remote sensing

MDPI

Article

Evaluating Scale Transferability and Observation-Based Calibration in High-Resolution PM2.5 Downscaling

Yiyang Jiang, Elsaid Mamdouh Mahmoud Zahran and Nicholas A. S. Hamm *

Faculty of Science and Engineering, University of Nottingham Ningbo China, Ningbo 315100, China; yiyang.jiang@nottingham.edu.cn (Y.J.); elsaid.zahran@nottingham.edu.cn (E.M.M.Z.)
* Correspondence: nicholas.hamm@nottingham.edu.cn

Highlights

What are the main findings?

- The scale-invariant assumption in PM2.5 statistical downscaling is only partially valid. PM2.5-predictor relationships change with spatial resolution, and sub-kilometre heterogeneity cannot be fully recovered from a smoothed 1 km source.
- Applying observation-based calibration improves the accuracy of PM2.5 statistical downscaling and recovers information not available in the smoothed 1-km source.

What are the implications of the main findings?

- High-resolution PM2.5 products should be evaluated not only by spatial resolution or internal validation against coarse-grid targets, but also by scale-transferability, ground-observation agreement, and spatial plausibility.
- Station-level cross-validated daily fine- resolution (100 m) maps of PM2.5 air pollution were produced for Zhejiang province, China, for 2020–2024 with spatial cross-validation results. These can be aggregated to different spatial and temporal units to support the required application.

Abstract

High-resolution PM2.5 mapping is increasingly required for exposure assessment and for assessing compliance with environmental policies. However, statistical downscaling from coarse gridded products assumes that PM2.5-predictor relationships are transferable across spatial resolutions. This study evaluated that assumption and developed a two-stage framework for 100 m PM2.5 mapping in Zhejiang Province, China. In Stage 1, the 1 km CHAP-PM2.5 product was downscaled with multi-source predictors using XGBoost and TabICL v2. In Stage 2, monitoring observations were integrated through daily bias correction and residual calibration. Scale transferability was assessed using coefficient-side similarity, output-side transfer tests, and a simulation-based recoverability experiment; calibration was evaluated using stratified and buffered site-based cross-validation. Scale transferability was only partial: predictor effects and transferred outputs became less stable as the resolution gap increased, and downscaling recovered only part of the known sub-kilometre heterogeneity. Stage 1 mainly enhanced spatial texture, whereas Stage 2 produced clearer improvements in agreement with ground observations and recovered much of the variability lost in the coarse-resolution data product. In cross-validation, the best calibrated product increased R^2 from 0.55 to 0.74 and reduced RMSE from 7.3 to 5.5 $\mu g/m^3$. This paper provides an extensive evaluation for 2020 and provides a series of quality-assured daily 100 m resolution PM2.5 maps for Zhejiang province for 2020–2024 which can be flexibly aggregated in space and time for different applications.

Check for updates

Academic Editor: Carmine Serio

Received: 2 July 2026
Revised: 20 August 2026
Accepted: 21 August 2026
Published: 25 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.

Keywords: PM2.5; high-resolution mapping; statistical downscaling; scale transferability; observation-based calibration; spatial cross-validation

1. Introduction

PM2.5 refers to fine particulate matter with an aerodynamic equivalent diameter of 2.5 μm or smaller. Because of its adverse effects on human and environmental health [1,2], China has established a national ground-based PM2.5 monitoring network through the China National Environmental Monitoring Centre (CNEMC). These monitoring stations provide reliable point-level observations, but their spatial coverage is limited and they cannot directly represent PM2.5 concentrations in areas without monitoring sites. To overcome this limitation, satellite remote sensing and related geospatial data have been used widely to estimate the spatial distribution of PM2.5 [3–7]. Based on aerosol optical depth (AOD), meteorological variables, land-use information, and ground observations, a number of gridded PM2.5 products have been developed [8–15], providing continuous spatial coverage beyond the monitoring network.

Although existing satellite-derived or data-fusion-based PM2.5 products have substantially improved spatial coverage, their spatial resolution often remains too coarse to represent local-scale variability in air pollution [16]. Many products are provided at kilometre-level resolution or coarser [9], meaning that sub-grid variation associated with roads, land-use contrasts, population density, urban morphology, and local emission patterns may be smoothed or lost. This limitation is particularly relevant for exposure assessment and local environmental analysis, where fine-scale spatial contrasts in PM2.5 are important [16]. Therefore, there is a need to refine existing coarse-resolution PM2.5 products into finer-resolution surfaces while retaining their large-scale spatial consistency.

Statistical downscaling offers a solution to this problem [16]. In a typical statistical downscaling framework, the relationship between a coarse-resolution target variable and auxiliary predictors aggregated to the same coarse scale is first modelled. The fitted relationship is then applied to predictors prepared at a finer spatial resolution to generate a high-resolution estimate [16,17]. This procedure assumes that the relationship learned at the coarse resolution is transferable to the finer resolution, namely that the fitted relationship is scale-invariant [17–20]. Although statistical downscaling has been widely used in remote sensing [16,21–23] and has also been applied to PM2.5 mapping [16,24–26], the validity of this scale-invariance assumption has usually not been addressed [16,18–20]. Beyond this scale-transferability issue, improving spatial resolution alone does not necessarily guarantee improved agreement with independent ground observations [16,18].

This study addresses these issues by developing and evaluating a two-stage framework for 100 m PM2.5 mapping in Zhejiang Province, China. In Stage 1, a 1 km PM2.5 product was downscaled to 100 m using multi-source auxiliary predictors and two tabular learning models, XGBoost and TabICL v2. In Stage 2, independent ground monitoring observations were used to calibrate Stage 1 products through daily bias correction and residual calibration. The study focuses on two research questions: (1) how can existing coarse-resolution PM2.5 products be improved using statistical downscaling and observation-based calibration? and (2) is the scale-invariance assumption valid, i.e., can models trained at coarse resolution be reliably applied to high-resolution PM2.5 predictors? By jointly assessing scale transferability, ground-observation agreement, and spatial plausibility, this study aims to clarify the conditions under which high-resolution PM2.5 downscaling can provide reliable added value beyond the original coarse-resolution product.

2. Conceptual Framework

Statistical downscaling is a common way to generate fine-resolution environmental surfaces [16]. It derives fine-resolution spatial information by first modelling the relationship between a coarser-resolution target and auxiliary predictors at the same scale and then applying the learned relationship to predictors available at finer spatial resolution [16–18,21,22,24,27–29]. For example, Yang et al. [16] improved Atmospheric Composition Analysis Group (ACAG) PM2.5 from 10 km to 300 m in China using statistical downscaling and Dinkelacker et al. [24] used a random forest model to downscale source-resolved CTM PM2.5 predictions from 36 km to 1 km in southwestern Pennsylvania. Beyond PM2.5, similar coarse-to-fine modelling strategies have been applied to other environmental variables. Tian et al. [28] trained a random forest model to characterize the relationship between coarse-resolution SMAP soil moisture and land-surface variables and subsequently applied the trained model to fine-resolution inputs to generate soil moisture estimates at 0.05° resolution. Hu et al. [29] established an XGBoost relationship between GOME-2 solar-induced chlorophyll fluorescence (SIF) and environmental predictors at 0.5° resolution and applied it at 0.05° resolution. This model-transfer procedure relies on the scale-invariance assumption [17–19]. This assumption is closely linked to the Modifiable Areal Unit Problem (MAUP), under which analytical results can change when spatial units or resolutions are modified [30]. The relevant concern here is the scale problem [30] or scale effect [19,31–33], because statistical downscaling explicitly transfers a model from a coarse support to a finer support [19]. Spatial aggregation smooths local variability, alters the distributions of PM2.5 and its predictors, and can change the strength or form of their relationship. Because the coarse-resolution target is already an areal average, fine-scale hotspots, road-related gradients, and local land-use contrasts may have been attenuated before model fitting. A model calibrated to this smoothed target may therefore be unable to recover the full sub-grid heterogeneity, and the additional detail it produces is itself subject to resampling uncertainty introduced when predictors are re-gridded between supports.

Another conceptual distinction separates resolution enhancement from observation-based calibration. A downscaled product may contain richer spatial texture than the original coarse product, yet greater spatial detail does not guarantee closer agreement with ground observations [16,18]: improved accuracy and improved spatial detail are distinct properties. High-resolution PM2.5 mapping therefore involves two related but different objectives, namely transferring a coarse gridded product to a finer resolution and calibrating the resulting product against independent ground measurements. Ground-observation integration differs fundamentally from downscaling because it introduces new information from monitoring sites rather than redistributing information already contained in the coarse product; it can correct systematic bias and local error that downscaling alone cannot address. This distinction motivates the two-stage design adopted in this study, in which Stage 1 performs statistical downscaling and Stage 2 uses ground observations to correct bias and residual error.

Validation should reflect these distinctions. In Stage 1, the modelling target is a spatially continuous gridded product with relatively uniform coverage, so the purpose of validation is to test whether the model reproduces the coarse-grid PM2.5-predictor relationship before transfer; conventional random cross-validation is effective when the available samples represent the prediction domain [34]. In Stage 2, however, the target is point-based ground-observed PM2.5 from unevenly distributed and potentially clustered monitoring stations. Randomly splitting individual observations may place records from the same or nearby stations in both the calibration and validation sets. Because nearby (both spatially and temporally) PM2.5 observations are often autocorrelated, this leakage reduces the effective independence of the validation data and may produce overly op-

timistic estimates of model performance [35–37]. Spatial cross-validation addresses this problem by defining validation folds according to geographic units, such as monitoring sites, spatial clusters, or spatial blocks, rather than by randomly assigning individual observations [35,38]. In this study, a spatial cross-validation framework was used: all observations from a validation site and its neighbours are excluded from model fitting, and predictions are evaluated at monitoring locations not used for observation-based calibration [39]. Together, these concepts form the logic of the study: statistical downscaling addresses whether a coarse product can be refined to a finer resolution; the MAUP and scale-invariance assumption address whether the transferred relationship is valid across supports, and site-based validation addresses whether observation-based calibration improves agreement with independent measurements.

3. Materials and Methods

3.1. Study Area

Zhejiang Province, located in eastern China, was selected as the study area, with a particular focus on Ningbo City. The province is predominantly characterised by mountainous and hilly terrain, with relatively limited plains mainly distributed in coastal and riverine areas [40]. Most urban areas are concentrated on these low-lying plains. The province therefore exhibits strong contrasts in terrain, urbanisation, population density, and emission-related land-use patterns, making it a suitable region for evaluating high-resolution PM2.5 downscaling. Figure 1 shows the location of Zhejiang Province, its topography, and the spatial distribution of the PM2.5 monitoring stations used in this study.

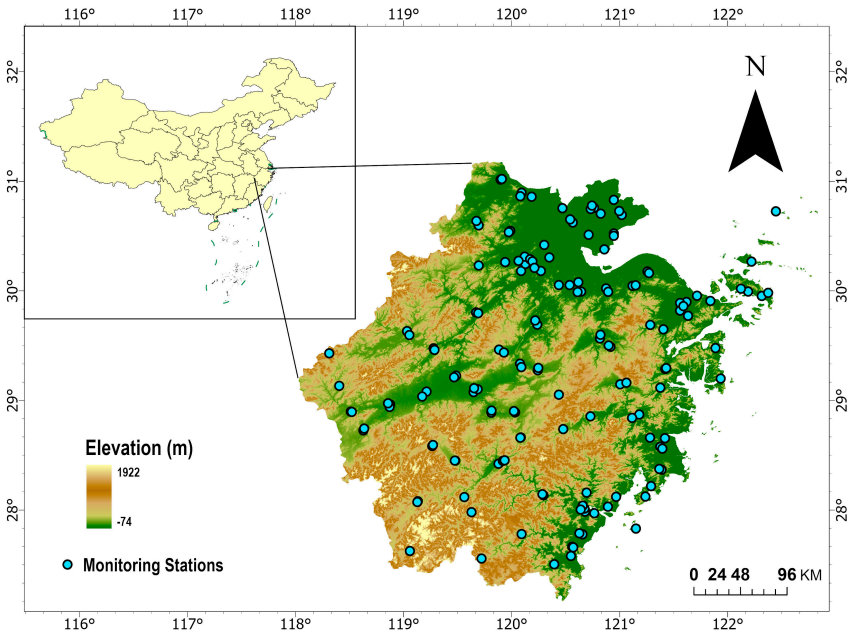


Figure 1. Study area. The figure shows the location of Zhejiang Province, terrain elevation, and PM2.5 monitoring stations.

3.2. Datasets

The datasets used in this study include ground-level PM2.5 observations, gridded PM2.5 estimates, remote-sensing and meteorological predictors, and socioeconomic auxiliary variables.

3.2.1. PM2.5 Site Observations

Ground-level PM2.5 observations were obtained from the provincial air-quality data product provided by the Qingyue Open Environment Data Centre (https://data.epmap.org/product/province_air, accessed on 31 March 2026). The dataset contains daily PM2.5 measurements from both national-controlled and provincial-controlled air-quality monitoring stations. For this study, only provincial-controlled stations were retained. National-controlled stations were excluded because the baseline gridded PM2.5 product used in this study was generated with support from national monitoring observations. Retaining these stations for validation could therefore reduce the independence of the assessment and overestimate agreement between the gridded product and ground observations. The procedure for excluding the national-controlled stations is described in Supplementary Material S1.

For each monitoring record, daily PM2.5 concentration, station name, longitude, and latitude were extracted. Records with missing or invalid PM2.5 values or coordinates were removed (details about data cleaning can be found in Supplementary Material S1.5). For the main calibration and cross-validation analysis, in each month, stations with fewer than 10 valid observation days were excluded to ensure sufficient temporal representation. In total, there are 98 provincial monitoring stations, including 62 stations (63%) over flat areas and 36 stations (37%) over rugged terrain areas.

3.2.2. Other Datasets

In this study, the PM2.5 product to be downscaled was the ChinaHighPM2.5 product from the ChinaHighAirPollutants (CHAP) dataset (CHAP-PM2.5) [11,12]. This product was selected because it provides daily gridded PM2.5 estimates across China with strong reported performance [9]. The dataset integrates ground observations, satellite remote sensing (i.e., DEM, land cover and nighttime light), meteorological variables, and other auxiliary information, and provides PM2.5 estimates at 1 km spatial resolution. In this study, the CHAP-PM2.5 product was used as the coarse-resolution target for Stage 1 statistical downscaling.

The datasets used in this study are summarised in Table 1. Full descriptions, including native resolution, temporal coverage, source, and the preprocessing, aggregation, and resampling rules applied to each layer, are provided in Supplementary Material S1.

Table 1. Datasets used for PM2.5 downscaling and observation-based calibration.

Dataset	Role in Study	Source	Native Resolution
Ground PM2.5	Stage 2 calibration/validation target	Qingyue Open Environment Data Centre	Point, daily
CHAP-PM2.5	Stage 1 coarse-resolution target	Wei et al. [11], Wei et al. [12],	1 km, daily
AOD	Predictor	Bai et al. [8]	1 km, daily
2 m temperature; precipitation	Predictor	ERA5-Land (ECMWF)	~9 km, hourly to daily
Land cover	Predictor	ESRI (Sentinel-2)	10 m, yearly
Nighttime light	Predictor	Chen et al. [41]	~500 m, yearly
Building height	Predictor	Wu et al. [42]	10 m, 2020
DEM	Predictor	SRTM (CGIAR-CSI)	~90 m, 2000
Road density	Predictor	OpenStreetMap	Vector to grid
Population density	Predictor	Chen et al. [43]	100 m
POI density	Predictor	Gaode (Amap) API	Vector to grid
Global Relief Classes (GRC)	For terrain identification	Yang et al. [40]	~30 m

3.3. Methodology

A two-stage workflow was developed to generate 100 m PM2.5 estimates by combining the 1 km CHAP-PM2.5 product, multi-source auxiliary predictors, and independent ground monitoring observations (Figure 2). Stage 1 focused on statistical downscaling. In this step, the relationships between CHAP-PM2.5 and auxiliary predictors were learned at the 1 km scale and then transferred to the 100 m predictor grid. Two tabular learning models, Extreme Gradient Boosting (XGBoost) [44] and TabICL v2 [45,46], were evaluated to examine whether different modelling strategies affected the accuracy and spatial structure of the downscaled product.

