Article

Not peer-reviewed version

Spatially Continuous PM10 Exposure Mapping in the Campania Region Using a Land-Use Random Forest Model: Integration of Monitoring Data, Geographic Predictors, ERA5 Reanalysis, and CHIMERE Model Output

Elena Chianese and Angelo Riccio *

Posted Date: 17 April 2026

doi: 10.20944/preprints202604.1265.v1

Keywords: air quality; PM_{10}; Land-Use Random Forest; LURF; exposure assessment; campania region; ERA5; CHIMERE; MODIS AOD; spatial mapping; dust transport

 doi:10.20944/preprints202604.1265.v1

Article

Spatially Continuous PM10 Exposure Mapping in the Campania Region Using a Land-Use Random Forest Model: Integration of Monitoring Data, Geographic Predictors, ERA5 Reanalysis, and CHIMERE Model Output

Elena Chianese and Angelo Riccio *

Department of Science and Technology, Parthenope University of Naples, Centro Direzionale, Isola C4, 80143 Naples, Italy
* Correspondence: angelo.riccio@uniparthenope.it; Tel.: +39-081-547-6613

Abstract

In this study we develop a Land-Use Random Forest (LURF) model for the Campania Region (southern Italy) that combines 2022 daily PM_{10} observations from 13 quality-controlled ARPA Campania stations with a rich set of spatial predictors to produce daily concentration maps at 1000 m $\times$ 1000 m resolution, from which annual statistics (mean, percentiles, and exceedances) are derived through temporal aggregation. The predictor space includes resident population, land-cover and imperviousness indicators, road-network metrics derived from OpenStreetMap, meteorological fields from the ERA5 reanalysis, satellite aerosol optical depth (AOD) from MODIS Terra and Aqua—scaled by ERA5 boundary-layer height (AOD/pbl)—daily mean PM_{10} from a nested CHIMERE simulation, and a binary categorical predictor (`IdDust`) flagging days affected by Saharan dust transport events. The hyperparameters for the LURF model are selected via a nested inner grid search; generalisation performance is assessed through a spatially aware leave-location-out cross-validation (LLO-CV) scheme, which prevents optimistic bias arising from spatial autocorrelation among neighbouring stations. Under LLO-CV, the LURF achieves $R^2 = 0.54$, RMSE $= 11.0$ μg m^{-3}, and MAE $= 8.0$ μg m^{-3}, against $R^2 = -1.11$, RMSE $= 23.6$ μg m^{-3}, and MAE $= 19.1$ μg m^{-3} for the raw CHIMERE output evaluated on the same observations. The inclusion of `IdDust` as a categorical covariate allows the Random Forest to partition the training distribution between dusty and non-dusty regimes, improving the representation of episodic high-PM_{10} events and reducing systematic underestimation at the upper tail of the concentration distribution. CTM-derived PM_{10} and ERA5 boundary-layer and pressure fields emerge as the dominant predictors, collectively accounting for the majority of explained variability, while `IdDust` ranks among the physically interpretable secondary predictors. The 1000 m maps highlight marked urban–rural contrasts, resolving hotspots in the Naples metropolitan area and along major motorway corridors that remain unresolved at typical CTM grid spacings. By embedding physically based CTM output, satellite aerosol diagnostics, and dust-event classification within a flexible machine-learning framework, the proposed approach offers a low-cost, operationally tractable tool for high-resolution PM_{10} exposure assessment in regions characterised by complex terrain and heterogeneous emission sources.

Keywords: air quality; PM_{10}; Land-Use Random Forest; LURF; exposure assessment; campania region; ERA5; CHIMERE; MODIS AOD; spatial mapping; dust transport

1. Introduction

Particulate matter remains one of the most consequential air pollutants for public health across European and global contexts [1,2]. Chronic exposure to elevated PM_{10} concentrations is linked to a

 doi:10.20944/preprints202604.1265.v1

broad spectrum of adverse outcomes, including cardiovascular and respiratory disease, premature mortality, and impaired agricultural productivity, with burdens that fall disproportionately on residents of densely built city centres and other high-exposure microenvironments [1–3]. Generating reliable, spatially resolved concentration estimates is therefore a prerequisite for epidemiological surveillance, health-impact assessment, and the implementation of the revised EU Air Quality Directive, which progressively tightens limit values towards World Health Organization guideline levels while demanding more granular exposure information [2,4]. The core challenge is that regulatory monitoring networks, although providing high-quality measurements at high temporal resolution, sample only a handful of locations per urban area. This sparse spatial footprint is insufficient to characterise the steep intra-urban gradients that develop downwind of traffic corridors, junctions, ports, or industrial zones—precisely the environments where human exposure is highest [5–7].

Land-use regression (LUR) modelling was originally conceived to bridge the gap between discrete observations and spatially continuous concentration fields [8]. The approach links pollutant measurements at monitoring sites with geospatial covariates such as population density, land-cover classes, road proximity, and meteorological context, and then projects these relationships across a fine-resolution grid. Over the past two decades, LUR methods have been widely adopted in multi-city European studies, most notably within the ESCAPE project [9,10], for mapping PM_{10}, $PM_{2.5}$, black carbon, and NO_2 at spatial resolutions suitable for cohort-based health analyses [8,11]. Building on these foundations, more recent developments have integrated satellite-derived aerosol optical depth (AOD) products and chemistry transport models (CTM) output into the LUR predictor space to improve coverage in regions with sparse ground monitoring and to better constrain background aerosol fields [12,13].

The Random Forest extension of LUR, hereafter referred to as Land-Use Random Forest (LURF), can accommodate non-linear interactions and higher-order dependencies among predictors that classical linear LUR cannot capture. In this framework, ensembles of regression trees are trained on high-dimensional spatial covariate vectors and pollutant observations, and the resulting model is used to generate gridded concentration fields at resolutions much finer than the underlying monitoring network. LURF and related ensemble-tree methods have been progressively adopted for mapping NO_2 and particulate matter across European settings [11,14–16], offering improved predictive performance and robustness relative to traditional linear formulations, particularly under the small-sample conditions typical of regulatory networks. In the Naples metropolitan area, for example, Chianese and Riccio [17] demonstrated the utility of the LURF approach for long-term NO_2 exposure assessment by combining citizen-science passive samplers with the ARPA Campania reference network, achieving annual mean predictions with an uncertainty compatible with European Commission recommendations [4]. The present work builds on this methodological foundation and extends it to PM_{10} at the full regional scale of Campania.

Mapping PM_{10} poses particular challenges because this pollutant is a chemically and physically heterogeneous mixture. Road traffic emissions, brake and tyre wear, carbonaceous combustion products, marine sea salt, crustal material, and secondary inorganic aerosols each exhibit distinct spatial footprints and contrasting seasonal dynamics. Source apportionment studies in Campania have documented the dominance of road traffic within the Naples urban core, while port and maritime emission signatures are more prominent at coastal stations [6,7,18]. Emission factor campaigns in Naples have further shown that even the inorganic ion fraction of PM—nitrate, sulphate, ammonium—retains spatial imprints linked to traffic intensity at the road-segment scale [6]. Collectively, these findings motivate the adoption of a modelling architecture capable of resolving multi-source spatial heterogeneity: this is precisely the role that can be effectively addressed by a LURF ensemble. Among the episodic sources that can dominate daily PM_{10} in the Mediterranean basin, long-range transport of Saharan mineral dust is particularly relevant: it can elevate ground-level concentrations by several tens of $\mu g\,m^{-3}$ within hours, affecting regulatory exceedance statistics and masking the local emission signal that the LURF is designed to resolve [19,20]. Explicitly flagging dust-affected days via the binary

 doi:10.20944/preprints202604.1265.v1

indicator `IdDust` allows the Random Forest to learn separate concentration–predictor relationships for dusty and non-dusty regimes, thereby reducing systematic underestimation on episodic high-PM_{10} days.

The LURF framework adopted here can be viewed as the spatial counterpart of multi-layer perceptron (MLP) networks previously applied to air-quality time series in Campania [21]. Rather than relying on temporal windows of past concentrations, it exploits the multivariate spatial structure embedded in land-use, road-network, meteorological, satellite, and CTM-derived predictors. The inclusion of meteorological data enriches the predictor space with information on boundary-layer dynamics, advection regimes, and moisture profiles that influence PM_{10} dilution and transport across time scales ranging from hours to seasons [17,22].

The novelty of this study lies in the explicit integration of high-resolution CTM output, satellite-derived aerosol indicators, and ERA5 reanalysis within a unified Random Forest framework, combined with a spatially explicit validation strategy and the derivation of probabilistic exposure metrics at regional scale. This hybrid use of CTM output is conceptually aligned with earlier European efforts that integrated CTM and satellite information into land-use models for large-scale PM and NO_2 mapping [12,13], as well as with the growing body of work combining meteorological data, satellite AOD, and machine learning for exposure estimation [22,23].

The paper is organised as follows. Section 2 describes the study area, monitoring data, and spatial predictor datasets, including the CTM configuration. Section 3 details the LURF architecture, predictor extraction, and cross-validation scheme. Results are presented in Section 4, which covers model performance, predictor importance, regional PM_{10} patterns, and comparison with CHIMERE. Section 5 discusses the findings in the context of the LUR/LURF literature, highlights innovative aspects of the LURF–CHIMERE framework, and explores implications for exposure and health-impact assessment, as well as limitations and future perspectives. Conclusions are summarised in Section 6.

2. Study Area and Data

2.1. Study Domain

The Campania Region (see Figure 1) lies along the eastern flank of the Tyrrhenian Sea in southern Italy and covers approximately 13 600 km^2 between the Apennine mountain chain to the north-east and the Gulf of Naples to the south-west. Its complex orography gives rise to a distinctive interplay between sea-breeze circulation and orographic channelling that strongly influences the accumulation and dispersion of primary and secondary aerosols [24]. The Naples metropolitan area—home to roughly 3.1 million inhabitants and among the most densely urbanised territories in southern Europe—constitutes the dominant PM source region within this domain. Surrounding municipalities contribute additional industrial, agricultural, and maritime emissions, making Campania an ideal, albeit demanding, test bed for spatially continuous PM exposure modelling.

