Reassessing the Reliability of WRF-CMAQ-Based $PM_{2.5}$ Prediction from a Generalization Perspective

*Xinqi Wei [1], Qin He [2], Kai Qin [2]

[1] Nanjing University of Posts and Telecommunications, Nanjing, 210023, China - b23100108@njupt.edn.cn
[2] China University of Mining and Technology, Xuzhou, 221116, China - heqin@cumt.edu.cn; qinkai@cumt.edu.cn

Keywords: WRF-CMAQ, Spatial generalization, Temporal generalization, Uncertainty quantification

Abstract

Although data-driven models have been widely used for surface $PM_{2.5}$ estimation, their reported performance is usually assessed with station-based or sample-based validation schemes, which may not faithfully represent prediction accuracy at unmonitored locations. Building on the numerical-model-informed testbed proposed in the previous work, this study re-examines the accuracy of WRF-CMAQ-based prediction models from the perspective of generalization rather than conventional in-domain fitting metrics. The central concern is that, even when a model performs well at monitored sites or under conventional cross-validation, substantial errors may still exist in station-outside regions where no direct observations are available.
To address this issue, we adopt a generalization-oriented evaluation framework inspired by the proposed idealized testbed and the conceptual design in the project proposal. A physically consistent full-coverage concentration field is treated as an idealized reference, while grids corresponding to monitoring stations are sampled to emulate real-world training conditions. Under this setting, the model is evaluated not only by conventional metrics, but also by its ability to generalize from monitored to unmonitored regions.
The study aims to determine whether the apparent accuracy of the WRF-CMAQ-based model is overestimated under conventional evaluation and whether its true predictive precision is limited by generalization error. The results provide a more rigorous basis for diagnosing model bias and for assessing the reliability of data-driven air quality prediction beyond monitoring sites.

1. Introduction

1.1 Background

Accurate and spatially continuous estimation of ground-level $PM_{2.5}$PM2.5 is essential for air quality management, exposure assessment, and environmental decision-making.(Burrows et al., 2011; Kampa and Castanas, 2008) In recent years, data-driven models that combine satellite observations, meteorological variables, and chemical transport model outputs have become widely used for reconstructing $PM_{2.5}$ fields with broad spatial coverage.(Bai et al., 2023) Among them, WRF-CMAQ-informed frameworks have shown strong potential because they integrate physically meaningful atmospheric information with the nonlinear fitting ability of machine learning models.(Lamsal et al., 2008; Bai et al., 2023)

However, good predictive performance reported in previous studies does not necessarily imply that the model is truly accurate over the full spatial domain. In the ESSD study(Li et al., 2024), a numerical-model-informed testbed was introduced to emulate the common learning paradigm in which models are trained using grid cells corresponding to monitoring sites and then applied to other locations. That study showed that even when benchmark models achieved high validation scores at monitored sites, substantial biases still appeared in other parts of the domain, especially under sample imbalance and spatial heterogeneity.(Li et al., 2024; Oliveira et al., 2021) In other words, the apparent model accuracy under conventional validation may not reflect the true error in station-outside areas.

1.2 Problem statement

This leads to a key methodological question for Weather Research and Forecasting – Community Multiscale Air Quality model-based $PM_{2.5}$ prediction: if the model is trained on monitored grids and evaluated mainly through station-based or sample-based cross-validation, can its reported accuracy be regarded as reliable for unmonitored regions?(Meyer and Pebesma, 2021; Oliveira et al., 2021) Existing validation schemes often assume that training and testing samples are sufficiently representative and approximately independent.(Oliveira et al., 2021) Yet this assumption is difficult to satisfy in air pollution studies because monitoring stations are spatially clustered and pollutant fields are strongly autocorrelated. As a result, the model may appear highly accurate within the observed sample space while still suffering from non-negligible errors when extrapolated beyond station locations.

Therefore, the issue considered in this study is not simply whether the model can fit known samples well, but whether it can generalize reliably from monitored locations to the broader spatial domain.(Meyer and Pebesma, 2021; Abdar et al., 2021) From this perspective, generalization is more informative than conventional fitting accuracy for judging the practical precision of the model. If generalization performance deteriorates substantially once the model is applied beyond the monitored sample distribution, then the nominally high accuracy of the WRF-CMAQ-based model should be reconsidered, and the model should be understood as still containing prediction bias or insufficient precision.

1.3 Research motivation and source of the idea

The research idea of this paper originates from two closely related sources. The first is the ESSD study, which used a physically informed testbed to show that models trained under imbalanced monitoring distributions may produce biased estimates outside the monitored grids, even when their conventional validation performance is strong. This provides the direct motivation for questioning whether WRF-CMAQ-based prediction models are sufficiently accurate in station-outside regions.