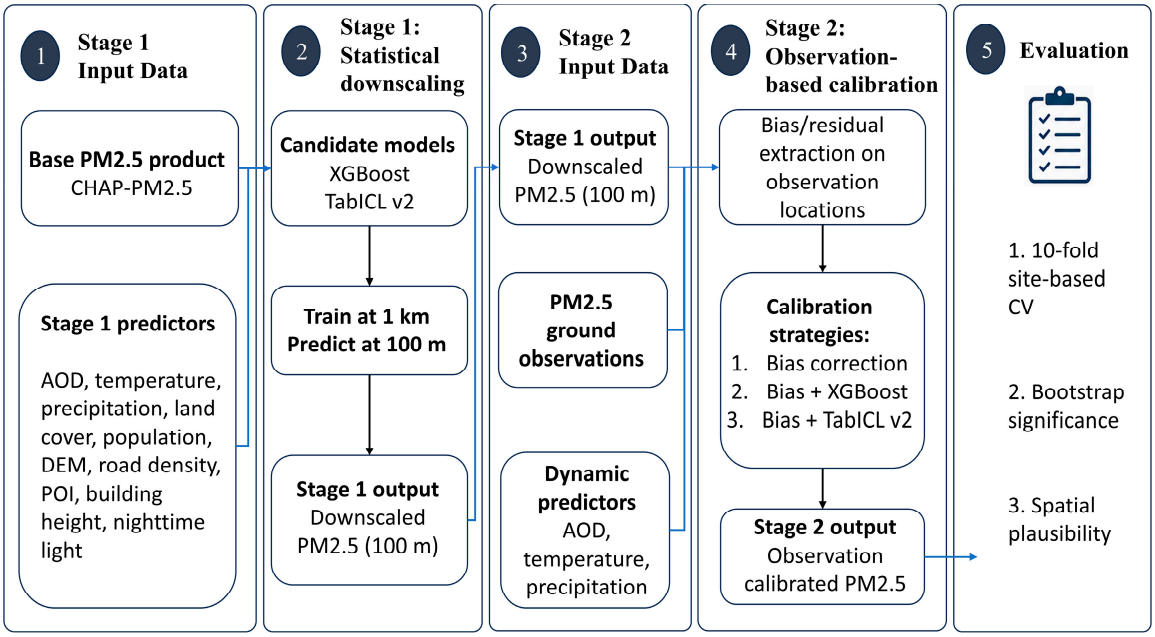


Figure 2. Flowchart for the downscaling process.

Stage 2 focused on observation-based calibration. The Stage 1 product was adjusted using ground monitoring observations through three correction strategies: simple daily bias correction, bias correction plus XGBoost residual modelling, and bias correction plus TabICL v2 residual modelling. These methods represent increasing levels of model flexibility and were compared to assess whether residual modelling provides robust improvement beyond simple bias correction.

This two-stage design separates resolution enhancement from ground-observation integration. Stage 1 evaluates whether the coarse-resolution PM2.5-predictor relationship can be transferred to a finer grid, whereas Stage 2 evaluates whether the resulting high-resolution product can be calibrated to independent ground observations. In both stages, possible spatial artefacts were evaluated [13,47]. Model configuration, hyperparameters, software and computer hardware used are provided in Supplementary Material S2.

3.3.1. Stage 1: PM2.5 Statistical Downscaling

Stage 1 generated a 100 m PM2.5 surface by transferring the relationship between the 1 km CHAP-PM2.5 product and auxiliary predictors to the 100 m grid. Let y^L denote the CHAP-PM2.5 and let X^L denote a set of predictors co-registered to the same coarse grid. We trained a nonlinear relationship $f(\cdot)$ using XGBoost and TabICL v2 to approximate:

$$y^L = f\left(X^L\right) + \varepsilon, \quad (1)$$

where ε denotes residual errors. XGBoost is a machine learning method that fits an additive model of decision trees by minimizing a regularized objective, enabling flexible nonlinear response functions and interactions among predictors [44]. In addition to XGBoost, TabICL v2 was used as an alternative Stage 1 learner. TabICL v2 was used as an in-context deep learning model, in which predictions are generated conditional on the training examples provided as context rather than through conventional task-specific parameter estimation [45]. In this study, TabICL v2 was evaluated as a flexible tabular foundation model for learning the coarse-resolution PM2.5-predictor relationship and transferring it to the 100 m predictor grid.

Predictor harmonization followed the rules described in Supplementary Material S1.5. After harmonization, predictors were provided at 1 km and 100 m resolutions. After fitting the model at 1 km resolution, the learned relationship was applied to the predictor stack prepared on the 100 m grid:

$$\hat{y}^H = f\left(X^H\right), \tag{2}$$

where X^H denotes the fine-grid predictor stack and $\hat{y}^H$ denotes the Stage 1 downscaled PM2.5 estimate. This procedure produced two Stage 1 products, one based on XGBoost and the other based on TabICL v2, which were then used as baseline high-resolution products for the Stage 2 observation-based calibration.

3.3.2. Stage 2: Integration of Ground Observations

Stage 2 was designed to integrate ground monitoring observations into the Stage 1 downscaled PM2.5 products. Although Stage 1 generates 100 m PM2.5 surfaces, these products remain constrained by the 1 km CHAP-PM2.5 product used as the training target. They may therefore inherit systematic bias from the original gridded product and may not fully agree with ground-level observations.

Let $y_{i,t}$ denote the observed PM2.5 concentration at monitoring site i on day t, and let $\hat{H}_{i,t}^{(m)}$ denote the Stage 1 estimate extracted at the corresponding monitoring location. The raw residual between the ground observation and the Stage 1 estimate was defined as:

$$e_{i,t}^{(m)} = y_{i,t} - \hat{H}_{i,t}^{(m)}, \tag{3}$$

where $e_{i,t}^{(m)}$ denotes the correction required to align the Stage 1 estimate with the ground observation at site i and day t.

Three Stage 2 calibration strategies were evaluated: (1) simple daily bias correction, (2) bias correction plus XGBoost residual modelling, and (3) bias correction plus TabICL v2 residual modelling. These methods represent increasing levels of model flexibility, from a purely additive daily adjustment to nonlinear machine learning residual correction.

Simple Daily Bias Correction

The first Stage 2 method was a simple daily bias correction. For each day, the mean residual (i.e., bias) between ground observations and Stage 1 estimates was calculated from the training monitoring sites:

$$b_t^{(m)} = \frac{1}{n_t}\sum_{i=1}^{n_t}\left(y_{i,t} - \hat{H}_{i,t}^{(m)}\right), \tag{4}$$

where $b_t^{(m)}$ denotes the daily bias term for the Stage 1 model m, and n_t denotes the number of available training observations on day t.

Let $\hat{H}_{g,t}^{(m)}$ denote the Stage 1 PM2.5 estimate for grid cell g on day t, generated by the Stage 1 model m, where m denotes either XGBoost or TabICL v2. The bias-corrected PM2.5 estimate for grid cell g was then calculated as:

$$\hat{C}_{g,t}^{(m)} = \hat{H}_{g,t}^{(m)} + b_t^{(m)}. \quad (5)$$

This method applied the same additive correction to all grid cells on a given day. Therefore, daily bias correction provided a conservative calibration strategy that corrects systematic underestimation or overestimation while avoiding new spatial artefacts from sparse station data.

Residual Modelling After Daily Bias Correction

The remaining residual after daily bias correction was defined as:

$$r_{i,t}^{(m)} = y_{i,t} - \left(\hat{H}_{i,t}^{(m)} + b_t^{(m)}\right), \quad (6)$$

where $r_{i,t}^{(m)}$ denotes the residual that remains after correcting the daily average bias. For the residual-modelling methods, $r_{i,t}^{(m)}$ was used as the response variable. All residual models were fitted separately by month. Within each month, the daily bias was first calculated for each date, and the remaining residuals were then pooled to train one month-specific residual model. Only daily varying predictors, including Stage 1 PM2.5, AOD, and meteorological variables, were used; static predictors would have identical values across dates while the corresponding residuals changed from day to day. The general model can be written as:

$$r_{i,t}^{(m)} = f_q\left(AOD_{i,t},\ Temperature_{i,t},\ Precipitation_{i,t},\ \hat{H}_{i,t}^{(m)}\right) + \varepsilon_{i,t}, \quad (7)$$

where $f_q(\cdot)$ denotes the residual model used by method q (i.e., XGBoost and TabICL v2) and $\varepsilon_{i,t}$ is the error term. The final calibrated estimate was then obtained as:

$$\hat{C}_{g,t}^{(m,q)} = \hat{H}_{g,t}^{(m)} + b_t^{(m)} + \hat{r}_{g,t}^{(m,q)}, \quad (8)$$

where $\hat{r}_{g,t}^{(m,q)}$ denotes the remaining residual estimated for grid cell g on day t by residual modelling method q, conditional on the Stage 1 product generated by model m.

Model Comparison and Final Product Selection

The three Stage 2 calibration strategies were applied separately to the two Stage 1 products generated by XGBoost and TabICL v2. This produced a set of candidate 100 m PM2.5 products with different levels of complexity. Their performance was evaluated using site-based cross-validation metrics, including R^2, RMSE, MAE, bias and slope, together with visual inspection of mapped spatial structure.

The final product was not selected solely according to the best validation metrics. Instead, model selection considered three criteria: improvement against ground observations, preservation of plausible spatial structure, and methodological parsimony. This was necessary because flexible residual models may improve station-level validation metrics while introducing unstable spatial artefacts when extrapolated to the full prediction grid [13,47]. The comparison therefore evaluated whether residual modelling provided robust added value beyond daily bias correction, or whether daily bias correction alone was sufficient for producing a stable observation-calibrated PM2.5 surface.

3.3.3. Validation

Different validation strategies were used for Stage 1 downscaling and Stage 2 calibration, because the two stages differ in both prediction target and spatial sampling structure. In Stage 1, as a within-domain reconstruction diagnostic, k-fold random cross-validation was used to determine whether the model can reproduce the CHAP-covariate relationship for held-out 1 km pixels drawn from the same daily mapping domain. In Stage 2, however, the target variable is ground-observed PM2.5 at unevenly distributed monitoring stations. Therefore, a stratified and buffered spatial cross-validation strategy was adopted for Stage 2.

In Stage 1, we chose 10-fold CV for the XGBoost model and 5-fold CV for the TabICL v2 model for computational reasons. Firstly, we calculated the model fitting R^2 and 10-fold CV R^2 to evaluate the goodness of fit and prediction ability of the model. Secondly, we evaluated the accuracy of the downscaled product. We used R^2, RMSE, MAE and bias as metrics, using ground PM2.5 as the observations and PM2.5 products as the estimations to evaluate the accuracy of CHAP-PM2.5 and Stage 1 PM2.5. The equations of the metrics above can be written as (9), (10), (11) and (12):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \tag{10}$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}, \tag{11}$$

$$bias = \frac{\sum \hat{y}_i - y_i}{n}, \tag{12}$$

where n represents the number of PM2.5 monitoring sites and i indexes the monitoring sites. y_i and $\hat{y}_i$ are the ith PM2.5 observation and estimation, respectively. $\overline{y}$ is the average PM2.5 observation at all sites.

In addition to pixel-level cross-validation at the training resolution, we assessed whether the final 100 m Stage 1 products preserved the coarse-scale information of the original CHAP-PM2.5 product. For each day in 2020 and for each Stage 1 model, the 100 m prediction raster was aggregated to the original 1 km CHAP-PM2.5 grid using the arithmetic mean. The original and aggregated values were compared using the bias and RMSE. Metrics were calculated by month. This analysis evaluated the coarse-grid consistency of the final 100 m maps rather than their agreement with independent ground observations. For each Stage 1 model, out-of-fold predictions for individual 1 km grid cells were also stratified by terrain type, and monthly R^2 and RMSE were calculated separately for flat and rugged terrain. The stratification methodology is described in Supplementary Material S3.

Stage 2 calibration was evaluated using terrain-stratified and buffered 10-fold site-based cross-validation. This design combined two components. First, the monitoring sites were stratified according to their surrounding terrain context to maintain representation of both flat and rugged environments in each validation fold. Second, spatial buffers were applied around the validation sites to control spatial dependence between the training and validation samples [35–37,39].

Terrain context was derived from the Global Relief Classes (GRC) dataset at 1 arcsec resolution [40]. At Level 2 of the GRC classification, codes 11–14 represent flat terrain at different elevation levels, whereas codes 21–25 represent rugged terrain at different relief

levels. These subclasses were combined into two broad terrain groups: flat and rugged terrain. The detailed procedure is described in Supplementary Material S3.

To select a buffer appropriate for the spatial configuration of the monitoring network, we adapted the spatial-distance CV framework of Li et al. [39]. For each candidate buffer, we calculated the mean distance from the held-out validation sites to their nearest retained training sites after spatial exclusion. The buffer was then selected by comparing this distance with the mean distance from prediction grid cells to their nearest monitoring sites in the complete network. Matching these two distances makes the prediction task represented by cross-validation comparable to that encountered during final mapping. Among the buffer distances evaluated, 2 km produced the closest match. We therefore report results for both 0 km (site-based cross-validation), as the grouped cross-validation baseline, and 2 km, as the buffered validation scenario designed to represent the spatial conditions of final prediction. We then extended this analysis to include 5 km, 10 km, 20 km and 30 km buffers to specifically evaluate the prediction accuracy at locations far from the observations. Details are provided in Supplementary Material S5.

Within each fold, the daily mean bias was calculated using only the retained training sites after buffer exclusion. For date t, the bias was calculated as the mean difference between ground observations and the corresponding Stage 1 estimates among the retained training sites. This training-derived daily bias was then applied to the validation observations from the same date. If no retained training observation was available for a particular validation date, the mean residual across all retained training records within that month was used as the fallback bias.

For the residual-modelling methods, the daily bias was first removed from the training residuals. The remaining residuals from the retained training sites were pooled within the corresponding month to fit one month-specific XGBoost or TabICL v2 model. The predictors were the Stage 1 PM2.5 estimate, AOD, temperature, and precipitation. The final prediction for each held-out observation was obtained by adding the Stage 1 estimate, the training-derived daily bias, and the predicted local residual. Thus, neither the daily bias nor the residual model used observations from the validation sites.

After completing all 10 folds, the out-of-fold predictions in each month were pooled to calculate R^2, RMSE, MAE, bias, and validation slope. These metrics were calculated for all validation observations and separately for sites in the flat-dominated and rugged-dominated terrain contexts.

3.3.4. Significance Test of Performance Improvements

To evaluate whether the performance improvements in Stage 2 were statistically significant, we applied a site-level bootstrap procedure to the pooled out-of-fold validation results in Stage 2. Because repeated daily observations from the same monitoring site are not independent, significance was not assessed by treating all station-day records as independent samples. Instead, bootstrap resampling was performed at the monitoring-site level. In each bootstrap iteration, monitoring sites were sampled with replacement within a given month. Whenever a site was selected, all pooled validation observations associated with that site were retained in the resampled dataset. These were used to calculate the validation metrics. This procedure was repeated 2000 times for each month, yielding empirical bootstrap distributions of the performance differences.

Statistical significance of the performance improvements was evaluated from the bootstrap distributions using two-sided bootstrap p-values, testing whether the improvement differed from zero. Improvements were considered statistically significant when the bootstrap p-value was less than 0.05, with significance levels reported as $p < 0.05$, $p < 0.01$, and $p < 0.001$.

3.3.5. Assessment of Scale Transferability

Models developed at coarse spatial resolutions may not be directly transferable at higher resolutions due to scale effects [30]. We assessed scale transferability using three complementary analyses: (1) coefficient-side: coefficient similarity, (2) output-side: prediction difference, and (3) simulation-based fine-scale recoverability. Each analysis was first conducted for the complete study area and was subsequently repeated separately for flat and rugged terrains.

Coefficient-Side: Coefficient Similarity Test

The coefficient-side analysis examined whether predictor effects remained stable when the same PM2.5-predictor relationship was estimated at different spatial resolutions. Multiple linear regression was used for this analysis because its coefficients provide an explicit and comparable description of model structure. The 1 km PM2.5 raster and predictor stack were aggregated to resolutions from 2 km to 8 km, and linear models were fitted at each resolution. Because aggregation changes the number of available grid cells, each model was fitted using a fixed sample size of 1000 pixels, and the sampling procedure was repeated 200 times to obtain coefficient distributions.

Similarity between coefficient distributions was quantified using the Bhattacharyya coefficient (BC), which measures the overlap between two probability distributions and ranges from 0 to 1 [48]. For each predictor, BC was calculated between the standardized coefficient distributions obtained at two resolutions. For variable j, denote the distribution of coefficients under resolution s_1, s_2 as $f_{j,\,s_1}$ and $f_{j,\,s_2}$ respectively, we have:

$$BC_j(s_1,\ s_2) = \int \sqrt{f_{j,s_1}(x) f_{j,s_2}(x)}dx,\ 0 \le BC_j \le 1. \tag{13}$$

To ensure that agreement on influential predictors contributed more strongly to the overall similarity score, predictor-specific BC values were weighted by their relative importance, measured using the Lindeman, Merenda and Gold (LMG) metric [49]. For each pair of resolutions, the mean LMG contribution of each predictor across the two resolutions was normalized so that the weights summed to one. The weighted coefficient similarity between resolutions s_1 and s_2 was calculated as:

$$BC_{weighted}(s_1,\ s_2) = \sum_{j=1}^{p} \overline{w_j}(s_1,\ s_2) BC_j(s_1,\ s_2), \tag{14}$$

where $\overline{w_j}(s_1,\ s_2)$ is the normalized average LMG weight of predictor j across the two resolutions. Higher weighted BC values indicate greater similarity in model structure across spatial resolutions, while lower values indicate weaker scale transferability in the estimated PM2.5-predictor relationship.

Output-Side: Test of Differences in Transferred Predictions

The output-side analysis evaluated whether models trained at coarser resolutions could reproduce the original 1-km PM2.5 field when applied back to the 1 km predictor grid.