 doi:10.20944/preprints202604.1265.v1

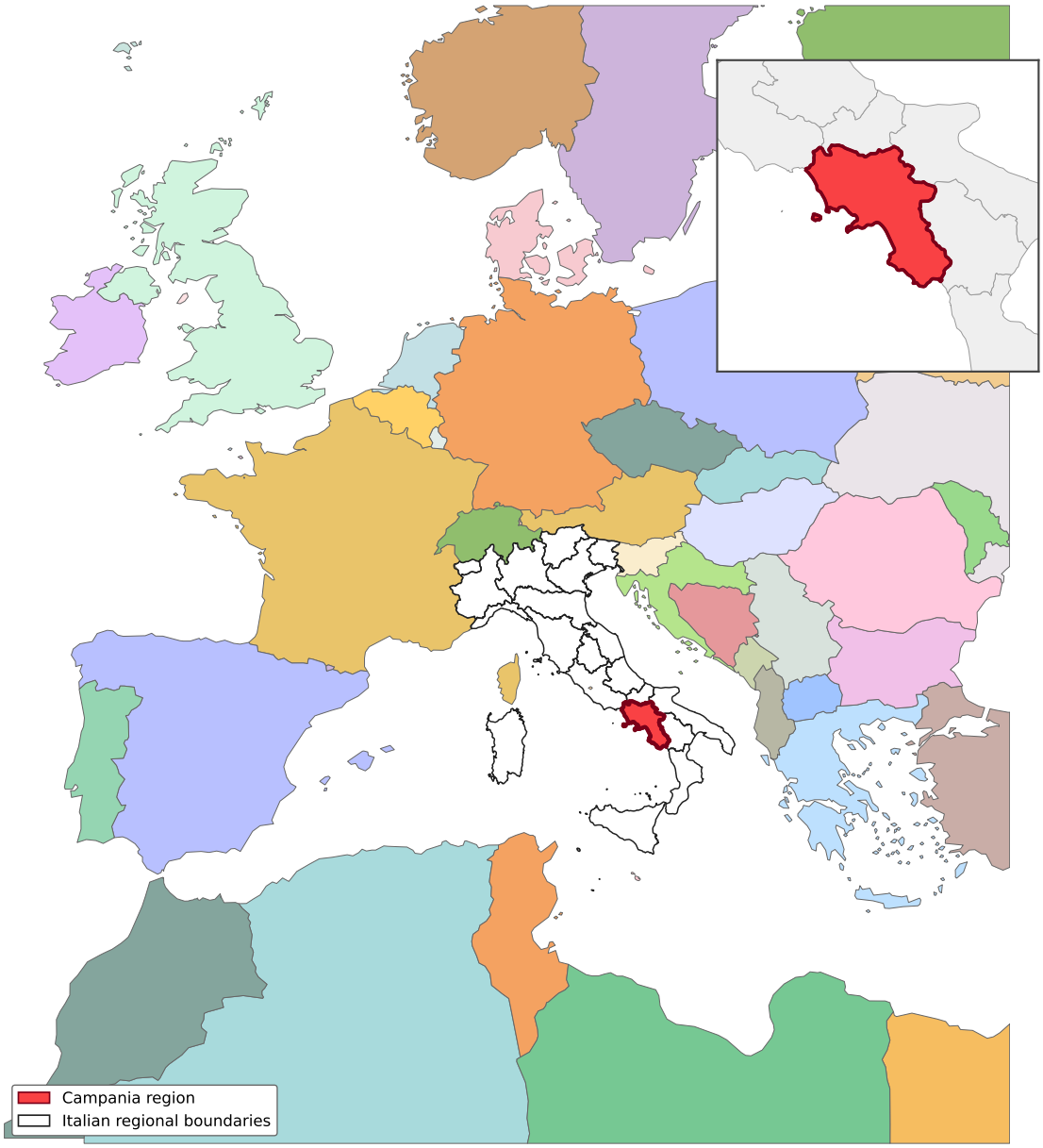

Figure 1. Map of European countries and Italian regions, with the Campania region outlined in bold black and highlighted in the upper-right inset.

2.2. Air Quality Monitoring Data

Observed daily mean PM_{10} concentrations were retrieved from the ARPA Campania regulatory monitoring network (see Figure 2) for the year 2022 [25]. The network comprises 42 stations spanning urban, suburban, traffic, industrial, and rural-background typologies, classified in accordance with the European Air Quality Directive (2008/50/EC); however, only 13 provided quality-controlled PM_{10} records and were retained in this study. The remaining 29 stations either do not measure PM_{10} (monitoring NO_2, O_3, or $PM_{2.5}$ only) or failed the quality-control threshold of at least 75% valid daily observations over the 2022 reference period [4], which is the minimum data-capture requirement prescribed by the EU Air Quality Directive for annual statistics. The 13 retained stations include 8 traffic, 3 industrial area, and 2 background site, providing a reasonable typological spread for training the LURF model across the range of exposure environments present in Campania. It is important to note that the limited number of monitoring stations (n = 13) may constrain the statistical representativeness of the training dataset, particularly in capturing the full range of spatial heterogeneity across rural and coastal environments. This limitation is partially mitigated by the inclusion of physically-based

 doi:10.20944/preprints202604.1265.v1

predictors (e.g., CTM output and meteorological fields), which provide additional spatial structure beyond the monitoring network.

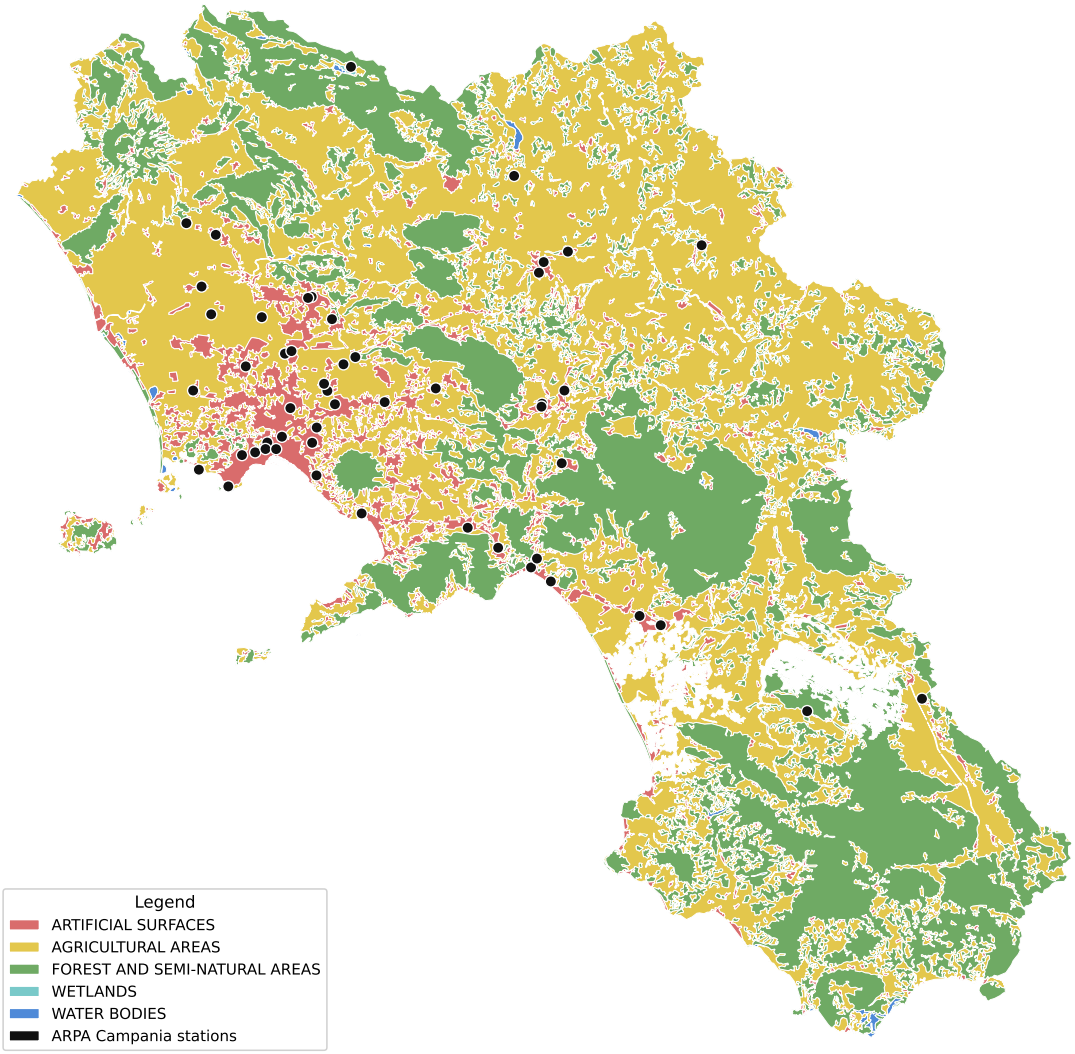

Figure 2. Locations of ARPA Campania monitoring stations (black dots), overlaid on the CORINE CLC map (level 1). The stations are unevenly distributed, with most concentrated in urban areas.

2.3. Spatial Predictors and ERA5 Reanalysis

Several categories of spatial predictors were assembled for each monitoring station within circular buffer radii of 20, 50, 100, 500, and 1000 m. Land-cover information was drawn from the Corine Land Cover 2018 layer (CLC 2018) [26], with individual classes grouped into three broad categories: urban fabric, agriculture and natural vegetation, and water bodies. Built-up density and imperviousness fraction were extracted from the Copernicus High Resolution Layers (HRL) at 10 m resolution [26]. Resident population counts were obtained from the Italian National Institute of Statistics (ISTAT) and disaggregated by census area. Road length and density by class were derived from the OpenStreetMap network (downloaded January 2026); primary, secondary, and motorway classes were retained separately to reflect their contrasting emission profiles [6]. Distance to the nearest motorway centreline was also included as a geometric predictor.

Meteorological predictors—2 m air temperature (T_{2m}), dew-point temperature (T_{dp}), surface pressure (p_s), 10 m zonal (u_{10}) and meridional (v_{10}) wind components, and planetary boundary-layer height (PBLH)—were derived from ERA5 global reanalysis at 0.25° ×0.25° horizontal resolution (approximately 28 km over the Mediterranean).

 doi:10.20944/preprints202604.1265.v1

2.4. Satellite Aerosol Optical Depth

Columnar AOD at 550 nm was retrieved from the Collection 6.1 MCD19A2 product (MAIAC Land Aerosol Optical Depth). Daily AOD values were quality-screened to retain only high-confidence retrievals and subsequently spatially averaged over a $0.1^\circ \times 0.1^\circ$ grid centred on each monitoring station. Daily mean AOD, combining morning Terra and afternoon Aqua overpasses, was then computed for each station and used as a predictor variable. The dual-instrument configuration partially reduces diurnal sampling bias and provides a more representative estimate of daily aerosol loading, particularly over the frequently cloud-covered Campania coastline.

In addition to raw AOD, a derived variable was constructed by scaling AOD with the planetary boundary-layer height (PBLH) obtained from ERA5, hereafter referred to as 'AOD/pbl'. This variable provides a proxy for near-surface aerosol concentration by accounting for vertical dilution processes within the boundary layer. While AOD represents the total columnar aerosol load, its relationship with surface PM_{10} depends strongly on the mixing depth of the atmosphere. For a given columnar aerosol burden, a deeper boundary layer implies greater vertical dispersion and thus lower surface concentrations, whereas shallow boundary-layer conditions favour pollutant accumulation near the ground. By normalising AOD with PBLH, the AOD/pbl indicator effectively incorporates this vertical mixing effect, enhancing the physical interpretability of satellite-derived aerosol information and improving its relevance as a predictor of ground-level PM_{10} variability. Similar approaches have been widely adopted in the literature to improve the linkage between columnar aerosol properties and ground-level concentrations, particularly under varying atmospheric stability conditions [27–30].

2.5. CHIMERE Chemical-Transport Model Output

Gridded PM_{10} concentration fields were extracted from a CHIMERE [31,32] CTM simulation configured with a three-domain nested architecture and run for the full year 2022. The European parent domain was discretised at 25 km horizontal resolution, encompassing the entire continent, while nested Italian and Campania domains were run at 5 km and 1 km resolution, respectively. The innermost domain covered the entire Campania region using a 390×382 grid-cell configuration, fully encompassing the ARPA Campania monitoring network. All domains shared 14 terrain-following vertical layers, starting at 20 m above ground level, with layer thickness increasing with altitude. Meteorological inputs were provided by the Weather Research and Forecasting (WRF) model, driven by NCEP initial and boundary conditions at $0.25^\circ \times 0.25^\circ$ resolution [33]. Anthropogenic and natural emissions for the European domain were derived from the CAMS-REG-ANT-v8.1 inventory [34], while emissions for the Italian and Campania nests were refined using high-resolution spatial proxies.

A multi-step downscaling procedure was implemented to redistribute emissions onto the target model grid. The downscaling procedure comprised two main stages: (i) sectoral mapping and (ii) spatial allocation. Emissions reported according to the NFR (Nomenclature For Reporting) classification were mapped onto SNAP (Selected Nomenclature for Air Pollution) sectors to enable sector-specific temporal and chemical profiles. Where a one-to-many correspondence existed between NFR and SNAP categories, emissions were split using activity-based weighting factors derived from auxiliary datasets (for example fuel consumption, industrial statistics, or emission factor ratios). Spatial disaggregation was then performed using high-resolution proxy datasets. Population density maps were used for residential combustion (SNAP 2), road-network density for traffic emissions (SNAP 7), land-use data for agricultural sources (SNAP 10), and industrial registries or point-source databases for large combustion and industrial sectors (SNAP 3 and SNAP 4). Emissions were allocated to each grid cell according to

$$E_i = E_{\text{tot}} \cdot \frac{P_i}{\sum_j P_j}, \tag{1}$$

where E_i is the emission assigned to grid cell i, E_{tot} is the total emission for a given sector and pollutant, and P_i is the value of the proxy field in cell i. The resulting emissions were formatted into NetCDF files compatible with the CHIMERE emission pre-processor (EMISURF).

 doi:10.20944/preprints202604.1265.v1

The full modelling chain was run on a distributed HPC infrastructure using in-house tools [35–37].