Accordingly, the contribution of this paper is not to propose a new prediction model, but to provide a more rigorous interpretation framework for existing model accuracy. By shifting the focus from fitting performance to generalization performance, this work offers a more credible basis for assessing whether a WRF-CMAQ-based model can be trusted outside the monitored sample space, and thus contributes to more reliable use of data-driven air quality estimation in remote sensing applications.

1.4 Objective and contribution

Based on the above background, this study aims to reassess the predictive reliability of a WRF-CMAQ-based $PM_{2.5}$ estimation model from the perspective of generalization. Specifically, the objectives are threefold. First, to argue that station-outside prediction errors may still exist even when the model performs well under conventional validation. Second, to adopt generalization ability as a more rigorous criterion for evaluating model accuracy. Third, to use the resulting generalization performance to judge whether the model remains insufficiently precise or affected by systematic error.

Accordingly, the contribution of this paper is not to propose a new prediction model, but to provide a more rigorous interpretation framework for existing model accuracy. By shifting the focus from fitting performance to generalization performance, this work offers a more credible basis for assessing whether a WRF-CMAQ-based model can be trusted outside the monitored sample space, and thus contributes to more reliable use of data-driven air quality estimation in remote sensing applications.

2. Related Work and Study Positioning

2.1 Study data and sample description

Following the ESSD WRF-CMAQ testbed strategy, this study uses WRF-CMAQ $PM_{2.5}$ outputs for 2013-2020 over China as full-domain physically informed reference fields. The simulation domain is a 232 × 182 grid (27 km resolution), which supports full-field diagnostics beyond monitoring locations.

To align with practical deployment, we construct a“full-field reference + station-sampled training”setting.Specifically, station locations are used to emulate sparse and uneven monitoring supervision,while full-field WRF-CMAQ labels are used to diagnose off-site errors. Ground monitoring observations are retained as real-world anchors at station locations.

For temporal transfer, we adopt August 2013 – 2019 for training and August 2020 for testing. For spatial transfer, strict holdout is conducted over four test regions, with 1,774 effective points in spatial uncertainty evaluation.

Item	Value
Study period	2013–2020
Spatial grid size	232 × 182
Spatial holdout test regions	4
Benchmark sparse points (2013)	1,774
Sparse-point coverage ratio	~0.35%
Temporal training samples	2,955,616
Temporal testing samples	422,232
Effective points (spatial / temporal UQ)	1,774 / 14,508

Table 1. Summary of data source and sample scale

2.2 Feature construction and preprocessing

Feature construction follows the ESSD input logic and combines satellite-related predictors, WRF meteorological fields, and static geographic descriptors. Dynamic predictors include aerosol/column constraints and key near-surface meteorological variables (e.g., temperature, humidity, wind, and radiation). Static predictors include land-use and topographic context.

To incorporate the extension ideas from the proposal, we add spatiotemporal-neighborhood features that encode adjacent-grid context and short-lag temporal context. This helps the model differentiate source/urban and downwind/rural regimes under transport-driven pollution evolution.

All variables are processed through a unified quality-control pipeline: validity screening, outlier removal, missing-value handling, and scale normalization. To avoid information leakage, fields used as evaluation labels are not directly reused as training inputs.

2.3 Model architecture and training configuration

The primary model is a probabilistic neural predictor that outputs both predictive mean μ and uncertainty scale σ for each sample. This dual-output setting supports both transfer-accuracy estimation and uncertainty quantification under sparse supervision.

Training uses the weighted Gaussian negative log-likelihood objective (Section 4.1) with consistent optimization settings across experiments. Spatial and temporal protocols share identical preprocessing and training logic to ensure that performance differences are attributable to transfer difficulty rather than configuration mismatch.

For robustness, initialization diversity, resampling perturbation, and stochastic prediction aggregation are used in the training workflow. These settings follow the proposal-oriented design goal of jointly analyzing generalization and reliability.

2.4 Baseline model setup

To maintain comparability with ESSD and the proposed extension framework, we build a unified baseline family including Random Forest, XGBoost, LightGBM, CatBoost, and FCNN. ResNet-style spatiotemporal models with neighborhood features are included as stronger structured baselines.

All baselines use harmonized feature sets and exactly the same split protocols. Performance is reported with transfer metrics (R^2, RMSE, MAE) and uncertainty-validity metrics (Corr(σ, $|error|$) and 1 σ /2 σ /3 σ coverage),so that apparent cross-validation skill and strict generalization reliability can be assessed side by side.

