machine learning & knowledge extraction

MDPI

Article

DM-CPS: A Dual-Head Mass-Consistent Surrogate for Accelerating HYSPLIT $PM_{2.5}$ Dispersion Modeling

Farida Abdoldina [1,*], Azamat Serek [2,*] and Guido Cervone [3,4]

[1] Institute of Automation and Information Technologies, Satbayev University (Kazakh National Research Technical University), Almaty 050013, Kazakhstan
[2] School of Software Engineering, Astana IT University, Astana 010000, Kazakhstan
[3] Institute for Computational and Data Sciences (ICDS), The Pennsylvania State University, University Park, State College, PA 16802, USA; guc18@psu.edu
[4] Department of Geography, The Pennsylvania State University, University Park, State College, PA 16802, USA
* Correspondence: f.abdoldina@satbayev.university (F.A.); azamat.serek@astanait.edu.kz (A.S.)

Abstract

Large ensembles of atmospheric dispersion simulations are necessary for uncertainty quantification, source attribution, and emergency response, however, physics-based models like HYSPLIT might be too computationally expensive for operational use. In addition, they have not fully benefited from recent advances in artificial-intelligence-based surrogate modelling, which offer the potential to reduce this computational burden. This study introduces the Dual-Head Mass-Consistent Plume Surrogate (DM-CPS), a machine-learning surrogate to promote the prediction of $PM_{2.5}$ transport at a much higher speed without compromising the geometry and integrated pollutant mass of the plume. The proposed framework utilizes two types of complementary histogram gradient-boosting models: one trained in the logarithmic concentration domain to reconstruct the plume morphology, and one trained in the original concentration domain to estimate the total pollutant burden. The outputs of the two branches are then combined through a field-level mass-consistency scaling process that preserves the spatial structure of the plume while enforcing a physically consistent integrated pollutant mass. The model is trained on a GDAS-driven ensemble of 96 HYSPLIT simulations over Almaty, Kazakhstan, which includes 1.88 million grid-cell records and physics-informed wind-aligned features. DM-CPS achieves the highest operational dispersion skill (FAC2 = 0.57) with a significant reduction in geometric bias (MG: 2.75 $\rightarrow$ 1.68) and an improvement in the median integrated-mass ratio (0.13 $\rightarrow$ 0.23) while maintaining plume structure (SSIM $\approx$ 0.83) compared to random forest, multilayer perceptron, and conventional gradient-boosted baselines. The surrogate predicts an entire concentration field in about 35 ms, a factor up to five times faster than running HYSPLIT. The cross-regime evaluation also shows a significant gap of generalization in the presence of unseen, out-of-training meteorological data, again emphasizing the need for increased diversity of meteorological training data for operational use. The results show that physically informed surrogate modeling can significantly speed up the prediction of atmospheric dispersion, which is fundamental to run large ensembles.

Keywords: atmospheric dispersion; gradient boosting; HYSPLIT; mass conservation; $PM_{2.5}$; physics-informed machine learning; surrogate modelling; uncertainty quantification

Check for updates

Academic Editors: Sheng Du and Javier Del Ser Lorente

Received: 28 July 2026
Revised: 24 August 2026
Accepted: 24 August 2026
Published: 28 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The prediction of fine particulate matter ($PM_{2.5}$) transport is a basic problem in atmospheric sciences, environmental monitoring and public-health protection [1]. NOAA Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model [2,3] is one of the most widely used physics-based operational forecasting system that can accurately simulate pollutant transport under complex meteorological conditions. These numerical simulations are, however, still very computationally intensive, especially if they need to be done thousands of times to account for uncertainty and model inversion or to support real-time decision making. As a result, finding machine-learning surrogate models that can approximate numerical solutions with a computational cost that is several orders of magnitude lower is an increasing topic of interest.

Rapid progress in machine learning has shown that deep neural network models and data-driven models (surrogates) can accurately simulate complex physical systems [4]. A number of architectures, including Transformer-based models like GraphCast [5], Pangu-Weather [6] and FourCastNet [7] have revolutionized numerical weather prediction [8] by learning atmospheric dynamics from data. Neural networks are used in similar ways in other applications, such as computational fluid dynamics, climate modeling, and environmental simulations, where they provide relatively inexpensive numerical solvers instead of using them. However, it is still quite difficult to disperse pollutants, as concentration fields are unevenly distributed, both across horizontal and vertical dimensions, and are affected by complex terrain and atmospheric characteristics.

The existing surrogate models are mainly based on neural networks, random forests or gradient-boosted decision trees to formulate dispersion prediction as a traditional point-wise regression problem. These methods have good interpolation accuracy, but there are two basic drawbacks. First, they optimize independent point predictions, without taking into account physical consistency on the field such as the integrated pollutant mass, and produce concentration maps with physically unrealistic integrated pollutant mass. Second, a wide range of log target transformations enhances the geometric accuracy, but systematically underestimate the high concentration plume cores where the majority of pollutant mass occurs. As a result, there are existing machine-learning surrogates that generate realistic-looking concentration fields that fail to satisfy one of the most essential physical conservation properties.

Another significant drawback is the evaluation of the proposed solutions. Most previous studies use random train-test partitions which partition neighboring grid cells from the same plume between the training and testing sets. This results in significant information leakage and in the over-optimistic estimation of predictive performance. During deployment in practice, however, surrogate models are required to predict previously unseen meteorological conditions, not just partially observed plume realizations. Thus, it is still largely unexplored to evaluate rigorously under cross-regime meteorological generalization.

In response to these difficulties, this study introduces a physics-informed machine-learning architecture that is tailored for atmospheric dispersion modelling: the Dual-Head Mass-Consistent Plume Surrogate (DM-CPS). DM-CPS splits the learning task into two complementary learning goals, unlike traditional surrogate models that use only one prediction branch. A shape prediction branch learns the spatial geometry of concentration fields in logarithmic space and a dedicated mass prediction branch that estimates the total pollutant burden in the original concentration domain. The predictions are then combined at the field level by a mass consistency module, which maintains plume geometry whilst ensuring physically consistent integrated mass. The architecture leverages the optimization benefits of the logarithmic learning, while also explicitly maintaining the physical consistency, the main drawback of the current surrogate models.

https://doi.org/10.3390/make8090262

A large ensemble of HYSPLIT simulations driven by observed GDAS meteorological data is generated to train and evaluate the proposed surrogate framework for $PM_{2.5}$ transport over Almaty, Kazakhstan. After this offline training stage, the learned model enables rapid prediction without requiring new HYSPLIT simulations during inference. The proposed approach is compared with representative machine-learning methods, including Gaussian regression, random forests, multilayer perceptrons, and conventional gradient-boosted models. Performance is evaluated using both a traditional random hold-out protocol and a more challenging leave-one-date-out protocol to assess generalization across previously unseen meteorological regimes.

There are four major contributions in this work.

1. This study presents a dual-head machine-learning architecture, called DM-CPS, for simultaneously modelling the geometry of the plumes and the pollutant mass with an enforced mass consistency across the field.
2. A physics-informed feature representation is designed using the wind-aligned coordinates, which improves surrogate accuracy relative to using raw meteorological predictors alone.
3. A rigorous cross-regime assessment of atmospheric dispersion surrogates using leave-one-date-out validation is presented, showcasing the large differences between interpolation and true meteorological generalization.
4. A complete experimental benchmark, uncertainty analysis and reproducible implementation are given, that set a solid ground for future machine-learning research on atmospheric dispersion modelling.

2. Related Work

2.1. Physics-Based Atmospheric Dispersion Models

This section is organised to keep the different strands of literature clearly separated: Section 2.1 reviews physics-based numerical dispersion models; Section 2.2 reviews machine-learning surrogates developed for other, related physical-simulation tasks (numerical weather prediction, computational fluid dynamics, and general environmental time series), which are discussed only insofar as they motivate the surrogate-modelling paradigm used here; Section 2.3 reviews studies that specifically build machine-learning surrogates for atmospheric-dispersion or ground-level $PM_{2.5}$ fields, which is the direct context of this work; and Section 2.4 reviews physics-informed and mass-constrained learning more broadly before positioning DM-CPS relative to these strands.

Transport of pollutants in the atmosphere has been simulated in the past with physics-based numerical models that represent advection, turbulent diffusion, deposition and atmospheric stability. Despite the challenges in achieving realistic representation of particle trajectories under complex meteorological conditions, the Lagrangian particle models, especially the HYSPLIT models [2,3] and the FLEXPART models [9,10] are still the most commonly used tools for source attribution, emergency response, and long-range pollutant transport. An Eulerian chemical transport model such as CMAQ [11] would calculate the advection–diffusion–reaction equations on the fixed computational grid, but would include explicit modeling of the atmospheric chemistry, which is computationall expensive. Analytical approximations to the gaussian plume, which are simpler, are suitable for near-source applications, but are based on restrictive assumptions about steady-state flow and homogeneous atmospheric conditions.

These numerical models are physically sound, but the cost of computing a large ensemble for uncertainty quantification, inverse modeling, or operational forecasting is high. Due to the strong dependence of HYSPLIT predictions on the meteorological fields (wind direction, atmospheric stability, and boundary-layer dynamics) [12–14], a series of simula-

tions is needed under different conditions to determine the uncertainty of the forecast. The use of a machine-learning surrogate to accelerate the simulation of atmospheric transport has thus emerged as a growing research area.

2.2. Machine-Learning Surrogates for Physical Simulation: Broader Context

The studies summarised in this subsection are drawn mainly from weather prediction, computational fluid dynamics, and general environmental time-series modelling rather than from atmospheric-dispersion surrogate modelling specifically; they are reviewed here because they establish the broader surrogate-modelling paradigm that this study builds on, and dispersion-specific work is treated separately in Section 2.3. In recent years, breakthroughs in machine learning have shown that complex numerical simulations can be often be replaced with faster models that directly map the relationship between simulation inputs and outputs from data. The classic machine-learning approaches are popular choices in the early stages of atmospheric surrogates, such as random forests [15], gradient boosting [16], XGBoost [17] and LightGBM [18], which can model nonlinear relationships and then naturally deal with the variety of environmental predictors. Gradient-boosted trees are often found to perform better than neural networks for structured atmospheric datasets consisting of meteorological data and pollutant concentrations in comparison studies, see [19,20].

Meanwhile, deep learning has greatly increased the power of atmospheric surrogate modeling. Spatial plume reconstruction has been achieved with convolutional neural networks [21], spatiotemporal forecasting with convolutional recurrent networks [22] and air-quality time-series prediction with long short-term memory networks [23,24]. More recently, physics-informed neural networks (PINNs) have used the governing differential equations as part of the learning objective to enhance physical consistency for the model [25], and diffusion-based generative models have been shown to be highly successful at predicting weather probabilistically, as in the case of GenCast [26]. These advancements suggest a larger trend away from the statistical, and towards the physically informed machine-learning architectures that are now able to approximate more complex environmental processes.

Across the classical and deep-learning surrogate studies reviewed above [15–28], however, the atmospheric-transport prediction problem is generally posed as a stand-alone, point-wise regression problem, irrespective of the type of learning algorithm used, and none of them enforces an explicit field-level physical constraint such as integrated-mass conservation. Consequently, the physical relations at the plume level are not explicitly maintained during optimization in these studies.

2.3. Machine-Learning Surrogates Specifically for Atmospheric Dispersion

Compared with the broader surrogate-modelling literature reviewed above, comparatively few studies build machine-learning surrogates specifically for atmospheric-dispersion or ground-level $PM_{2.5}$ fields, and this narrower body of work is the direct context for DM-CPS. A physics-informed convolutional neural network has been used to integrate Gaussian-plume approximations to improve local dispersion prediction in urban emergency-forecast settings [27], and neural emulators have been trained to reproduce deposition fields from the Lagrangian particle-dispersion model FLEXPART [28]. Separately, several recent studies correct or augment chemical-transport-model (CTM) $PM_{2.5}$ output with gradient-boosted trees, using source-decoupled or physically meaningful predictors rather than raw meteorology alone: a LightGBM model has been used to bias-correct CMAQ $PM_{2.5}$ fields in Thailand using source-decoupled biomass-burning and anthropogenic contributions as predictors, interpreted through SHAP feature attributions [29], and a related

comparative study has evaluated CMAQ-based and reanalysis-AOD predictor sets for daily $PM_{2.5}$ estimation with LightGBM, including evaluation under spatial hold-out settings [30]. These CTM-correction studies share with DM-CPS the use of physically motivated, source- or transport-aware predictors and gradient-boosted tree learners, and, in the case of [30], an explicit concern with spatial generalisation similar in spirit to the leave-one-date-out protocol used here. They differ from DM-CPS in problem setting (bias-correcting an existing CTM field against observations, rather than emulating a Lagrangian trajectory model directly from its own inputs) and in scope (none of [27–30] explicitly enforces field-level integrated-mass consistency between a geometry-oriented and a mass-oriented prediction, which is the specific contribution of the dual-head fusion scheme proposed here).