2.6. Identification of Saharan Dust Days

Mineral dust transported from the Saharan Desert constitutes a recurrent and scientifically well-documented episodic perturbation to PM_{10} levels across the Mediterranean basin [19]. To represent this source explicitly in the LURF framework, a binary categorical variable `IdDust` was constructed for each day (s).

A day is labelled `IdDust = 1` (dust-affected) only when all three of the following criteria are satisfied simultaneously:

1. **Coarse-fraction dominance:** $\frac{PM_{10} - PM_{2.5}}{PM_{10}} > 0.5$, indicating that super-micron crustal particles account for more than half of the total suspended aerosol mass. This threshold is consistent with the aerosol-type discrimination approach developed for the Mediterranean by Barnaba and Gobbi [38], who showed that Saharan dust events are characterised by a marked shift towards coarse-mode aerosol relative to background continental or maritime conditions.
2. **Elevated AOD:** daily mean AOD at 550 nm > 0.3, derived from the MODIS Terra/Aqua MAIAC product described in Section 2.4. An AOD of 0.3 represents a conservative upper bound for the background Mediterranean aerosol climatology and has been widely used as an indicator of enhanced aerosol loading during dust intrusion episodes [38].
3. **PM_{10} exceedance:** observed daily mean PM_{10} at station s exceeds the station-specific annual 75th percentile $Q_{75}(\mathrm{PM}_{10}, s)$, ensuring that the flag is restricted to genuine concentration peaks rather than moderate aerosol-loading situations in which crustal background is non-negligible but not dominant.

Days satisfying the three criteria were further cross-checked against the operational ground-level dust surface concentration maps produced by the WMO Barcelona Dust Regional Center (BDRC; http://dust.aemet.es), operated jointly by the Agencia Estatal de Meteorología (AEMET) and the Barcelona Supercomputing Center (BSC) under WMO designation [20]. The BDRC provides daily dust forecasts and analysis maps at $0.5^\circ \times 0.5^\circ$ resolution for the North African, Mediterranean, and European domain; agreement between the derived `IdDust` flag and the BDRC dust surface concentration fields (daily maximum $> 50\,\mu\mathrm{g\,m^{-3}}$ over Campania) was required to confirm each candidate dust episode.

$PM_{2.5}$ data used in criterion (1) were available from a subset of the ARPA Campania network for the year 2022. At stations without co-located $PM_{2.5}$ measurements, the coarse-fraction criterion was evaluated using the network-median $PM_{2.5}/PM_{10}$ ratio from stations where both species were measured on the same day, providing a conservative estimate of the coarse fraction. For stations and days where this imputation could not be applied with sufficient confidence (fewer than five co-measured station pairs on a given day), the `IdDust` flag was set to missing and that record was excluded from training.

3. Methods

3.1. Land-Use Random Forest Framework

The LURF model follows the framework validated for long-term NO_2 mapping in Naples [17] and is adapted here for regional daily PM_{10} prediction. The modelling unit is the station–day pair (s, t), where s indexes the monitoring location and t the calendar day: each training record comprises a predictor vector $\mathbf{x}_{s,t}$—including daily-resolved ERA5 meteorological fields, CHIMERE PM_{10}, MODIS AOD when available, the binary categorical variable `IdDust`, and time-invariant land-use and road-network descriptors—and the corresponding observed daily mean PM_{10} concentration $y_{s,t}$. The resulting `IdDust` indicator was encoded as an integer binary variable ($0 =$ non-dusty, $1 =$ dusty) and appended to the predictor vector $\mathbf{x}_{s,t}$.

 doi:10.20944/preprints202604.1265.v1

3.2. Predictor Selection and Hyperparameter Optimisation

Model generalisability was assessed using a spatially structured cross-validation design defined at the station level and applied to the station–day dataset; thus, when station (s) was assigned to the test fold, all associated daily observations (s,t) were excluded from model fitting and used exclusively for validation, following a leave-location-out cross-validation (LLO-CV) scheme [39,40]. This approach is consistent with recent recommendations for spatially structured validation in machine-learning models applied to environmental data [41,42].

Hyperparameters were optimised within a nested cross-validation framework. For each outer training set, an inner station-based cross-validation was used to perform a grid search over the hyperparameters, namely the number of predictor variables randomly sampled at each split (*mtry*), the minimum number of observations in terminal nodes (*min.node.size*), and the number of trees in the ensemble (*num.trees*). The final model was then refitted on the full outer training set and evaluated on the corresponding held-out stations. Overall performance was quantified by aggregating predictions across all outer folds. The grid search was conducted over the following sets: $mtry = \{2,5,10\}$, $min.node.size = \{3,5,10\}$ and $num.trees = \{500,800,1000\}$. Candidate configurations were ranked using the root-mean-square error (RMSE), and the optimal set was selected by minimising RMSE. The most frequently selected hyperparameters across outer folds were $mtry = 5$, $min.node.size = 3$ and $num.trees = 500$.

3.3. Spatial Prediction and Uncertainty Mapping

Following hyperparameter selection, the model was refitted on the complete calibration dataset and applied to a 1000 m prediction grid covering Campania. For each grid cell g and day t, the mean prediction across the decision trees provides the daily PM_{10} estimate $\hat{y}_{g,t}$. Annual statistics are then computed as temporal aggregates of daily predictions at each grid cell. In particular, the annual mean PM_{10} concentration at grid cell g is defined as:

$$\hat{Y}_g^{\text{annual}} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_{g,t}$$

where T is the number of valid prediction days. Additional annual indicators (e.g., percentiles or exceedance counts) are derived consistently from the daily prediction time series $\{\hat{y}_{g,t}\}_{t=1}^{T}$.

The exceedance probability at each grid cell g is defined as the fraction of the $B = 500$ individual tree predictions whose temporally aggregated annual mean exceeds a regulatory threshold τ:

$$\hat{P}_g^{\text{exc}}(\tau) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left[\frac{1}{T}\sum_{t=1}^{T}\hat{y}_{g,t}^{(b)} > \tau\right], \tag{2}$$

where $\hat{y}_{g,t}^{(b)}$ is the daily PM_{10} estimate of tree b at grid cell g on day t, and $\mathbf{1}[\cdot]$ is the indicator function. This formulation is applied to two regulatory thresholds: $\tau_{40} = 40\,\mu\text{g}\,\text{m}^{-3}$, corresponding to the current EU annual limit value under Directive 2008/50/EC, and $\tau_{20} = 20\,\mu\text{g}\,\text{m}^{-3}$, corresponding to the stricter limit value introduced by Directive (EU) 2024/2881 to be attained by 1 January 2030 [43]. This formulation propagates ensemble dispersion directly from the daily prediction scale to the annual regulatory indicator, without assuming a parametric distribution for the tree-level predictions. It is, in this sense, a non-parametric analogue of probabilistic forecasting based on ensemble spread [44]. It should be noted, however, that daily predictions from the same tree share ERA5 and CHIMERE inputs across consecutive days, so the inter-tree spread on the annual mean primarily reflects *structural* uncertainty arising from the ensemble composition (bootstrap resampling, random feature selection) rather than the *sampling* uncertainty associated with temporal variability. The resulting exceedance probabilities should therefore be interpreted as a relative measure of prediction confidence across grid cells, rather than as frequentist exceedance probabilities in a strict climatological sense.

 doi:10.20944/preprints202604.1265.v1

4. Results

4.1. Cross-Validation Performance and Error Structure

The predictive performance of the models was evaluated against observed PM_{10} concentrations (from ARPA Campania daily mean in 2022) using standard error metrics and a decomposition of the mean squared error (MSE) into systematic and random components. The coefficient of determination (R^2) was computed in a predictive framework, comparing model outputs against observations.

The CTM output exhibits poor agreement with observations, with $R^2 = -1.11$, RMSE $= 23.6\ \mu g\ m^{-3}$, and MAE $= 19.1\ \mu g\ m^{-3}$ (see Figure 3). Here R^2 is computed as the predictive coefficient of determination, $R^2 = 1 - SS_{res}/SS_{tot}$, where SS_{tot} is the total variance of the observations about their mean; a negative value indicates that the model performs worse than a simple mean-based predictor. Decomposition of the MSE reveals that a substantial fraction of the error is systematic, as evidenced by the shallow slope of the regression line and the strong underestimation of high PM_{10} concentrations. This indicates that the CTM fails to reproduce both the magnitude and variability of observed concentrations, likely due to unresolved local emission sources and the spatial smoothing inherent to kilometre-scale grid discretisation.

In contrast, the LURF model shows a marked improvement, with $R^2 = 0.54$, RMSE $= 11.0\ \mu g\ m^{-3}$, and MAE $= 8.0\ \mu g\ m^{-3}$. The regression slope is closer to unity, indicating a substantial reduction in systematic bias. Although the predictive performance ($R^2 = 0.54$) represents a substantial improvement over the CTM baseline, it also indicates that a significant fraction of variance remains unexplained, likely due to unresolved local emission sources, measurement representativeness errors, and the intrinsic stochasticity of particulate matter dynamics. The MSE decomposition suggests that the remaining error is predominantly random, reflecting residual variability not captured by the predictor set. Some dispersion persists, particularly at higher concentration levels, indicating that extreme events remain only partially resolved.

A systematic inspection of the scatter plot (Figure 3, right panel) reveals a noteworthy clustering of observations on `IdDust` days (red-filled circles, defined in Section 2.6) in the upper-right region of the diagram, corresponding to observed PM_{10} values well above 50 $\mu g\ m^{-3}$. On these days, LURF predictions tend to fall below the 1:1 line, indicating systematic underestimation that is larger in magnitude than the residuals observed for non-dusty days at comparable concentration levels. This bias is physically interpretable: during Saharan dust intrusions, the dominant PM_{10} source is long-range transported crustal aerosol, a source regime that is morphologically and chemically distinct from the traffic, biomass-combustion, and secondary aerosol signals that govern the non-dusty training population. The inclusion of the binary `IdDust` flag in the predictor vector $\mathbf{x}_{s,t}$ partially mitigates this effect by enabling the Random Forest to route tree splits into a dust-specific sub-space, reducing the systematic underestimation relative to a model trained without any dust-episode identifier; the residual bias visible in Figure 3 on `IdDust` days therefore represents the irreducible component that cannot be recovered from the available predictor set—a finding that motivates the future replacement of the binary flag with a continuous dust surface concentration field (Section 5).