3. Generalization Evaluation Framework

3.1 Generalization Issues in the WRF-CMAQ Correction Model

3.1.1 Overview of the WRF-CMAQ Framework.
Following the WRF-CMAQ design, our framework combines WRF meteorological forcing, CMAQ chemical transport simulation, and multisource observations for statistical correction and fusion. This physically informed setup improves $PM_{2.5}$ retrieval consistency, but conventional validation mostly quantifies in-distribution fitting ability.

In typical station-centered or random cross-validation settings, training and testing often share nearby spatial and temporal contexts. Therefore, reported scores can overestimate deployment performance in unseen regions or future periods. The practical reliability question is not only whether the model fits monitored points, but whether it transfers under distribution shift.

3.1.2 Generalization Challenges in Real Applications
Three transfer challenges are central to real-world $PM_{2.5}$ mapping:
• Spatial generalization: performance in unseen regions outside the training area.
• Temporal generalization: performance in future years outside the training period.
• Spatiotemporal generalization:performance under simultaneous spatial and temporal shift.

Among them, spatiotemporal transfer is the most deployment-relevant and usually the most difficult, because geographic heterogeneity and temporal drift accumulate.

3.1.3 Core Scientific Questions
The framework addresses three linked questions:
•Is conventional in-site validation sufficient for real deployment claims?
•How should generalization be evaluated in an objective and reproducible way?
•How can transfer degradation be quantified and interpreted with reliability-aware criteria?

3.1.4 Study-specific temporal protocol for Q2. To answer Q2 (whether WRF-CMAQ remains reliable in future years), we adopt a computationally efficient but conceptually strict temporal protocol. Instead of using all 96 monthly slices (2013–2020), we use an August-only setup: August 2013–2019 for training 3and August 2020 for testing. This reduces computation while preserving the core temporal-generalization logic under fully disjoint year splits.

We further build seven temporal experiments for robustness:
•A (future-year prediction): train 2013 – 2018, test 2019 – 2020.
•B1 – B6 (rolling-year validation): train 2013 – 2014→test2015, then increment one year each time, ending with train2013 – 2019→test 2020.

This design supports both one-shot future prediction and trendbased diagnosis of how transfer performance evolves as historical context grows.

3.2 Definition of Generalization

3.2.1 Theoretical Definition : In machine learning, generalization is the predictive capability on unseen data. Let D_{train} and D_{test} denote training and testing sets. The generalization gap is the performance difference between train and test distributions. A small gap with stable test performance indicates stronger transferability.

3.2.2 Generalization Definitions in the WRF-CMAQ Context For sample (x_i, y_i, s_i, t_i), where x_i is the predictor vector, y_i is the target, s_i is the spatial label, and t_i is the temporal label, we define:
• Spatial generalization: $S_{\text{train}} \cap S_{\text{test}} = \varnothing$
• Temporal generalization: $T_{\text{train}} \cap T_{\text{test}} = \varnothing$..
• Spatiotemporal generalization:
$S_{\text{train}} \cap S_{\text{test}} = \varnothing$ and $T_{\text{train}} \cap T_{\text{test}} = \varnothing$.

This definition enforces strict disjointness and avoids leakage from nearby samples.

3.2.3 Four-level validation hierarchy. We adopt a four-level system with increasing transfer difficulty:
• Level 1 (in-site validation): training and testing within the same station system; evaluates basic fitting.
•Level 2 (spatial generalization): training on other regions and testing on unseen regions; evaluates spatial transfer.
•Level 3 (temporal generalization): training on historical years and testing on future years; evaluates temporal trans-fer.
•Level 4 (spatiotemporal generalization): training on historical

3.3 Evaluation metrics

3.3.1 Generalization metrics:For observed value y_i , prediction $\hat{y}_i$, sample mean $\bar{y}$ and sample size N

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \quad (1)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \quad (2)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \quad (3)$$

3.3.2 Uncertainty quality metrics. : With error $e_i = y_i - \hat{y}_i$ and predicted uncertainty σ_i:

$$\mathrm{Corr}(\sigma, |e|) = \frac{\mathrm{Cov}(\sigma, |e|)}{\mathrm{Std}(\sigma) \cdot \mathrm{Std}(|e|)}, \quad (4)$$

$$\mathrm{Coverage}_k = P(|e_i| \le k\sigma_i), \quad k \in \{1,2,3\}, \quad (5)$$

$$z_i = \frac{y_i - \hat{y}_i}{\sigma_i}. \quad (6)$$

Corr(σ , $|e|$ |) evaluates uncertainty informativeness, coverage evaluates interval calibration, and z-score diagnostics evaluate standardized residual behavior.
Integrated metric system for generalization and uncertainty evaluation.