For each day, the 1-km predictor stack and the 1-km PM2.5 target were first aggregated to a sequence of training resolutions from 2 km to 8 km. At each training resolution s, an XGBoost model was trained using the aggregated predictors and the aggregated PM2.5 target. The fitted model was then applied back to the original 1-km predictor stack, generating a transferred 1 km prediction from a model trained at resolution s. The signed transfer difference was defined as:

$$D_{s,t}(i) = \hat{Y}_{s\to 1\mathrm{km},t}(i) - Y_{1\mathrm{km},t}(i), \tag{15}$$

where $Y_{1\mathrm{km},t}(i)$ is the original 1 km PM2.5 value for pixel i on day t, and $\hat{Y}_{s\to 1\mathrm{km},t}(i)$ is the corresponding prediction at the same 1 km pixel generated by the model trained at resolution s. A value of $D_{s,t}(i)$ close to zero indicates that the model trained at resolution s can reproduce the original 1 km PM2.5 value at that location, whereas larger positive or negative values indicate increasing local transfer error.

For each day and training resolution, the signed transfer differences were summarised using their variance. Monthly mean variances were then calculated to evaluate how transfer sensitivity changed with training resolution. In addition, the transferred predictions were compared with the original 1 km PM2.5 product using R^2, RMSE, and regression slope. These metrics assessed overall agreement, error magnitude, and preservation of the PM2.5 dynamic range, respectively.

Simulation-Based Assessment of Sub-Kilometre Heterogeneity Recoverability

Direct validation of the 1 km-to-100 m transfer was not possible because no spatially continuous 100 m PM2.5 reference surface was available. Therefore, a simulation-based recoverability experiment was conducted to evaluate whether the Stage 1 downscaling framework could recover sub-kilometre heterogeneity under controlled conditions.

A synthetic 100 m PM2.5 truth was generated using the Stage 1 predictors on 1st January 2020, including AOD, temperature, precipitation, DEM, population density, building height, nighttime light, road length, and POI density. The synthetic truth combined predictor-driven effects, weak nonlinear and interaction terms, a broad spatial trend, and a smoothed residual component to represent fine-scale spatial heterogeneity not fully explained by the predictors. Details are provided in Supplementary Material S4.

The synthetic 100 m truth was then aggregated to 1 km using mean aggregation and treated as the coarse-resolution target. The predictors were also aggregated to 1 km to construct the training dataset. The downscaling models were trained using the 1 km synthetic target and the aggregated predictors and were then applied back to the original 100 m predictors. This design mirrors the main Stage 1 downscaling setting, but with the advantage that the 100 m PM2.5 field is known.

Five approaches were compared. Uniform disaggregation assigned each 1 km value directly to all 100 m pixels within the corresponding 1 km cell, representing a no-heterogeneity baseline. Single-predictor redistribution used the high-resolution predictor most strongly correlated with the 1 km synthetic target to redistribute each 1 km value within its corresponding 1 km cell while preserving the cell mean. Three model-based approaches were then tested: linear regression, XGBoost, and TabICL v2. All three were trained at 1 km and applied to the 100 m predictor grid.

The recoverability of fine-scale heterogeneity was evaluated using four metrics: heterogeneity explained variance (HEV), variability ratio (VR), r_{anom}, and r_{sd}. For each 100 m pixel i, the within-cell heterogeneity was defined as:

$$H_i^{truth} = Y_i^{truth} - \overline{Y}_{c(i)}^{truth}, \tag{16}$$

$$H_i^{pred} = Y_i^{pred} - \overline{Y}_{c(i)}^{pred}, \tag{17}$$

where H_i^{truth} and H_i^{pred} denote the within-cell heterogeneity of the true and predicted 100 m pixels, respectively. $c(i)$ denotes the 1 km cell to which pixel i belongs, and $\overline{Y}_{c(i)}$ denotes the corresponding within-cell mean. Based on these heterogeneities, four indicators were used.

First, heterogeneity explained variance (HEV) was defined as:

$$HEV = 1 - \frac{\sum_i (H_i^{pred} - H_i^{truth})^2}{\sum_i (H_i^{truth})^2}, \tag{18}$$

which measures the proportion of within-cell heterogeneity variance recovered by the prediction. Analogous to the coefficient of determination R^2, HEV represents the proportion of variance explained by the prediction; however, it is calculated for within-cell deviations from the respective coarse-cell means and therefore specifically measures the proportion of sub-grid heterogeneity variance recovered by the model.

Second, the variance ratio (VR) was calculated as:

$$VR = \frac{Var\left(H_i^{pred}\right)}{Var\left(H_i^{truth}\right)}, \tag{19}$$

which quantifies whether the amplitude of the recovered heterogeneity is attenuated ($VR < 1$), preserved ($VR \approx 1$), or exaggerated ($VR > 1$).

Third, the Pearson correlation between truth and predicted heterogeneities was calculated as:

$$r_{anom} = cor\left(H_i^{truth},\ H_i^{pred}\right), \tag{20}$$

which evaluates how well the spatial pattern of within-cell heterogeneity is reproduced.

Fourth, for each 1 km cell c, the strength of within-cell heterogeneity was summarised by the standard deviation (SD) of the heterogeneities:

$$SD_c^{truth} = SD\left(H_c^{truth}\right),\ SD_c^{pred} = SD\left(H_c^{pred}\right), \tag{21}$$

and the Pearson correlation between these quantities was computed as:

$$r_{sd} = cor\left(SD_c^{truth},\ SD_c^{pred}\right), \tag{22}$$

which evaluates whether the model correctly identifies which 1 km cells are internally more heterogeneous than others.

This experiment does not directly demonstrate that real PM2.5 satisfies scale invariance below 1 km. It is a controlled recoverability assessment with a known synthetic truth. Rather, it provides a controlled test of whether the Stage 1 framework can recover predictor-supported sub-kilometre heterogeneity when such heterogeneity is known to exist in the fine-scale truth but has been hidden in the aggregated coarse-resolution PM2.5 products.

Terrain-Stratified Implementation of the Scale-Transferability Analyses

To examine whether scale transferability differed between terrain types, the three analyses described above were repeated separately for flat and rugged areas. Terrain classification followed the GRC dataset and the definitions described in Section 3.3.3 and the methodology in Supplementary Material S3.

Terrain stratification reduced the number of valid pixels available for model fitting, particularly at coarser resolutions. Therefore, in the coefficient-side analysis, the number of sampled grid cells in each repetition was reduced from 1000 to 400, and the maximum spatial resolution considered was reduced from 8 km to 6 km. The sampling procedure was still repeated 200 times. All other settings, including the regression specification, coefficient standardisation, LMG-based importance weighting, and calculation of the Bhattacharyya coefficient, remained unchanged. The settings of the output-side and simulation experi-

ments were also identical to those used in the corresponding full-domain analyses, apart from the terrain-specific masking and separate model fitting.

3.3.6. Prediction Intervals

To assess predictive uncertainty in the final calibrated PM2.5 product, prediction intervals were estimated for the Stage 2 remaining residual, following the same correction structure as the Stage 2 point-prediction model (i.e., Stage 1 PM2.5 + daily bias + remaining residual). Quantile prediction [50] was applied to the local residual, and the resulting residual interval was added back to the Stage 1 estimate and the fold-specific daily mean residual to obtain the final PM2.5 prediction interval. The 0.05 and 0.95 quantiles were used to construct a nominal 90% raw quantile interval. Details are provided in the Supplementary Material S13.

3.3.7. PM2.5 Aggregation for Exposure Assessment

A key advantage of the 100 m PM2.5 product over the original 1 km product is that it enables finer-scale aggregation for small spatial units and supports more detailed assessment of population exposure within those units. PM2.5 exposure was therefore further summarised for Shounan Street, Yinzhou District and the adjacent University of Nottingham Ningbo China (UNNC) in Ningbo using both area-weighted and population-weighted aggregation. The 100 m gridded population dataset for China in 2020 from Chen et al. [43] was used as the common aggregation grid. For each spatial unit, the original 1 km CHAP-PM2.5 product and the Stage 2 100 m product were aligned to the same 100 m population grid using nearest-neighbour resampling. The 1 km CHAP-PM2.5 values were assigned to the corresponding 100 m population-grid cells, while the Stage 2 values were matched to the nearest 100 m population-grid cells. To ensure a fair comparison, only 100 m cells that were fully located within the target spatial unit and had valid PM2.5 values in both products were retained.

Area-weighted PM2.5 was calculated as the arithmetic mean of PM2.5 values across the retained 100 m cells, because all retained cells were treated as equal-area grid cells. Population-weighted PM2.5 was calculated by weighting each retained cell by its resident population. The difference between the area-weighted and population-weighted estimates was used to quantify the effect of within-unit population distribution on estimated PM2.5 exposure. Differences between the Stage 2 product and the original CHAP-PM2.5 product were then calculated to evaluate how the calibrated high-resolution surface affected exposure estimates in the selected spatial units. The detailed demonstration is provided in Supplementary Material S10.

4. Results

4.1. Validation of the CHAP-PM2.5 Product

The original CHAP-PM2.5 product was validated in Zhejiang Province in 2020 using daily ground-based observations (Figure 3). The product shows good agreement with observations (R^2 = 0.73, RMSE = 7.96 μg/m^3, MAE = 5.58 μg/m^3). The scatterplot also shows a systematic pattern where the fitted regression line has a slope smaller than 1, suggesting that high concentrations are underestimated while low concentrations are slightly overestimated. This is common for PM2.5 prediction models [16,26,51]. This implies a tendency toward smoothing, where extreme PM2.5 values are compressed toward the mean. The intercept of around 8 further indicates a baseline offset, meaning even when observed PM2.5 is low, the product still predicts a moderate background concentration. Together, these features suggest that the coarse-resolution PM2.5 product captures general spatial trends but tends to underestimate pollution peaks and overpredict cleaner conditions.

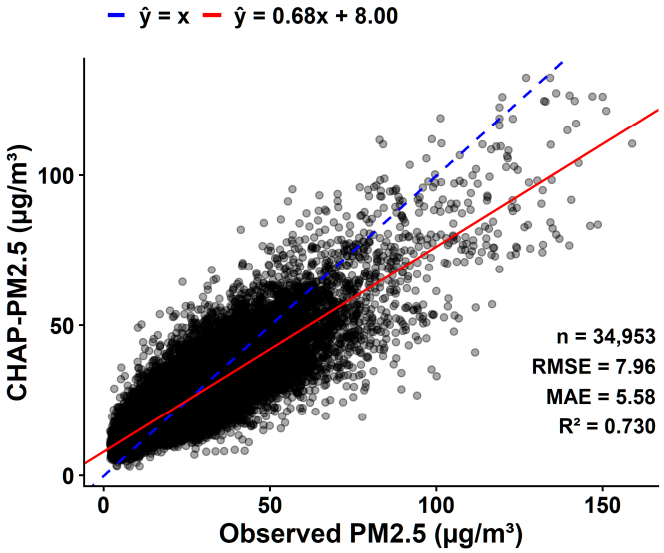


Figure 3. Validation of the CHAP-PM2.5 product using ground observations in Zhejiang Province.

4.2. Scale-Invariance Assumption

4.2.1. Coefficient-Side: Coefficient Similarity Test

Figure 4 presents the monthly cross-scale stability of PM2.5-predictor relationships, measured using the LMG-weighted Bhattacharyya coefficient of standardized regression coefficients. Overall, coefficient-distribution similarity decreases as the spatial-resolution gap increases. Adjacent or near-adjacent resolution pairs show the highest similarity, including 1 km–2 km, 2 km–3 km, and 3 km–4 km, with mean similarities of 94.8%, 96.8%, and 94.1%, respectively. In contrast, larger-gap comparisons show lower similarities, such as 1 km–8 km, 2 km–8 km, and 4 km–8 km, with mean values of 55.0%, 60.8%, and 69.5%, respectively.

Importantly, the results indicate that cross-scale similarity is affected by both the absolute and relative differences between spatial resolutions. For example, among resolution pairs with a doubling relationship, the mean similarity decreases from 94.8% for 1 km–2 km to 89.9% for 2 km–4 km and 69.5% for 4 km–8 km, suggesting that larger absolute differences can reduce similarity even when the relative scale ratio is the same. The examples with a difference of 1 km all show a very high similarity above 90%, with low variability. The wider shaded bands for larger-gap pairs further suggest greater day-to-day variability in cross-scale similarity. These findings suggest that PM2.5-predictor relationships are more transferable between similar spatial supports, while the absolute difference is more important than the relative difference in determining the similarity.

4.2.2. Output-Side: Model Prediction Difference Test

The output-side transfer experiment provided consistent evidence for the scale effect as shown in Figure 5. In 2020, the monthly mean variance of the transfer difference increased monotonically with training scale, indicating that models trained at coarser resolutions became progressively less capable of reproducing the original 1-km PM2.5 field when transferred back to the 1-km predictors. This result is consistent with the weighted-BC analysis above: as the scale gap increases, not only does coefficient similarity decline, but the transferred outputs themselves also deviate increasingly from the fine-resolution target. The monotonic increase in transfer variance across scales therefore provides output-side support for the conclusion that the scale-invariance assumption weakens as the difference between training scale and application scale becomes larger. Monthly differences in transfer sensitivity were also evident, as the magnitude of transfer variance differed substantially

among months. The highest transfer variance and transfer sensitivity were observed in December and January. These two months also have the widest PM2.5 concentration ranges, suggesting that transfer sensitivity is likely influenced by monthly PM2.5 variation and the amount of fine-scale information lost during spatial aggregation.

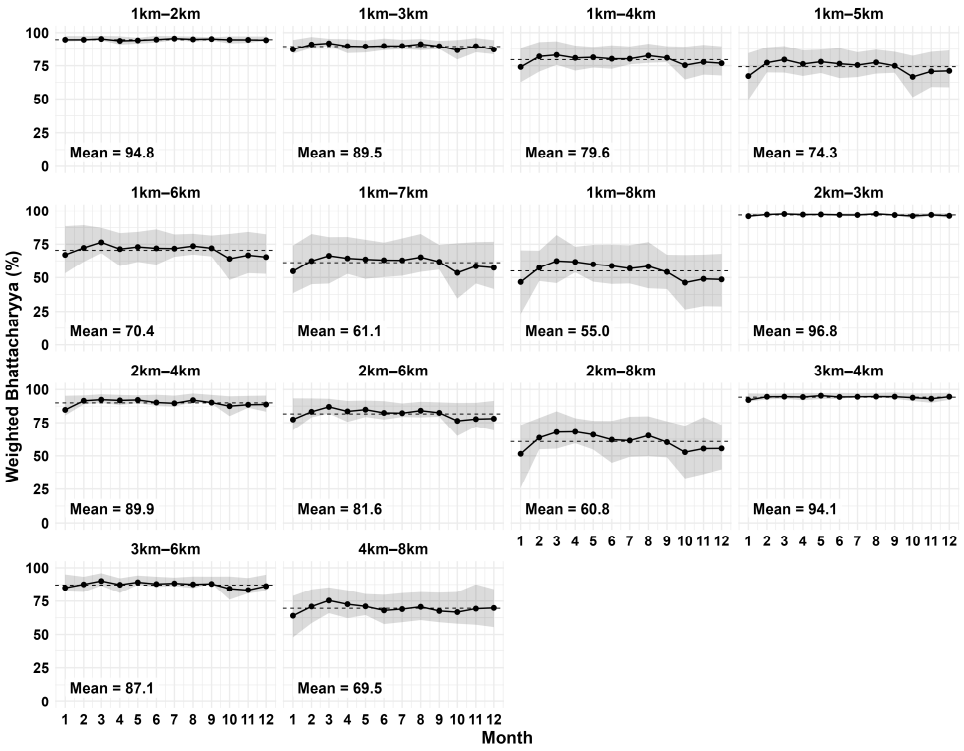


Figure 4. Monthly cross-scale stability of PM2.5-predictor relationships between spatial resolutions (2020). Points represent the monthly mean of daily similarity values; the shaded band represents the within-month interquartile range of daily similarity values. The dashed horizontal line shows the overall mean similarity for each resolution pair.

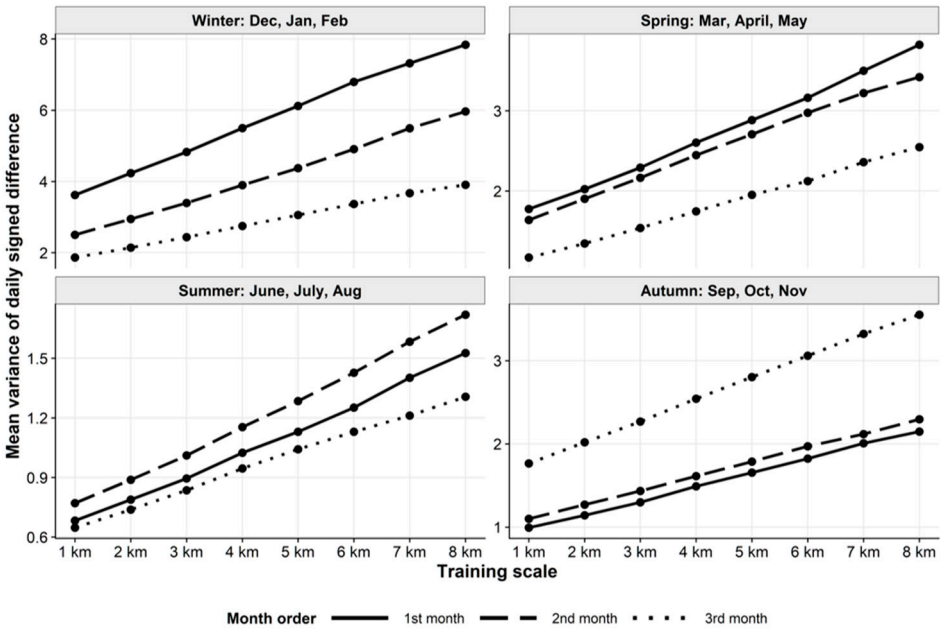


Figure 5. The monthly mean variance of the transfer difference in 2020. Month order means the order of months in the corresponding season (e.g., March, April, May).