No marked over-prediction at the most polluted sites is evident, a differential bias frequently observed in linear LUR applications but largely suppressed in tree-based ensembles [13,45]. These performance figures are comparable to those reported in recent machine-learning-based PM mapping studies integrating satellite, meteorological, and land-use predictors at regional and continental scales [14–16,46], while being more conservative than values obtained from naive cross-validation approaches that ignore spatial autocorrelation [39,40].

 doi:10.20944/preprints202604.1265.v1

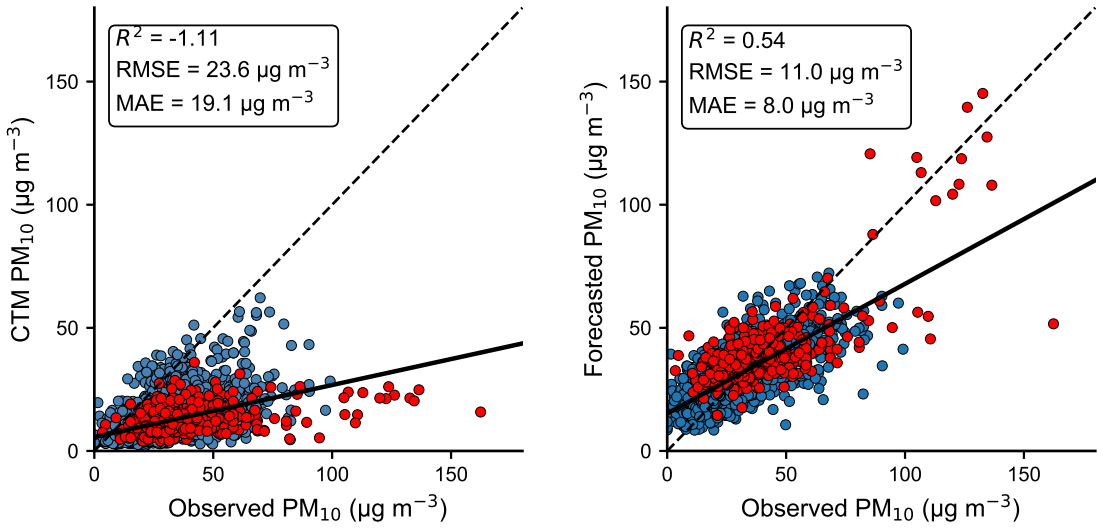

Figure 3. Observed versus predicted daily mean PM_{10} concentrations at ARPA Campania monitoring stations. The left panel shows the comparison with raw CHIMERE output ($R^2 = -1.11$, RMSE $= 23.6\ \mu g\,m^{-3}$, MAE $= 19.1\ \mu g\,m^{-3}$), whereas the right panel reports the LLO-CV results for the LURF model ($R^2 = 0.54$, RMSE $= 11.0\ \mu g\,m^{-3}$, MAE $= 8.0\ \mu g\,m^{-3}$). The dashed line denotes the 1:1 relationship, and the solid line indicates the least-squares regression fit. Observations on dust days (`IdDust = 1`) are plotted as red-filled circles.

Residual diagnostics (Figure 4) reveal a near-Gaussian error distribution that is centred on quasi zero bias $= 0.61\ \mu g\,m^{-3}$, indicating the practical absence of systematic model bias. The distribution is, however, negatively skewed (skewness $= -1.29$), and a Shapiro–Wilk test formally rejects strict normality ($p < 0.001$). This departure from Gaussian behaviour is driven by a tail of large negative residuals at the highest observed concentrations (above $\sim 80\ \mu g\,m^{-3}$), suggesting a tendency to underestimate some extreme PM_{10} episodes, possibly linked to the limited number of training samples at such concentrations or to episodic transport events that are not fully captured by the daily mean CTM predictor. Inspection of the scatter diagram (Figure 3, right panel) reveals that a disproportionate fraction of the largest negative residuals corresponds to observations on `IdDust` days (red-filled circles), confirming that Saharan dust transport episodes are the primary driver of upper-tail underestimation.

 doi:10.20944/preprints202604.1265.v1

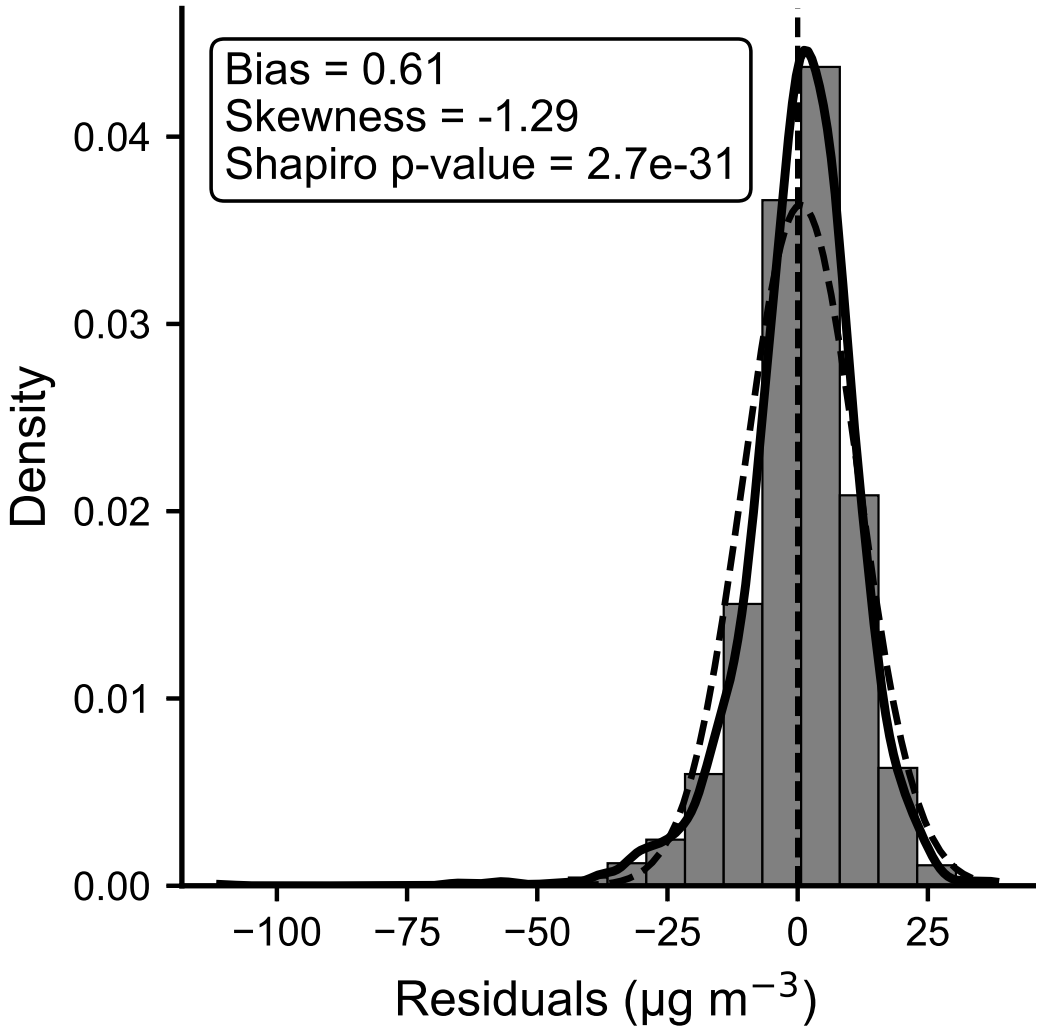

Figure 4. Residual diagnostics for LURF-predicted daily mean PM_{10} at ARPA Campania monitoring sites (skewness $= -1.29$, Shapiro–Wilk $p - value < 0.001$). The dashed line represents the Gaussian fit to residuals. The histogram indicates an error distribution centred on zero with a slight negative skew driven by underestimation at the highest concentrations.

Heteroscedasticity is present in a mild form (see Figure 5). In particular, the spread of residuals becomes moderately wider and somewhat more asymmetric in the upper range of predictions, with a limited number of outliers. This pattern suggests that model uncertainty is not constant across the prediction domain and tends to increase with PM_{10} levels. Such behaviour is consistent with the intrinsic variability of particulate matter, where higher concentrations are often associated with episodic events and local emission heterogeneity that are more difficult to capture with statistical models. The observed heteroscedasticity indicates that model errors are dominated by random variability at low concentrations, whereas both random and systematic components contribute at higher concentrations, leading to reduced predictive reliability in the upper tail of the distribution.

Taken together, these diagnostics indicate that the predictor set captures the dominant emission patterns and meteorological modulation within Campania and that remaining errors are compatible with typical regulatory PM_{10} measurement uncertainty and the representativeness error of fixed-site monitors in complex urban environments [3,8].

While the comparison against raw CTM output highlights the added value of the LURF framework, future work should also consider intermediate benchmarks (e.g., linear LUR or reduced-predictor

 doi:10.20944/preprints202604.1265.v1

models) to more explicitly quantify the contribution of each predictor group and to disentangle the respective roles of statistical learning and physical modelling.

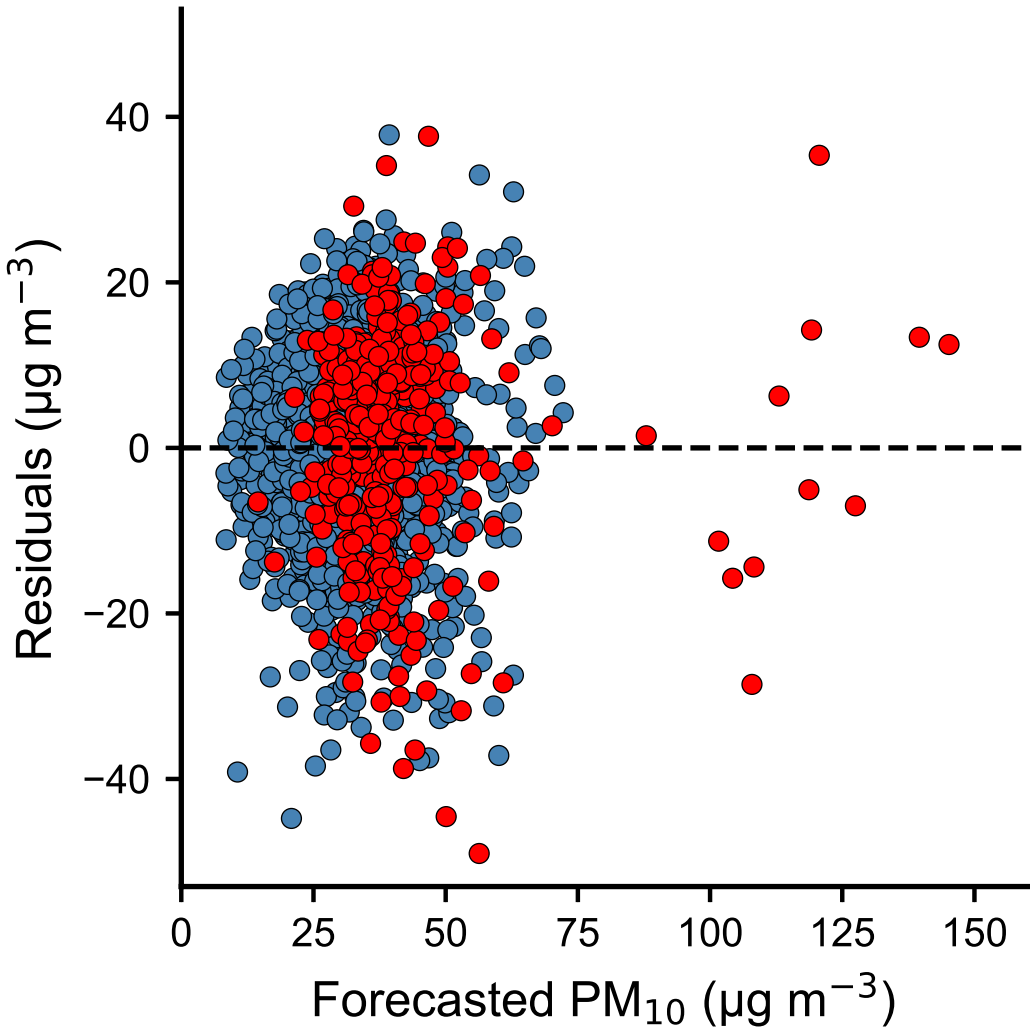

Figure 5. Residuals (observed minus predicted daily mean PM_{10}, $\mu g\,m^{-3}$) plotted against LURF-predicted values ($\mu g\,m^{-3}$) at ARPA Campania monitoring sites. The dashed horizontal line marks zero residual. Increasing spread at higher predicted concentrations reflects mild heteroscedasticity in the LURF error structure. Forecasts on dust days are plotted as red-filled circles.