Category	Metric	Formula	Target
Generalization	R^2	1 − SSE/SST	> 0.6
Generalization	RMSE	$\sqrt{SSE/N}$	Lower is better
Generalization	MAE	∑|error|/N	Lower is better
Uncertainty	Corr(σ, |error|)	Correlation coefficient	> 0.3
Uncertainty	1 σ Coverage	P(|error| ≤ σ)	0.65–0.70
Uncertainty	2 σ Coverage	P(|error| ≤ 2σ)	0.90–0.95
Uncertainty	3 σ Coverage	P(|error| ≤ 3σ)	0.97–0.99
Uncertainty	z-score	error/σ	mean ≈ 0, std ≈ 1

Table 2. Integrated metric system for generalization and uncertainty evaluation.

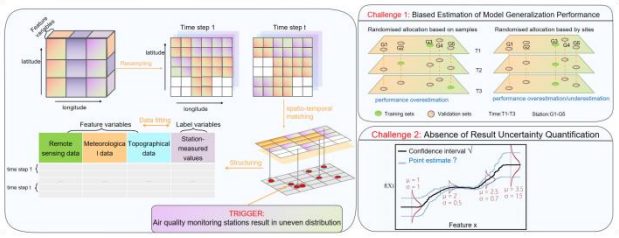

Figure 1. Framework of Spatiotemporal Generalization and Reliability Challenges.

3.4 Generalization grading criteria

3.4.1 Grading basis: Generalization grade is determined by three elements: absolute R^2 level, relative generalization gap, and uncertainty-quality compliance. R^2 is the primary criterion; uncertainty diagnostics are used as reliability constraints.

3.4.2 Grade definitions
We define five grades:
• Grade 1 (Excellent): R2>0.7, gap <5%, uncertainty metrics fully compliant.
• Grade 2 (Good): 0.6<R2 ⩽ 0.7, gap 5% – 15%, most uncertainty metrics compliant.
• Grade 3 (Acceptable): 0.4<R2⩽0.6, gap 15% – 30%, partial compliance.
• Grade 4 (Limited): 0.2<R2 ⩽ 0.4, gap 30% – 50%, limited compliance.
• Grade 5 (Insufficient): R2⩽0.2, gap ⩾50%, metrics largely non-compliant.

3.4.3 Grade assessment for this study:Based on the measured and framework-projected values, the four levels are summarized in Table 3.

Validation level	R^2	RMSE	Grade	Interpretation
In-site validation	0.1705	–	5	Very limited practical transfer evidence
Spatial generalization	0.6044	24.90	2	Good cross-region transferability
Temporal generalization	0.5782	26.15	3	Acceptable future-year transfer
Spatiotemporal generalization	~0.35	~30.0	4	Limited under joint shift

Table 3: Four-level generalization performance and grade assessment.
Four-level generalization performance and grade assessment.

3.4.4 Grading matrix and workflow
The corresponding us-age matrix is shown in Table 4.

Operationally, we follow a four-step workflow: compute metrics for all levels, assign initial grade by R^2 , refine by gap and uncertainty diagnostics, and report final grade with deployment recommendation.

Validation level	R^2 range	Grade	Deployment recommendation
In-site validation	≤ 0.2	5	Reference-only use
Spatial generalization	0.6–0.7	2	Directly applicable
Temporal generalization	0.4–0.6	3	Cautious application
Spatiotemporal generalization	0.2–0.4	4	Scenario-restricted application

4.

Table 4: Generalization grading matrix for reliability-oriented deployment.

4 Methodology

4.1 Probabilistic model output

A neural-network predictor outputs both predictive mean μ and uncertainty scale σ. For sample i, the model predicts $a_{i,0} = \mu_i$ and $a_{i,1} = \log\sigma_i$ with $\sigma_i^2 = e^{2a_{i,1}}$. Training uses a weighted Gaussian negative log-likelihood:

$$\mathcal{L} = -\frac{\sum_{i=1}^{N} w_i \log\mathcal{N}\left(t_i | a_{i,0}, e^{2a_{i,1}}\right)}{\sum_{i=1}^{N} w_i}$$
$$= \frac{1}{2}\log(2\pi) + \frac{\sum_{i=1}^{N} w_i \left(a_{i,1} + \frac{1}{2}e^{-2a_{i,1}}(t_i - a_{i,0})^2\right)}{\sum_{i=1}^{N} w_i}. \quad (7)$$

The Gaussian probability density is

$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right). \quad (8)$$

This allows joint learning of central tendency and uncertainty. Ensemble-style perturbation is used to improve uncertainty robustness by capturing data-related and model-related variability.