Figure 6 compares the transferred predictions with the original 1 km CHAP-PM2.5 target after models trained at different spatial resolutions were projected back to the 1 km grid. A consistent scale-dependent pattern is observed across all months. As the training scale becomes coarser from 1 km to 8 km, R^2 gradually decreases, RMSE increases, and the regression slope declines. This indicates that models trained at coarser resolutions become progressively less able to reproduce the original 1 km PM2.5 field [19,20,23]. The decline in R^2 reflects a weakening agreement between the transferred predictions and the 1 km target, while the increase in RMSE indicates larger prediction discrepancies. More importantly, the decreasing slope suggests that coarse-trained models increasingly compress the dynamic range of the 1 km PM2.5 field, meaning that high values tend to be underestimated and low values tend to be overestimated. Although the absolute performance differs among months, the direction of change is consistent, showing that the loss of prediction-target agreement is systematic rather than month-specific. These results provide output-side evidence that PM2.5-predictor relationships are sensitive to training resolution and that models calibrated at coarser scales cannot fully preserve the magnitude and variability of the original 1 km product when transferred back to a finer grid.

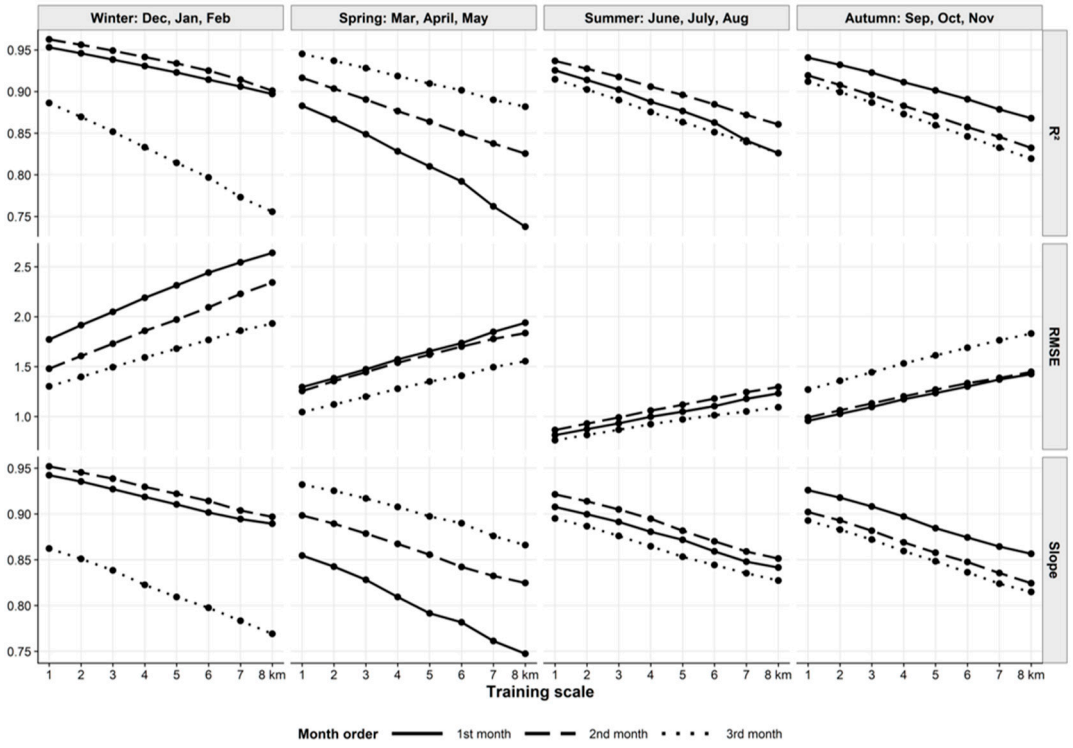


Figure 6. Effect of training scale on the agreement between transferred predictions and the original 1 km CHAP-PM2.5 target. Models were trained at spatial resolutions from 1 km to 8 km and then projected back to the 1 km predictor grid.

4.2.3. Simulation-Based Recoverability of Sub-Kilometre Heterogeneity

The comparison of five methods showed clear differences (Figure 7, Table 2). Uniform disaggregation, by construction, produced no within-cell heterogeneity and therefore yielded zero values for all heterogeneity-recovery metrics. The single-predictor redistribution baseline introduced strong sub-grid variation, but this variation was largely unrealistic. This is reflected by its strongly negative HEV and extremely inflated variance ratio, despite only weak agreement with the true heterogeneity pattern. This indicates that redistributing the 1 km mean according to a single high-resolution covariate can severely over-amplify local heterogeneity and generate largely spurious fine-scale structure.

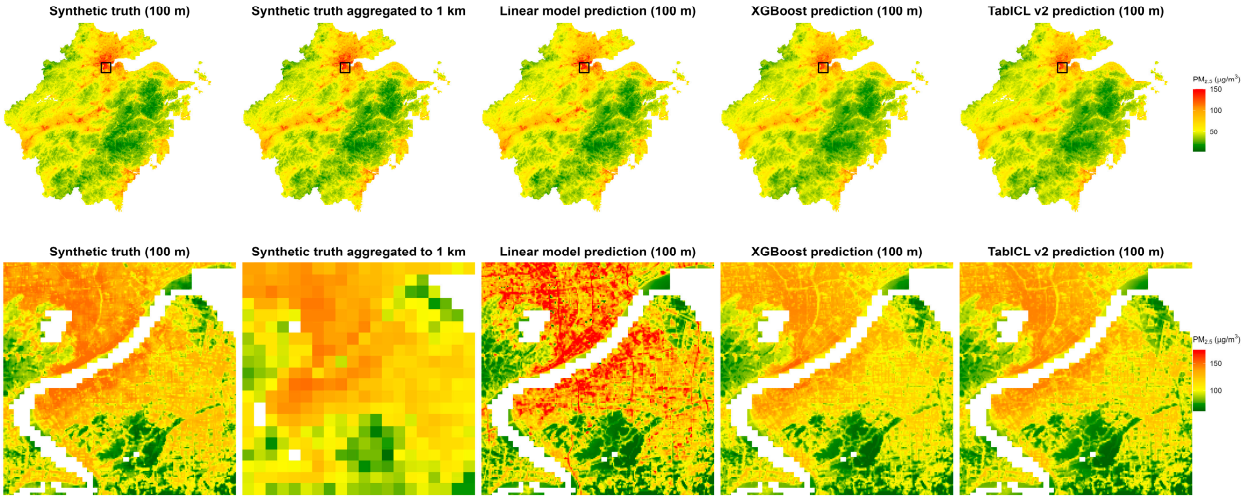


Figure 7. Regional and zoomed comparison of the synthetic 100 m truth, aggregated 1 km truth, and 100 m predictions from the linear model, XGBoost, and TabICL v2. The lower zoomed maps cover central Hangzhou.

Table 2. Comparison for simulation-based sub-kilometre heterogeneity recoverability.

Method	HEV	VR	r_{anom}	r_{sd}
Uniform	0	0	0	0
Single-predictor (road)	−167.07	177.34	0.39	0.17
Linear model	0.50	1.05	0.75	0.9
XGBoost	0.73	0.50	0.87	0.95
TabICL v2	0.74	0.50	0.87	0.95

The linear model recovered part of the within-cell heterogeneity. Its VR was close to one (1.05), indicating that the overall magnitude of within-cell variability was similar to that of the synthetic truth. However, this does not necessarily imply more accurate heterogeneity recovery. Compared with XGBoost and TabICL v2, the linear model had lower HEV and r_{anom}, suggesting weaker agreement with the true fine-scale heterogeneity pattern. The zoomed maps in Figure 7 also show that the linear model produced locally scattered high and low values, which are likely related to linear extrapolation when the model trained at 1 km was applied to the 100 m predictor stack. Thus, the linear model reproduced a relatively large amount of variability, but part of this variability appears spatially unstable rather than accurately aligned with the true 100 m pattern.

The two nonlinear models achieved the strongest overall recoverability. XGBoost obtained HEV = 0.73, $r_{anom} = 0.87$, and $r_{sd} = 0.95$, while TabICL v2 produced very similar results, with HEV = 0.74, $r_{anom} = 0.87$, and $r_{sd} = 0.95$. These results indicate that both models recovered the spatial pattern of within-cell heterogeneity more accurately than the uniform, single-predictor, and linear-model baselines. Their VR values were approximately 0.50, indicating that the amplitude of the recovered sub-grid variability was attenuated. This attenuation is expected because the models were trained using the aggregated 1 km synthetic target rather than the original 100 m truth. During aggregation, part of the 100 m within-cell variance is smoothed out and is no longer explicitly present in the training target. As a result, a model calibrated at 1 km can still use high-resolution predictors to recover the relative spatial arrangement of fine-scale heterogeneity, but it has limited ability to reconstruct the full variance that was removed by spatial averaging. In other words, XGBoost and TabICL v2 captured the spatial pattern of fine-scale heterogeneity

more reliably than the baselines, but their predictions remained compressed relative to the true 100 m field.

Figure 7 provides a visual comparison of these results. The upper row shows the regional-scale maps, while the lower row shows the zoomed area over central Hangzhou. In the regional maps, all model-based predictions broadly reproduce the large-scale spatial gradients in the synthetic truth. However, the differences become clearer in the zoomed maps. The aggregated 1 km truth shows the expected coarse and blocky structure, whereas the 100 m predictions recover finer spatial detail. The linear model introduces more local speckled variation and isolated extreme patches, consistent with its larger VR. By contrast, the XGBoost and TabICL v2 predictions are visually smoother but more consistent with the main spatial organisation of the synthetic 100 m truth, including the broad locations of high-value zones, surrounding lower-value areas, and several locally connected linear and patch-like features. These visual patterns are consistent with the heterogeneity-based metrics: Stage 1 downscaling can recover a substantial part of the sub-kilometre spatial pattern, but the full amplitude of fine-scale variability is not completely restored.

4.2.4. Terrain-Stratified Differences in Scale Transferability

The coefficient-side analysis showed a consistent scale-dependent decline in both terrain types (Table 3). The mean LMG-weighted BC decreased as the difference between the two spatial resolutions increased in both flat and rugged terrain. However, rugged terrain had a higher mean weighted BC than flat terrain for all ten resolution pairs examined. The terrain difference was relatively small between similar resolutions but became more pronounced for larger scale gaps. Thus, across the kilometre-scale resolution contrasts examined here, the estimated coefficient distributions were more stable, and therefore less sensitive to changes in the spatial resolution, in rugged terrain than in flat terrain.

Table 3. Mean LMG-weighted Bhattacharyya coefficient for flat and rugged terrain across the km-scale resolution pairs. The difference Rugged − Flat is expressed as a percentage.

Resolution Pair (km)	Flat (%)	Rugged (%)	Difference
1–2	94.1	97.8	3.7
1–3	91.4	96.2	4.8
1–4	86.3	91.4	5.1
1–5	80.5	87.7	7.2
1–6	74.3	87.6	13.3
2–3	96.3	97.6	1.3
2–4	92.2	94.1	1.9
2–6	82.3	90.7	8.4
3–4	94.7	96.6	1.9
3–6	85.6	94.0	8.4

The output-side results showed that model performance deteriorated with increasing training scale in both terrain types, but the rate of deterioration was greater in flat terrain (Figure 8). Consequently, although flat terrain retained a higher absolute R^2 at every training scale and a higher slope at most scales, the advantage decreased as the training resolution became coarser. Flat terrain had lower RMSE and MAE at 1 km, but these errors exceeded those in rugged terrain from approximately 2 km onwards. Bias remained small in both terrain types, particularly up to 6 km, indicating that the principal deterioration concerned local prediction-target agreement and dynamic-range preservation rather than a strong overall shift in the mean. These within-terrain changes indicate that, across the 1–8 km range examined, transferred outputs in flat terrain were more sensitive to reduced training resolution. Nevertheless, differences in model performance among terrain types

should be interpreted cautiously. RMSE and MAE are scale-dependent and should be considered relative to the terrain-specific mean concentration and standard deviation.

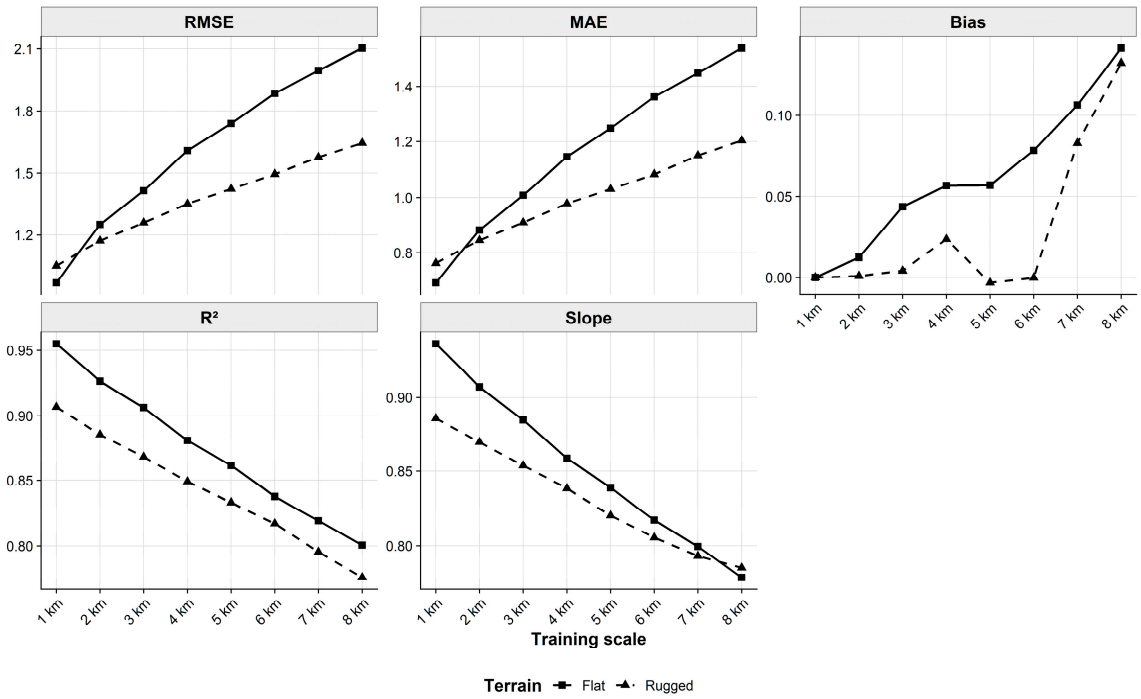


Figure 8. Terrain-stratified output-side transfer performance across training resolutions. Terrain-specific XGBoost models were trained at resolutions from 1 to 8 km and evaluated against the original 1 km CHAP-PM2.5 target within the corresponding 1 km terrain stratum. Points show the mean daily RMSE, MAE, bias, R^2, and regression slope in 2020.

The simulation experiment showed a different terrain contrast in the recovery of sub-grid heterogeneity (Table 4). Both nonlinear models achieved higher HEV, VR, and r_{anom} in flat terrain than in rugged terrain, indicating substantially stronger recovery of fine-scale heterogeneity in flat areas. In contrast, r_{sd} differed little between the two terrain types and was slightly higher in rugged terrain, indicating that the nonlinear models were similarly capable of identifying the relative spatial pattern of 1 km cells with higher or lower within-cell variability. However, the substantially lower VR in rugged terrain indicates that the magnitude of this variability was less fully recovered, even though its relative spatial distribution was well captured. The linear model did not follow the same pattern, showing higher HEV and correlation metrics in rugged terrain. The terrain contrast in recoverability was therefore model- and metric-dependent in this experiment. Nevertheless, for the nonlinear models used in Stage 1, the stronger recovery of within-cell heterogeneity in flat terrain was qualitatively consistent with the higher R^2 observed in this terrain type in the output-side experiment.

Taken together, the three lines of evidence show that scale transferability was partial and terrain-dependent. The coefficient-side analysis indicates that PM2.5–predictor relationships were not strictly invariant across kilometre-scale spatial supports and changed more substantially as the scale gap increased. The output-side test further shows that, when models trained at increasingly coarse resolutions were projected back onto the 1 km grid, spatial detail and dynamic range were progressively lost. Although these two analyses do not directly test the sub-kilometre domain, they demonstrate that scale invariance is already unstable across kilometre-scale supports and therefore should not be assumed for the transfer from 1 km to 100 m. Nevertheless, the simulation experiment demonstrates that coarse-resolution training does not imply a complete inability to generate sub-grid

variation. Nonlinear Stage 1 models reconstructed a substantial proportion of fine-scale heterogeneity and recovered spatial patterns absent from the aggregated 1 km target. However, because spatial aggregation had already removed part of the local variance and extremes from the training target, the range of predicted 100 m values remained compressed relative to the synthetic truth.

Table 4. Terrain comparison for simulation-based sub-km heterogeneity recoverability.