4.2. Predictor Importance and Role of CHIMERE and ERA5

Figure 6 reports the relative importance of the main predictors entering the LURF model. CHIMERE PM_{10} fields, summarised at the monitoring locations, emerge as the single most influential predictor, accounting for $\approx 15.9 \pm 1.4\%$ of total relative importance. ERA5 mean sea-level pressure (`mslp`), 2 m air temperature (`t2m`), 10 m wind speed components (`u10` and `v10`), 2 m dew-point temperature (`d2m`), and planetary boundary-layer height (`pbl`) together contribute a further $\approx 40\%$, indicating that large-scale and boundary-layer meteorological conditions exert a strong control on the spatial variability of daily mean PM_{10} across the region. The AOD/pbl indicator ($13.5 \pm 1.7\%$) ranks as the second most influential predictor, confirming that satellite aerosol products—once adjusted for vertical mixing—effectively capture background aerosol loading over both coastal and inland areas. In contrast, purely demographic or static land-use descriptors such as resident population show comparatively low importance, suggesting that within the range of conditions sampled by the ARPA

 doi:10.20944/preprints202604.1265.v1

network, PM_{10} gradients are driven more by atmospheric dispersion and regional background than by population density alone.

However, it is important to note that the prominent role of CHIMERE as a predictor should be interpreted with caution, as it reflects both its physical relevance in representing regional background concentrations and its statistical correlation with observed PM_{10}. In this sense, the LURF model operates as a data-driven bias-correction and downscaling tool rather than an entirely independent predictive framework.

The binary categorical indicator `IdDust` (Section 2.6) appears among the secondary predictors in Figure 6, with a mean relative importance of approximately 3–5% (the precise value varies across outer folds due to the limited number of dust-affected days in the 2022 dataset). Despite this modest overall weight, its inclusion produces a disproportionate improvement in the representation of the upper tail of the concentration distribution: on `IdDust = 1` days, the LURF is able to route tree splits towards a coarser-aerosol, dust-driven regime, reducing the negative residuals that would otherwise arise from attempting to explain Saharan dust peaks with meteorological and traffic-based predictors alone. This behaviour is consistent with the known tendency of tree-based regressors to exploit categorical regime indicators for distributional partitioning [47], and aligns with the residual analysis of Section 4.1, where the largest negative residuals are concentrated precisely at the highest observed PM_{10} concentrations—events that predominantly correspond to dust episodes.

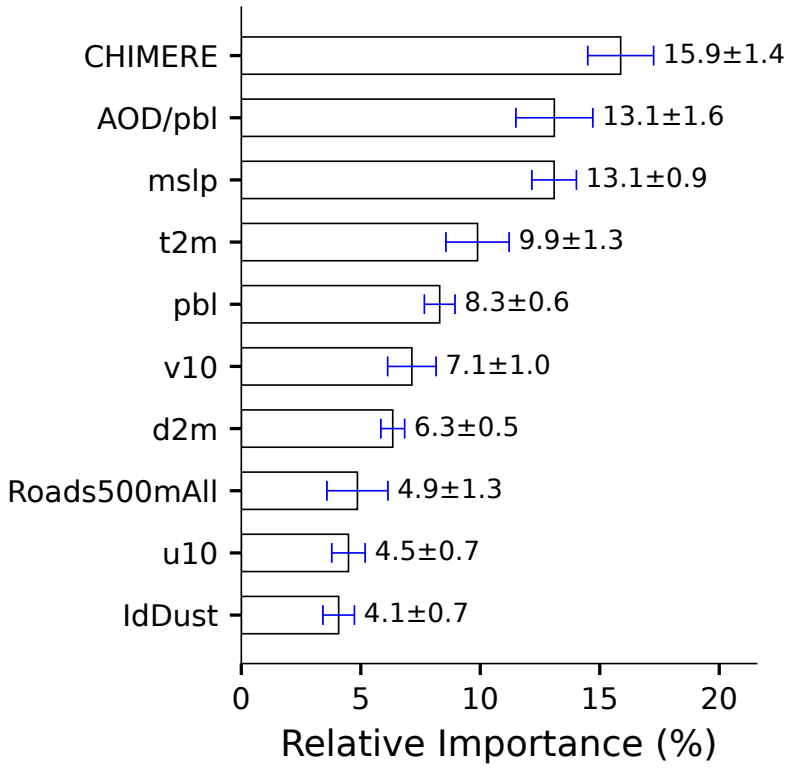

Figure 6. Relative importance of the top 10 predictors in the LURF model. Bars represent the mean relative importance (%) across outer-fold resampling, while error bars indicate the corresponding standard deviation, reflecting variability due to data partitioning. Importance is computed as the decrease in predictive performance (RMSE-based) when each predictor is randomly permuted. The AOD/pbl indicator is derived by scaling MODIS AOD with ERA5 boundary-layer height, accounting for vertical mixing effects.

4.3. Regional PM_{10} Patterns at 1000 m Resolution

The annual maps presented in this section are derived by temporal aggregation of daily LURF predictions, as described in Section 3.3, rather than being directly predicted as annual quantities.

 doi:10.20944/preprints202604.1265.v1

4.3.1. LURF Annual Mean Map

The LURF-predicted annual mean PM_{10} (Figure 7) field reveals a pronounced urban–rural gradient over the Campania region. Peak concentrations—approaching or locally exceeding the EU annual limit value of 40 $\mu g/m^3$—are confined to the Naples metropolitan core, main motorway corridors, and selected periurban industrial zones in the northern Campania plain. By contrast, elevated Apennine terrain, forested inland areas, and the southwestern coast remains around 30 $\mu g/m^3$. The spatial pattern is coherent with source-apportionment evidence for the Campania domain, which attributes high PM_{10} burden in the urban core primarily to road traffic, brake and tyre wear, and secondary inorganic aerosol [6,7,18]. Elevated concentrations also delineate the Port of Naples and the adjacent coastal infrastructure, where maritime emissions superimpose on road-traffic signals in an area with sparse monitoring coverage.

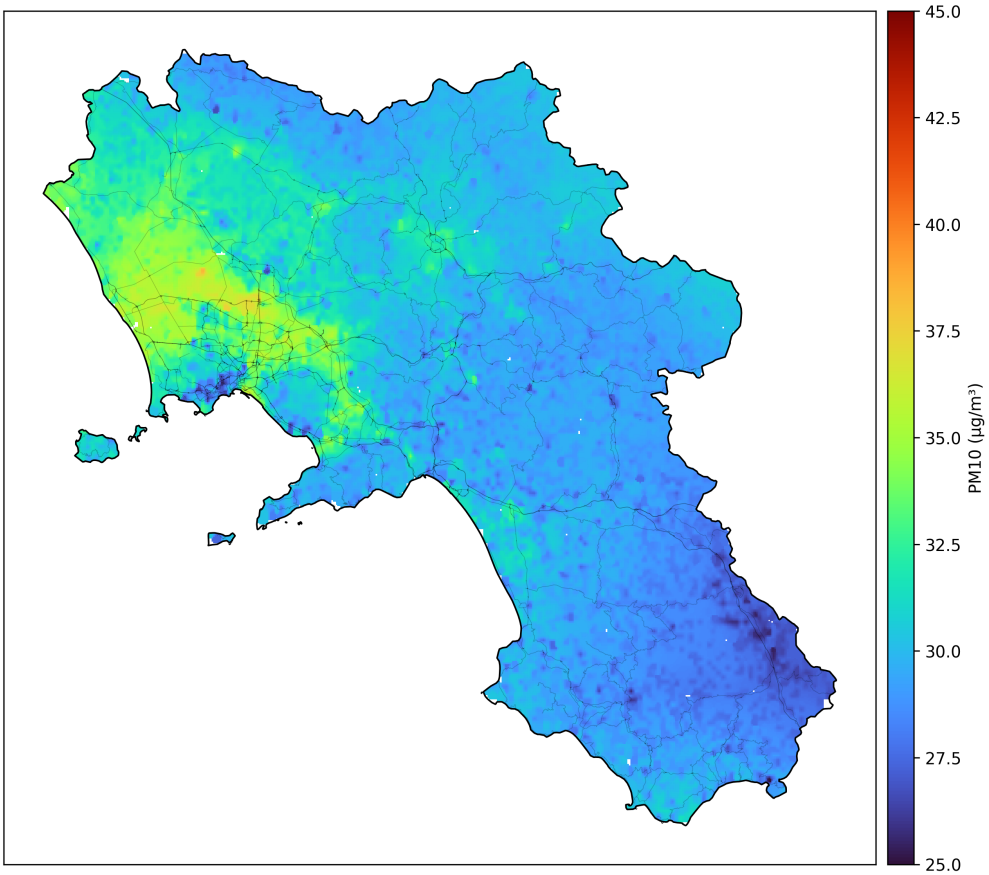

Figure 7. Spatial distribution of LURF-predicted annual mean PM_{10} concentrations across the Campania Region at 1 km resolution. Main roads are overlaid as thin lines.

4.3.2. CTM Annual Mean Map

The corresponding CHIMERE annual mean field (Figure 8) exhibits markedly different characteristics. Predicted concentrations span approximately 0–20 $\mu g/m^3$ across the domain, falling substantially below observed annual means at traffic and urban stations. The map shows a broad, spatially smooth hotspot centred on the Naples conurbation and its northern plain, reflecting the 1 km grid discretisation of the innermost CTM domain; however, sub-kilometre gradients along individual motorway corridors and port access roads are absent. Furthermore, concentrations in rural and mountainous areas are systematically overestimated relative to the LURF field, a pattern consistent with the tendency of Eulerian transport models to spread emission plumes over coarser grid cells and to underrepresent the steep near-road enhancements resolved by the LURF [3,37].

 doi:10.20944/preprints202604.1265.v1

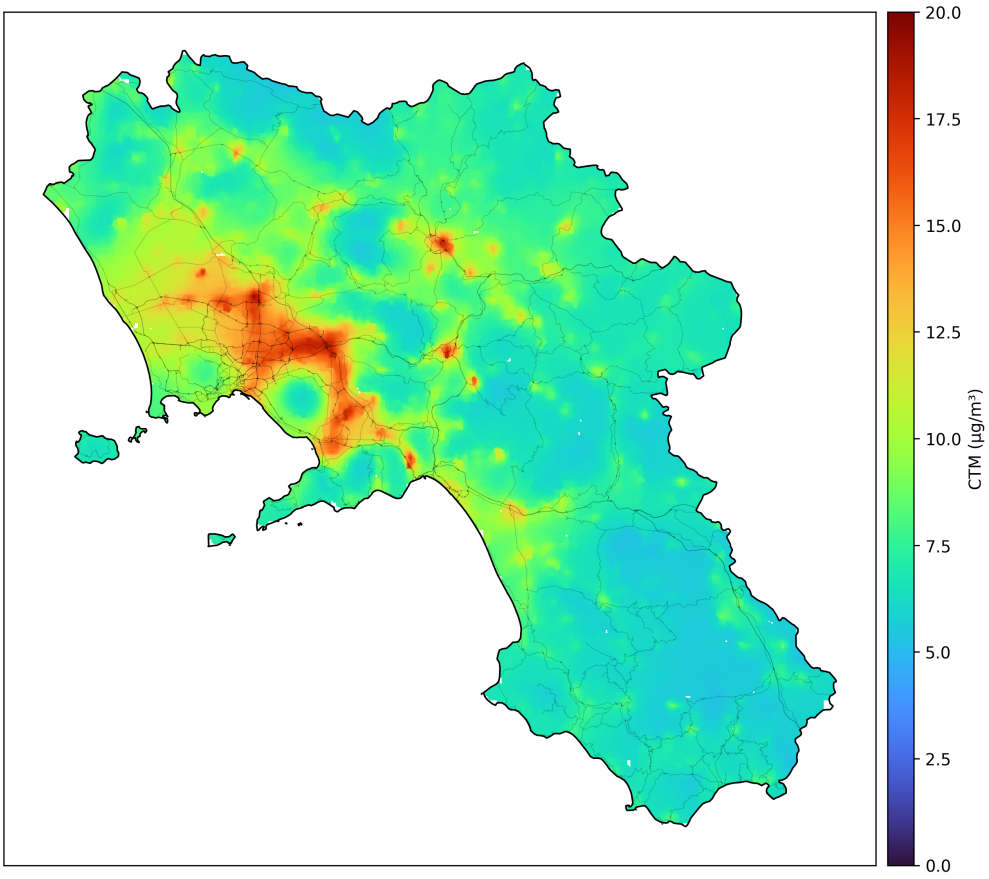

Figure 8. Spatial distribution of CTM-predicted annual mean PM_{10} concentrations across the Campania Region at 1 km resolution. Main roads are overlaid as thin lines.