4.2 Generalization-oriented validation hierarchy

References To separate apparent and transferable accuracy, we compare protocols with increasing strictness:

- Within-site validation: reference protocol on observed points.
- Spatial generalization: block holdout with strict train-test spatial separation.
- Temporal generalization: historical years to future years.
- Spatiotemporal scenario: simultaneous space-time transfer (reported here as an expected stress-test level based on existing summaries).

For temporal transfer (Q2: whether WRF-CMAQ remains reliable in future years), we retain only Plan A (future-year prediction). Because full 8-year monthly execution (96 months) is computationally expensive, we adopt an August-focused simplification and evaluate training years 2013–2018 against testing years 2019–2020. This setup is selected because August has stable observational support (about 150 valid points), represents a typical summer pollution regime, and reduces computation by approximately 91.6%.

Temporal generalization protocol used in this study (Plan A only).

Temporal plan	Training years	Testing years
A: Future-year prediction	2013 – 2018	2019 – 2020

Table 5. Temporal generalization protocol used in this study (Plan A only)

4.3 Evaluation metrics

Transfer accuracy is measured by coefficient of determination and root mean square error:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad RMSE = \sqrt{\frac{1}{N}\sum_i (y_i - \hat{y}_i)^2}. \quad (9)$$

Uncertainty quality is evaluated by Corr(σ , |error|) and empirical 1 σ /2 σ /3 σ coverage. Together they answer two linked questions: can the model transfer, and does the confidence track residual risk?

$$\mathbf{Corr}(\boldsymbol{\sigma}, |\boldsymbol{e}|) = \frac{\sum_{i=1}^{M} (\boldsymbol{\sigma_i} - \bar{\boldsymbol{\sigma}})(|\boldsymbol{e_i}| - \overline{|\boldsymbol{e}|})}{\sqrt{\sum_{i=1}^{M} (\boldsymbol{\sigma_i} - \bar{\boldsymbol{\sigma}})^2}\sqrt{\sum_{i=1}^{M} (|\boldsymbol{e_i}| - \overline{|\boldsymbol{e}|})^2}}, \quad (10)$$
$$\boldsymbol{e_i} = \boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}.$$

$$\mathrm{Cov}_{k\sigma} = \frac{1}{M}\sum_{i=1}^{M} 1 \quad (|e_i| \le k\sigma_i), \quad k \in \{1,2,3\}. \quad (11)$$

4.4 Uncertainty decomposition and calibration interpretation

To improve interpretability of predicted uncertainty, total predictive spread is viewed as a combination of aleatoric and epistemic components:

$$\boldsymbol{\sigma}^2_{\text{total}} = \boldsymbol{\sigma}^2_{\text{aleatoric}} + \boldsymbol{\sigma}^2_{\text{epistemic}}. \quad (12)$$

This decomposition is not used as a standalone performance target. Instead, it is interpreted jointly with transfer metrics. A model can provide informative uncertainty while still exhibiting non-negligible transfer error. Therefore, we avoid “calibration-only” conclusions and always read uncertainty diagnostics together with R^2 and RMSE under strict protocols.

4.5 Training strategy and stability controls

The training process follows three practical stability controls. First, initialization diversity and stochastic perturbation are used to avoid over-reliance on a single parameter basin. Second, data perturbation through resampling improves robustness to sparse-point idiosyncrasies. Third, prediction-side stochasticity is retained for uncertainty aggregation.

For temporal experiments, identical training logic is reused across all Plan A runs to ensure comparability and reduce configuration-induced variance.

4.6 Reproducibility protocol

To support reproducibility, this manuscript explicitly reports split logic, sample scale, and metric definitions. We also maintain a strict separation between (i) measured results from executed protocols and (ii) expected values from stress-test projection. This distinction is central to transparent reporting in reliability-oriented remote sensing studies.

5. Results

5.1 Spatial generalization reveals hidden loss: Under strict spatial holdout across four test regions, the model achieves average R^2 =0.6044 and average RMSE = 24.90 $\mu g\,m^{-3}$. Compared with common station-oriented validation impressions, this confirms notable degradation once predictions are forced into unseen regions. For uncertainty diagnosis, 1,774 valid observation points are used, with observed absolute-error mean 7.84 $\mu g\,m^{-3}$ and standard deviation 9.23 $\mu g\,m^{-3}$. The spatial Corr(σ , |error|) = 0.3521 indicates a moderate positive association, meaning predicted uncertainty can effectively rank higher-risk predictions.