Terrain	Method	Best Single Predictor	HEV	VR	r_{anom}	r_{sd}
Flat	Uniform	/	0	0	0	0
Flat	Single predictor	Building Height	−79.30	90.27	0.58	−0.24
Flat	Linear model	/	0.54	1.13	0.78	0.77
Flat	XGBoost	/	0.85	0.65	0.93	0.93
Flat	TabICL v2	/	0.86	0.67	0.94	0.94
Rugged	Uniform	/	0	0	0	0
Rugged	Single predictor	AOD	−31.78	32.58	0.07	0.17
Rugged	Linear model	/	0.62	0.85	0.80	0.92
Rugged	XGBoost	/	0.65	0.34	0.85	0.94
Rugged	TabICL v2	/	0.67	0.38	0.85	0.94

Terrain did not uniformly increase or reduce scale transferability. Instead, different components of transferability exhibited different terrain-dependent patterns: coefficient distributions were more stable across scales in rugged areas, whereas nonlinear models recovered a larger proportion of sub-grid heterogeneity in flat areas. Importantly, the greater coefficient stability observed in rugged areas did not translate into uniformly better output-side performance or stronger nonlinear heterogeneity recovery. Coefficient stability, reproduction of the 1 km target, and recovery of within-cell variation therefore represent distinct aspects of scale transferability.

4.3. Validation of Stage 1

To evaluate the consistency of the Stage 1 predictions, these were aggregated and compared to the original CHAP-PM2.5 product (Supplementary Material Section S6). This yielded a bias less than 0.1 $\mu g/m^3$ and an RMSE between 0.5 and 2 $\mu g/m^3$. While Stage 1 is not a mass-preserving model, the aggregated values were consistent with the original CHAP-PM2.5 product.

Figure 9 shows the validation results of the XGBoost-based Stage 1 downscaling model against the original 1 km CHAP-PM2.5 product in 2020. The full figure including both model fitting and 10-fold random cross-validation can be found in Supplementary Material Figure S15. Overall, the 10-fold cross-validation results show strong agreement with the CHAP-PM2.5 target, although the scatter becomes slightly more dispersed than in the model-fitting panels (Figure S15). This difference is expected, because the validation predictions are generated for held-out pixels that were not used for model fitting. Compared with the fitting results, cross-validation shows modest reductions in R^2 and slightly greater spread around the 1:1 line, indicating some loss of predictive accuracy when the model is applied to unseen samples. Nevertheless, the overall validation performance remains high in every month, with slopes still close to 1 and bias values remaining very small. Taken together, these results suggest that the XGBoost-based Stage 1 model has a strong ability to learn and generalize the coarse-grid PM2.5-predictor relationship. A comparison of the flat and rugged terrain showed that their RMSE (difference < 0.01 $\mu g/m^3$) and R^2 (difference < 0.01) were very similar (Supplementary Material S7, Figure S8).

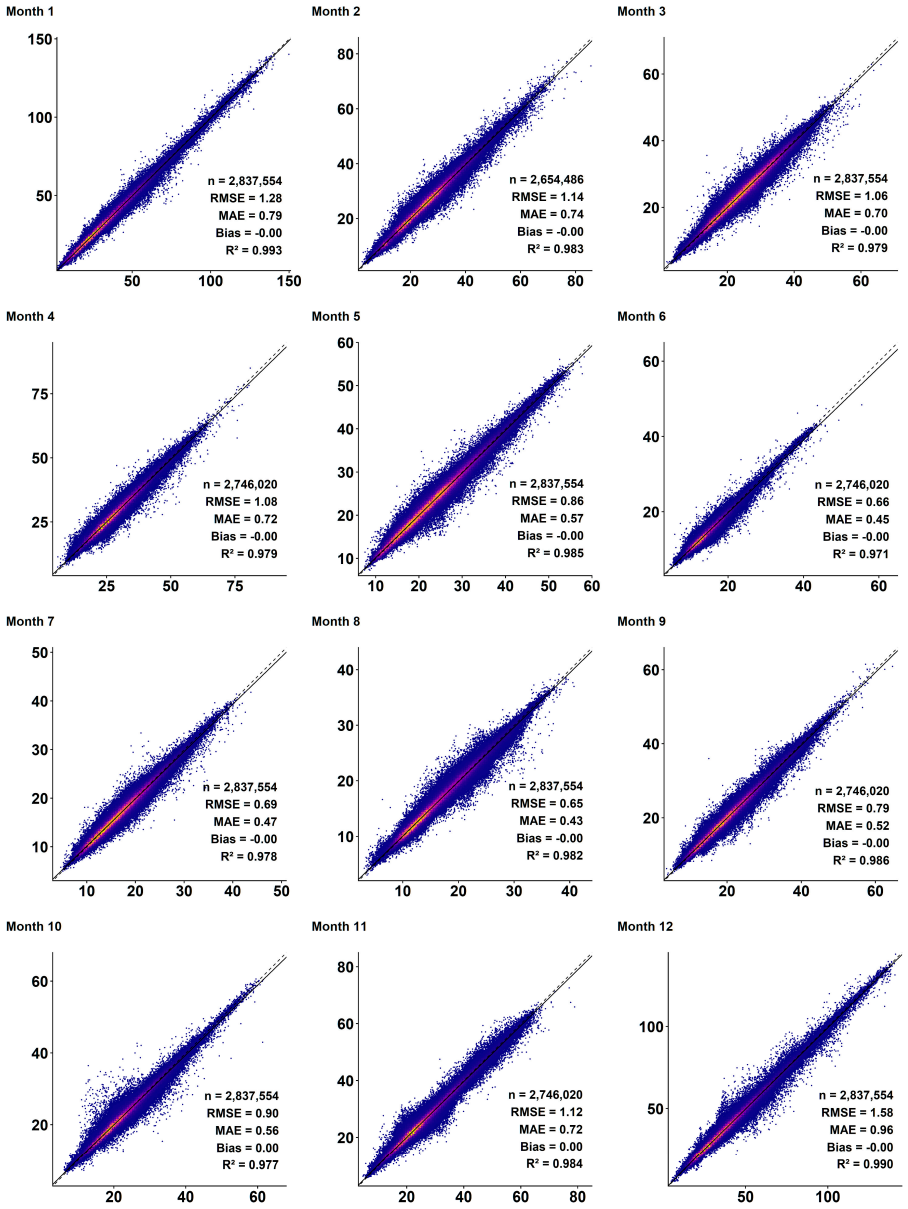


Figure 9. Monthly 10-fold random cross-validation of the Stage 1 XGBoost vs. original 1 km CHAP-PM2.5. Density-coloured scatterplots show monthly out-of-fold predictions for 2020.

Note that this Stage 1 validation does not represent an independent assessment against ground observations. Instead, it evaluates how well the downscaling model can reproduce the original CHAP-PM2.5 raster used as the training target. Therefore, the purpose of Figure 9 is to assess the internal consistency and transferability of the Stage 1 model with respect to the coarse-resolution PM2.5 product, rather than to demonstrate improvement in agreement with monitoring-site observations. The latter is examined separately in the following Stage 1 product validation against ground PM2.5 observations.

For computational reasons, the Stage 1 validation for TabICL v2 was presented in a slightly different form. Unlike XGBoost, TabICL v2 is computationally much more expensive (approximately 40 min versus 1 min for one day prediction, details in Supplementary Material S2). As a result, repeating the validation procedure over large pixel-level datasets is substantially more time-consuming. Therefore, TabICL v2 was evaluated using 5-fold random cross-validation instead of 10-fold cross-validation, in order to reduce computational burden while still providing out-of-fold predictions for all validation samples. In addition, only the cross-validation scatterplot was reported for TabICL v2.

Figure 10 shows the 5-fold cross-validation performance of the TabICL v2-based Stage 1 model against the 1 km CHAP-PM2.5 target. Across all months, the predicted values are tightly concentrated around the 1:1 line, with very high cross-validated R^2 values and low RMSE and MAE. The agreement remains strong in both low-concentration months and high-concentration months, indicating that TabICL v2 can effectively reconstruct the spatial and temporal variation in the original CHAP-PM2.5 product at the 1 km scale. Compared with the XGBoost-based Stage 1 validation, the TabICL v2 scatterplots generally show tighter agreement with the CHAP-PM2.5 target and slightly stronger validation metrics, even though TabICL v2 was evaluated using 5-fold rather than 10-fold cross-validation. Because 5-fold cross-validation uses a smaller proportion of the data for training in each fold than 10-fold cross-validation, this result suggests that TabICL v2 has strong generalisation ability and may provide better Stage 1 reproduction of the CHAP-PM2.5 target than XGBoost. A comparison of the flat and rugged terrain showed that their RMSE (difference < 0.01 μg/m^3) and R^2 (difference < 0.01) were very similar (Figure S8).

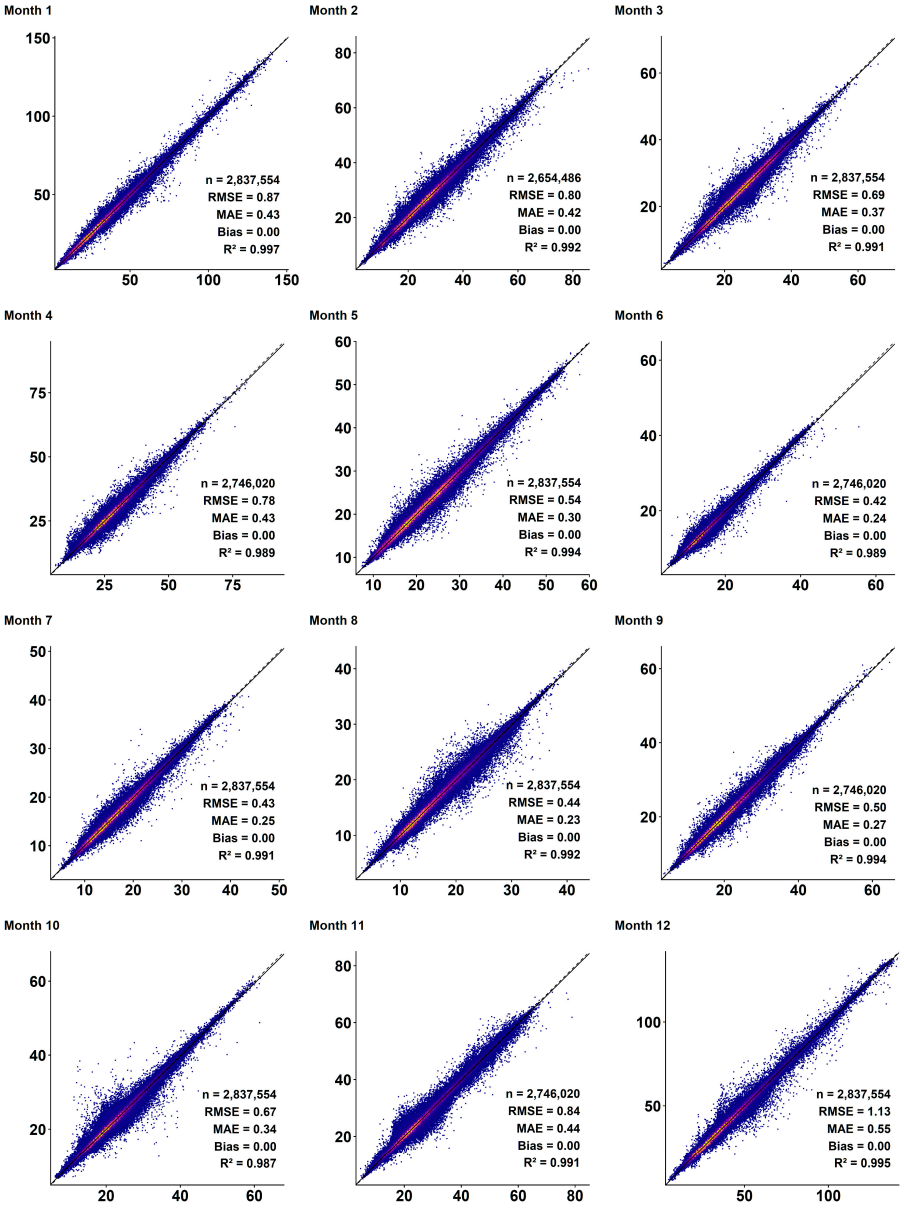


Figure 10. Monthly 5-fold cross-validation of the Stage 1 TabICL v2 vs. original 1 km CHAP-PM2.5. Density-coloured scatterplots show monthly out-of-fold predictions for 2020.

4.4. Validation of Stage 2

Stage 2 was evaluated from three complementary perspectives: its month-specific improvement over the CHAP-PM2.5 under the primary 2 km exclusion buffer, its sensitivity to increasing validation-to-training separation (using increasingly larger buffers), and its performance across flat and rugged terrain contexts.

4.4.1. Monthly Improvements Under the 2 km Exclusion Buffer

Under the 2 km training-site exclusion buffer, all three Stage 2 calibration methods improved R^2, RMSE, and slope error relative to the original CHAP-PM2.5 product in every month and for both Stage 1 sources (Figure 11). The bias + TabICL v2 method consistently produced the largest improvements, followed generally by bias plus XGBoost, whereas bias-only calibration produced the smallest gains. The magnitude of improvement varied seasonally. Reductions in RMSE were particularly pronounced in months with relatively large errors in the original CHAP product, while significant improvements in R^2 were also evident during months with lower PM2.5 concentrations. The similar monthly patterns obtained from the two Stage 1 sources indicate that the benefits of Stage 2 calibration were not strongly dependent on the Stage 1 model. Detailed performance metrics and significance test results for 0 and 2 km buffered CV are presented in Tables S9 and S10.

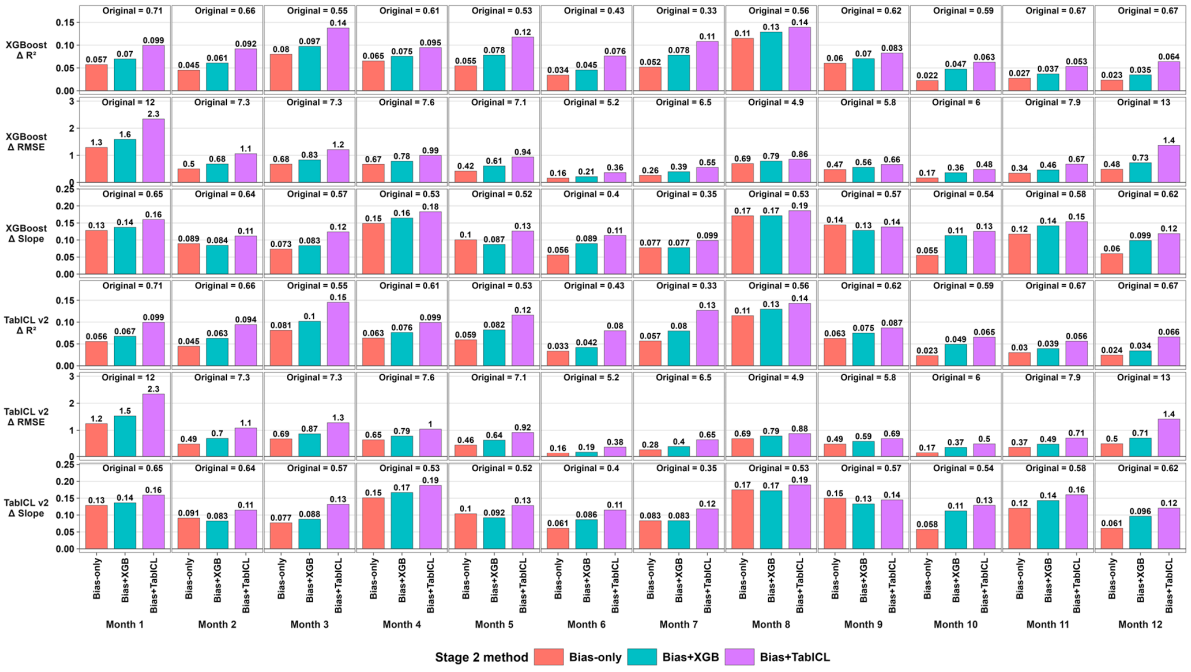


Figure 11. Monthly improvements in Stage 2 validation performance relative to the original CHAP-PM2.5 under a 2 km training-site exclusion buffer. Rows show changes in R^2, RMSE, and slope error for the two Stage 1 sources, and columns represent the 12 months. Values above the bars show the improvement relative to CHAP-PM2.5. Positive values consistently indicate improvement.

4.4.2. Sensitivity to Training-Site Exclusion Distance

Stage 2 performance weakened as the training-site exclusion buffer increased from 0 to 30 km (Figure 12). This decline was steepest between 0 and 2 km and was most evident for bias plus TabICL v2. For exclusion distances greater than 5 km, the decline in performance is shallow. Nevertheless, the Stage 2 methods continued to improve R^2, RMSE, MAE, and regression slope relative to the CHAP-PM2.5 product and Stage 1 predictions even under the 30 km buffer. The patterns were nearly identical for the two Stage 1 sources. Increasing the exclusion buffer progressively increased the separation between validation and training

sites (Table S2). The mean distance to the nearest training site varied from 13 km for the 0 km buffer to 41 km for the 30 km buffer (2 km: 19 km; 5 km: 24 km; 10 km: 26 km; 20 km: 33 km). The nearest-pair residual correlation was also reduced (Figure S9), with nearly identical patterns for the two Stage 1 sources. This reduction in residual dependence was consistent with the attenuation of Stage 2 predictive performance at larger buffer distances (see Supplementary Material S8 and Figure S9). Further analysis showed that 30% of the province lay within 13 km of a monitoring station, 72% within 24 km and 10% lay more than 33 km from a monitoring station (Table S3).