4.3.3. Exceedance Probability Maps

Figure 9 shows the exceedance probability $\hat{P}_g^{\mathrm{exc}}(\tau)$ (equation 2) evaluated at two regulatory thresholds: $\tau_{20} = 20\,\mu\mathrm{g\,m^{-3}}$ (left panel), corresponding to the 2030 annual limit value under Directive (EU) 2024/2881 [43], and $\tau_{40} = 40\,\mu\mathrm{g\,m^{-3}}$ (right panel), corresponding to the current limit under Directive 2008/50/EC.

For the current τ_{40} threshold, probabilities above 0.5 are confined to a compact zone encompassing the Naples metropolitan core (Municipality of Naples, northern plain cluster, and the western arc of the Campi Flegrei district) and isolated hotspots along motorways, with the majority of the Campania domain remaining below 0.2, consistent with annual means in the 25–35 $\mu\mathrm{g\,m^{-3}}$ range.

For the stricter 2030 threshold τ_{20}, the spatial extent of high-probability cells expands markedly: probabilities above 0.5 cover not only the dense urban core but also a substantial fraction of the peri-urban plain extending northwards, reflecting predicted annual means well above 20 $\mu\mathrm{g\,m^{-3}}$ over much of the metropolitan area. This result underscores the challenge that the Campania region—and southern Italian urban areas more broadly—will face in achieving compliance with the forthcoming 2030 standard, a finding consistent with the gap documented by national-scale modelling studies [22].

 doi:10.20944/preprints202604.1265.v1

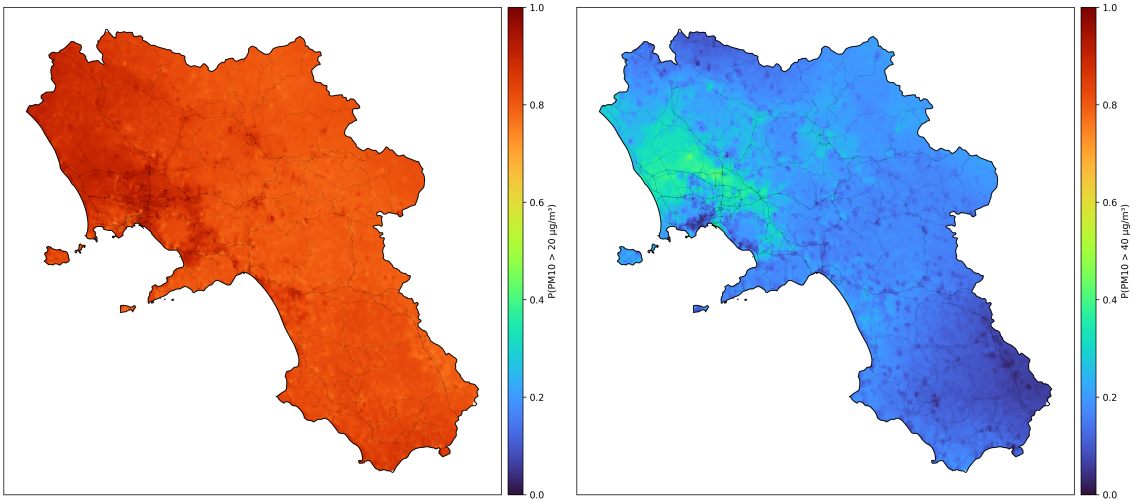

Figure 9. Ensemble-based exceedance probability $\hat{P}_g^{\mathrm{exc}}(\tau)$ (equation 2) for the LURF-predicted annual mean PM_{10} over the Campania Region at 1000 m resolution. *Left*: probability of exceeding the 2030 EU annual limit value of 20 $\mu g\,m^{-3}$ (Directive (EU) 2024/2881). *Right*: probability of exceeding the current EU annual limit value of 40 $\mu g\,m^{-3}$ (Directive 2008/50/EC). Main roads are overlaid as thin lines.

The inter-tree standard deviation of the LURF ensemble (Figure 10) provides a spatially resolved measure of prediction uncertainty. Uncertainty is relatively low in the well-monitored urban core and along the main motorway axes, and increases towards sparsely monitored mountainous and coastal areas, mirroring patterns reported in European-scale LURF and hybrid LUR/CTM studies [13,45].

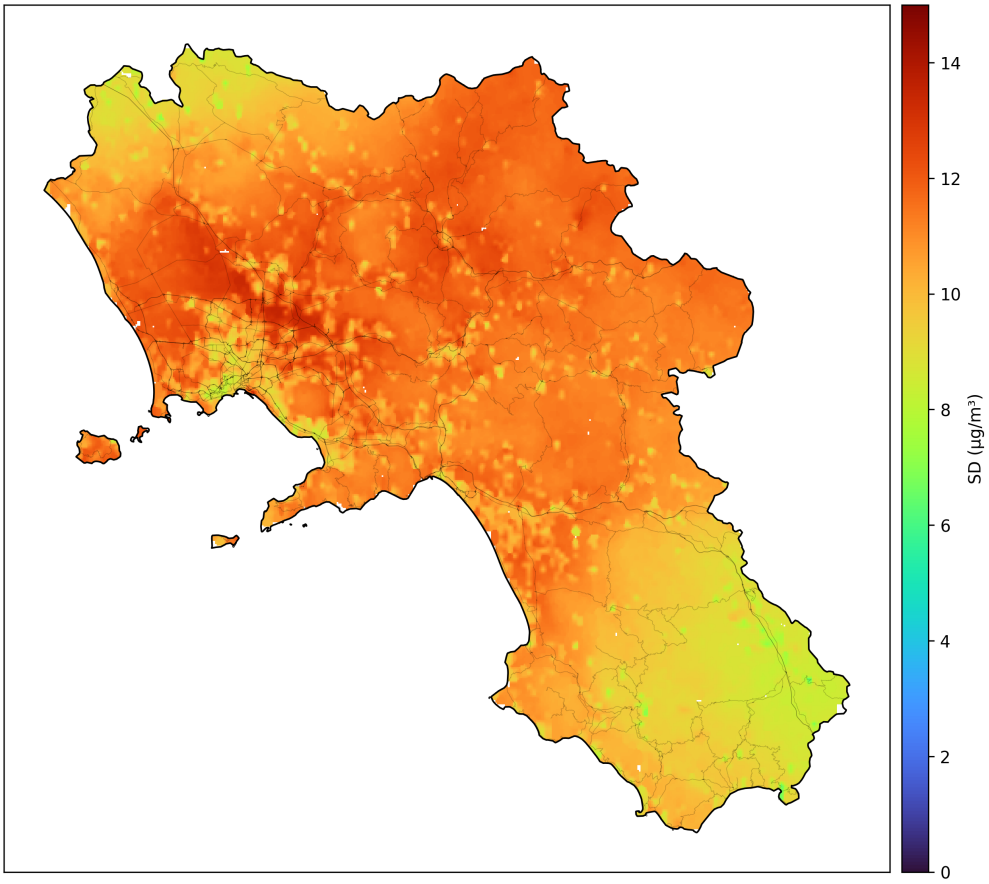

Figure 10. Standard deviation of LURF-predicted annual mean PM_{10} (across 500 decision trees) at 1000 m resolution. Higher values indicate larger ensemble spread and, therefore, greater predictive uncertainty.

 doi:10.20944/preprints202604.1265.v1

5. Discussion

5.1. Positioning and Innovative Aspects of the LURF–CHIMERE Framework

The Campania LURF model extends the growing family of land-use-based machine-learning frameworks for air-quality exposure assessment. Compared with classical LUR studies developed within the ESCAPE project and subsequent multi-city efforts [8,9,48], the present work features four main distinguishing elements: (i) a Random Forest algorithm designed to handle high-dimensional, collinear predictor spaces; (ii) the embedding of CTM-derived PM_{10} fields as a covariate, thereby fusing deterministic and statistical information; (iii) the explicit classification of Saharan dust transport days via the binary indicator `IdDust`, constructed using the $PM_{10}/PM_{2.5}$ coarse-fraction criterion and MODIS AOD thresholds and validated against WMO Barcelona Dust Regional Center operational products [20]; and (iv) the application to a regional domain characterised by complex coastal and orographic meteorology and a heterogeneous mix of urban, industrial, and rural environments. Similar hybrid frameworks combining land-use regression, satellite observations, and chemical transport modelling have been successfully applied across North America, Europe, and Asia [14,22,49–51], reinforcing the general applicability of the approach adopted here.

Recent reviews emphasise that machine-learning extensions of LUR tend to outperform linear LUR in terms of predictive skill, particularly where monitoring is sparse and spatial structure is complex [22, 45]. The performance metrics reported in Section 4.1 are comparable to those of other European applications combining ERA5 reanalysis, satellite retrievals, and ground-based measurements, but the present framework differs in the explicit integration of a regional CTM field at 1 km resolution. This positions the study close to the "hybrid" LUR/CTM/satellite models developed for NO_2 and $PM_{2.5}$ at continental scales [12,13], while pushing spatial resolution down to 1000 m over a single, meteorologically complex region.

Several aspects of this study extend the current state of the art more specifically. A first distinguishing element is the hybrid use of a nested CTM as a predictor at regional scale: rather than relying on a generic continental CTM output, the present work integrates CHIMERE simulations from a dedicated three-domain nested configuration (25, 5, and 1 km) specifically tuned for Italy and Campania. The nested CHIMERE field provides a physically based background PM_{10} distribution that LURF refines to 1000 m resolution, effectively implementing a CTM-informed statistical downscaling strategy. This hybrid architecture separates the problem into two complementary tasks: the CTM encodes regional-scale emission budgets and long-range transport dynamics, while the Random Forest learns the local transfer function that maps coarse grid-cell concentrations onto the fine-scale emission heterogeneity captured by road-network, land-use, and boundary-layer predictors—a decomposition analogous to the "physical prior + statistical correction" paradigm advocated by de Hoogh et al. [13] and Fania et al. [22]. Closely related to this is the joint exploitation of ERA5, MODIS AOD, and CTM output within a single predictor space: the combination of ERA5 meteorological diagnostics, satellite-based AOD from the Terra and Aqua platforms, and CHIMERE concentration fields provides a rich, multi-source description of atmospheric processes across scales, fused with land-use and traffic indicators. This degree of integration of reanalysis, satellite, and CTM information within a tree-based LURF framework has so far received limited attention in the PM_{10} literature.

A further methodological contribution is the generation of probabilistic exceedance maps (Figure 9, defined formally in equation 2), an output layer absent in most comparable studies. By propagating ensemble uncertainty through to two regulatory indicators—the current EU limit of 40 $\mu g\,m^{-3}$ and the forthcoming 2030 limit of 20 $\mu g\,m^{-3}$ under Directive (EU) 2024/2881 [43]—the framework provides regulators and epidemiologists with spatially resolved risk products rather than point estimates, conceptually analogous to probabilistic weather forecasting [44] and constituting a significant step towards uncertainty-aware air-quality assessment. Finally, the physical plausibility of the entire framework is underpinned by the rich body of local process knowledge available for the Campania domain, which has been the subject of detailed emission-factor, source-apportionment, and meteorological studies [6,7,18,24]. Embedding these insights into predictor design and interpretation illustrates how

 doi:10.20944/preprints202604.1265.v1

domain-specific, process-based understanding can guide and constrain the use of machine learning in air-quality modelling.