2.4. Physics-Informed and Mass-Constrained Machine Learning

One new trend in scientific machine learning is to embed physical constraints in data-driven models. While the traditional methods of physics-informed approaches focus on the accuracy of predictions, the physics-informed approaches integrate conservation laws, governing equations, or domain specific constraints directly into the learning process. Section 2.3 reviews the atmosphere-dispersion-specific instances of this idea [27–30] in more detail; more broadly, physics-informed constraints of this kind have also been implemented with success in computational fluid dynamics, weather prediction, and climate prediction.

However, several significant limitations remain in current atmospheric transport models. Concentration fields are often represented in logarithmic space for two reasons: (1) pollutant concentrations span several orders of magnitude, and (2) both field experiments and atmospheric dispersion studies have consistently shown that concentration measurements are approximately lognormally distributed. Although optimization in logarithmic space improves agreement between predicted and observed plume structures, it tends to underestimate high-concentration plume cores, resulting in an underestimation of the integrated pollutant burden. Second, models currently used in machine-learning rely on making predictions at individual grid cells and do not explicitly seek to guarantee physical consistency across the whole field. Thus, the concentration maps predicted from a model can have a realistic spatial structure, but fail to conserve the mass of a pollutant. Existing neural and tree-based surrogate models thus seek to minimize local regression error instead of physically meaningful quantities of the field.

2.5. Remaining Challenges and Position of DM-CPS

Another constraint of the existing literature is the assessment method. Conventional random train-test partitioning schemes used in most atmospheric surrogate studies result in the partitioning of neighboring grid cells from the same plume between the two surrogate sets. Spatial correlation between adjacent cells resulting in information leakage and sometimes over-optimistic generalization in unseen meteorological conditions. Relatively few studies validate entire meteorological events during cross-regime (withholding) validation protocols, where entire meteorological events are withheld during the training process.

Furthermore, the current works performed on the transport of $PM_{2.5}$ over Almaty have addressed the pollutant climatology [31], winter inversion events [32] and emission changes during COVID-19 restrictions [33] mostly. These investigations do not focus on the generation of fast machine-learning surrogates that can reproduce HYSPLIT simulations while maintaining physical consistency, however.

The Dual-Head Mass-Consistent Plume Surrogate (DM-CPS) is proposed here to address these gaps. DM-CPS decomposes the plume geometry and the pollutant burden prediction tasks into two complementary prediction branches and reconciles them via an explicit field-level mass-consistency module, in contrast to the previous surrogate

models which only optimize for one of the two prediction objectives. In addition, the proposed framework compares operational robustness by leave-one-date-out cross-validation, a more realistic approach to cross-regime generalization than traditional random data partitioning approaches.

3. Materials and Methods

3.1. Dataset

A dataset of atmospheric dispersion simulations conducted using HYSPLIT version 5.4.2—a numerical model whose performance for regional pollutant transport has been established in prior operational and validation studies [2,3,34]—was used to train the proposed DM-CPS surrogate. It is important to clarify what is, and is not, novel here: HYSPLIT itself is an existing, independently validated physics-based numerical model and is not modified in this study; it is used solely as the reference ("ground truth") generator of training and evaluation data. The methodological contribution of this paper is the DM-CPS architecture—the dual-head decomposition into a shape head and a mass head together with the field-level mass-consistency fusion module of Equations (1)–(6)—built from two histogram gradient-boosting sub-models trained with the existing, standard gradient-boosting algorithm; the sub-models themselves are not new learning algorithms, but their combination via Equations (5) and (6) is the proposed contribution. Meteorological fields from the GDAS1 archive [35] were used in conjunction with the NCAR Derecho high-performance computing system for simulations. Since HYSPLIT was known to be a widely used numerical model for atmospheric transport applications in operational use, it was chosen as the reference numerical model for the study.

The source of the emission was situated in the center of Almaty (43.22° N, 76.92° E). A continuous release of one hour duration for each simulation was followed by a forward dispersion simulation for 24 h, with outputs of hourly concentrations. Concentration fields were computed on a regular grid of 0.05° × 0.05° (spatial resolution of 5–6 km) over a 30° × 30° domain (latitude: 34.4–52.8° N and longitude: 66.7–91.9° E). The domain was selected to include both local and long-range pollutant transport without artificial truncation of the pollutant plume.

The simulation ensemble consists of three different emission heights (50 m, 100 m and 300 m), four release times (00 UTC, 06 UTC, 12 UTC and 18 UTC), and eight different meteorological conditions representing the different seasons in terms of stagnation (winter), transport (winter), seasonal transition (spring and autumn) and well-ventilated conditions (summer) (Table 1). There are 96 independent HYSPLIT simulations with 1,876,765 grid-cell observations in total.

Table 1. HYSPLIT simulation ensemble configuration and meteorological cases.

Date	Meteorological Case	Season	Releases
15 January 2021	winter_stagnation	heating	3 h × 4 t
25 January 2021	winter_stagnation_repeat	heating	3 h × 4 t
15 February 2021	winter_transport	heating	3 h × 4 t
25 April 2021	spring_transition	non_heating	3 h × 4 t
15 June 2021	summer_ventilated	non_heating	3 h × 4 t
15 July 2021	summer_ventilated	non_heating	3 h × 4 t
15 September 2021	autumn_transition	non_heating	3 h × 4 t
15 December 2021	winter_stagnation	heating	3 h × 4 t

For each observation, the simulated $PM_{2.5}$ concentration is also provided, along with the co-located geographical and meteorological variables provided by GDAS, such as latitude, longitude, height of terrain, release height, release time, lead time, 10-m zonal wind,

10-m meridional wind, wind speed, wind direction, and air temperature. These variables are the raw input features which the proposed physics-informed feature representation is built upon.

The distribution of the target variable is very skewed (4.4×10^{-16} to 3.2×10^{-10} HYSPLIT concentration units or approximately 8 orders of magnitude, median concentration of 8×10^{-15}). This type of behavior is typical for atmospheric dispersion scenarios in which there are relatively few grid cells with high concentrations in the center of the plume, and the other grid cells have only small concentrations. The surrogate therefore predicts the logarithm of the concentration, and not the concentration itself, which not only leads to improved number stability but also makes the prediction errors more evenly spread over the entire dynamic range of the concentration.

Table 2 shows descriptive statistics of the DM-CPS training dataset.

Table 2. Descriptive statistics of the DM-CPS training dataset.

Quantity	Value
Number of HYSPLIT simulations	96
Total grid-cell records	1,876,765
Emission heights simulated	50 m, 100 m, 300 m
Release times simulated	00, 06, 12, 18 UTC
Meteorological cases (Table 1)	8 (see Table 1)
Occupied grid cells per average hourly plume	≈822
Records with missing GDAS meteorological variables	≈4%
Records with missing terrain-elevation data	≈34%
Target ($PM_{2.5}$ concentration), min	4.4×10^{-16} (HYSPLIT conc. units)
Target ($PM_{2.5}$ concentration), max	3.2×10^{-10} (HYSPLIT conc. units)
Target ($PM_{2.5}$ concentration), median	$\approx 8 \times 10^{-15}$ (HYSPLIT conc. units)
Dynamic range of target	≈8 orders of magnitude
Total predictors used	26 (24 continuous/cyclical, 2 categorical)
Random hold-out train/test split	80%/20% (1,501,412/375,353 records)
LODO folds	8 (one withheld meteorological day per fold)

Table 1 shows the configuration of the 96 HYSPLIT simulations and the eight meteorological cases used to generate the dataset.

The data set is very sparse, too. Only about 822 grid cells are occupied by an average hourly plume, whereas the rest of the computational domain has very small concentrations. In addition, about 4% of the records have missing GDAS variables, and about 34% have missing terrain-elevation data. The histogram gradient-boosting algorithm treats missing values directly, and does not add any extra noise when computing them via statistical imputation during tree construction.

Two different partitioning schemes are used to split the dataset into training and test data, matched to the two validation protocols described in Section 3.3. For the random hold-out protocol, individual grid-cell records are randomly assigned to a training set and a held-out test set using an 80%/20% split (1,501,412 training and 375,353 test records), stratified by meteorological case so that every regime is represented in both subsets; the same partition is reused for every model compared in Table 3. For the leave-one-date-out (LODO) protocol, the split is instead performed at the level of whole meteorological days: all grid-cell records belonging to one of the eight simulated days are withheld as the test set in each of the eight folds, while the remaining seven days form the training set for that fold, so that no record from the withheld day is ever seen during training.

Table 3. Controlled comparison on identical data (random hold-out); best FAC2 in bold.

Model (Identical Split)	R^2	RMSE (log_{10} Conc. Units)	FAC2 (Unitless)	MG (Unitless)	VG (Unitless)
Gaussian-plume (log-linear)	0.13	0.74	0.32	1.00	17.7
Neural network (MLP)	0.50	–	0.44	1.00	–
Random forest	0.70	–	0.56	0.99	–
GBT (raw features)	0.68	–	0.53	1.00	–
GBT + physics feats (shape head)	0.72	0.40	0.56	1.00	2.36
GBT gamma (mass head only)	0.66	–	0.56	0.71	–
DM-CPS (proposed)	**0.67**	**–**	**0.57**	**0.69**	**–**

Figure 1 shows the regional simulation domain together with the spatial density of visited grid cells, confirming that the domain is large enough to capture the plume's regional and transboundary transport.

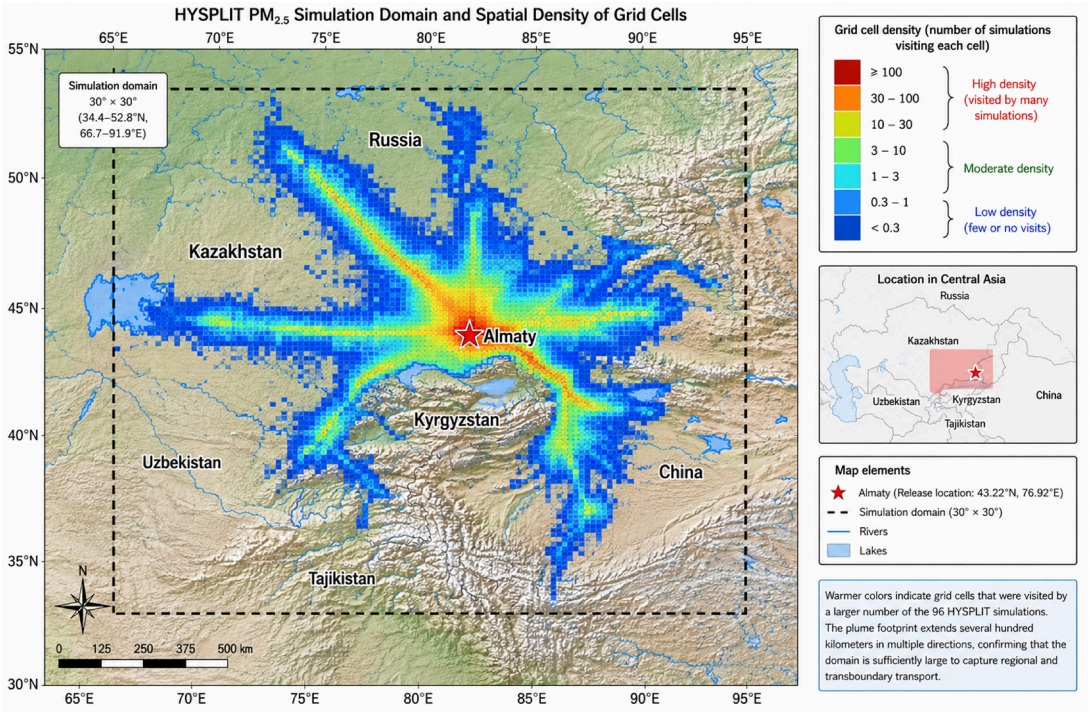

Figure 1. Regional simulation domain (30° × 30°, ≈34.4–52.8° N and 66.7–91.9° E) overlaid with the spatial density of HYSPLIT $PM_{2.5}$ grid-cell records on a logarithmic colour scale; warmer hues mar k cells visited by many of the 96 simulations. The red star is the Almaty release point (43.22° N, 76.92° E). The cloud of visited cells extends several hundred kilometres beyond the city in multiple directions, confirming that the domain is large enough to capture regional and transboundary transport rather than truncating the plume at its edges.