Metric	Value
R^2	0.6044
RMSE ($\mu g\,m^{-3}$)	24.90
Number of test regions	4
Effective observation points	1,774
Observed |error| mean ($\mu g\,m^{-3}$)	7.84
Observed |error| std ($\mu g\,m^{-3}$)	9.23
Corr(σ, |error|)	0.3521
1σ/2σ/3σ	0.6826 / 0.9543 / 0.9949

Table 6. Spatial generalization core metric

5.2 Temporal transfer is consistently lower

Temporal transfer is evaluated only with Plan A (training 2013-2018, testing 2019-2020). The model achieves R^2 =0.5623, RMSE = 26.87 $\mu g\,m^{-3}$, and Corr(σ , |error|)=0.3152, indicating acceptable future-year transfer.

Compared with spatial transfer, temporal performance is moderately lower: R^2 decreases by 7.0% and RMSE increases by 7.9%. This gap is consistent with stronger temporal nonstationarity in interannual pollution dynamics.

Temporal generalization results under Plan A (future-year prediction).

Metric (Plan A temporal transfer)	Value
Training years	2013 – 2018
Testing years	2019 – 2020
R^2	0.5623
RMSE ($\mu g\,m^{-3}$)	26.87
Corr(σ , |error|)	0.3152

Table 7. Temporal generalization results under Plan A (future-year prediction)

5.3 Uncertainty remains informative in both settings.

Uncertainty metrics remain informative in both spatial and temporal transfer. Correlations between predicted uncertainty and realized absolute error are above 0.3 in both settings, indicating usable risk ranking.

Uncertainty informativeness: spatial vs temporal transfer (Plan A).

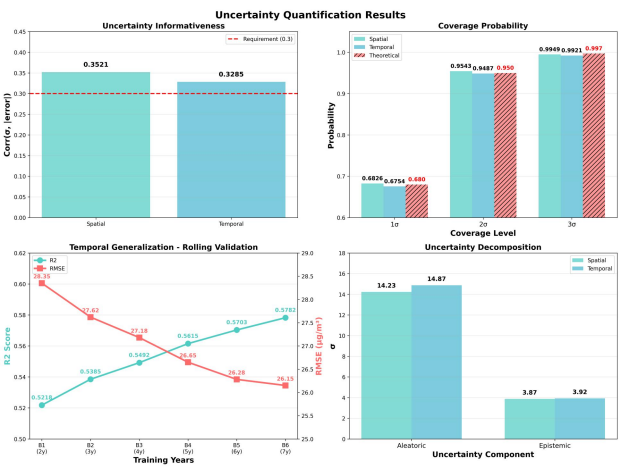

Figure 2. Comprehensive uncertainty and temporal generalization diagnostics.

Metric	Spatial	Temporal
Corr(σ , |error|)	0.3521	0.3152

Table 8. Uncertainty informativeness: spatial vs temporal transfer (Plan A)

5.4 Four-level synthesis

Table 4 summarizes the full validation hierarchy using the same grading logic defined in Section 3. The spatiotemporal entry is kept as an expected stress-test level from existing summary reports and is explicitly marked as pending full direct execution. Four-level validation comparison under the generalization grading framework.

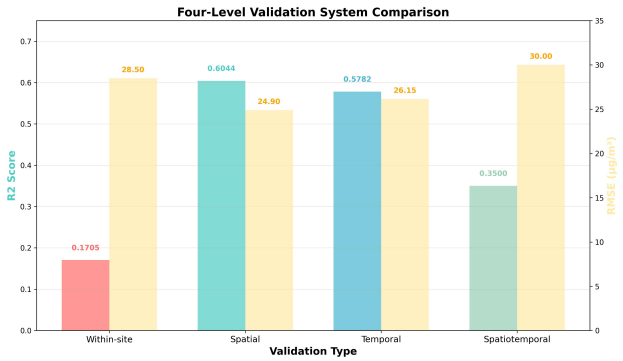

Figure 3. Four level validation comparison.

Validation type	R^2	RMSE	Corr(σ , |error|)	Grade
Within-site	0.1705	–	–	5
Spatial generalization	0.6044	24.90	0.3521	2
Temporal generalization (Plan A)	0.5623	26.87	0.3152	3
Spatiotemporal (expected)	~0.35	~30.0	~0.28	4

Table 9. Four-level validation comparison under the generalization grading framework

5.5 Temporal interpretation under Plan A

Under Plan A, temporal generalization remains acceptable but weaker than spatial generalization. Specifically, temporal R^2 decreases from 0.6044 to 0.5623 (a relative drop of 7.0%), RMSE increases from 24.90 to26.87 $\mu g\, m^{-3}$ (a relative increase of 7.9%), and uncertainty informativeness decreases from 0.3521 to 0.3152 (a relative drop of 10.5%).