5.2. Implications for Exposure and Health-Impact Assessment

The exceedance-probability map (Figure 9) shows that the probability values above 0.5 are spatially colocalised with the highest population density in the Naples metropolitan area, implying that health-burden estimates based on raw CTM output would systematically underestimate the fraction of the population exposed to concentrations above the EU limit value. This has direct implications for health-burden and environmental-justice assessments, in line with recent work on high-resolution urban air-pollution mapping and its use in environmental-justice analyses [5,52]. The spatially resolved uncertainty estimates provided by the LURF ensemble further enable propagation of exposure uncertainty through to health-impact calculations, an aspect often treated only qualitatively in deterministic CTM studies [3,53].

5.3. Limitations, Critical Assessment, and Future Perspectives

Despite its strengths, the present framework has several limitations that merit explicit discussion. The first concerns network sparsity and spatial representativeness. The model is trained on 13 regulatory stations, the majority of which are located in or near the Naples metropolitan area; this uneven spatial coverage introduces a structural asymmetry whereby the LURF performs well in the well-observed urban domain but relies on extrapolation elsewhere. As correctly reflected by the ensemble uncertainty maps (Figure 10), predictive confidence decreases markedly over Apennine terrain and along the southern coastal fringe. The exceedance-probability map (Figure 9) should therefore be interpreted with caution in areas of high ensemble uncertainty, where low exceedance probabilities may partly reflect insufficient training signal rather than genuinely clean air conditions. A more balanced station network—including rural background, coastal, and high-altitude sites—would be essential to consolidate the representativeness of regional exposure estimates.

A second limitation concerns temporal scope. The present study is based on a single calendar year (2022) and produces annual-mean output only. While this is appropriate for regulatory compliance purposes, it precludes the characterisation of intra-annual variability that is ecologically and epidemiologically relevant. The Campania region is subject to marked seasonal contrasts: summer sea-breeze circulation promotes recirculation of polluted air masses in the Naples bay, while autumn and winter cold-pool episodes lead to intense stagnation and biomass-combustion episodes [18,24]. A single annual-mean map, whether LURF-based (Figure 7) or CTM-based (Figure 8), compresses this temporal complexity into a single spatial field, potentially masking seasonal exposure peaks of public-health relevance. Extending the framework to monthly or seasonal maps by incorporating time-stratified ERA5 and CHIMERE predictors is therefore an important next step.

A third, methodologically more subtle, limitation concerns the dual role of CHIMERE in this study: it enters the LURF both as a predictor variable and as the deterministic benchmark against which LURF performance is assessed. This design is intentional—the LURF is explicitly conceived as a statistical post-processor that learns to correct and downscale CTM output using local observational and geospatial information, in a manner conceptually analogous to model output statistics (MOS) approaches widely used in numerical weather prediction [44,54]. As a consequence, the performance gain of LURF over raw CHIMERE should be interpreted as the added value of data-driven downscaling and bias correction (Figure 7 versus Figure 8), rather than as evidence that the statistical model can operate independently of the CTM. Critically, this design also implies that systematic biases in CHIMERE—for instance, an underestimation of secondary inorganic aerosol formation during stagnation episodes or an inaccurate representation of maritime emission plumes in coastal grid cells—will propagate into the LURF predictions to a degree that depends on how consistently those biases manifest at the monitoring-station locations included in the training set.

A further limitation, specific to the `IdDust` categorical predictor, concerns the availability and spatial representativeness of $PM_{2.5}$ measurements within the ARPA Campania network. Because

 doi:10.20944/preprints202604.1265.v1

$PM_{2.5}$ is not measured at all 13 stations retained in this study, the coarse-fraction criterion $(PM_{10} - PM_{2.5})/PM_{10} > 0.5$ required to assign `IdDust = 1` must be partially imputed from the network median at the co-measuring stations on the same day (see Section 2.6). This imputation introduces an additional source of uncertainty for the stations and days where $PM_{2.5}$ is absent. Moreover, the binary `IdDust` classification captures only the most intense dust episodes and does not account for moderate or partially dust-contaminated days that fall below the combined threshold.

A final limitation concerns the interpretability of the Random Forest. While ensemble tree methods excel at capturing non-linear associations, they do not provide the same level of parametric interpretability as classical regression models. Variable-importance metrics (Figure 6) and partial-dependence plots provide some insight, but causal inference regarding individual predictors must remain grounded in process-based understanding and complementary modelling approaches [8,45]. In particular, the high importance attributed to CHIMERE (Figure 6) could partly reflect collinearity between the CTM field and other predictors (ERA5, AOD/pbl) that are derived from the same meteorological forcing used to drive CHIMERE itself.

6. Conclusions

This study presented a hybrid Land-Use Random Forest (LURF) model for spatially continuous PM_{10} exposure mapping across the Campania Region at 1000 m resolution, trained on 2022 daily observations from 13 ARPA Campania regulatory stations and a high-dimensional predictor space comprising land-use, road-network, ERA5 meteorological, MODIS AOD, and CHIMERE CTM fields.

Under spatially aware leave-location-out cross-validation, the LURF achieves a twofold reduction in error relative to the raw CHIMERE output (see Section 4.1). Residual diagnostics confirm the practical absence of systematic bias, while mild heteroscedasticity at the highest concentrations points to residual difficulties in representing episodic PM_{10} events. Predictor importance analysis identifies CHIMERE PM_{10} as the single most influential covariate ($\sim$18% of total importance), followed by ERA5 pressure and boundary-layer fields (collectively $\sim$40%) and the AOD/pbl satellite composite, confirming that atmospheric dispersion and regional background dominate PM_{10} variability over the region at the daily time scale. The binary categorical variable `IdDust` ranks among the secondary predictors and contributes to reducing systematic underestimation on episodic Saharan dust days, which account for some of the largest negative residuals observed under LLO-CV.

The 1000 m annual mean maps resolve intra-urban gradients along motorway corridors and in the Naples metropolitan core that remain invisible at typical CTM grid spacings. Population-weighted exposure exceeds the simple area-averaged CTM estimate, with direct implications for health-burden and environmental-justice assessments. The probabilistic exceedance maps (Figure 9, equation 2) further show that, against the current EU limit of 40 $\mu g\,m^{-3}$, probabilities above 0.5 are confined to the Naples metropolitan core and selected motorway corridors; against the forthcoming 2030 limit of 20 $\mu g\,m^{-3}$ under Directive (EU) 2024/2881 [43], high-probability cells expand across the entire metropolitan area and peri-urban plain, highlighting the scale of the compliance challenge that Campania—and comparable southern European urban regions—will face in the coming years.

Overall, the proposed framework demonstrates how the integration of physically-based and data-driven approaches can enhance exposure assessment in data-sparse regions, while highlighting the need for careful interpretation of hybrid model outputs and continued investment in monitoring infrastructure.

Future developments should focus on a continuous dust-loading proxy—for example, the CHIMERE or CAMS dust surface concentration field—as a supplement or alternative to the binary flag, allowing the model to represent a continuum of dust influence rather than a hard classification. The extension of this framework to $PM_{2.5}$ and the incorporation of time-varying ERA5 and CHIMERE predictors at monthly resolution will enable seasonal exposure maps. Testing the transferability of the trained LURF to other Italian regions, via domain-adaptation strategies, would further consolidate this

 doi:10.20944/preprints202604.1265.v1

approach as a versatile, low-cost complement to deterministic CTM modelling for regional air-quality management.

Author Contributions: Conceptualisation, A.R. and E.C.; methodology, A.R. and E.C.; software, A.R.; validation, A.R. and E.C.; formal analysis, E.C.; investigation, E.C.; data curation, E.C.; writing—original draft preparation, E.C. and A.R.; writing—review and editing, A.R.; project administration, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Scripts used to implement the R-based workflow to train and validate the LURF model is publicly available at https://doi.org/10.5281/zenodo.19496301 under GPL-3 license. ARPA Campania monitoring data are directly available (https://dati.arpacampania.it/dataset/dati-monitoraggio-qualita-aria). ERA5 reanalysis data are freely accessible from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). MODIS AOD products are distributed by NASA Earthdata (https://earthdata.nasa.gov). Corine Land Cover and Copernicus HRL data are available at https://land.copernicus.eu. Processed LURF outputs will be made available upon reasonable request to the corresponding author.

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

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 Disease Study 2015. *The Lancet* **2017**, *389*, 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.
2. World Health Organization. WHO global air quality guidelines: particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Technical report, World Health Organization, Geneva, 2021. https://www.who.int/publications/i/item/9789240034228. Last accessed: 1 April 2026.
3. Solazzo, E.; Riccio, A.; Van Dingenen, R.; Valentini, L.; Galmarini, S. Evaluation and uncertainty estimation of the impact of air quality modelling on crop yields and premature deaths using a multi-model ensemble. *Science of the Total Environment* **2018**, *633*, 1437–1452. https://doi.org/10.1016/j.scitotenv.2018.03.317.
4. European Commission. Directive of the European Parliament and of the Council on Ambient Air Quality and Cleaner Air for Europe, 2022. Accessed 1 April 2026.
5. Apte, J.S.; Manchanda, C. High-resolution urban air pollution mapping. *Science* **2024**, *385*, 380–385. https://doi.org/10.1126/science.adq3678.
6. Riccio, A.; Chianese, E.; Tirimberio, G.; Prati, M. Emission factors of inorganic ions from road traffic: A case study from the city of Naples (Italy). *Transportation Research Part D: Transport and Environment* **2017**, *54*, 239–249. https://doi.org/10.1016/j.trd.2017.05.008.
7. Riccio, A.; Chianese, E.; Monaco, D.; Costagliola, M.; Perretta, G.; Prati, M.; Agrillo, G.; Esposito, A.; Gasbarra, D.; Shindler, L.; et al. Real-world automotive particulate matter and PAH emission factors and profile concentrations: Results from an urban tunnel experiment in Naples, Italy. *Atmospheric Environment* **2016**, *141*, 379–387. https://doi.org/10.1016/j.atmosenv.2016.07.006.
8. Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* **2008**, *42*, 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057.
9. Eeftens, M.; Beelen, R.; de Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dëljus, E.; Dons, E.; Heinrich, J.; et al. Development of land use regression models for $PM_{2.5}$, $PM_{2.5}$ absorbance, PM_{10} and PM_{coarse} in 20 European study areas; results of the ESCAPE project. *Environmental Science & Technology* **2012**, *46*, 11195–11205. https://doi.org/10.1021/es301948k.
10. Wang, M.; Beelen, R.; Basagaña, X.; Becker, T.; Cesaroni, G.; de Hoogh, K.; Dedele, A.; Declercq, C.; Dimakopoulou, K.; Eeftens, M.; et al. Evaluation of land use regression models for NO_2 and particulate matter in 20 European study areas: the ESCAPE project. *Environmental Science & Technology* **2013**, *47*, 4357–4364. https://doi.org/10.1021/es305129t.
11. Wolf, K.; Cyrys, J.; Harciníková, T.; Gu, J.; Kusch, T.; Hampel, R.; Schneider, A.; Peters, A. Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution

 doi:10.20944/preprints202604.1265.v1

in Augsburg, Germany. *Science of the Total Environment* **2017**, *579*, 1531–1540. https://doi.org/10.1016/j.scitotenv.2016.11.160.