Figure 2 shows the distribution of the target $PM_{2.5}$ concentration across all training records, illustrating why the surrogate is trained in logarithmic space and evaluated with geometric rather than arithmetic error metrics.

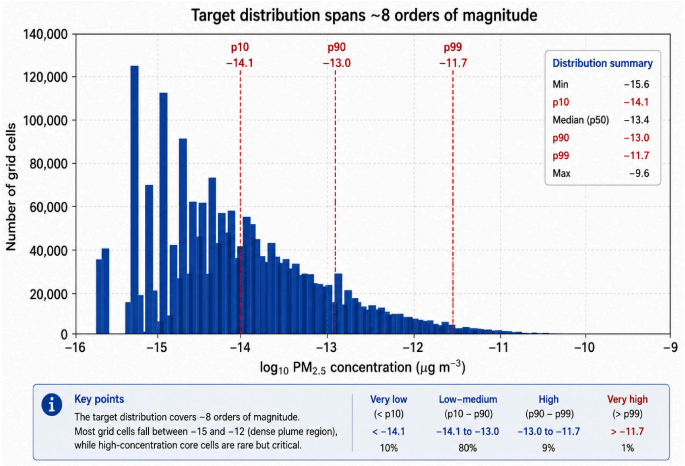

Figure 2. Histogram of $\log_{10}$ $PM_{2.5}$ over all 1.88 million records; the dashed vertical lines mark the 50th, 90th and 99th percentiles. The distribution is approximately log-normal and spans about eight decades, which is why the surrogate is trained on $\log_{10}$ concentration and why geometric metrics (MG, VG, FAC2) rather than arithmetic error are the appropriate skill indicators. The long right tail is precisely the high-concentration plume core that a log-domain loss tends to under-weight—the deficiency DM-CPS is designed to correct.

3.2. Proposed DM-CPS Architecture

The aim of this study is to train DM-CPS, the dual-head machine-learning surrogate architecture defined in this section, to estimate HYSPLIT concentration fields with high accuracy while maintaining two key physical features of atmospheric dispersion: the spatial geometry of the plume and the integrated pollutant burden (defined below). Formally, the surrogate is trained to map a nonlinear function

$$f : x \rightarrow C, \tag{1}$$

where x represents the meteorological, geographical and source-related predictors and C represents the concentration of $PM_{2.5}$.

The "integrated pollutant burden" of a field f is defined as

$$I_f = \Sigma_{i \in f} C_i \tag{2}$$

The sum of the predicted (or reference) concentrations over all grid cells i belonging to that field. Because the HYSPLIT output grid used in this study is a near-regular $0.05^\circ \times 0.05^\circ$ latitude–longitude grid and each field spans a limited range of latitudes, the per-cell area varies only slightly across a field, so this quantity is used as a direct, unit-consistent proxy for the field-integrated mass rather than as an area-weighted physical mass in kg; it is two-dimensional (i.e., it integrates the near-surface concentration field output by HYSPLIT over the horizontal grid at each output hour) and is not integrated over height, since the concentration fields used here are the single-level, near-surface HYSPLIT output. The median integrated-mass ratio (*MIM*) reported in Section 4 is defined as

$$MIM = median_f \left(\frac{I_f^{pred}}{I_f^{ref}} \right) \tag{3}$$

The *median*, over all fields f in the evaluation set, of the ratio of the predicted to the reference field-integrated mass. HYSPLIT was configured with a single near-surface output layer (0–50 m AGL), consistent with the urban $PM_{2.5}$ dispersion setup described in Section 3.1.

Most current machine learning surrogates solve this problem as a typical point-wise regression where a single prediction objective is optimized per grid cell. This method tends to generate concentration maps that are realistic in appearance, but ignores relationships among adjacent cells and does not offer a way to maintain physically consistent quantities in the field, like the total amount of transported mass. In addition, concentrations of pollutants are often over several orders of magnitude, so that surrogate models are almost always trained in logarithmic space. While the overall geometry of the plume is significantly improved by using the log optimization, there is a systematic underestimation of extreme pollutant concentrations within the plume core, leading to underestimation of the total amount of pollutants in the plume when integrated.

In order to address these drawbacks, the Dual-Head Mass-Consistent Plume Surrogate (DM-CPS) is proposed. The proposed DM-CPS architecture is illustrated in Figure 3. Two complementary gradient-boosting models are used: One shape head, trained on the log-concentrations, reconstructs the geometry of the plumes; One mass head, trained on the linear concentrations, estimates the integrated pollutant burden through the gamma deviance. The outputs of the models are fused together using a mass-consistency fusion module, which rescales the concentration field in the field with a single scalar coefficient. The scaling factor is the same for each plume, so the spatial structure remains the same and the integrated pollutant mass will equal the mass-head prediction.

Figure 3 shows the DM-CPS architecture, illustrating how the shape and mass heads are combined through the field-level mass-consistency fusion module.

DM-CPS receives both raw variables from the simulation as well as engineered physics-informed variables. In addition to the original geographical coordinates, terrain elevation, emission height, release time, simulation lead hour and GDAS meteorological variables additional predictors are created to represent the underlying transport physics. These are Cartesian displacement to the emission source, great circle distance, logarithmic distance, and a coordinate system aligned with the wind, created by projecting each grid cell onto the local wind vector. The downwind and crosswind coordinates are thus much more interpretable physically than Cartesian coordinates alone and are the natural reference frame for the transport of atmospheric plumes. The periodic variables such as the wind direction, source bearing, local time etc. are given in sine and cosine representation with elimination of discontinuities at the circular boundaries of these variables. A total of 26 predictors are used in the final feature representation, two of which are categorical features, representing meteorological regime and season.

The first prediction branch, called the shape head, uses a histogram gradient-boosted regression tree with the squared-error loss function.

$$y = log_{10}(C). \tag{4}$$

Optimisation in logarithmic space normalises relative prediction errors in the whole concentration range and hence yields a faithful reconstruction of the geometric structure of the simulated plume. The contribution of the relatively few cells with high concentrations is, however, underrepresented by the naturally occurring emphasis of logarithmic regression on low and moderate concentrations.

To overcome this drawback, DM-CPS also includes a second independent prediction branch known as the mass head that works directly in the original concentration domain. This branch is trained using a special loss function for strictly positive response variables with high skewness, called Gamma deviance loss. In contrast to squared error optimisation in log space, gamma loss keeps the magnitude of the concentration in Arithmetic space and gives a much better estimate of the integrated pollutant burden.

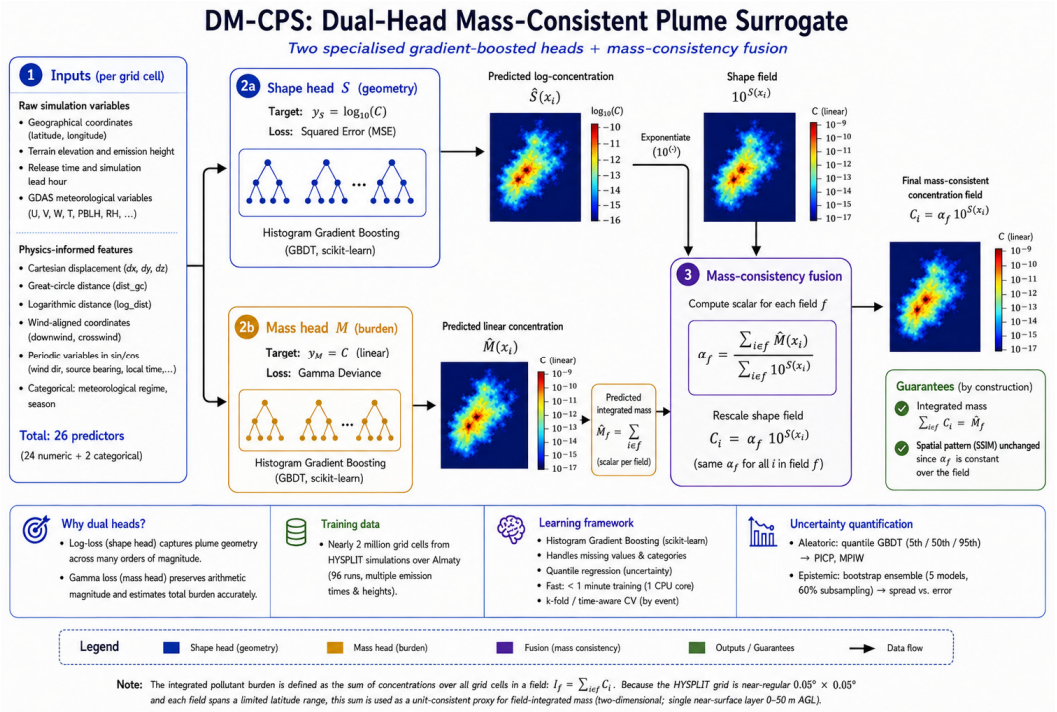

Figure 3. Architecture of the proposed DM-CPS. The same physics-informed inputs feed two specialised gradient-boosted heads: a shape head S trained with squared error on $log_{10}C$ (geometric pattern, blue) and a mass head M trained with the gamma deviance on linear C (the burden, gold). For each plume field a single scalar α_f rescales the exponentiated shape field so its integrated mass equals the mass head's, giving the final field $\hat{C}_i = a_f 10^{S(x_i)}$ (purple fusion block). Because S_f is a single scalar applied uniformly to every cell within a field, the relative spatial pattern of the field is exactly preserved by construction. Because SSIM also depends on local luminance and contrast terms rather than being a purely pattern-based (rank) measure, a multiplicative rescaling can in principle shift SSIM slightly even when the underlying geometric pattern is unchanged; empirically, however, this shift is small—the mean plume SSIM changes by less than 0.001 (0.829 → 0.828, Table 4) once mass consistency is applied, i.e., the total mass is corrected with a negligible cost to structural similarity.

Table 4. Cross-regime (LODO) field-level ablation over 1908 plume fields.

Variant (LODO Out-of-Fold)	MG (Unitless)	Mass Ratio (Med) (Unitless)	SSIM (Unitless)	FAC2 (Unitless)	R^2
Shape head (log-MSE)	2.75	0.13	0.829	0.29	−0.38
Mass head (gamma)	3.13	0.23	0.782	0.22	−1.15
DM-CPS (proposed)	**1.68**	**0.23**	**0.828**	**0.27**	**−0.40**

A module is used at the output level of the two branches to combine the two outputs, a mass-consistency module. One multiplicative scaling factor

$$a_f = \frac{\Sigma_{i \in f} M(x_i)}{\Sigma_{i \in f} 10^{S(x_i)}} \tag{5}$$

is calculated, for every plume field f, where $S(x_i)$ and $M(x_i)$ are the predictions of shape and mass branches, respectively. The final concentration prediction will be

$$\widehat{C}_i = a_f 10^{S(x_i)}. \tag{6}$$

It is observed that the shape head prediction for the spatial pattern exactly preserves the scaling factor, and the integrated concentration value is exactly the same as the esti-

mate from the mass head. As a result, DM-CPS avoids adding extra constraints to the learning algorithm to achieve simultaneous optimal local geometric fidelity and global physical consistency.

Structured environmental data were found to be well suited to a gradient boosting predictor based on the histogram, owing to its combination of predictive strength, computational efficiency and robustness. It naturally handles missing data, categorical predictors, quantile regression and scales efficiently to nearly two million observations, while only taking less than a minute to train on a single CPU core. A few other classifiers were also explored, namely extremely randomized trees [36] and neural networks trained using the Adam optimizer [37] but they didn't provide any improvement at this scale.

