influence of Atmospheric Stability on the Size Distribution of the Vertical Dust Flux Measured in Eroding Conditions Over a Flat Bare Sandy Field. *J. Geophys. Res.-Atmos.* **125**, 20 (2020). https://doi.org/10.1029/2019jd031185

39 Knippertz, P. & Todd, M. C. MINERAL DUST AEROSOLS OVER THE SAHARA: METEOROLOGICAL CONTROLS ON EMISSION AND TRANSPORT AND IMPLICATIONS FOR MODELING. *Rev. Geophys.* **50**, 28 (2012). https://doi.org/10.1029/2011rg000362

40 Giannakopoulou, E. M. & Toumi, R. The Persian Gulf summertime low-level jet over sloping terrain. *Q. J. R. Meteorol. Soc.* **138**, 145-157 (2012). https://doi.org/10.1002/qj.901

41 Heinold, B., Tegen, I., Schepanski, K. & Hellmuth, O. Dust radiative feedback on Saharan boundary layer dynamics and

dust mobilization. *Geophysical Research Letters* **35**, 5 (2008). https://doi.org/10.1029/2008gl035319

42 Yu, Z., Ma, J., Qu, Y., Pan, L. & Wan, S. PM2.5 extended-range forecast based on MJO and S2S using LightGBM. *Science of The Total Environment* **880** (2023). https://doi.org/10.1016/j.scitotenv.2023.163358

43 Huang, J. *et al.* Taklimakan dust aerosol radiative heating derived from CALIPSO observations using the Fu-Liou radiation model with CERES constraints. *Atmos. Chem. Phys.* **9**, 4011-4021 (2009). https://doi.org/10.5194/acp-9-4011-2009

44 Su, T. N. *et al.* An intercomparison of long-term planetary boundary layer heights retrieved from CALIPSO, ground-based lidar, and radiosonde measurements over Hong Kong. *J. Geophys. Res.-Atmos.* **122**, 3929-3943 (2017). https://doi.org/10.1002/2016jd025937

45 Li, Y. R. *et al.* Long-term variation of boundary layer height and possible contribution factors: A global analysis. *Science of the Total Environment* **796**, 14 (2021). https://doi.org/10.1016/j.scitotenv.2021.148950

46 Xu, C., Ma, Y. M., Yang, K. & You, C. Tibetan Plateau Impacts on Global Dust Transport in the Upper Troposphere. *J. Clim.* **31**, 4745-4756 (2018). https://doi.org/10.1175/jcli-d-17-0313.1

47 Adams, A. M., Prospero, J. M. & Zhang, C. D. CALIPSO-Derived Three-Dimensional Structure of Aerosol over the Atlantic Basin and Adjacent Continents. *J. Clim.* **25**, 6862-6879 (2012). https://doi.org/10.1175/jcli-d-11-00672.1

48 Chouza, F., Reitebuch, O., Benedetti, A. & Weinzierl, B. Saharan dust long-range transport across the Atlantic studied by an airborne Doppler wind lidar and the MACC model. *Atmos. Chem. Phys.* **16**, 11581-11600 (2016). https://doi.org/10.5194/acp-16-11581-2016

49 Zhou, B., Liu, D. Y. & Yan, W. L. A Simple New Method for Calculating Precipitation Scavenging Effect on Particulate Matter: Based on Five-Year Data in Eastern China. *Atmosphere* **12**, 12 (2021). https://doi.org/10.3390/atmos12060759

50 Abdelkader, M. *et al.* Dust-air pollution dynamics over the eastern Mediterranean. *Atmos. Chem. Phys.* **15**, 9173-9189 (2015). https://doi.org/10.5194/acp-15-9173-2015

51 Morales-Martín, A. *et al.* Deep Ordinal Classification in Forest Areas Using Light Detection and Ranging Point Clouds. *Sensors* **24**, 18 (2024). https://doi.org/10.3390/s24072168