12. Vienneau, D.; de Hoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Western European land use regression incorporating satellite- and ground-based measurements of NO_2 and PM_{10}. *Environmental Science & Technology* **2013**, *47*, 13555–13564. https://doi.org/10.1021/es403089q.
13. de Hoogh, K.; Chen, J.; Gulliver, J.; Hoffmann, B.; Stafoggia, M.; Künzli, N.; Kloog, I.; Vineis, P.; Brunekreef, B.; Hoek, G.; et al. Development of West-European $PM_{2.5}$ and NO_2 land use regression models incorporating satellite-derived and chemical transport modelling data. *Environmental Research* **2016**, *151*, 1–10. https://doi.org/10.1016/j.envres.2016.07.005.
14. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M.B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; et al. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environment International* **2019**, *130*, 104909. https://doi.org/10.1016/j.envint.2019.104909.
15. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes Jr, M.G.; Estes, S.M.; Quattrochi, D.A.; Puttaswamy, S.J.; et al. Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sensing of Environment* **2014**, *140*, 220–232. https://doi.org/10.1016/j.rse.2013.08.032.
16. Chen, G.; Li, S.; Knibbs, L.D.; Hamm, N.A.; Cao, W.; Li, T.; Guo, J.; Ren, H.; Abramson, M.J.; Guo, Y. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. *Science of the Total Environment* **2018**, *636*, 52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251.
17. Chianese, E.; Riccio, A. Long-term variation in exposure to NO_2 concentrations in the city of Naples, Italy: Results of a citizen science project. *Science of the Total Environment* **2024**, *931*, 172799. https://doi.org/10.1016/j.scitotenv.2024.172799.
18. Sirignano, C.; Riccio, A.; Chianese, E.; Ni, H.; Zenker, K.; D'Onofrio, A.; Meijer, H.A.; Dusek, U. High contribution of biomass combustion to $PM_{2.5}$ in the city centre of Naples (Italy). *Atmosphere* **2019**, *10*, 451. https://doi.org/10.3390/atmos10080451.
19. Vogel, F.; Putero, D.; Bonasoni, P.; Cristofanelli, P.; Zanatta, M.; Marinoni, A. Saharan dust transport event characterization in the Mediterranean atmosphere using 21 years of in-situ observations. *Atmospheric Chemistry and Physics* **2025**, *25*, 15453–15468. https://doi.org/10.5194/acp-25-15453-2025.
20. WMO Barcelona Dust Regional Center. Daily Dust Products—Ground-level dust concentration maps. http://dust.aemet.es, 2023. Operated by AEMET and BSC on behalf of WMO. Accessed 1 April 2026.
21. Chianese, E.; Camastra, F.; Ciaramella, A.; Landi, T.; Staiano, A.; Riccio, A. Spatio-temporal learning in predicting ambient particulate matter concentration by multi-layer perceptron. *Ecological Informatics* **2019**, *49*, 54–61. https://doi.org/10.1016/j.ecoinf.2018.12.001.
22. Fania, A.; Monaco, A.; Pantaleo, E.; Maggipinto, T.; Bellantuono, L.; Cilli, R.; Lacalamita, A.; La Rocca, M.; Tangaro, S.; Amoroso, N.; et al. Estimation of daily ground level air pollution in Italian municipalities with machine learning models using Sentinel-5P and ERA5 data. *Remote Sensing* **2024**, *16*, 1206. https://doi.org/10.3390/rs16071206.
23. Shtein, A.; Kloog, I.; Schwartz, J.; Silibello, C.; Michelozzi, P.; Gariazzo, C.; Viegi, G.; Forastiere, F.; Karnieli, A.; Just, A.C.; et al. Estimating daily $PM_{2.5}$ and PM_{10} over Italy using an ensemble model. *Environmental Science & Technology* **2019**, *54*, 120–128. https://doi.org/10.1021/acs.est.9b04279.
24. Hernández-Ceballos, M.; Rubino, M.; Sirignano, C.; Chianese, E.; Riccio, A. The cause-effect relationship between synoptic and local wind patterns and PM_{10} concentrations in the complex-orography urban area of Naples (Italy). *City and Environment Interactions* **2025**, *27*, 100200. https://doi.org/10.1016/j.cacint.2025.100200.
25. Agenzia Regionale per la Protezione dell'Ambiente della Campania. Dati monitoraggio Qualità dell'Aria, 2026. https://dati.arpacampania.it/dataset/dati-monitoraggio-qualita-aria. Data set. Last accessed: 1 April 2026.
26. Copernicus Land Monitoring Service. CORINE Land Cover, 2018. https://land.copernicus.eu/en/. Last accessed: 1 April 2026.
27. Hoff, R.M.; Christopher, S.A. Remote sensing of particulate pollution from space: Have we reached the promised land? *Atmospheric Environment* **2009**, *43*, 4257–4266. https://doi.org/10.1016/j.atmosenv.2009.05.005.

 doi:10.20944/preprints202604.1265.v1

28. Gupta, P.; Christopher, S.A. Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research: Atmospheres* **2009**, *114*, D14205. https://doi.org/10.1029/2008JD011496.
29. Ma, Z.; Hu, X.; Sayer, A.M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-based spatiotemporal trends in PM2. 5 concentrations: China, 2004–2013. *Environmental Health Perspectives* **2015**, *124*, 184. https://doi.org/10.1289/ehp.1409481.
30. Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating ground-level $PM_{2.5}$ by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophysical Research Letters* **2017**, *44*, 11–985. https://doi.org/10.1002/2017GL075710.
31. Rouil, L.; Honoré, C.; Vautard, R.; Beekmann, M.; Bessagnet, B.; Malherbe, L.; Meleux, F.; Dufour, A.; Elichegaray, C.; Flaud, J.M.; et al. PREV'AIR: An operational forecasting and mapping system for air quality in Europe. *Bulletin of the American Meteorological Society* **2009**, *90*, 73–84. https://doi.org/10.1175/2008BAMS2390.1.
32. Mailler, S.; Menut, L.; Khvorostyanov, D.; Valari, M.; Couvidat, F.; Siour, G.; Turquety, S.; Briant, R.; Tuccella, P.; Bessagnet, B.; et al. CHIMERE-2017: From urban to hemispheric chemistry-transport modeling. *Geoscientific Model Development* **2017**, *10*, 2397–2423. https://doi.org/10.5194/gmd-10-2397-2017.
33. NOAA Physical Sciences Laboratory. NCEP/NCAR Reanalysis Data, 2026.
34. Kuenen, J.; Dellaert, S.; Visschedijk, A.; Jalkanen, J.P.; Super, I.; Denier van der Gon, H. CAMS-REG-v4: a state-of-the-art high-resolution European emission inventory for air quality modelling. *Earth System Science Data* **2022**, *14*, 491–515. https://doi.org/10.5194/essd-14-491-2022.
35. Giunta, G.; Montella, R.; Mariani, P.; Riccio, A. Modeling and computational issues for air/water quality problems: A grid computing approach. *Nuovo Cimento della Società Italiana di Fisica C* **2005**, *28*, 215–224. https://doi.org/10.1393/ncc/i2005-10184-3.
36. Montella, R.; Giunta, G.; Riccio, A. Using grid computing based components in on demand environmental data delivery. In Proceedings of the Second Workshop on Use of P2P, GRID and Agents for the Development of Content Networks, 2007, pp. 81–86. Accessed 1 April 2026.
37. Riccio, A.; Ciaramella, A.; Giunta, G.; Galmarini, S.; Solazzo, E.; Potempski, S. On the systematic reduction of data complexity in multimodel atmospheric dispersion ensemble modeling. *Journal of Geophysical Research: Atmospheres* **2012**, *117*. https://doi.org/10.1029/2011JD016503.
38. Barnaba, F.; Gobbi, G.P. Aerosol seasonal variability over the Mediterranean region and relative impact of maritime, continental and Saharan dust particles over the basin from MODIS data in the year 2001. *Atmospheric Chemistry and Physics* **2004**, *4*, 2367–2391. https://doi.org/10.5194/acp-4-2367-2004.
39. 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. https://doi.org/https://doi.org/10.1111/ecog.02881.
40. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N.; et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* **2020**, *11*, 4540. https://doi.org/10.1038/s41467-020-18321-y.
41. Meyer, H.; Reudenbach, C.; Hengl, T.; Katurji, M.; Nauss, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* **2018**, *101*, 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001.
42. Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Arroita, G. blockCV: An R package for generating spatially or environmentally separated folds for *k*-fold cross-validation of species distribution models. *Methods in Ecology and Evolution* **2019**, *10*, 225–232. https://doi.org/10.1111/2041-210X.13107.
43. European Parliament and Council of the European Union. Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe. https://eur-lex.europa.eu/eli/dir/2024/2881/oj, 2024. Official Journal of the European Union.
44. Wilks, D.S. *Statistical Methods in the Atmospheric Sciences*, 3rd ed.; Academic Press: Oxford, 2011.
45. Ma, X.; Zou, B.; Deng, J.; Gao, J.; Longley, I.; Xiao, S.; Guo, B.; Wu, Y.; Xu, T.; Xu, X.; et al. A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: A perspective from 2011 to 2023. *Environment International* **2024**, *183*, 108430. https://doi.org/10.1016/j.envint.2024.108430.
46. Xue, T.; Zheng, Y.; Geng, G.; Xiao, Q.; Meng, X.; Wang, M.; Li, X.; Wu, N.; Zhang, Q. Estimating spatiotemporal variation in $PM_{2.5}$ concentrations using satellite data and machine learning. *Atmospheric Chemistry and Physics* **2019**, *19*, 10409–10424. https://doi.org/10.5194/acp-19-10409-2019.

 doi:10.20944/preprints202604.1265.v1

47. Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32. https://doi.org/10.1023/A:1010933404324.
48. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.Y.; Künzli, N.; Schikowski, T.; Marcon, A.; et al. Development of NO_2 and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe–the ESCAPE project. *Atmospheric Environment* **2013**, *72*, 10–23. https://doi.org/10.1016/j.atmosenv.2013.02.037.
49. van Donkelaar, A.; Martin, R.V.; Brauer, M.; Kahn, R.; Levy, R.; Verduzco, C.; Villeneuve, P.J. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. *Environmental Science & Technology* **2015**, *50*, 3762–3772. https://doi.org/10.1021/acs.est.5b05833.
50. Kloog, I.; Chudnovsky, A.A.; Just, A.C.; Nordio, F.; Koutrakis, P.; Coull, B.A.; Lyapustin, A.; Wang, Y.; Schwartz, J. A new hybrid spatio-temporal model for estimating daily multi-year $PM_{2.5}$ concentrations across northeastern USA. *Atmospheric Environment* **2014**, *95*, 581–590. https://doi.org/10.1016/j.atmosenv.2014.07.014.
51. Just, A.C.; Wright, R.O.; Schwartz, J.; Coull, B.A.; Baccarelli, A.A.; Tellez-Rojo, M.M.; Moody, E.; Wang, Y.; Lyapustin, A.; Kloog, I. Using high-resolution satellite aerosol optical depth to estimate daily PM2. 5 geographical distribution in Mexico City. *Environmental Science & Technology* **2015**, *49*, 8576–8584. https://doi.org/10.1021/acs.est.5b00859.
52. Southerland, V.A.; Brauer, M.; Mohegh, A.; Hammer, M.S.; Van Donkelaar, A.; Martin, R.V.; Apte, J.S.; Anenberg, S.C. Global urban temporal trends in fine particulate matter ($PM_{2.5}$) and attributable health burdens: estimates from global datasets. *The Lancet Planetary Health* **2022**, *6*, e139–e146. https://doi.org/10.1016/S2542-5196(21)00350-8.
53. Solazzo, E.; Riccio, A.; Kioutsioukis, I.; Galmarini, S. Pauci ex tanto numero: reduce redundancy in multi-model ensembles. *Atmospheric Chemistry and Physics* **2013**, *13*, 8315–8333. https://doi.org/10.5194/acp-13-8315-2013.
54. Glahn, H.R.; Lowry, D.A. The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology* **1972**, *11*, 1203–1211. https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.