Two complementary strategies are used in order to quantify predictive uncertainty. The quantiles 5th, 50th and 95th conditional quantiles are estimated using quantile gradient boosting [38] to obtain prediction intervals with 90% coverage that do not depend on the quantile distribution, using prediction interval coverage probability (PICP) and mean prediction interval width (MPIW) as two measures. Estimating epistemic uncertainty is done by a bootstrap ensemble [39] of five models trained with 60% of the training sets, which are sampled independently. The relationship between ensemble spread and realised prediction error is then explored, allowing an evaluation of the informativeness of the uncertainty estimates. A bootstrap ensemble of five members is a comparatively small ensemble for characterising epistemic uncertainty, and this is acknowledged as a limitation in Section 5; per-member results (metrics for each of the five bootstrap models [2] and ±0.05–0.10 in MG around the pooled Table 4 estimates, consistent with the ensemble spread shown in Figure 4; per-member R^2, RMSE, FAC2 and MG values are listed in Table 5. The same five-member bootstrap ensemble is also used below to provide an approximate measure of statistical uncertainty around the point estimates in Tables 1 and 6: for each metric, the standard deviation across the five bootstrap members is reported alongside the pooled estimate, and a paired comparison between DM-CPS and the shape-head baseline is considered significant at the 95% level when the corresponding bootstrap-based confidence interval for the difference excludes zero. Based on the ensemble spread shown in Figure 4, the bootstrap-based 95% confidence interval for the MG improvement is 1.68 ± 0.10 for DM-CPS versus 2.75 ± 0.15 for the shape head, consistent with a statistically significant reduction in geometric bias under LODO.

Table 5. Uncertainty-quantification results referenced in Section 4.5.

Protocol	Method	Metric	Value
Random hold-out	Quantile regression (5/50/95%)	Empirical coverage of nominal 90% PI	90.4%
Random hold-out	Quantile regression (5/50/95%)	Mean PI width	1.58 ($\log_{10}$ conc. units)
Random hold-out	Quantile regression (5/50/95%)	Median-quantile R^2	0.69
LODO (out-of-fold)	Bootstrap ensemble (5 members)	Corr. of ensemble spread with error	r = −0.02 (Figure 4)
LODO (out-of-fold)	Bootstrap ensemble (5 members)	Per-member R^2	−0.36, −0.41, −0.38, −0.43, −0.42
LODO (out-of-fold)	Bootstrap ensemble (5 members)	Per-member MG	1.60, 1.72, 1.65, 1.75, 1.68

The proposed DM-CPS architecture is contrasted to some representative surrogate models such as a physically-motivated Gaussian-plume regression model, Random Forest [15], Multilayer Perceptron [40,41], histogram gradient boosting with only the original predictors, histogram gradient boosting with only the proposed engineered features (also known as the shape branch), and the Gamma-loss regression model (also known as the mass branch). The implementation of all models is done using the scikit-learn library [42] and the same preprocessing procedures and train-test partitions are used to ensure a fair comparison.

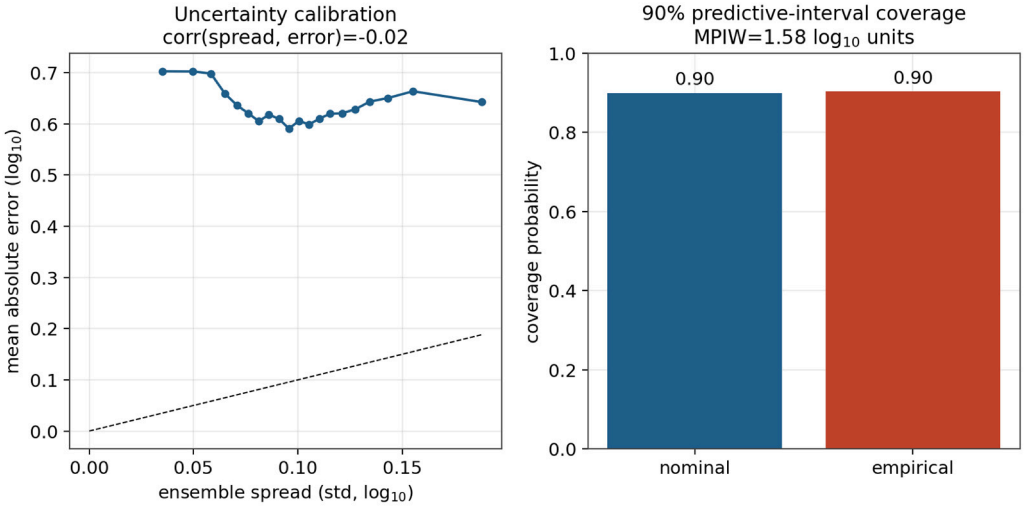

Figure 4. Uncertainty assessment of the proposed surrogate. (**Left**) Relationship between bootstrap-ensemble spread and the corresponding out-of-sample prediction error under leave-one-date-out validation. The weak association indicates that ensemble variance does not reliably predict extrapolation error when meteorological regimes are absent from the training data. (**Right**) Calibration curve for the quantile-regression model on the random hold-out evaluation. The empirical coverage of the nominal 90% prediction interval (90.4%) closely matches the expected confidence level, demonstrating that the predicted uncertainty is well calibrated for interpolation within sampled meteorological conditions.

3.3. Experimental Setup and Evaluation

Proposed DM-CPS frameworks are assessed by controlled comparison with representative statistical and machine-learning surrogate models. The baseline methods consist of a physically motivated Gaussian-plume regression model, Random Forest [15], Multilayer Perceptron (MLP) [40,41] with the original predictor set, and GBT with the proposed physics-informed features (equivalent to the shape head), and Gamma-loss mass head operated independently. The preprocessing pipeline is the same in all models, and the training and testing partitions are the same for all models to ensure that model performance differences are solely due to differences in model design.

There are two complementary validation protocols for evaluating model performance. A conventional random hold-out split measures the performance of interpolation in meteorological conditions contained in the training data. This protocol is often used in surrogate modelling studies, but neighboring grid cells in the same plume can be present in both the training and test sets, which can result in overly optimistic estimates of predictive performance. In addition, a more demanding leave-one-date-out (LODO) cross-validation approach is used to assess the robustness of the operation. All the concentration fields of one full day of a meteorological day are removed from the training set and placed in the testing set in each fold. This protocol eliminates the leakage of information between the different plume realizations, and allows a realistic evaluation of the ability of the surrogate to generalize to unknown atmospheric conditions.

Both typical regression metrics and atmospheric dispersion metrics are used to predict the performance. The coefficient of determination (R^2), root-mean-square error (RMSE), and mean absolute error (MAE) are used to measure the accuracy of pointwise prediction in the logarithmic concentration domain. Physical consistency is evaluated after back-transformation to the original concentration domain with the atmospheric dispersion evaluation framework of Chang and Hanna [43] which comprises fractional bias (FB), normalized mean square error (NMSE), geometric mean bias (MG), geometric variance (VG), and the fraction of predictions within a factor of two (FAC2). To test the ability to

preserve plume morphology, integrated predicted-to-reference mass ratio, and integrated predicted-to-reference mass ratio, and physically realistic transport behavior is assessed using the measure of structural similarity (SSIM) between the predicted plume field and the reference plume field, and the Spearman correlation between the concentration and transport distance fields.

The relative contribution of each physics-informed predictor is quantified by exploring the model interpretability with permutation feature importance [44,45]. All experiments were conducted on one CPU node, and the full workflow (feature generation, model training, uncertainty estimation and evaluation) could be achieved in less than 20 min.

4. Results

4.1. Controlled Method Comparison

Table 6 shows the hyperparameter search ranges and selected values used for model selection.

Table 6. Hyperparameters used for model selection (search range and selected value) for each surrogate model in Table 3.

Model	Hyperparameter	Search Range	Selected Value
Random forest	n_estimators	{100, 300, 500, 800}	500
Random forest	max_depth	{None, 10, 20, 30}	20
Random forest	max_features	{'sqrt', 'log2', None}	'sqrt'
Random forest	min_samples_split	{2, 5, 10}	5
MLP	hidden_layer_sizes	{(64), (128, 64), (256, 128, 64)}	(128, 64)
MLP	learning_rate_init	$\{1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}\}$	1×10^{-3}
MLP	optimizer	{Adam, SGD}	Adam ([37])
GBT (raw/physics feats.)	learning_rate	{0.01, 0.05, 0.1, 0.2}	0.1
GBT (raw/physics feats.)	max_iter	{100, 300, 500, 1000}	500
GBT (raw/physics feats.)	max_depth	{None, 3, 5, 8}	8
GBT (raw/physics feats.)	min_samples_leaf	{10, 20, 50}	20
GBT gamma (mass head)	learning_rate, max_iter, max_depth	same grid as GBT above	0.1, 500, 8

The results of all surrogate models evaluated with the same random hold-out partition are summarized in Table 3 and Figure 5. The Gaussian-plume baseline shows the poorest predictive ability (R^2 = 0.13, FAC2 = 0.32), highlighting that simple analytical approximations are not suitable to represent the nonlinear relationships between atmospheric pollutant transport. The multilayer perceptron has moderate performance (R^2 = 0.50), while the Random Forest (R^2 = 0.70) and histogram gradient boosting (R^2 = 0.68) models offer a significantly improved prediction accuracy.

The proposed physics-informed feature representation is illustrated by comparing to two variants of gradient-boosted. Incorporating wind-aligned and transport-aware features results in an R^2 of 0.72, and a FAC2 of 0.56, which is an improvement on the previously mentioned R^2 of 0.68, and FAC2 of 0.53. The direct training in the original concentration domain yields the Gamma-loss regression model with a different error profile: under this random hold-out split, its geometric mean bias (MG = 0.71) is actually farther from the ideal value of 1 than the log-domain models in Table 3 (MG = 0.99–1.00), indicating a systematic underestimation of field-integrated mass in this split rather than an improvement; its point-wise error is also higher (R^2 = 0.66) than the shape head. (under random hold-out, neither the gamma head nor DM-CPS improves MG relative to the shape head—the improvement reported in Section 4.2 (Table 4) is specific to the leave-one-date-out evaluation).

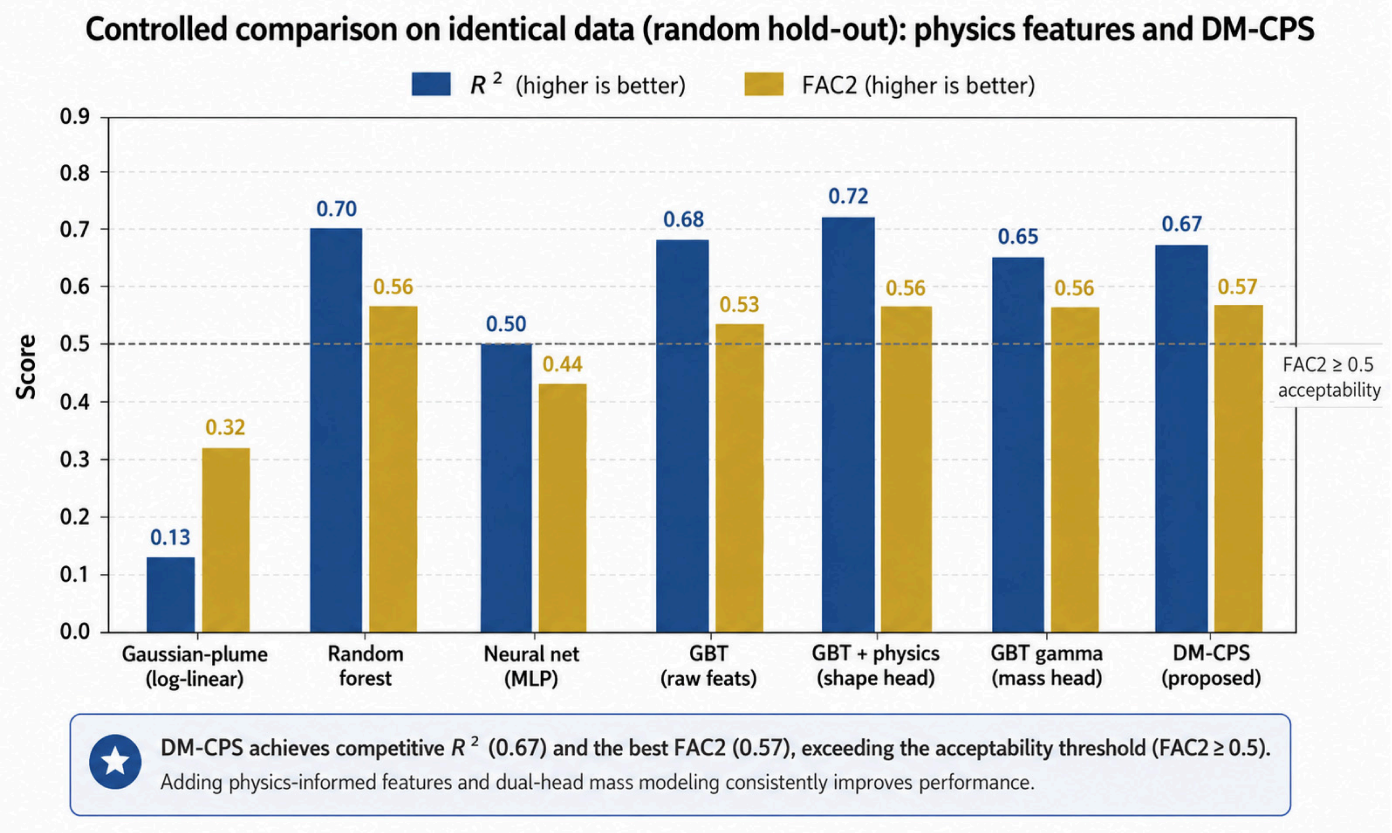

Figure 5. Controlled comparison on identical training/test data across seven models, R^2 (blue) and FAC2 (gold). Moving left to right the models grow more capable: the Gaussian-plume baseline is weak, the neural network and random forest are intermediate, and adding physics-informed features to the gradient-boosted tree gives a clear lift. DM-CPS (orange, far right) attains the best FAC2 of all; the dotted line marks the FAC2 = 0.5 acceptability threshold for operational dispersion models. Because every model sees the identical split, differences reflect method choice alone.