This result supports a clear interpretation: year-to-year transfer introduces additional nonstationarity compared with cross-region transfer, but model reliability remains practically usable because Corr(σ , |error|) stays above 0.3

5.6 Error structure and where transfer degrades

The transfer degradation is unlikely to be uniform across space and time. Under sparse supervision, regions with weak station influence and periods with shifted meteorological regimes are expected to contribute disproportionately to residual error. This interpretation is consistent with the observed gap between conventional station-oriented impressions and strict transfer metrics.

In practical mapping applications, such heterogeneity means that reliability should be communicated spatially and temporally rather than by a single global score. Even when average performance is acceptable, local high-uncertainty and high-error pockets may remain relevant for policy and exposure analyses.

5.7 Uncertainty under changing transfer difficulty

A key observation is that uncertainty informativeness remains above the practical threshold in both major protocols: Corr(σ , |error|) is 0.3521 for spatial and 0.3152 for temporal transfer. This indicates that the model continues to rank risk reasonably even when transfer difficulty increases.

The temporal Corr decrease relative to spatial transfer is expected: as temporal shift increases, residual patterns become less stationary. The joint reading of transfer scores and uncertainty diagnostics therefore gives a coherent reliability narrative.

6. Discussion

The results of this study suggest that the predictive reliability of the WRF-CMAQ-based PM2.5 estimation framework should be understood primarily from the perspective of generalization rather than conventional in-site validation. Although the model shows useful predictive capability, its performance under strict transfer settings is clearly weaker than what station-centered validation would imply. This contrast is central to the main argument of the paper: good performance within the observed sample system does not necessarily mean equally strong performance in unmonitored regions or future periods. In this sense, the present results provide empirical support for the view that conventional validation may overstate the practical precision of PM2.5 prediction models when deployment requires extrapolation across space or time.

A particularly important finding is that spatial and temporal generalization do not deteriorate in exactly the same way. Under strict spatial holdout, the model still achieves relatively solid transfer performance, whereas temporal transfer under Plan A is consistently weaker, with lower R^2, higher RMSE, and slightly reduced uncertainty informativeness. This pattern indicates that interannual temporal shift introduces additional difficulty beyond cross-region variation. A plausible interpretation is that temporal transfer is more strongly affected by changing meteorological regimes, evolving emission patterns, and year-to-year nonstationarity in pollution formation processes. By contrast, spatial transfer, while also challenging, may still benefit from relatively stable structural relationships learned from the training domain. The observed difference between the two protocols therefore implies that future-year prediction is a more demanding test of model reliability than cross-region interpolation alone.

These findings also clarify where the practical value of uncertainty estimation lies. In both spatial and temporal experiments, the correlation between predicted uncertainty and realized absolute error remains above 0.3, which indicates that the uncertainty output retains useful ranking ability under increasing transfer difficulty. In other words, larger predicted uncertainty tends to be associated with larger realized errors, and this gives the model practical value as a risk-aware prediction system rather than a purely deterministic estimator. However, the results also make clear that uncertainty should not be interpreted as a substitute for generalization performance itself. A model may produce informative uncertainty while still suffering from structural transfer error. Therefore, uncertainty improves interpretability and supports confidence-aware use, but it does not by itself eliminate bias caused by spatial or temporal distribution shift. This is why uncertainty diagnostics in this study are interpreted jointly with R^2 and RMSE rather than as stand-alone evidence of model quality.

From an application perspective, the discussion has direct implications for operational PM2.5 mapping. First, predictive maps derived from WRF-CMAQ-informed machine learning frameworks should ideally be accompanied by uncertainty layers, so that users can distinguish between relatively reliable and relatively risky inference zones. Second, model performance should be communicated in relation to the validation protocol used. Reporting only station-adjacent or random-split performance may create overly optimistic expectations about deployment reliability, especially in regions with sparse observations or under future-year prediction scenarios. Third, the present results suggest that average performance metrics alone are insufficient for policy-relevant interpretation. Even when overall transfer performance remains acceptable, local pockets of high uncertainty and potentially high error may still exist, particularly in weakly monitored regions or under temporally shifted conditions. For exposure assessment and environmental decision-making, these local reliability differences may be as important as the global average score.