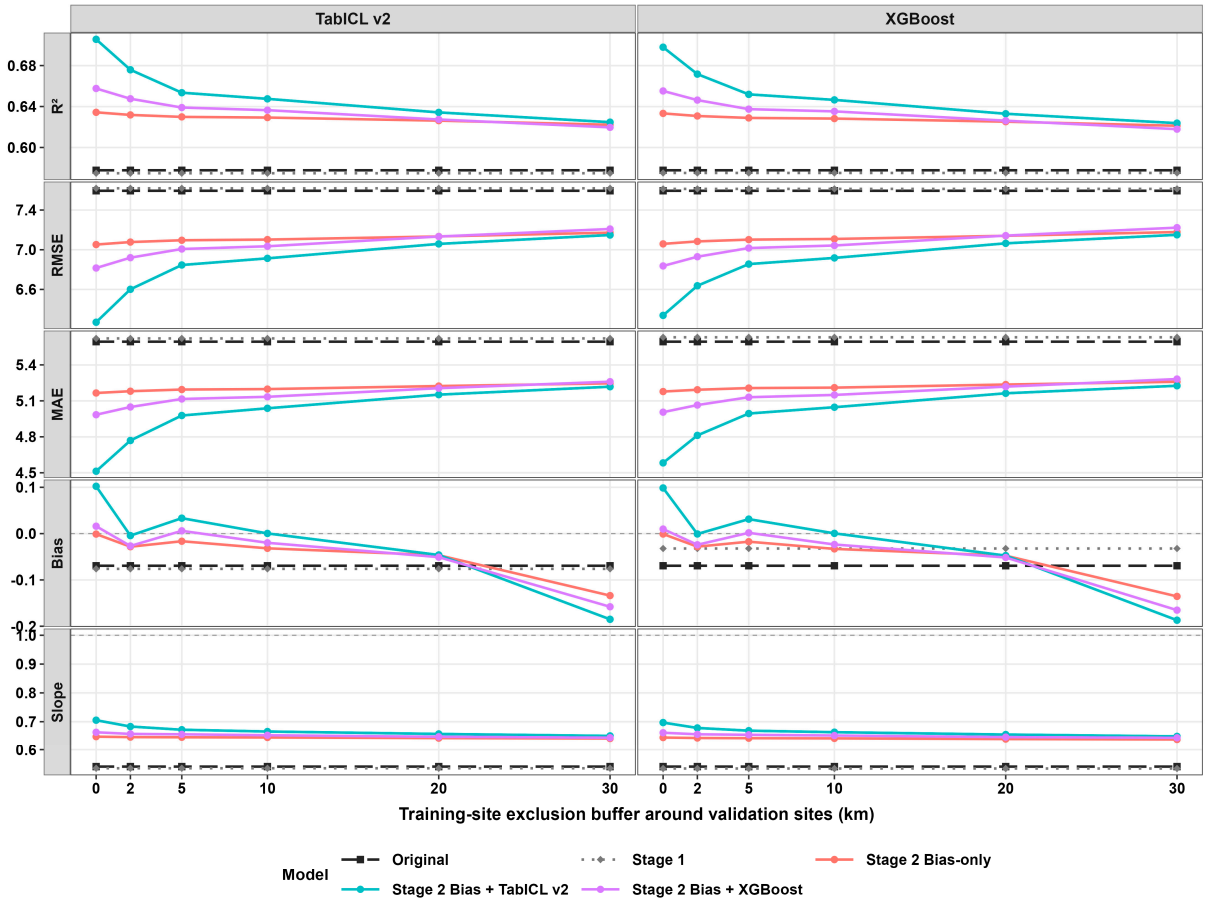


Figure 12. Annual mean sensitivity (2020) of Stage 2 validation performance to the training-site exclusion buffer. Columns show results based on TabICL v2 and XGBoost Stage 1 predictions, while rows show validation metrics for the different Stage 2 models.

4.4.3. Terrain-Stratified Validation

The terrain-stratified analysis revealed a metric-dependent contrast between flat and rugged areas (Figure 13). Across most models and months, flat areas exhibited higher R^2, smaller absolute bias, and slopes closer to unity, indicating a stronger ability to reproduce the observed concentration range. In contrast, rugged areas generally exhibited lower RMSE and MAE in absolute concentration units. These patterns were consistent between the two Stage 1 sources. Stage 2 calibration reduced, but did not completely eliminate, the terrain-related differences observed for the CHAP-PM2.5 product and Stage 1 predictions.

The apparently contrasting results for R^2 and the absolute error metrics can be explained partly by differences in the distribution of observed PM2.5 between the two terrain classes. Table 5 shows that the annual mean (2020) observed PM2.5 concentration was 26.8 μg/m^3 at flat sites and 22.0 μg/m^3 at rugged sites, while the corresponding standard deviations were 14.1 and 11.2 μg/m^3, respectively. Although the Stage 2 RMSE was smaller

in rugged areas, the normalized RMSE was larger in rugged areas. The lower absolute RMSE in rugged areas was partly attributable to their smaller observed mean and concentration range. Relative to the magnitude of the observed PM2.5 variation, prediction errors were larger in rugged terrain. For both terrains, Stage 2 modelling led to improved accuracy compared to Stage 1. This pattern is consistent for all months (see Table S6).

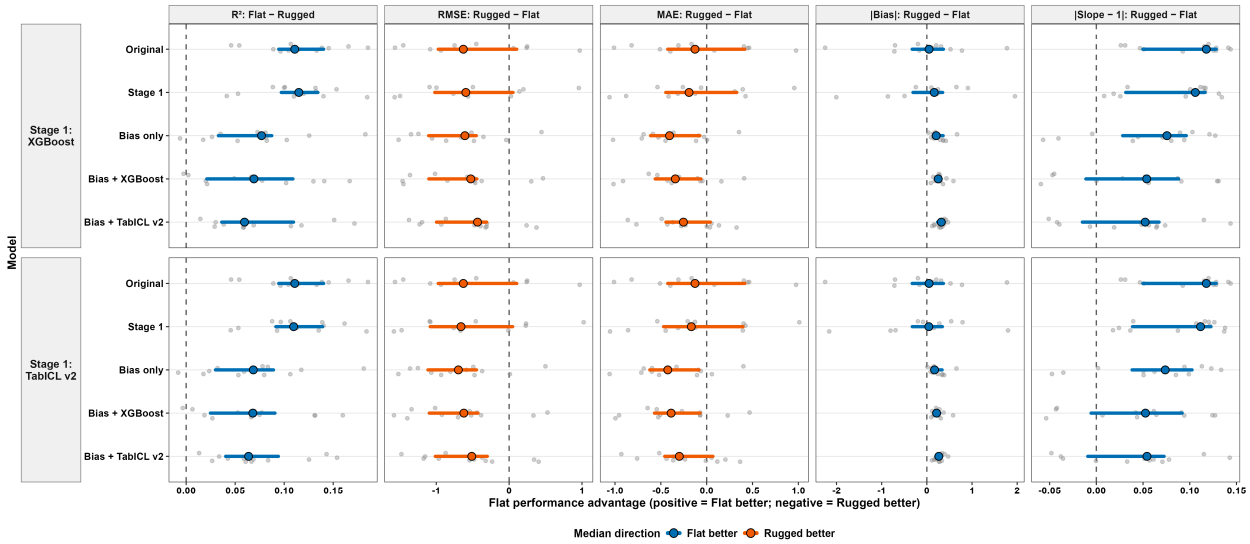


Figure 13. Paired monthly differences in validation performance between flat and rugged terrain at a 2 km training-site exclusion buffer. Grey points represent monthly paired Flat-Rugged validation metrics. Coloured points and horizontal bars represent the median and interquartile range across the 12 months. Positive values indicate better performance in flat terrain and negative values indicate better performance in rugged terrain. Blue (orange) points indicate that the median difference favours flat (rugged) terrain. Positive values consistently indicate better performance in flat terrain. For detailed results used to produce this figure, see Supplementary Material S9.

Table 5. Terrain-specific PM2.5 observed at monitoring stations and Stage 1 and Stage 2 RMSEs. TabICL v2 was used in Stage 1 and bias + TabICL v2 was used in Stage 2. The observed mean, observed SD, and RMSE are annual values (2020), expressed in $\mu g/m^3$. RMSE/SD is dimensionless.

Terrain	Sites, n	Observed PM2.5 Mean	Observed PM2.5 SD	Stage 2 RMSE	Stage 2 RMSE/SD	Stage 1 RMSE	Stage 1 RMSE/SD
Flat	62	26.79	14.07	7.15	0.51	8.23	0.59
Rugged	36	22.03	11.17	6.54	0.59	7.64	0.68

4.5. Mapping Results

Figure 14 compares the original 1 km CHAP-PM2.5 product with the Stage 1 downscaled products generated by XGBoost and TabICL v2 on 15 January 2020 to illustrate the spatial behaviour of different Stage 1 downscaling methods. This date was selected as a representative daily example for visual comparison because it showed clear regional PM2.5 gradients as well as visible local spatial variation in the zoom-in areas. At the full-domain scale, both Stage 1 products preserve the large-scale spatial pattern of the original CHAP-PM2.5 field. In the finest zoom-in view, both Stage 1 products introduce sub-kilometre spatial detail that is absent from the original 1 km product, which remains visibly blocky after zooming. The XGBoost product shows sharper local contrasts reflecting the covariate data, whereas the TabICL v2 product shows smoother transitions between neighbouring pixels. In terms of the local geographical setting (bottom panels), Ningbo

Lishe International Airport and the surrounding rural and agricultural areas are located in the lower-left part of the domain, mountainous terrain and a small lake occupy the upper part, and the urban centre of Haishu District lies on the right-hand side of the domain. PM2.5 concentrations are relatively low in the lower-left and upper parts of the domain but increase toward the more urbanized area on the right. This spatial distribution is consistent with the expected contrast between rural or mountainous environments and the densely urbanized city. The corresponding satellite image, which provides additional land-surface context, is presented in Supplementary Material S12 (Figure S11).

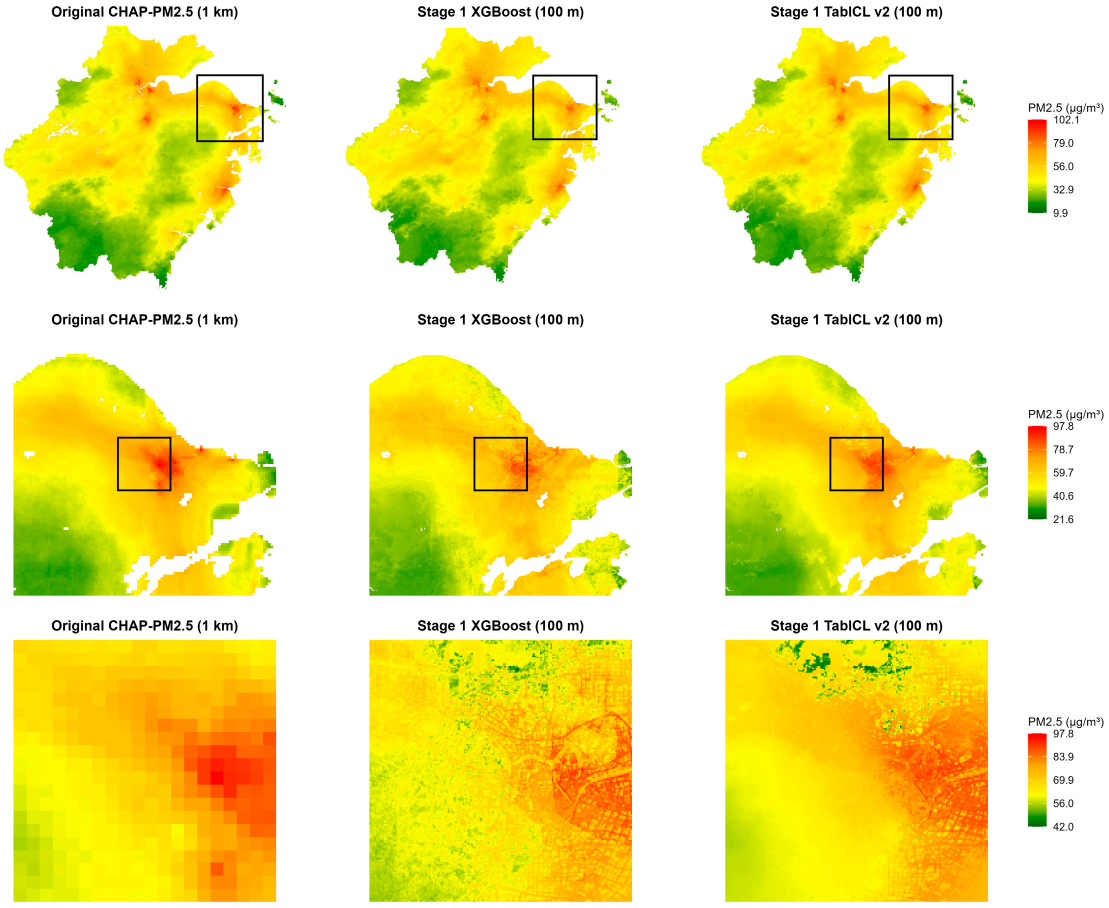


Figure 14. Comparison of the original CHAP-PM2.5 product and Stage 1 downscaled products on 15 January 2020. The first row shows the full study area, the second row shows a regional zoom-in view of the Ningbo area and the third row shows a local zoom-in view. Black rectangles indicate the zoom-in regions. The mid-range covers 100 × 100 km and the zoom-in covers 20 × 20 km.

Figure 15 compares the Stage 1 downscaled products and the corresponding Stage 2 observation-calibrated products on 15 January 2020. The panels compare two Stage 1 products and six Stage 2 calibrated products. In the Stage 2 panel labels, the first model name denotes the Stage 1 downscaling model used to generate the 100 m PM2.5 surface, while the final model name denotes the residual model used in Stage 2. Overall, Stage 2 correction substantially changes the concentration level of the Stage 1 products. Compared with the two Stage 1 maps, all Stage 2 products show lower PM2.5 concentrations over large parts of the study area. The daily bias correction on this date mainly applied a negative adjustment to the Stage 1 products. The bias-only correction largely preserves the broad spatial pattern of the corresponding Stage 1. In contrast, residual modelling modified the local spatial pattern more strongly. The XGBoost residual correction produces visible patch-like artefacts, especially in the zoom-in panels, regardless of whether the

Stage 1 input was generated by XGBoost or TabICL v2. Similar artefacts have been reported by other authors [47] who have performed XGBoost predictive modelling and by authors who have used similar tree-based machine learning methods [13,52]. Details of Stage 2 artefacts are presented in Supplementary Material S11. The TabICL v2 residual correction produces a smoother and more spatially coherent surface, with fewer fragmented artefacts than the XGBoost residual correction, while still adjusting the Stage 1 product beyond a simple additive bias shift. The zoom-in panels further show that the final spatial pattern is influenced by both the Stage 1 downscaling model and the Stage 2 residual model. Overall, the figure demonstrates that Stage 2 calibration improves the observation-constrained product but can also alter spatial structure.

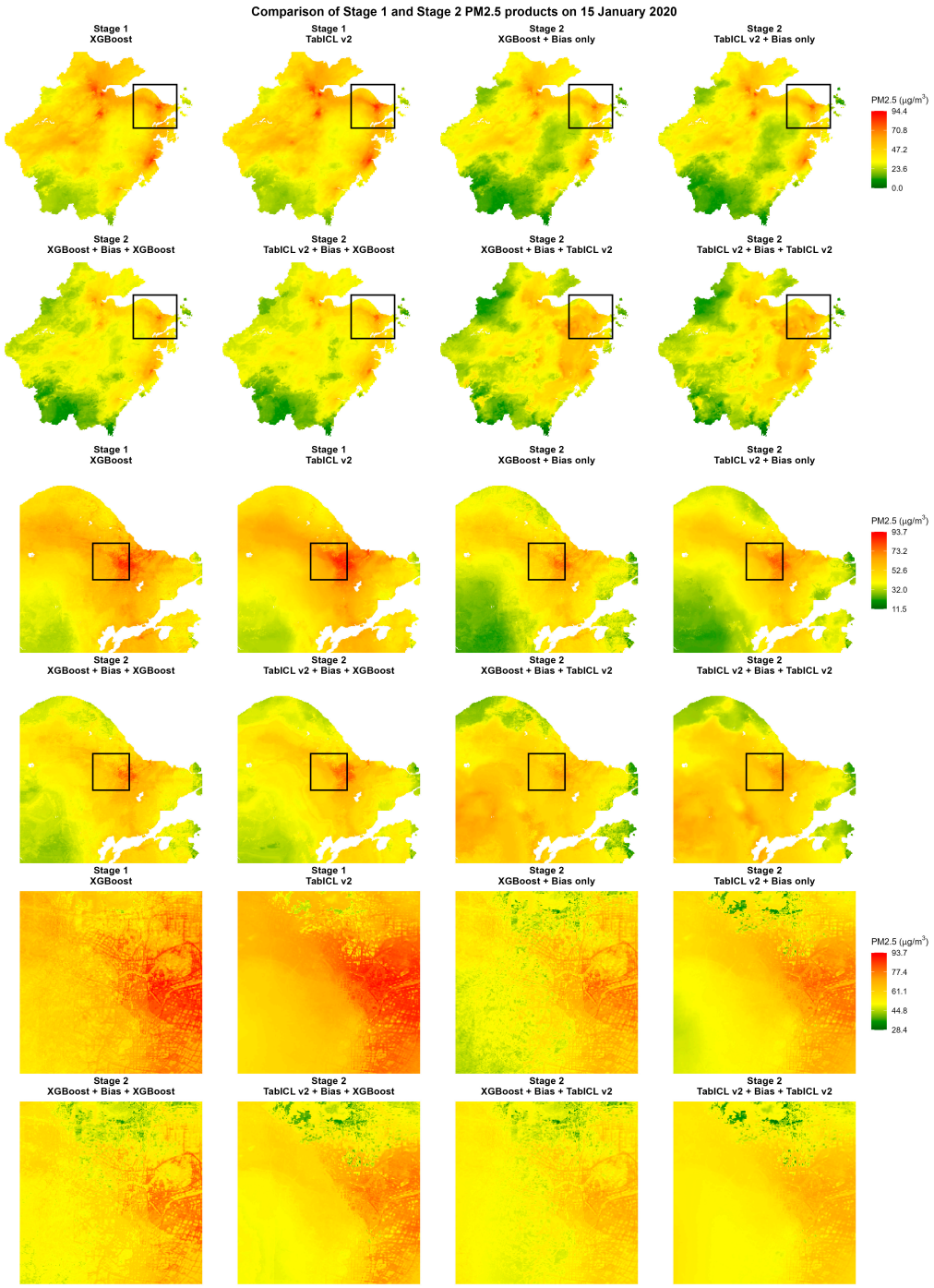


Figure 15. Mapping comparison of Stage 1 downscaling and Stage 2 observation-based calibration methods on 15 January 2020. Other details as for Figure 14.