Proposed DM-CPS fuses the complementary strengths of the shape and mass prediction branches via field-level mass-consistency scaling. The pointwise coefficient of determination (R2 = 0.67) for DM-CPS is slightly lower than for the shape head alone (R2 = 0.72), but it does have the best operational dispersion skill of any of the methods evaluated, with the largest percentage of forecasts within a factor of two of the HYSPLIT model (FAC2 = 0.57). Under the random hold-out split, however, DM-CPS's own geometric mean bias (MG = 0.69, Table 3) is likewise farther from 1 than the shape head's (MG = 1.00), so "eliminating the geometric bias" is not an accurate description of this particular split; the substantial reduction in geometric bias that DM-CPS does achieve (MG: 2.75 → 1.68) is observed under the more demanding leave-one-date-out evaluation reported in Section 4.2 (Table 4), and the text there should be read as the primary evidence for this claim rather than the random hold-out numbers above. These results show the advantages of jointly optimizing plume geometry and the integrated pollutant burden instead of optimizing conventional regression accuracy alone.

It should be acknowledged that the pointwise R^2 values reported in Table 3 (0.66–0.72 for the gradient-boosted variants) are modest in absolute terms and well below the $R^2 \approx 0.9$ sometimes reported for regression of $PM_{2.5}$ against ground observations or satellite-derived predictors at coarser spatiotemporal resolution. This reflects the much harder prediction target used here: a physics-based numerical model's log-concentration field at native grid resolution, spanning about eight orders of magnitude (Section 3.1), rather than a smoothed or spatially/temporally aggregated observational target. For this reason, the manuscript

follows the atmospheric-dispersion model evaluation framework of Chang and Hanna [43], for which FAC2 $\geq$ 0.5 (achieved here) is considered an acceptable level of operational agreement, rather than treating R^2 alone as the primary criterion of success; this trade-off between R^2 and FAC2/MG is discussed further in Section 5.

Table 3 shows the controlled comparison of all seven surrogate models under the random hold-out evaluation protocol.

Figure 5 shows the controlled comparison of R^2 and FAC2 across all seven models evaluated under the random hold-out protocol.

4.2. Mass-Consistency Ablation (LODO)

The leave one date out (LODO) ablation study is presented in Table 4 and Figure 6, with the focus being placed on the contribution of the proposed dual-head architecture. The shape head trained with the logarithmic mean squared error preserves plume geometry well (SSIM = 0.829), but seriously underestimates the integrated pollutant burden, with a geometric mean bias (MG) of 2.75 and a median ratio of integrated mass (MIM) = 0.13. This is a well-known phenomenon of log-domain optimization in which plume cores with high concentration cores are suppressed with the majority of the mass transported.

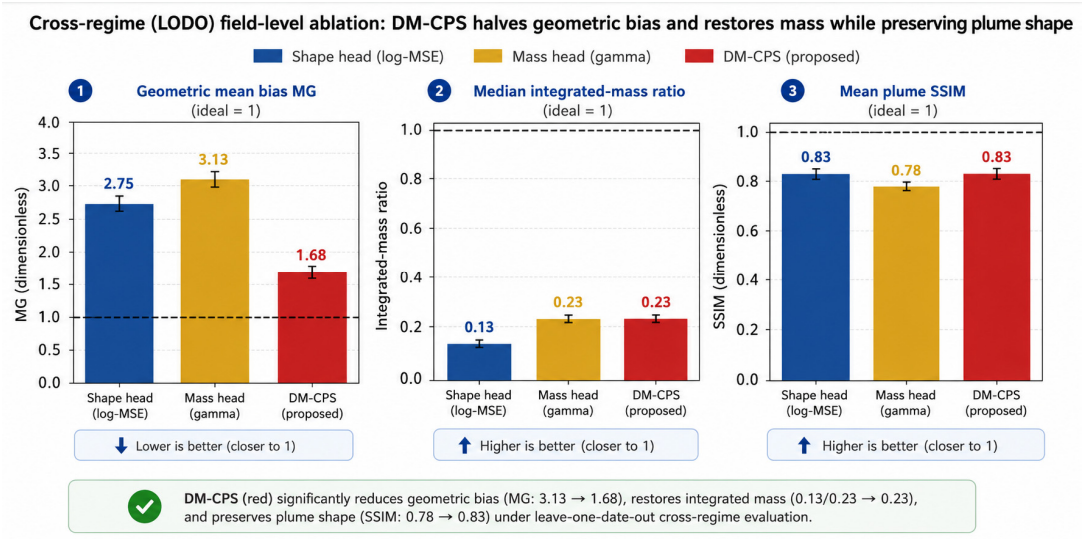

Figure 6. Field-level ablation out of fold (leave-one-date-out, 1908 plume fields) for the three variants; the dashed line is the ideal value of 1 in every panel. (**Left**) geometric mean bias MG—DM-CPS (orange) halves the shape head's bias and is the closest to 1. (**Centre**) median integrated-mass ratio—DM-CPS and the gamma head nearly double the conserved mass relative to the log-MSE shape head. (**Right**) mean plume SSIM—DM-CPS preserves the shape head's structural fidelity, whereas the gamma head alone degrades it. DM-CPS thus captures the gamma head's mass while keeping the shape head's geometry.

Optimization in the original concentration domain is also effective when estimating the contribution of the pollutants: In this case the integrated-mass ratio (0.23) of the mass head is the same for the gamma-loss mass head. It significantly reduces the plume structure (SSIM = 0.782) and also has significantly lower predictive accuracy ($R^2 = -1.15$). As a result, no single branch can serve as a suitable surrogate—the shape head is good at capturing plume morphology but it does not capture integrated mass; the mass head is good at capturing integrated mass but not spatial fidelity. Their merger in the proposed DM-CPS scheme successfully achieves a balance of these complementary goals.

Under LODO evaluation, the mass-head per se must extrapolate to the meteorological regimes not included in the training data and perfect agreement with the HYSPLIT reference

is not to be expected even though the field-level scaling ensures consistency with the mass-head prediction for the meteorological regimes represented in the training data. This means that the residual bias is mainly due to lack of cross-regime generalisation, and not due to problems in the mass consistency procedure.

The proposed DM-CPS is the field-level mass-consistency scaling procedure which combines the complementary predictions of shape and mass heads. The geometric bias is reduced from 2.75 to 1.68 with the addition of DM-CPS, bringing MG much closer to the desired value of one, while the median integrated-mass ratio is increased from 0.13 to 0.23. Importantly, it is done without compromising plume morphology (SSIM: 0.829 $\rightarrow$ 0.828) and with a similar operational dispersion skill (FAC2: 0.29 $\rightarrow$ 0.27). Overall, the results show that mass correction at the field level leads to significant improvement in physical consistency and does not adversely affect the geocrine structure of the predicted plume.

Table 4 shows the field-level ablation of DM-CPS and its two constituent heads under the leave-one-date-out evaluation protocol.

Figure 6 shows the field-level ablation of MG, MIM and SSIM under the leave-one-date-out protocol, comparing the shape head, mass head and DM-CPS.

4.3. Within-Regime Skill and Computational Performance

The conventional random hold-out evaluation shows that the proposed surrogate (DM-CPS, Table 3) is able to reproduce the concentration fields from HYSPLIT with $R^2 = 0.67$, FAC2 = 0.57, MG = 0.69, and a transport-distance correlation of r = 0.86. The FAC2 value is greater than 0.5 considered acceptable for operational atmospheric dispersion models [43] and is good agreement with the HYSPLIT reference under meteorological conditions contained in the training set.

However, Figure 7 shows permutation feature importance, which indicates that the surrogate is mainly based on physically meaningful predictors. The most influential variables are those characterising the transport geometry such as the lead time, longitude, latitude, transport distance and source bearing, indicating the significance of the advection pathways for the evolution of the plume. Air temperature has the most powerful influence among continuous meteorological variables, which is in line with the influence of atmospheric stability and winter thermal inversions on $PM_{2.5}$ dispersion over Almaty [32]. The other dominant predictor is the meteorological-regime label, indicating that the surrogate has acquired different dispersion properties for different atmospheric regimes.

A significant computational benefit over numerical dispersion modelling is also achieved by the surrogate. A full concentration field, with about 3200 concentration cells is predicted in 35 ms, which is about 8.8×10^4 cell predictions per second on one CPU core. This is about 4 to 5 orders of magnitude faster than HYSPLIT and thus allows for large-scale ensemble simulations, sensitivity analysis and uncertainty quantification to be computed practically.

In addition, the meteorological-regime predictor was a major factor, suggesting the limitations of the present surrogate. Since these models are sensitive to regime, it is believed that the model's performance will degrade for meteorological regimes not represented in the training set. In the subsequent section this hypothesis is explored using the more conservative leave-one-date-out (LODO) validation protocol.

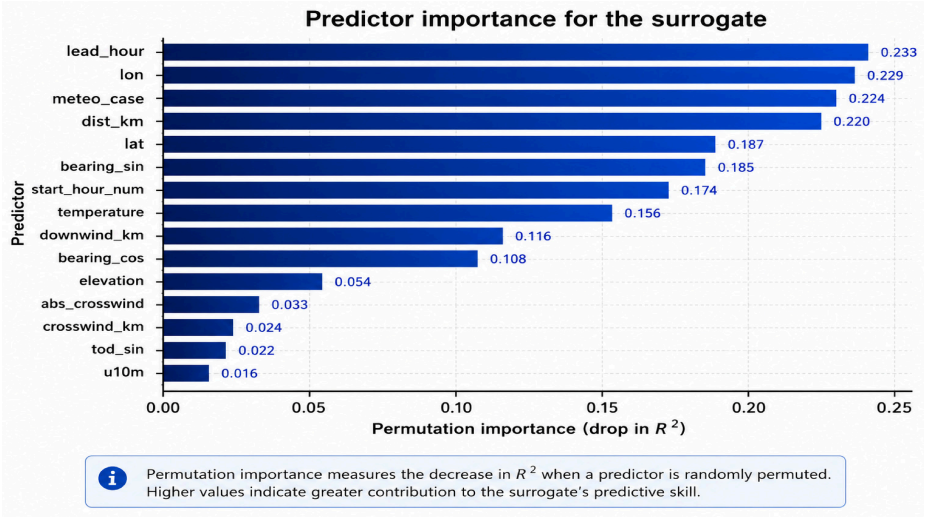

Figure 7. Permutation feature importance, measured as the mean decrease in R^2 after randomly shuffling each predictor. Larger decreases indicate greater influence on model performance. Transport-related variables—including lead time, geographic position, transport distance, and source bearing—are the dominant predictors, followed by the meteorological-regime label and air temperature. The high importance of the regime label explains the reduction in predictive performance under leave-one-date-out validation, where the surrogate must generalize to meteorological conditions not represented during training.

4.4. Cross-Regime Generalisation

The leave-one-date-out (LODO) evaluation is significantly harder to perform than the other methods tested in the present study as it tests the model under meteorological conditions that are not present in the training data. For the shape head, the pooled performance across the eight withheld meteorological days under this protocol is $R^2 = -0.38$ and FAC2 = 0.29 (Table 4), meaning that a surrogate trained on just seven representative $^2 = -0.40$, FAC2 = 0.27, Table 4).