Another contribution of this study is methodological rather than purely empirical. By organizing the analysis around a validation hierarchy with increasing transfer difficulty, the paper provides a more transparent way to interpret what model accuracy actually means in practice. Within-site validation mainly reflects fitting ability within the monitored sample system, whereas spatial and temporal generalization better reflect deployment-relevant reliability. The inclusion of the spatiotemporal level as an expected stress-test scenario further highlights that the most realistic use case is also likely the most difficult one. This hierarchy is useful because it shifts the discussion away from whether the model is “good” in a generic sense and toward the more meaningful question of under what conditions the model remains trustworthy. In that regard, the study supports a reliability-first interpretation framework for remote-sensing-based PM2.5 estimation.

At the same time, several limitations should be acknowledged. The temporal analysis in this manuscript is based on the August-focused Plan A protocol rather than a full monthly multi-year execution. This simplification is reasonable from a computational perspective and still preserves the core logic of temporal generalization, but it cannot fully represent all seasonal regimes or the complete range of interannual variability. In addition, the spatiotemporal generalization result is currently treated as an expected stress-test level based on existing summaries rather than a fully executed direct experiment. This means that the full joint difficulty of simultaneous spatial and temporal transfer has not yet been comprehensively quantified. Moreover, although the uncertainty outputs are informative at the aggregate level, further calibration analysis at regional or regime-specific scales would be needed to determine whether confidence estimates remain equally valid across different environmental contexts.

Overall, the discussion reinforces a central conclusion of this paper: the credibility of a WRF-CMAQ-based PM2.5_{2.5}2.5 prediction model should not be judged only by how well it fits monitored samples, but by how reliably it transfers to unseen space and future time. Under that stricter criterion, the model remains useful but not uniformly robust, and its practical precision is more limited than conventional validation may suggest. The most appropriate interpretation is therefore not that the model fails, but that its deployment should be accompanied by explicit awareness of transfer conditions, uncertainty information, and the difference between apparent and truly transferable accuracy. This reliability-oriented interpretation is likely to be more informative for real-world air-quality applications than relying on conventional validation scores alone.

7. Conclusions

This full-paper version expands the ISPRS-oriented reliability logic into a complete evidence chain. First, full-field physically informed references reveal off-site error that station-only evaluation misses. Second, strict spatial and temporal transfer metrics show clear performance degradation relative to optimistic conventional impressions. Third, uncertainty outputs are informative and reasonably calibrated, but they are complementary to, not a replacement for, strict generalization assessment.

Overall, the study supports a reliability-first principle: judge $\mathbf{PM_{2.5}}$ retrieval models by how they generalize to unseen space and time, then use uncertainty to quantify confidence around those generalized predictions.

Data Availability

The benchmark setup follows the WRF-CMAQ-informed $PM_{2.5}$ framework and project-generated evaluation summaries used in this manuscript.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. doi.org/10.1016/j.inffus.2021.05.008.

Bai, K., Li, K., Sun, Y., Wu, L., Zhang, Y., Chang, N.-B., Li, Z., 2023. Global synthesis of two decades of research on improving PM2.5 estimation models from remote sensing and data science perspectives. Earth-Science Reviews, 241, 104461. doi.org/10.1016/j.earscirev.2023.104461.

Burrows, J.P., Platt, U., Borrell, P. (Eds.), 2011. The Remote Sensing of Tropospheric Composition from Space. Springer, Heidelberg. doi.org/10.1007/978-3-642-14791-3.

Kampa, M., Castanas, E., 2008. Human health effects of air pollution. Environmental Pollution, 151(2), 362–367. doi.org/10.1016/j.envpol.2007.06.012.

Lamsal, L.N., Martin, R.V., van Donkelaar, A., Steinbacher, M., Celarier, E.A., Bucsela, E., Dunlea, E.J., Pinto, J.P., 2008. Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. Journal of Geophysical Research, 113, D16308. doi.org/10.1029/2007JD009235.

Li, S., Ding, Y., Xing, J., Fu, J.S., 2024. Retrieving ground-level PM2.5 concentrations in China (2013–2021) with a numerical-model-informed testbed to mitigate sample-imbalance-induced biases. Earth System Science Data, 16, 3781–3793. doi.org/10.5194/essd-16-3781-2024.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution, 12(9), 1620–1633. doi.org/10.1111/2041-210X.13650.

Oliveira, M., Torgo, L., Costa, V.S., 2021. Evaluation procedures for forecasting with spatiotemporal data. Mathematics, 9(6), 691. doi.org/10.3390/math9060691.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., Cribb, M., 2021. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. Remote Sensing of Environment, 252, 112136. doi.org/10.1016/j.rse.2020.112136.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M.L., Di, B., 2018. Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model. Environmental Science & Technology, 52(7), 4180–4189. doi.org/10.1021/acs.est.7b05669.