4.6. Area-Weighted and Population-Weighted Aggregation

Figure 16 illustrates the local-scale differences between the original CHAP-PM2.5 product and the final Stage 2 product in two selected small areas in Ningbo. The Stage 2 product shown is the TabICL v2 + bias + TabICL v2 configuration (either Stage 1 model could be used; we chose TabICL v2 for this demonstration). The two areas were selected because they contain contrasting land-use and population settings. The central part of Shounan Street contains Ningbo's Southern Business District, which has a dense concentration of businesses and housing, whereas its southwestern part is dominated by suburban farmland. The University of Nottingham Ningbo China (UNNC) area has the UNNC campus in the centre, residential neighbourhoods to the west, and farmland to the southeast. In Shounan Street, the Stage 2 map reveals fine-scale variation that is not visible in the original 1 km CHAP-PM2.5 product, including a low-concentration area over farmland in the southwest, moderately higher concentrations to the north in residential land, and a narrow high-concentration strip corresponding to a road crossing the farmland. In the UNNC area, the Stage 2 product similarly distinguishes lower concentrations over farmland in the southeast, elevated concentrations along a road to the southwest of the farmland, and relatively higher concentrations in the northwestern school and residential area. These land-use-related spatial contrasts are largely absent in the original CHAP-PM2.5 product, demonstrating the added value of the calibrated 100 m surface for representing intra-urban PM2.5 heterogeneity, thereby addressing both bias and spatial-resolution limitations.

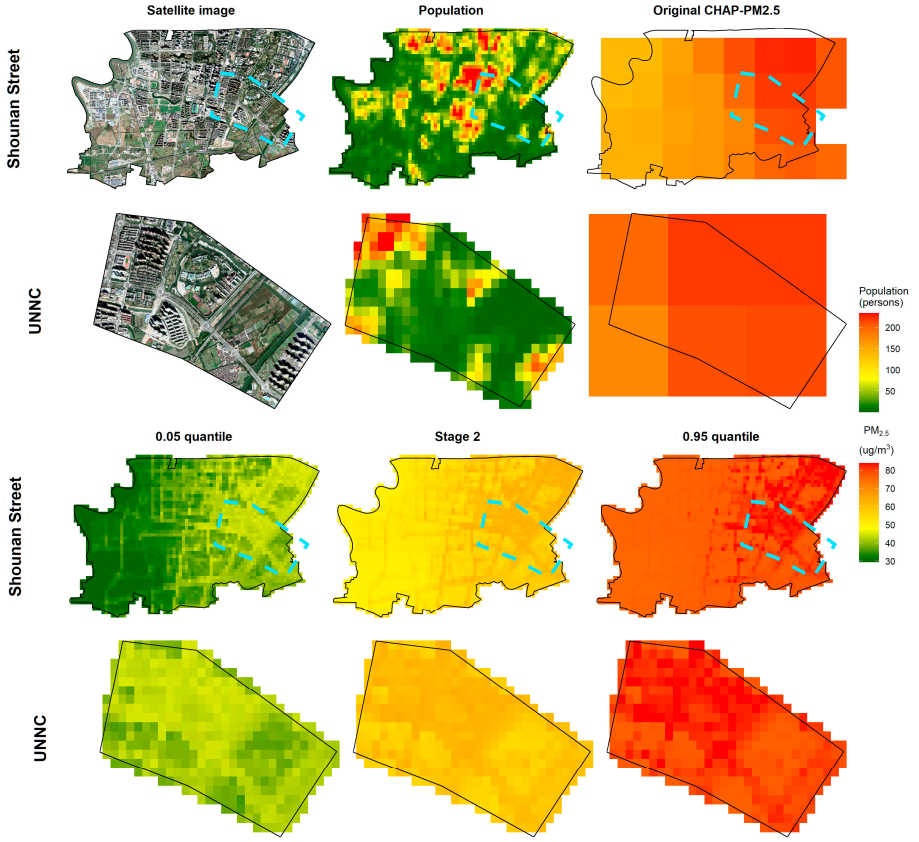


Figure 16. Comparison of satellite imagery, gridded population, the original CHAP-PM2.5 product, the Stage 2 PM2.5 prediction, and the 0.05 and 0.95 quantile prediction surfaces for Shounan Street and the UNNC area in Ningbo on 15 January 2020. Satellite imagery was obtained from WorldView-3. The Stage 2 product combines Stage 1 TabICL v2 and Stage 2 bias + TabICL v2 local residual modelling. The dashed blue line indicates the UNNC area.

Table 6 shows that the Stage 2 product produced consistently lower area-weighted and population-weighted PM2.5 exposure estimates than the original CHAP-PM2.5 product in both Shounan Street and the UNNC area, i.e., CHAP-PM2.5 overestimated the PM2.5 concentration. The difference between population-weighted and area-weighted estimates was comparatively small, indicating that population weighting introduced a more moderate adjustment than the change from the original 1 km surface to the calibrated 100 m surface. Nevertheless, the direction of this adjustment varied by spatial unit and product, reflecting differences in how population distribution overlapped with local PM2.5 patterns. At both the street and UNNC scale, population-weighted estimates were higher than area-weighted estimates for Stage 2 products, suggesting that population clusters overlapped with areas of relatively high PM2.5 concentrations. These results suggest that the Stage 2 product can affect small-area exposure estimates while also supporting more detailed population-weighted aggregation.

Table 6. Area-weighted and population-weighted PM2.5 exposure estimates for Shounan Street and the University of Nottingham Ningbo China (UNNC) area on 15 January 2020.

Spatial Unit	Product	Area-Weighted PM2.5	Population-Weighted PM2.5	Difference
Shounan Street	CHAP-PM2.5 1 km	71.2	72.7	−1.5
Shounan Street	Stage 2 PM2.5 100 m	60.8	62.5	−1.7
UNNC	CHAP-PM2.5 1 km	78.5	77.6	0.9
UNNC	Stage 2 PM2.5 100 m	67.1	69.1	−2.0

5. Discussion

5.1. Scale Transferability Is Partial and Terrain-Dependent

A central question in this study is whether the PM2.5-predictor relationship learned at a coarse resolution can be reliably transferred to a finer grid. The results suggest that this scale-invariance assumption is only partially valid. The coefficient-side comparison showed that predictor effects changed as spatial resolution became coarser, while the output-side transfer test showed that models trained at coarser resolutions became less able to reproduce the original 1 km PM2.5 field. The simulation experiment further showed that model-based downscaling could recover part of the known sub-kilometre heterogeneity, but the recovered variability was incomplete and model-dependent.

The different components of transferability exhibited distinct terrain-dependent patterns. Across the kilometre-scale contrasts, coefficient distributions were more stable in rugged terrain, whereas output-side performance deteriorated more rapidly with increasing training scale in flat terrain. Nevertheless, flat terrain generally retained higher R^2 values and allowed the nonlinear models to recover a larger proportion of sub-grid heterogeneity. The comparable r_{sd} values between terrain types further indicate that the models were similarly able to identify which 1-km cells had relatively higher or lower within-cell variability, despite recovering their magnitude less effectively in rugged terrain. Thus, greater coefficient stability did not necessarily translate into better output agreement or heterogeneity recovery, and no terrain class consistently performed better across all transferability metrics.

Here, the term "terrain-dependent" refers to systematic differences associated with terrain classes rather than an independent causal effect of terrain itself. Indeed, rugged terrain tends to be associated with lower, more spatially homogeneous PM2.5 values. These can potentially be adequately represented by coarser resolution data, compared to the more heterogeneous PM2.5 over flatter areas [53]. In contrast, the finer resolution data is needed to represent local variability over flat areas. Flat and rugged areas represent

different data environments, characterised by differences in PM2.5 concentration levels and variability and in the ranges and spatial gradients of model covariates. The observed patterns are therefore more directly attributable to the scale-specific predictor and response distributions, the stability of their relationships across spatial supports, and the loss of local variance and extremes through aggregation. Terrain provides a meaningful contextual classification for these differences but should not itself be interpreted as the sole or direct cause of variation in scale transferability.

These findings imply that Stage 1 downscaling should be interpreted as predictor-supported spatial refinement rather than as a full reconstruction of true 100 m PM2.5 variability. The 1 km CHAP-PM2.5 product contains useful coarse resolution information, and high-resolution predictors can help redistribute this information to a finer grid. However, because the training target is already aggregated and spatially smoothed, the downscaled product remains constrained by the information content of the original product. Therefore, scale transferability should be evaluated explicitly, especially when downscaled products are intended for local exposure assessment or fine-scale environmental analysis.

In this study, the target resolution was close to the absolute resolution of the CHAP-PM2.5 product, limiting the effect of the change of scale, while the simulation experiment highlights the limitations of this approach in terms of the smoothed Stage 1 prediction.

5.2. Resolution Enhancement and Observation-Based Calibration Are Distinct Tasks

The results also show that increasing the spatial resolution does not necessarily improve agreement with ground observations. Stage 1 transfers the coarse-scale CHAP-PM2.5-predictor relationship to the 100 m predictor grid, but it is trained to reproduce CHAP-PM2.5 rather than independent monitoring data. As a result, the Stage 1 products inherit the bias structure and smoothing characteristics of the original product; hence, Stage 1 products showed station-level validation performance similar to the original CHAP-PM2.5 product, despite providing finer spatial detail.

Stage 2 addresses a different problem. By incorporating independent monitoring observations through daily bias correction and residual modelling, it directly targets the discrepancy between gridded estimates and observed PM2.5. The consistent improvements after Stage 2 indicate that observation-based calibration is required when the objective is not only to increase spatial resolution, but also to reduce bias and to recover the information lost in the coarser resolution product as shown by the improved slope in Tables S9 and S10. This distinction is important because a visually detailed PM2.5 map may not reflect the correct range of values. Downscaling and calibration can therefore be treated as related but analytically distinct steps. This goes beyond the method presented by Yang et al. [16].

5.3. Stage-Specific Model Selection Should Balance Accuracy Gains, Spatial Plausibility, and Computational Efficiency

Model selection considered the balance among validation accuracy gains, spatial plausibility, and computational efficiency. Daily bias correction provided a conservative benchmark, improving predictions by adjusting the daily concentration level while preserving the spatial pattern inherited from the Stage 1 product. Its strong performance indicates that day-specific mean bias accounted for a substantial proportion of the disagreement between CHAP-PM2.5 and the provincial monitoring observations (Figure 11).

At Stage 1, TabICL v2 achieved slightly better cross-validation performance than XGBoost. However, the subsequent Stage 2 cross-validation results were nearly identical regardless of whether the XGBoost- or TabICL-based Stage 1 product was used. Given this limited downstream accuracy gain and the substantial difference in computational cost (approximately 1 min for XGBoost compared with 40 min for TabICL v2), XGBoost is suitable for Stage 1 modelling.

The balance differed at Stage 2. XGBoost produced conspicuous spatial artefacts in the residual-correction surfaces, whereas TabICL v2 did not show comparable artefacts and achieved appreciably better validation performance. Although TabICL v2 required approximately ten times as much computation time as XGBoost (10 min vs. 1 min) at this stage, its gains in validation accuracy and lack of artefacts justified its selection for the final Stage 2 residual correction. Details of computation times are given in Section S2.3.

Overall, we recommend using XGBoost for Stage 1 combined with the bias + TabICL v2 for Stage 2. This was the approach used for the 2021–2024 mapping.

5.4. Implications for High-Resolution PM2.5 Product Development

The two-stage framework used in this study provides a rigorous way to separate three questions that are often conflated in PM2.5 downscaling. First, can a coarse-resolution PM2.5 product be refined to a finer grid using covariates? Second, does the refined product reflect the distribution of values from independent ground observations? Third, does the final mapped product remain spatially plausible? The results show that these three objectives do not necessarily improve together. Stage 1 enhanced spatial resolution but did not substantially improve station-level accuracy. Stage 2 improved agreement with monitoring observations and could alter the spatial structure of the final product.

This has two implications. First, high-resolution PM2.5 products should not be evaluated only by their nominal spatial resolution. A 100 m product may still inherit the smoothing and bias structure of its coarse-resolution training target. Second, validation metrics should not be the only basis for product selection. In applications such as exposure assessment, urban environmental analysis, or health-risk estimation [54], spatial artefacts may affect local interpretation even when cross-validation metrics improve. A defensible high-resolution PM2.5 product should therefore balance scale-transfer validity, observation-level accuracy, and spatial plausibility.

5.5. Limitations and Future Work

Several limitations should be noted. First, no spatially continuous 100 m PM2.5 reference surface was available, meaning that the real 1 km-to-100 m transfer could not be directly validated. The simulation experiment provided controlled evidence on fine-scale recoverability; however, it is limited in scope and does not fully reflect hypothesized real PM2.5. Future work could extend this analysis by using an ensemble of synthetic PM2.5 surfaces to evaluate whether downscaling models can consistently recover fine-scale spatial heterogeneity under different controlled simulation settings.

Accordingly, the visual evidence presented in this study should be interpreted as a diagnostic for identifying potential spatial artefacts. Visual inspection can help identify obvious inconsistencies or potential artefacts, but it cannot establish whether individual local PM2.5 gradients are physically realistic.

Second, Stage 2 calibration relied on provincial monitoring stations, which were unevenly distributed and relatively sparse in rugged areas. Terrain-stratified analyses and spatially buffered site-based cross-validation were conducted, emphasizing predictions remain more uncertain in unmonitored rugged regions and other areas far from monitoring stations. Future work could evaluate how distance to monitoring stations (Figure S4) could be combined with the buffer-based cross-validation to provide local accuracy metrics. It could also use additional independent monitoring datasets [12], probability sampling [34], or spatial uncertainty mapping to better quantify prediction uncertainty.

Third, XGBoost and TabICL v2 are both tabular models. Other approaches, such as those based on geostatistics, provide an explicit model of the spatial structure for both downscaling and calibration [55]. However, these require denser datasets than were

available for this study. Further, TabICL v2 introduced computational challenges, especially for repeated cross-validation and daily full-grid mapping. Future work could explore more efficient context selection, spatial modelling and hybrid workflows that combine conservative daily bias correction with flexible residual modelling.

The reported uncertainty remains conditional on the CHAP-PM2.5 input and the downstream fitted models, as uncertainty arising from the original CHAP-PM2.5 processing and its dependence on downstream errors was not explicitly propagated. Although Stage 2 may reduce inherited bias at locations represented by monitoring observations, equivalent correction in unmonitored areas cannot be guaranteed. Future studies could compare the two-stage workflow with a wider range of modelling strategies.

Despite these limitations, this study shows that high-resolution PM2.5 downscaling should be evaluated as more than a resolution-enhancement problem. The results indicate that scale transferability is partial but the agreement with observations is improved by incorporating ground observations in a 2-stage approach. This combined perspective provides a more careful and principled framework for developing and selecting high-resolution PM2.5 products. Finally, a key advantage of our approach is that the 100 m, 1-day predictions may be flexibly aggregated to larger units in space and time as required by the application.

6. Conclusions

This study developed and evaluated a two-stage framework for high-resolution PM2.5 downscaling in Zhejiang Province. In Stage 1, the 1 km CHAP-PM2.5 product was statistically downscaled to 100 m using auxiliary predictors and two tabular learning models, XGBoost and TabICL v2. In Stage 2, ground observations were integrated through daily bias correction and residual calibration using XGBoost and TabICL v2. This design allowed resolution enhancement and observation-based calibration to be evaluated as two distinct modelling tasks.