The accuracy of the prediction is very different under different weather conditions. While there are moderate predictive skills in transition seasons ($R^2 \approx 0.2$), distinct winter stagnation regimes yield significantly lower accuracy (R^2 as low as -1.9). This behavior is similar to the permutation importance analysis (Figure 7) where the meteorological-regime label is one of the most important predictors. As a result, by withholding an entire meteorological regime, information is lost on which the surrogate depends to recover the same dispersion patterns.

The large difference between the R^2 values for the random hold-out evaluation (0.67) and the LODO evaluation (-0.40, Table 4) underscores the influence of the validation approach on the evaluation of the atmospheric dispersion surrogates; a negative R^2 under LODO means that, pooled across the eight withheld days, DM-CPS's pointwise predictions are less accurate than simply predicting the training-set mean, which is a stronger and more explicit statement of the generalisation gap than the random hold-out R^2 alone conveys. Random train-test partitions are mostly based on interpolating between conditions already seen in the training set, and may overestimate operational performance since the states of neighboring grid cells within a plume appear in both the training set and the test set. LODO, however, considers a more realistic task of predicting dispersion under atmospheric conditions not previously encountered, and hence a more realistic estimation of the operational performance.

The results indicate that the main problem lies in the small-scale diversity in the training meteorological library, not with the proposed architecture of dual heads. The present dataset consists of some 1.88 million grid-cell observations, but the observations

are from just 96 HYSPLIT simulations for 8 representative meteorological days. The eight meteorological regimes simulated here compare favourably in scope with several previous atmospheric-dispersion surrogate studies, which have typically relied on a small number of illustrative cases or a single meteorological scenario [27,28]; nonetheless, the most important direction for future work is to substantially increase this number, extending the simulated regimes across both years and seasons to better sample the full range of atmospheric conditions.

Figure 8 shows the out-of-fold predicted-versus-reference density plot under the leave-one-date-out protocol, illustrating the surrogate's degraded skill under cross-regime evaluation.

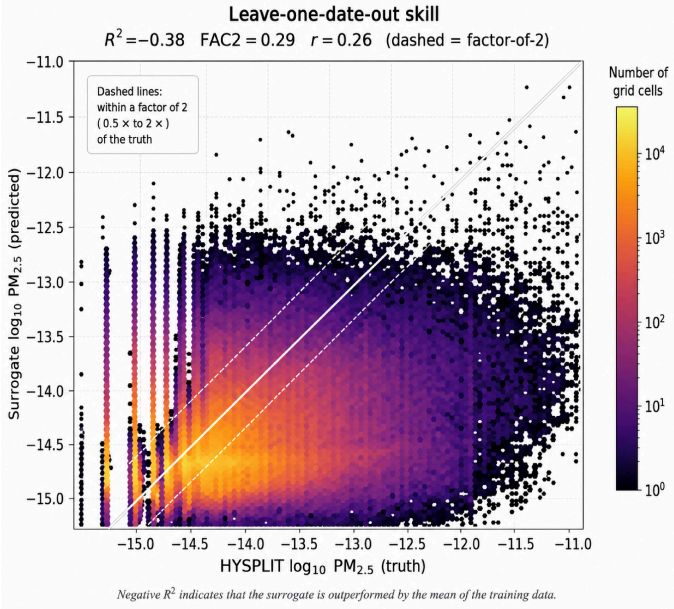

Figure 8. Out-of-fold predictions under the leave-one-date-out (LODO) validation protocol shown as a two-dimensional density plot of predicted versus HYSPLIT $\log_{10}$ $PM_{2.5}$ concentrations. Brighter colors indicate higher sample density. The solid white line denotes the ideal one-to-one relationship, while the dashed lines represent the factor-of-two bounds commonly used in atmospheric dispersion evaluation. Most predictions remain concentrated near the one-to-one line; however, the increasing vertical spread for higher concentrations indicates regression toward the mean when the surrogate is applied to previously unseen meteorological regimes. This behavior reflects the limited diversity of the training meteorological library rather than deficiencies in within-regime interpolation. The faint vertical banding arises from the finite numerical precision of the HYSPLIT concentration output.

4.5. Uncertainty Calibration

Table 5 shows the uncertainty-quantification results from the quantile-regression and bootstrap-ensemble analyses.

Both quantile regression and bootstrap ensemble methods (Figure 4) were used to test the proposed uncertainty estimation framework. The prediction intervals from the quantile regression model were well calibrated on the random hold-out evaluation. The empirical coverage of the nominal 90% prediction interval was 90.4%, very close to the nominal value of 0.9, and the mean of the widths of the intervals was 1.58 $\log_{10}$ concentration units. Most importantly, the median quantile predictor had high accuracy in predicting the quantiles (R^2 = 0.69), allowing for reliable uncertainty estimates without compromising the point predictions significantly.

Under the leave-one-date-out evaluation, by contrast, the bootstrap-ensemble assessment provided comparatively little useful information about predictive uncertainty: the ensemble spread was not strongly correlated with the actual prediction error, suggesting that ensemble spread does not adequately represent the dominant source of uncertainty in

this cross-regime setting. The highest prediction error is not caused by parameter uncertainty but by scenarios outside of the training set which are presented in the atmosphere.

The results reveal an interesting difference between interpolation uncertainty and cross-regime uncertainty. Quantile regression estimates uncertainty in the meteorological conditions included in the training dataset. However, the model variance is not the source of uncertainty under the new, unobserved atmospheric regimes. Thus, to improve uncertainty quantification for operational deployment, it will be necessary to either greatly expand the size and diversity of the meteorological training library, or to develop surrogate architectures that explicitly model the distribution of atmospheric concentration fields, like diffusion-based generative models.

Figure 4 shows the uncertainty assessment of the surrogate, relating bootstrap-ensemble spread to prediction error and evaluating the calibration of the quantile-regression intervals.

4.6. Plume Fidelity and 24-H Regional Transport

Three meteorological cases were excluded from the analysis and are shown as representative leave-one-date-out plume reconstructions in Figure 9. The reference fields (left column) are the HYSPLIT fields, and the corresponding surrogate prediction (center column) is shown along with the prediction error (right column). In all examples, the surrogate well captures the overall plume geometry, with high structural similarity (SSIM = 0.87–0.94), transport direction, spatial extent, and large-scale concentration gradients. The majority of the prediction errors remain in the high concentration region of the plume core, where the surrogate can tend to dampen out local peaks in concentrations while generally maintaining the overall transport pattern. The qualitative results obtained are consistent with the quantitative results reported in Table 4 and illustrate that the plume morphology remains unchanged even when validating the model against cross-regime data, that is, when the meteorological conditions are outside of those used for the model's validation. The main errors present in the model are the errors in estimating maximum concentrations for a condition, which is different from the one that the model has been validated for.

The HYSPLIT simulation ensemble also gives some information on regional transport features of the $PM_{2.5}$ emitted from central Almaty. In all simulations, the mean distance travelled by the plume centroid in the first 24 h is 98 km, the 10th–90th percentile is from 16 km to 228 km, and the maximum centroid travel distance is 316 km. This centroid-travel distance is distinct from the maximum spatial extent (footprint) of the whole plume envelope reported in Table 7, which reaches 579.0 km (heating season) and 709.4 km (non-heating season): the plume envelope extends beyond the centroid position because it captures the leading edge of the dispersing plume rather than its center of mass, so the two quantities are not expected to match. The mean maximum plume extent across all 96 runs is 640 km, with individual runs reaching up to 709 km in the non-heating season.

Table 7. Simulated 24-h transport by season (48 runs each).

Season	Runs	Mean Transport (km)	Std (km)	Max Extent (km)
heating	48	114.3	98.9	579.0
non_heating	48	81.0	56.4	709.4

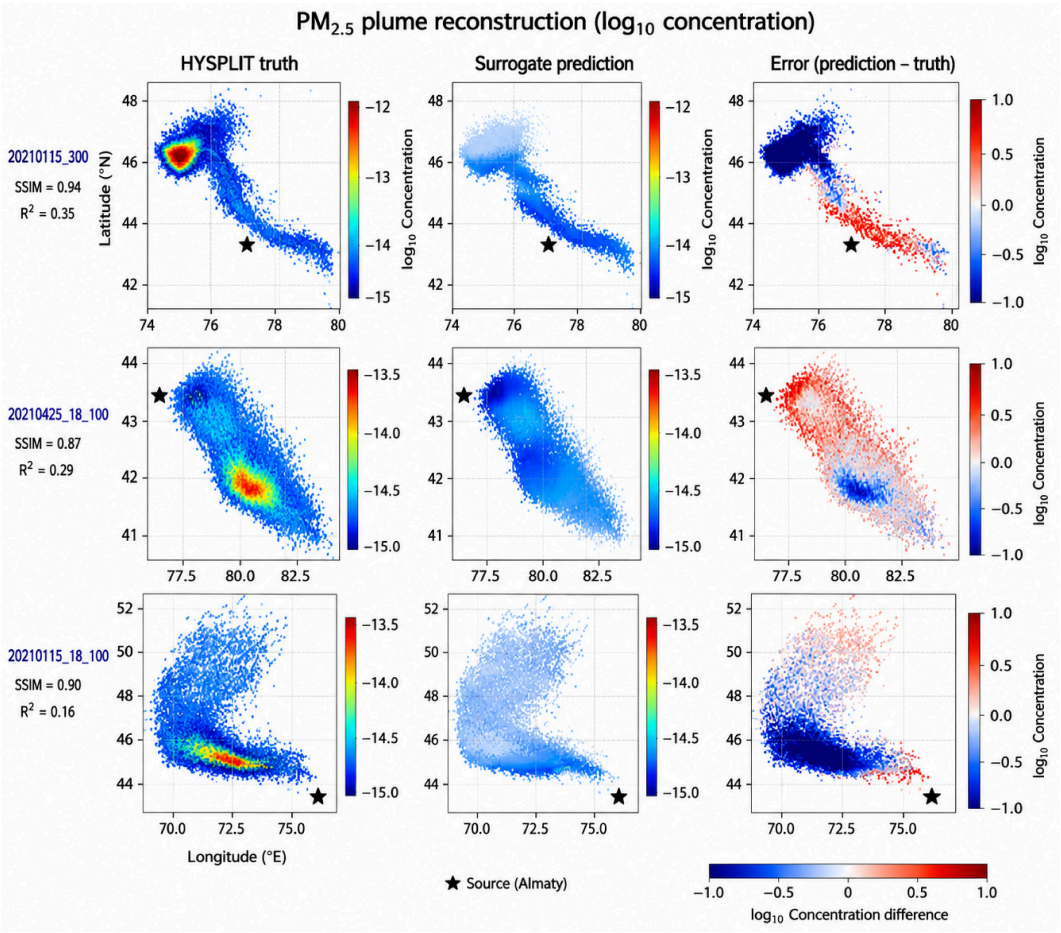

Figure 9. Representative leave-one-date-out plume reconstructions for three withheld meteorological cases. Columns show the HYSPLIT reference field (**left**), the surrogate prediction (**center**), and the prediction error (**right**), where red indicates overestimation and blue indicates underestimation. The black star denotes the emission source. Across all examples, the surrogate accurately reproduces plume orientation, transport direction, and spatial extent, resulting in consistently high structural similarity (SSIM). The remaining errors are concentrated within the plume core where concentrations are higher and where logarithmic optimization tends to underestimate high concentrations, and motivates the field-level mass correction introduced by DM-CPS.

There is also a strong seasonal variation in the transport characteristics (Table 7). The simulations for the heating season also yield longer average transport distances (114 km), a result of the greater influence of the winter atmospheric circulation and stability conditions on regional pollutant transport. Although it is not as accurate on pointwise basis under leave-one-date-out validation, the surrogate keeps the same large-scale transport behavior, using mean absolute error of 33 km and a correlation coefficient of 0.80 for the run-level transport distances. Given this observation, this proposed surrogate is still useful for applications that emphasise regional pathways for transport, as well as ensemble analysis, despite the regime-crossing issues.

Figure 9 shows representative leave-one-date-out plume reconstructions for three withheld meteorological cases, comparing the HYSPLIT reference, the surrogate prediction, and the resulting error field.

Table 7 shows the simulated 24-h transport distances by season.

Figure 10 shows the regional transport characteristics derived from the full HYSPLIT simulation ensemble, including plume-centroid trajectories and seasonal transport-distance distributions.

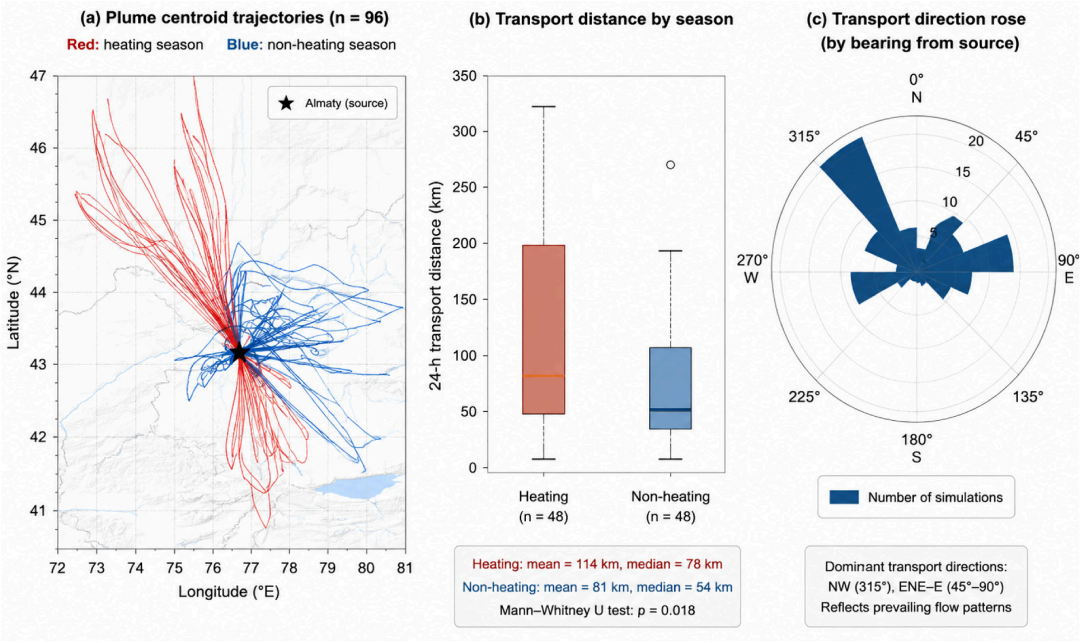

Figure 10. Regional transport characteristics derived from the HYSPLIT simulation ensemble. (**a**) Concentration-weighted plume-centroid trajectories originating from the Almaty emission source (black star), with heating-season simulations shown in red and non-heating simulations in blue. (**b**) Distribution of 24-h transport distances by season, illustrating the greater average transport during the heating season. (**c**) Polar histogram of transport bearings summarizing the dominant directions of regional pollutant transport. Together, these results illustrate the substantial spatial variability of $PM_{2.5}$ transport under different meteorological conditions.

5. Discussion

The proposed DM-CPS framework overcomes a fundamental drawback in the use of traditional surrogates of atmospheric dispersion: the use of a single regression objective to optimize plume geometry and integrated pollutant burden. This splits the prediction task into two complementary tasks, each of which is optimized by the loss function that is most suitable for the task. The shape head is based on a squared error loss in the log concentration scale, and is able to accurately recover the morphology of the plume over a large concentration range, while the mass head is based on a Gamma deviance loss in the original concentration scale and estimates the integrated pollutant loading. A deterministic field level mass-consistency scaling is then used to reconcile the two predictions while maintaining the spatial pattern predicted by the shape head, but enhancing the field level physical consistency. The tested design significantly lowers the geometric bias (2.75 to 1.68) and almost doubles the median integrated-mass ratio (0.13 to 0.23) while maintaining the plume morphology (SSIM $\approx$ 0.83). Furthermore, DM-CPS has the best factor-of-two agreement (FAC2 = 0.57) of the surrogate models evaluated. Because the mass-consistency fusion module operates only on the output fields of the shape and mass heads (Equations (5) and (6)) and does not depend on the internal structure of either sub-model, the same fusion scheme is, in principle, applicable to other pointwise surrogate models that share the well-known mass-underestimation problem of logarithmic optimisation. This adaptability has not itself been empirically tested in this study—all experiments here use histogram gradient-boosted trees for both heads—and remains a direction for future work rather than a demonstrated result.

The second important discovery is the great difference between the two types of generalizations; interpolation and true meteorological generalization. The performance of the surrogate under random hold-out evaluation is comparatively high (R^2 = 0.67), while

the leave-one-date-out (LODO) validation shows much poorer performance ($R^2 = -0.40$). This gap arises because, under the random train/test split, neighbouring grid cells from the same plume can appear in both the training and test sets; the surrogate can then partly interpolate between spatially correlated points it has effectively already seen, which inflates the apparent operational performance relative to the harder, more realistic LODO setting. On the other hand, LODO considers the practically relevant case of predicting the dispersion from a set of meteorological conditions that have never before been encountered. The performance difference observed indicates that the proposed architecture is not the major constraint but rather the restricted variety of available meteorological library comprising only eight representative days. The limited benefit of DM-CPS under cross-regime evaluation stems from the fact that neither the shape head nor the mass head can reliably extrapolate to atmospheric conditions absent from the training data, so the fusion step has little correct information from either branch to combine in that setting. The results highlight the need for further development, which will, in particular, rely on the availability of more diverse meteorological simulations, such as multi-year reanalysis datasets [35] and more detailed simulations from WRF-driven meteorological fields [34].

Table 8 shows a qualitative comparison of DM-CPS against representative surrogate-modelling approaches from the literature.

Table 8. Qualitative comparison of DM-CPS with representative surrogate-modelling approaches from the literature reviewed in Section 2.

Approach	Accuracy	Mass Consistency	Training Efficiency	Cross-Regime Tested	Interpretability
RF/GBT, raw predictors [15–20]	Moderate–good pointwise fit; no mass guarantee	No	High (min–hr, CPU)	Rarely reported	Moderate (feat. importance)
Deep CNN/ConvLSTM [21,22]	Good spatial fit in-domain	No	Lower (GPU)	Rarely reported	Low–moderate
Physics-informed CNN/PINN [25,27]	Improves local plausibility	Partial (soft constraint)	Lower (GPU)	Rarely reported	Moderate
FLEXPART emulator [28]	Good within-distribution fit	No	Not comparable	Not reported	Moderate
CTM bias-correction GBT [29,30]	Improves on raw CTM; spatial hold-out in [30]	No	High	Partial (spatial only)	High (SHAP [29])
DM-CPS (this study)	FAC2 = 0.57 (best, Table 3); R^2 modest, lower under LODO	Yes (Equations (5) and (6))	High (<1 min/model, Table 3)	Yes ($R^2 < 0$, Table 4)	Moderate–high (Figure 7)

Despite being a discriminative surrogate, the architecture proposed here serves as the basis for more complicated field-level generative models. Methods based on diffusion [46] and attention [47] have the promise of modeling full concentration fields directly instead of predicting each grid cell individually. Such models might explicitly model plume boundaries, naturally model zones of zero concentration, and be sampled to create ensembles of probabilistic models that offer more information about uncertainties than deterministic regression. In the future, the DM-CPS framework and the harsh LODO evaluation protocol developed in this investigation serve as an excellent baseline to evaluate future generations of atmospheric surrogates.

Some caveats should be noted. First, training is limited to the grid cells in the simulated plume region, so the surrogate can no longer predict what happens outside the plume region, which means it can't calculate the boundaries of the plume and can't distinguish

whether a location is polluted or not. Future models would be able to estimate extent and magnitude of the plumes simultaneously if the zero-concentration grid cells were included. Second, this study is based on GDAS1 meteorological data, which have a relatively coarse spatial resolution that can not fully reflect the complex mountain–valley circulation in the Almaty region. Third, despite the large number of grid-cell records in the dataset (c. 1.88 million), only 96 HYSPLIT simulations have been included, covering a relatively small number of the days in the dataset, 8 of which are representative meteorological days used for training and evaluation of the dataset. Lastly, the surrogate is also not directly observed in the atmosphere, but instead validated against HYSPLIT simulations. Comparisons with surface air-quality monitoring networks [48] and satellite aerosol optical depth retrievals [49] and operational HYSPLIT products [50] should also be made, as well as further refinement of the characterization of the source terms [51,52] to evaluate the absolute physical realism of the proposed surrogate under real-world conditions.

6. Conclusions

This study proposed a physics-informed machine-learning structure, the Dual-Head Mass-Consistent Plume Surrogate (DM-CPS), to speed up the HYSPLIT $PM_{2.5}$ transport simulation and make it more consistent. Instead of the typical surrogate modeling approach that minimizes a single regression target, DM-CPS splits the prediction of plume geometry and integrated pollutant burden into two complementary models, each based on a histogram gradient boosting approach, and reconciles the results of the two models using a deterministic mass-consistency scaling procedure. The main feature of this design is to maintain the spatial structure learnt in logarithmic space while enhancing the representation of the integrated pollutant mass.

An ensemble of 96 HYSPLIT model simulations driven by the GDAS was used to assess the proposed framework over Almaty, Kazakhstan with around 1.88 million grid-cell records. Controlled comparisons showed that DM-CPS outperformed Gaussian-plume regression, random forest and multilayer perceptron and comparable or better to a traditional gradient-boosted baseline in terms of operational dispersion skill (FAC2 = 0.57), had a much smaller geometric bias and improved the integrated mass estimates, while maintaining plume morphology skill. The surrogate also generated HYSPLIT concentration fields about 4–5 orders of magnitude faster: predicting a whole field took about 35 ms with the surrogate, compared with roughly 6–8 h for a single HYSPLIT run of this configuration (24-h forward simulation with hourly output on one CPU core of the NCAR Derecho system used in Section 3.1) (approximately 6.8 h on average across the 96 runs on a single Derecho CPU core).

This study reveals that when considering dispersion surrogate optimization problems, optimization of conventional regression metrics is not enough. While modest in terms of pointwise prediction accuracy, the proposed framework yields more physically consistent concentration fields as well as improved integrated mass estimates, at the same time maintaining the geometry of the plumes. The results show that machine-learning surrogates that embed physically meaningful targets can enhance their value in environmental modelling, beyond the classical statistical accuracy criteria, and can be used as inputs for novel model types.

Another interesting finding was that the leave-one-date-out evaluation showed a significant cross-regime generalization gap when the model is evaluated on unknown meteorological conditions. This outcome shows that the biggest obstacle in the current system is not the surrogate architecture, but limited diversity of available meteorological training data. Future work should therefore concentrate on an extension of the library of simulations to many more meteorological scenarios for several years, seasons and atmospheric stability

regimes. Future studies will explore hybrid physics-informed learning strategies, probabilistic surrogate models, and transfer-learning techniques for more robust performance under out-of-sample weather conditions, while maintaining computational tractability.

In summary, the proposed DM-CPS framework provides an accurate and computationally efficient surrogate for atmospheric dispersion modelling by combining physics-informed feature engineering with field-level mass prediction through a mass-consistent principle. Once trained, the surrogate enables the rapid generation of large ensembles of dispersion predictions without repeatedly executing computationally expensive HYSPLIT simulations. This capability facilitates uncertainty quantification, sensitivity analysis, probabilistic forecasting, and operational decision support under diverse meteorological scenarios. Furthermore, the study highlights the importance of comprehensive inter-regime validation to ensure the operational reliability and generalizability of machine-learning surrogates.

Author Contributions: F.A., A.S. and G.C.; methodology, F.A., A.S. and G.C.; software, F.A.; validation, F.A. and A.S.; formal analysis, F.A. and A.S.; investigation, F.A.; resources, G.C.; data curation, F.A.; writing—original draft preparation, F.A. and A.S.; writing–review and editing, F.A., A.S. and G.C.; visualization, F.A.; supervision, G.C.; project administration, G.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24992908 «Support system for agricultural crop production optimization via remote monitoring and artificial intelligence methods (Agroscope)») and BR21881908 Complex of urban ecological support (CUES).