The results show that the scale-invariance assumption underlying statistical downscaling is only partially valid. The coefficient-side comparison and output-side transfer test showed that PM2.5-predictor relationships and model outputs vary across kilometre-scale spatial supports, indicating that scale invariance should not be assumed for the 1 km-to-100 m transfer. The simulation experiment further suggests that, under controlled synthetic conditions, coarse-resolution training cannot fully recover sub-kilometre heterogeneity. Stage 1 downscaling can introduce predictor-supported fine-scale spatial detail, but it remains constrained by the information content and smoothing characteristics of the original CHAP-PM2.5 product.

Validation against ground observations further shows that increasing spatial resolution alone does not necessarily improve station-level accuracy. The Stage 1 products generated by XGBoost and TabICL v2 showed validation performance similar to the original CHAP-PM2.5 product. Stage 2 corrects for bias and, by incorporating ground-based observations, recovers information that was not available in the coarse resolution product. TabICL v2 residual correction achieved the best results in Stage 2. The use of XGBoost in Stage 1 retains spatial detail and is less computationally demanding than TabICL v2.

The final Stage 2 result is a set of quality-assured daily maps at 100 m resolution, together with 90% prediction intervals for Zhejiang province in China. While extensive validation is provided for 2020, we have also made these maps for the years 2021–2024, using a combination of XGBoost (Stage 1) + bias + TabICL v2 (Stage 2). We have provided the validation results for 2021–2024 as Supplementary Material. While many applications might not require such high-resolution maps, these can be aggregated flexibly in space and time to suit requirements.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs18172876/s1, Supplementary Material-4444879.pdf (detailed implementation and results supporting the main manuscript) and PM2.5_2021-24_validation.xlsx (validation for 2021–2024).

Author Contributions: Conceptualization, Y.J. and N.A.S.H.; methodology, Y.J., E.M.M.Z. and N.A.S.H.; validation, Y.J. and N.A.S.H.; formal analysis, Y.J. and N.A.S.H.; data curation, Y.J.; writing—original draft preparation, Y.J.; writing—review and editing, Y.J., E.M.M.Z. and N.A.S.H.; visualization, Y.J. and E.M.M.Z.; supervision, E.M.M.Z. and N.A.S.H.; funding acquisition, N.A.S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This project is supported by the Ningbo Natural Science Foundation (Project ID 2022J173).

Data Availability Statement: The sources for the data used in this study are provided in the Methods section and in the Supplementary Materials. The interested reader can follow those instructions to obtain the data. The produced PM2.5 maps are available from the authors. The methods used for the cross-validation, scale invariance analysis and downscaling are described in the Supplementary Materials and as R Notebooks (https://zenodo.org/records/21866561 (accessed on 11 August 2026)).

Acknowledgments: This project is supported by Ningbo Natural Science Foundation (Project ID 2022J173) and by a PhD scholarship from the University of Nottingham Ningbo China to Y.J. (23GCT_NH). During the preparation of this manuscript, the authors used ChatGPT (OpenAI, GPT-5.5 Thinking) for assistance with code development and English editing. The authors reviewed and edited all AI-generated outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACAG	Atmospheric Composition Analysis Group
AOD	Aerosol optical depth
BC	Bhattacharyya coefficient
CGIAR-CSI	Consultative Group on International Agricultural Research-Consortium for Spatial Information
CHAP	ChinaHighAirPollutants
CNEMC	China National Environmental Monitoring Centre
CV	Cross-validation
DEM	Digital elevation model
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis v5 (ERA5-Land)
GRC	Global Relief Classes
HEV	Heterogeneity explained variance
LMG	Lindeman, Merenda and Gold relative-importance metric
MAE	Mean absolute error
MAUP	Modifiable areal unit problem
PM2.5	Fine particulate matter (aerodynamic diameter $\leq$ 2.5 μm)
POI	Point of interest
RMSE	Root mean square error
SD	Standard deviation
TabICL v2	Tabular in-context learning model, version 2
UNNC	University of Nottingham Ningbo China
VR	Variance ratio
XGBoost	Extreme gradient boosting

References

1. Cohen, A.J.; Brauer, M.; Burnett, R.; Anderson, H.R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona, L.; Dandona, R.; et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *Lancet* **2017**, *389*, 1907–1918. [CrossRef]
2. Chen, J.; Hoek, G. Long-term exposure to PM and all-cause and cause-specific mortality: A systematic review and meta-analysis. *Environ. Int.* **2020**, *143*, 105974. [CrossRef]
3. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.S.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. *Sci. Total Environ.* **2018**, *636*, 52–60. [CrossRef]
4. van Donkelaar, A.; Martin, R.V.; Spurr, R.J.D.; Burnett, R.T. High-Resolution Satellite-Derived PM2.5 from Optimal Estimation and Geographically Weighted Regression over North America. *Environ. Sci. Technol.* **2015**, *49*, 10482–10491. [CrossRef]
5. Stafoggia, M.; Bellander, T.; Bucci, S.; Davoli, M.; de Hoogh, K.; De' Donato, F.; Gariazzo, C.; Lyapustin, A.; Michelozzi, P.; Renzi, M.; et al. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. *Environ. Int.* **2019**, *124*, 170–179. [CrossRef]
6. Handschuh, J.; Erbertseder, T.; Baier, F. Systematic Evaluation of Four Satellite AOD Datasets for Estimating PM2.5 Using a Random Forest Approach. *Remote Sens.* **2023**, *15*, 2064. [CrossRef]
7. Wang, F.; Yao, S.; Luo, H.; Huang, B. Estimating High-Resolution PM2.5 Concentrations by Fusing Satellite AOD and Smartphone Photographs Using a Convolutional Neural Network and Ensemble Learning. *Remote Sens.* **2022**, *14*, 1515. [CrossRef]
8. Bai, K.; Li, K.; Ma, M.; Li, K.; Li, Z.; Guo, J.; Chang, N.B.; Tan, Z.; Han, D. LGHAP: The Long-term Gap-free High-resolution Air Pollutant concentration dataset, derived via tensor-flow-based multimodal data fusion. *Earth Syst. Sci. Data* **2022**, *14*, 907–927. [CrossRef]
9. Di, Y.; Gao, X.; Liu, H.; Li, B.; Sun, C.; Yuan, Y.; Ni, Y. Accuracy assessment on eight public PM2.5 concentration datasets across China. *Atmos. Environ.* **2024**, *338*, 120799. [CrossRef]
10. Bai, Y.; Wu, L.; Qin, K.; Zhang, Y.; Shen, Y.; Zhou, Y. A Geographically and Temporally Weighted Regression Model for Ground-Level PM2.5 Estimation from Satellite-Derived 500 m Resolution AOD. *Remote Sens.* **2016**, *8*, 262. [CrossRef]
11. Wei, J.; Li, Z.; Cribb, M.; Huang, W.; Xue, W.; Sun, L.; Guo, J.; Peng, Y.; Li, J.; Lyapustin, A.; et al. Improved 1 km resolution PM2.5 estimates across China using enhanced space–time extremely randomized trees. *Atmos. Chem. Phys.* **2020**, *20*, 3273–3289. [CrossRef]
12. Wei, J.; Li, Z.; Lyapustin, A.; Sun, L.; Peng, Y.; Xue, W.; Su, T.; Cribb, M. Reconstructing 1-km-resolution high-quality PM2.5 data records from 2000 to 2018 in China: Spatiotemporal variations and policy implications. *Remote Sens. Environ.* **2021**, *252*, 112136. [CrossRef]
13. Chi, Y.; Zhan, Y.; Wang, K.; Ye, H. Spatial Distribution of Multiple Atmospheric Pollutants in China from 2015 to 2020. *Remote Sens.* **2023**, *15*, 5705. [CrossRef]
14. Li, L. A Robust Deep Learning Approach for Spatiotemporal Estimation of Satellite AOD and PM2.5. *Remote Sens.* **2020**, *12*, 264. [CrossRef]
15. Yu, X.; Xi, M.; Wu, L.; Zheng, H. Spatiotemporal Weighted for Improving the Satellite-Based High-Resolution Ground PM2.5 Estimation Using the Light Gradient Boosting Machine. *Remote Sens.* **2023**, *15*, 4104. [CrossRef]
16. Yang, Q.; Yuan, Q.; Li, T.; Yue, L. Mapping PM2.5 concentration at high resolution using a cascade random forest based downscaling model: Evaluation and application. *J. Clean. Prod.* **2020**, *277*, 123887. [CrossRef]
17. Liu, S.; Zhao, X.; Zhang, F.; Qiu, A.; Chen, L.; Huang, J.; Chen, S.; Zhang, S. Spatial Downscaling of NPP-VIIRS Nighttime Light Data Using Multiscale Geographically Weighted Regression and Multi-Source Variables. *Remote Sens.* **2022**, *14*, 6400. [CrossRef]
18. Jiang, H.; Shen, H.; Li, H.; Lei, F.; Gan, W.; Zhang, L. Evaluation of Multiple Downscaled Microwave Soil Moisture Products over the Central Tibetan Plateau. *Remote Sens.* **2017**, *9*, 402. [CrossRef]
19. Zhou, J.; Liu, S.; Li, M.; Zhan, W.; Xu, Z.; Xu, T. Quantification of the Scale Effect in Downscaling Remotely Sensed Land Surface Temperature. *Remote Sens.* **2016**, *8*, 975. [CrossRef]
20. Guo, Z.; Cheng, L.; Chang, L.; Li, S.; Li, Y. A Modified Method for Reducing the Scale Effect in Land Surface Temperature Downscaling at 10 m Resolution. *Remote Sens.* **2024**, *16*, 3908. [CrossRef]
21. Xu, S.; Zhao, Q.; Yin, K.; He, G.; Zhang, Z.; Wang, G.; Wen, M.; Zhang, N. Spatial Downscaling of Land Surface Temperature Based on a Multi-Factor Geographically Weighted Machine Learning Model. *Remote Sens.* **2021**, *13*, 1186. [CrossRef]
22. Liang, M.; Zhang, L.; Wu, S.; Zhu, Y.; Dai, Z.; Wang, Y.; Qi, J.; Chen, Y.; Du, Z. A High-Resolution Land Surface Temperature Downscaling Method Based on Geographically Weighted Neural Network Regression. *Remote Sens.* **2023**, *15*, 1740. [CrossRef]
23. Pu, R.; Bonafoni, S. Reducing Scaling Effect on Downscaled Land Surface Temperature Maps in Heterogenous Urban Environments. *Remote Sens.* **2021**, *13*, 5044. [CrossRef]
24. Dinkelacker, B.T.; Garcia Rivera, P.; Marshall, J.D.; Adams, P.J.; Pandis, S.N. High-resolution downscaling of source resolved PM2.5 predictions using machine learning models. *Atmos. Environ.* **2023**, *310*, 119967. [CrossRef]

25. Nguyen, P.D.M.; Phan, A.H.; Ngo, T.X.; Ho, B.Q.; Pham, T.V.; Nguyen, T.T.N. Mapping of high-resolution daily particulate matter (PM2.5) concentration at the city level through a machine learning-based downscaling approach. *Environ. Monit. Assess.* **2025**, *197*, 94. [CrossRef]
26. Shetty, S.; Hamer, P.D.; Stebel, K.; Kylling, A.; Hassani, A.; Berntsen, T.K.; Schneider, P. Daily high-resolution surface PM2.5 estimation over Europe by ML-based downscaling of the CAMS regional forecast. *Environ. Res.* **2025**, *264*, 120363. [CrossRef]
27. de Hoogh, K.; Héritier, H.; Stafoggia, M.; Künzli, N.; Kloog, I. Modelling daily PM2.5 concentrations at high spatio-temporal resolution across Switzerland. *Environ. Pollut.* **2018**, *233*, 1147–1154. [CrossRef]
28. Tian, J.; Lu, H.; Yang, K.; Qin, J.; Zhao, L.; Zhou, J.; Xu, Y. Spatiotemporal Downscaling of SMAP Soil Moisture with an Integrated Machine Learning and Land Data Assimilation Framework. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 21722–21734. [CrossRef]
29. Hu, C.; Xie, P.; Hu, Z.; Li, A.; Feng, H. The Downscaled GOME-2 SIF Based on Machine Learning Enhances the Correlation with Ecosystem Productivity. *Remote Sens.* **2025**, *17*, 2642. [CrossRef]
30. Openshaw, S. *The Modifiable Areal Unit Problem*; Geo Books: Norwich, UK, 1984.
31. Shang, K.; He, Z.; Sun, Y.; Wang, L.; Shou, Y. Scale and Zoning Effects in the Association between Built Environment Indicators and PM2.5/PM10 Concentrations: A case study of Gulou District, Nanjing. *Int. Rev. Spat. Plan. Sustain. Dev.* **2025**, *13*, 4–21. [CrossRef]
32. Yang, D.; Meng, F.; Liu, Y.; Dong, G.; Lu, D. Scale Effects and Regional Disparities of Land Use in Influencing PM2.5 Concentrations: A Case Study in the Zhengzhou Metropolitan Area, China. *Land* **2022**, *11*, 1538. [CrossRef]
33. Hamm, N.A.S.; Atkinson, P.M.; Milton, E.J. A per-pixel, non-stationary mixed model for empirical line atmospheric correction in remote sensing. *Remote Sens. Environ.* **2012**, *124*, 666–678. [CrossRef]
34. Wadoux, A.M.J.C.; Heuvelink, G.B.M.; de Bruin, S.; Brus, D.J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* **2021**, *457*, 109692. [CrossRef]
35. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
36. de Bruin, S.; Brus, D.J.; Heuvelink, G.B.M.; van Ebbenhorst Tengbergen, T.; Wadoux, A.M.J.C. Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecol. Inform.* **2022**, *69*, 101665. [CrossRef]
37. Pohjankukka, J.; Pahikkala, T.; Nevalainen, P.; Heikkonen, J. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 2001–2019. [CrossRef]
38. Just, A.C.; Arfer, K.B.; Rush, J.; Dorman, M.; Shtein, A.; Lyapustin, A.; Kloog, I. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. *Atmos. Environ.* **2020**, *239*, 117649. [CrossRef]
39. Li, T.; Shen, H.; Zeng, C.; Yuan, Q. A Validation Approach Considering the Uneven Distribution of Ground Stations for Satellite-Based PM2.5 Estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1312–1321. [CrossRef]
40. Yang, X.; Li, S.; Ma, J.; Chen, Y.; Zhou, X.; Li, F.; Xiong, L.; Zhou, C.; Tang, G.; Meadows, M.E. A typology of global relief classes derived from digital elevation models at 1 arcsec resolution. *Earth Syst. Sci. Data* **2025**, *17*, 4331–4350. [CrossRef]
41. Chen, Z.; Yu, B.; Yang, C.; Zhou, Y.; Yao, S.; Qian, X.; Wang, C.; Wu, B.; Wu, J. *The Global NPP-VIIRS-Like Nighttime Light Data (Version 2) for 1992–2025*; Harvard Dataverse: Cambridge, MA, USA, 2020; Volume 10. [CrossRef]
42. Wu, W.-B.; Ma, J.; Banzhaf, E.; Meadows, M.E.; Yu, Z.-W.; Guo, F.-X.; Sengupta, D.; Cai, X.-X.; Zhao, B. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. *Remote Sens. Environ.* **2023**, *291*, 113578. [CrossRef]
43. Chen, Y.; Xu, C.; Ge, Y.; Zhang, X.; Zhou, Y. A 100-m gridded population dataset of China's seventh census using ensemble learning and big geospatial data. *Earth Syst. Sci. Data* **2024**, *16*, 3705–3718. [CrossRef]
44. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016*; ACM: New York, NY, USA, 2016; pp. 785–794.
45. Qu, J.; Holzmüller, D.; Varoquaux, G.; Le Morvan, M. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv* **2026**, arXiv:2602.11139. [CrossRef]
46. Qu, J.; Holzmüller, D.; Varoquaux, G.; Le Morvan, M. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. In Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, Vancouver, BC, Canada, 13–19 July 2025; pp. 50817–50847.
47. Xu, M.; Li, Q.; Tian, S.; Kuang, S.; Li, T. Reconstruction of Land Surface Temperature Based on EATC Constraints and Spatially Adaptive Residual Correction: A Case Study of the Qinghai–Tibet Engineering Corridor. *Remote Sens.* **2026**, *18*, 2254. [CrossRef]
48. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **1967**, *15*, 52–60. [CrossRef]

49. Lindeman, R.H.; Merenda, P.F.; Gold, R.Z. *Introduction to Bivariate and Multivariate Analysis*; Scott, Foresman and Company: Glenview, IL, USA, 1980.
50. Romano, Y.; Patterson, E.; Candes, E. Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3543–3553.
51. He, Q.; Huang, B. Satellite-based mapping of daily high-resolution ground PM2.5 in China via space-time regression modeling. *Remote Sens. Environ.* **2018**, *206*, 72–83. [CrossRef]
52. Hutengs, C.; Vohland, M. Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sens. Environ.* **2016**, *178*, 127–141. [CrossRef]
53. Hamm, N.A.S. An Entropy-based Method to Evaluate the Appropriate Spatial Resolution. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2024**, *48*, 227–232. [CrossRef]
54. Punger, E.M.; West, J.J. The effect of grid resolution on estimates of the burden of ozone and fine particulate matter on premature mortality in the USA. *Air Qual. Atmos. Health* **2013**, *6*, 563–573. [CrossRef]
55. Atkinson, P.M. Downscaling in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *22*, 106–114. [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.