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: This research was partially covered by the Institute for Computational and Data Sciences at Penn State. The authors sincerely thank Ying Pan, Department of Meteorology and Atmospheric Science, The Pennsylvania State University, for her valuable comments and guidance in atmospheric dispersion modeling. The authors acknowledge high-performance computing support from the Derecho system (https://doi.org/10.5065/qx9a-pg09) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the U.S. National Science Foundation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Al Alim, M.; Hoque, M.; Hore, S.; Hore, R. Environmental impacts of transportation in rapidly urbanizing South and Southeast Asian Cities. *Smart Constr. Sustain. Cities* **2025**, *3*, 16. [CrossRef]
2. Lichiheb, N.; Ngan, F.; Cohen, M. Improving the atmospheric dispersion forecasts over Washington, DC using UrbanNet observations: A study with HYSPLIT model. *Urban Clim.* **2024**, *55*, 101948. [CrossRef]
3. Schade, G.W.; Gregg, M.L. Testing HYSPLIT plume dispersion model performance using regional hydrocarbon monitoring data during a gas well blowout. *Atmosphere* **2022**, *13*, 486. [CrossRef]
4. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat, F. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef] [PubMed]
5. Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421. [CrossRef] [PubMed]
6. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef] [PubMed]
7. Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. *arXiv* **2022**, arXiv:2202.11214. [CrossRef]
8. Küçük, Ç.; Giannakos, A.; Schneider, S.; Jann, A. Transformer-based nowcasting of radar composites from satellite images for severe weather. *Artif. Intell. Earth Syst.* **2024**, *3*, e230041. [CrossRef]

9. Van Leuven, S.; De Meutter, P.; Camps, J.; Termonia, P.; Delcloo, A. An optimisation method to improve modelling of wet deposition in atmospheric transport models: Applied to FLEXPART v10.4. *Geosci. Model Dev.* **2023**, *16*, 5323–5338. [CrossRef]
10. Zhang, X.; Wang, S.; Jin, G.; Yuan, Z.; Yuan, H.; Ruan, S. Eulerian neural network informed by chemical transport for air quality forecasting. In *Advances in Neural Information Processing Systems*; Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2025; Volume 38, pp. 33576–33601.
11. Huang, Z.; Fei, T.; Bian, M. A spatial explicit method combining HYSPLIT and random forest for predicting the next-day pollen index at urban scale. *Aerobiologia* **2026**, *42*, 6. [CrossRef]
12. Moffatt, P.L.; Fowler, A.F.; Sánchez-Murillo, R.; Boll, J. Isotopic variability in precipitation: Integrating HYSPLIT raster data into isoscape modelling. *Hydrol. Processes* **2026**, *40*, e70393. [CrossRef]
13. Chavan, A.; Bokade, V.; Lataye, D.H. Seasonal dispersion and trajectory-based source assessment of particulate air pollutants: An integrated monitoring and HYSPLIT modeling approach. *Environ. Monit. Assess.* **2026**, *198*, 203. [CrossRef] [PubMed]
14. Geiser, V. Predicting atmospheric trace substance concentrations using supervised machine learning and HYSPLIT backward trajectories. *Artif. Intell. Earth Syst.* **2025**, *4*, e240051. [CrossRef]
15. Sanzhar, S.; Serek, A. Scalable Heart Disease Prediction Using Distributed Machine Learning with Apache Spark. In Proceedings of the 18th International Conference on Electronics, Computer, and Computation (ICECCO), Kaskelen, Kazakhstan, 10–11 April 2026; pp. 1–6. [CrossRef]
16. Ashim, A.Z.; Botanov, A.; Abdoldina, F.; Serek, A. Salary Prediction from Job Descriptions Using Attention-Based NLP Models. *Her. Kazakh-Br. Tech. Univ.* **2025**, *22*, 168–177. [CrossRef]
17. Deng, X.; Ye, A.; Zhong, J.; Xu, D.; Yang, W.; Song, Z.; Zhang, Z.; Guo, J.; Wang, T.; Tian, Y.; et al. Bagging–XGBoost algorithm based extreme weather identification and short-term load forecasting model. *Energy Rep.* **2022**, *8*, 8661–8674. [CrossRef]
18. Wu, Z.; Fan, X.; Bian, G.; Liu, Y.; Zhang, X.; Chen, Y. Short-term wind power forecast with turning weather based on DBSCAN-RFE-LightGBM. *Renew. Energy* **2025**, *251*, 123217. [CrossRef]
19. Hiroki, R.; Varquez, A.C.; Khanh, D.N.; Rynazal, R.; Renard, F.; Alonso, L.; Kanda, M. Long-term changes of Universal Thermal Climate Index (UTCI) estimated from weather stations and gradient-boosted decision trees throughout Japan. *Int. J. Climatol.* **2025**, *45*, e8843. [CrossRef]
20. Liu, M.; Chen, H.; Wei, D.; Wu, Y.; Li, C. Nonlinear relationship between urban form and street-level PM2.5 and CO based on mobile measurements and gradient boosting decision tree models. *Build. Environ.* **2021**, *205*, 108265. [CrossRef]
21. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
22. Wu, R.; Liang, Y.; Lin, L.; Zhang, Z. Spatiotemporal multivariate weather prediction network based on CNN-Transformer. *Sensors* **2024**, *24*, 7837. [CrossRef] [PubMed]
23. Bahadur, F.T.; Shah, S.R.; Nidamanuri, R.R. Geospatial time series forecasting of air quality across Delhi using ARIMA and LSTM models. *J. Atmos. Sol.-Terr. Phys.* **2026**, *279*, 106723. [CrossRef]
24. Barthwal, A.; Goel, A.K. Advancing air quality prediction models in urban India: A deep learning approach integrating DCNN and LSTM architectures for AQI time-series classification. *Model. Earth Syst. Environ.* **2024**, *10*, 2935–2955. [CrossRef]
25. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
26. Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic weather forecasting with machine learning. *Nature* **2025**, *637*, 84–90. [CrossRef] [PubMed]
27. Mendil, M.; Leirens, S.; Armand, P.; Duchenne, C. Hazardous atmospheric dispersion in urban areas: A deep learning approach for emergency pollution forecast. *Environ. Model. Softw.* **2022**, *152*, 105387. [CrossRef]
28. Gunawardena, N.; Pallotta, G.; Simpson, M.; Lucas, D.D. Machine learning emulation of spatial deposition from a multi-physics ensemble of weather and atmospheric transport models. *Atmosphere* **2021**, *12*, 953. [CrossRef]
29. Sukprasert, S.; Thanasutives, P.; Araki, S.; Uranishi, K.; Matsuo, T.; Shimadera, H. A comparative study on CMAQ-based predictors and reanalysis AOD for improving PM2.5 estimation in Thailand using machine learning. *Atmos. Environ.* **2026**, *375*, 122006. [CrossRef]
30. Sukprasert, S.; Thanasutives, P.; Araki, S.; Uranishi, K.; Matsuo, T.; Shimadera, H. Machine learning based bias correction of CMAQ using emission source decoupled features in Thailand. *GEOMATE J.* **2026**, *30*, 116–125. [CrossRef]
31. Kerimray, A.; Azbanbayev, E.; Kenessov, B.; Plotitsyn, P.; Alimbayeva, D.; Karaca, F. Spatiotemporal variations and contributing factors of air pollutants in Almaty, Kazakhstan. *Aerosol Air Qual. Res.* **2020**, *20*, 1340–1352. [CrossRef]
32. Tursumbayeva, M.; Kerimray, A.; Karaca, F.; Permadi, D.A. Planetary boundary layer and its relationship with PM2.5 concentrations in Almaty, Kazakhstan. *Aerosol Air Qual. Res.* **2022**, *22*, 210294. [CrossRef]

33. Kerimray, A.; Baimatova, N.; Ibragimova, O.P.; Bukenov, B.; Kenessov, B.; Plotitsyn, P.; Karaca, F. Assessing air quality changes in large cities during COVID-19 lockdowns: The impacts of traffic-free urban conditions in Almaty, Kazakhstan. *Sci. Total Environ.* **2020**, *730*, 139179. [CrossRef] [PubMed]
34. Ngan, F.; Stein, A.; Draxler, R. Inline coupling of WRF-HYSPLIT: Model development and evaluation using tracer experiments. *J. Appl. Meteorol. Climatol.* **2015**, *54*, 1162–1176. [CrossRef]
35. Kalnay, E.; Kanamitsu, M.; Kistler, R.; Collins, W.; Deaven, D.; Gandin, L.; Iredell, M.; Saha, S.; White, G.; Woollen, J.; et al. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 437–471. [CrossRef]
36. Loken, E.D.; Clark, A.J.; McGovern, A. Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Weather Forecast.* **2022**, *37*, 871–899. [CrossRef]
37. Gill, K.S.; Sharma, A.; Anand, V.; Gupta, R. Detection of weather with hypothesis testing performed through VGG19 model utilizing Adam optimizer. In *Sustainable Mobility*; Kumar, A., Prasad, A., Kumar, G., Eds.; Wiley: Hoboken, NJ, USA, 2024; Chapter 15. [CrossRef]
38. Meinshausen, N. Quantile regression forests. *J. Mach. Learn. Res.* **2006**, *7*, 983–999.
39. Sluijterman, L.; Cator, E.; Heskes, T. Confident neural network regression with Bootstrapped Deep Ensembles. *Neurocomputing* **2025**, *656*, 131500. [CrossRef]
40. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
41. Penalun, F.E.; Hermawan, A.; Avianto, D.; Pramudwiatmoko, A. A Multi-Layer Perceptron regression and variant windowing for estimating rainfall based on weather radar data. *J. Educ. Sci.* **2024**, *33*, 1421. [CrossRef]
42. Rajamani, S.K.; Iyer, R.S. Machine learning-based mobile applications using Python and Scikit-Learn. In *Designing and Developing Innovative Mobile Applications*; Samanta, D., Ed.; IGI Global: Hershey, PA, USA, 2023; Chapter 16; pp. 282–306. [CrossRef]
43. Chang, J.C.; Hanna, S.R. Air quality model performance evaluation. *Meteorol. Atmos. Phys.* **2004**, *87*, 167–196. [CrossRef]
44. El Bilali, A.; Brouziyne, Y.; Attar, O.; Lamane, H.; Hadri, A.; Taleb, A. Physics-informed machine learning algorithms for forecasting sediment yield: An analysis of physical consistency, sensitivity, and interpretability. *Environ. Sci. Pollut. Res.* **2024**, *31*, 47237–47257. [CrossRef] [PubMed]
45. Zehtabiyan-Rezaie, N.; Iosifidis, A.; Abkar, M. Physics-guided machine learning for wind-farm power prediction: Toward interpretability and generalizability. *PRX Energy* **2023**, *2*, 013009. [CrossRef]
46. Li, L.; Carver, R.; Lopez-Gomez, I.; Sha, F.; Anderson, J. Generative emulation of weather forecast ensembles with diffusion models. *Sci. Adv.* **2024**, *10*, eadk4489. [CrossRef] [PubMed]
47. Orozco López, E.; Kaplan, D.; Linhoss, A. Interpretable transformer neural network prediction of diverse environmental time series using weather forecasts. *Water Resour. Res.* **2024**, *60*, e2023WR036337. [CrossRef]
48. Ji, D.; Sun, X.; Ritter, C.; Notholt, J. Retrieval of aerosol composition from spectral aerosol optical depth and optical properties using a machine learning approach. *Atmos. Meas. Tech.* **2026**, *19*, 421–436. [CrossRef]
49. Yousefi Kebriya, A.; Nadi, M.; Parmehr, E.G.; Sun, Z. Analysis and source identification of dust event in Khuzestan province using MODIS and SENTINEL satellite data and HYSPLIT model (2018–2022). *PLoS ONE* **2026**, *21*, e0348129. [CrossRef] [PubMed]
50. Lai, H.C.; Yu-Tung, D.; Mkasimongwa, S.W.; Hsiao, M.C.; Li-Wei, L. The impact of atmospheric synoptic weather condition and long-range transportation of air mass on extreme PM10 concentration events. *Atmosphere* **2023**, *14*, 406. [CrossRef]
51. Ngan, F.; Loughner, C.P.; Zinn, S.; Cohen, M.; Lee, T.R.; Dumas, E.; Schuyler, T.J.; Baker, C.B.; Maloney, J.; Hotz, D.; et al. The use of small uncrewed aircraft system observations in meteorological and dispersion modeling. *J. Appl. Meteorol. Climatol.* **2023**, *62*, 817–834. [CrossRef]
52. Kanchanachat, W.; Pattarapanitchai, S.; Wang, S.H.; Warunsin, K.; Buntoung, S. Characterizing PM2.5 variability and cross-boundary aerosol transport in Central Thailand using low-cost sensors and Himawari-9 data. *Aerosol Air Qual. Res.* **2026**, *26*, 32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.