engineering proceedings

MDPI

Proceeding Paper

Hierarchical Bayesian Spatial Framework for Air Quality Imputation in Sparse Sensor Networks †

Ioannis Konstantinidis [1,*] and Gerasimos Pylarinos [2]

[1] Ministry of Education, High School of Andritsaina, P.O. Box 27061 Andritsaina, Greece
[2] Ministry of Education, Junior High School of Vartholomio, P.O. Box 27050 Andritsaina, Greece; pylarinosphys@gmail.com
* Correspondence: ykon@protonmail.com
† Presented at the 2nd International Conference on Engineering, Technology, and Applied Science Innovations (ICETASI 2026), Zakynthos, Greece, 24–26 September 2026.

Abstract

Ground-based air quality monitoring networks in Southern Europe are often sparse, leaving large areas, such as rural plains, mountain ranges, and island groups, unmonitored. Standard spatial interpolation methods like Inverse Distance Weighting or Kriging struggle under these conditions, since they cannot easily incorporate physical terrain or provide reliable uncertainty estimates. This paper presents a Hierarchical Bayesian Spatial framework to address the sparse sensor coverage. We model the annual mean concentration of particulate matter ($PM_{2.5}$) across Greece using data from the European Environment Agency (EEA) e-Reporting network for 2023. We combine ground measurements with terrain elevation from digital elevation models. The continuous spatial surface is discretized using a Finite Element mesh via Stochastic Partial Differential Equations (SPDE). We fit the model deterministically using Integrated Nested Laplace Approximations (INLA). Adding terrain elevation and a sum-to-zero spatial constraint resolved parameter baseline confounding. The model fit improved significantly, increasing the marginal log-likelihood from -41.97 to -39.87. The model achieved an R^2 of 0.91 and a Root Mean Square Error (RMSE) of 2.83 $\mu g/m^3$. The global national baseline concentration was thus estimated at 17.74 $\mu g/m^3$ (95% Credible Interval: [11.34, 23.88]). Elevation indicated a negative effect ($\beta_{elev} = -7.24$ $\mu g/m^3$ per km), with an 89.2% posterior probability of reducing $PM_{2.5}$. Spatial autocorrelation persisted across a regional scale of $\varrho \approx 297$ km. Full posterior variance mapping identified Crete, the Ionian and the Aegean Seas, and Eastern Thrace as zones of maximum predictive uncertainty. These findings provide a quantitative baseline for prioritizing physical sensor deployment in undersampled regions.

Keywords: Bayesian Statistics; R-INLA; SPDE; air quality; uncertainty quantification; spatial interpolation; environmental engineering; $PM_{2.5}$

Check for updates

Academic Editor: Konstantinos Poirazidis

Published: 13 November 2026

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

1. Introduction

1.1. Context and Problem Statement

Atmospheric fine particulate matter ($PM_{2.5}$) poses severe health risks because its small size allows it to penetrate deep into the respiratory system and enter the bloodstream, triggering cardiovascular and pulmonary diseases. The World Health Organization (WHO) sets the recommended annual safety limit for $PM_{2.5}$ at 5 $\mu g/m^3$. However, urban centers across the Mediterranean frequently exceed this threshold [1,2]. Although reliable monitoring requires dense ground station networks, the high cost of installation and maintenance

leads to highly uneven spatial coverage. Thus, most sensors are heavily concentrated in metropolitan areas, while rural, coastal, and island regions remain severely undersampled.

1.2. Geography and Sparse Coverage in Greece

Greece's fragmented topography complicates environmental monitoring. The country features a long mountain range (such as the Pindus), fragmented coastal plains, and hundreds of inhabited islands across the Ionian and Aegean Seas. Official air quality monitoring stations in Greece are clustered in major urban centers, mainly in Athens and Thessaloniki. Vast geographic stretches, namely the Ionian Islands, Crete, Epirus, Thessaly's interior, and Thrace, lack physical sensors entirely. Interpolation across these unmonitored regions is complicated by complex topographical forcing. For instance, mountain ridges can obstruct regional advection, while coastal thermal inversions frequently trap aerosols in basin areas. Consequently, simple spatial averages become highly unreliable.

1.3. Limitations of Traditional Interpolation

Engineers and environmental data scientists frequently apply classical interpolation techniques to map aerosol concentrations across sparse ground networks [1,3]:

1. Inverse Distance Weighting (IDW): Calculates values based purely on Euclidean proximity, ignoring terrain barriers, atmospheric physics [1], and local emission dynamics.
2. Standard Spatial Kriging: Assumes spatial stationarity across the domain. It requires dense point samples and struggles in regions where stations are highly clustered [1,3].
3. Deterministic Machine Learning (e.g., Random Forests, XGBoost): Accurately models non-linear local relationships where dense data exist [1], but fails to provide true Bayesian credible intervals, leading to overconfident point predictions in unobserved territories [4,5].

1.4. Research Objectives and Contributions

This study introduces a Bayesian spatial approach to overcome sparse sensor coverage. The main contributions of this study are:

- Methodological Integration: We combine sparse ground measurements with terrain elevation inside a Hierarchical Bayesian Spatial framework.
- Stochastic Partial Differential Equatio (SPDE) Finite Element Mesh: We use *fmesher* to build a triangulated spatial mesh over Greece. This converts continuous Gaussian fields into computationally efficient Gaussian Markov Random Fields (GMRFs).
- Baseline Confounding Resolution: We show how applying a sum-to-zero spatial constraint (*constr* = TRUE) and informative priors prevents parameter confounding between the global intercept and spatial random fields.
- Uncertainty Quantification (UQ): We produce national prediction maps alongside posterior standard deviation maps. This highlights high-uncertainty zones where new sensors should be deployed.
- Spatial Cross-Validation (LOSO-CV): We implement Leave-One-Station-Out Cross-Validation to evaluate out-of-sample predictive generalizability in unobserved territories, demonstrating an out-of-sample MAE of 3.18 $\mu g/m^3$ and 95% credible interval coverage of 81.8% across background stations.

2. Related Work

2.1. Deterministic vs. Probabilistic Spatial Interpolation

Spatial interpolation methods have evolved from deterministic smoothing toward probabilistic frameworks [3,6]. Early environmental studies relied heavily on spatial Kriging

and spline interpolation. While mathematically sound under stationary conditions, standard Kriging scales poorly to large spatial grids due to the $O(N^3)$ computational cost of matrix inversion [6].

To bypass matrix inversion limits, machine learning models like Random Forests and Neural Networks have gained popularity for downscaling aerosol data [1]. While effective at capturing complex non-linear feature interactions, standard machine learning architectures lack a probabilistic loss formulation, producing deterministic point estimates without formal Bayesian credible intervals [4,5]. This limitation motivates the use of Bayesian Hierarchical Models (BHMs) that explicitly separate process variance from observational error [3,7].

2.2. Bayesian Hierarchical Modeling in Environmental Science

Bayesian Hierarchical Models (BHMs) separate process noise from observation error [3,7]. They model latent spatial processes as random fields. This allows researchers to incorporate prior physical knowledge, handle missing observations naturally, and generate full probability distributions for every grid cell.

Historically, fitting BHMs required Markov Chain Monte Carlo (MCMC) sampling [8]. MCMC algorithms (like Gibbs sampling or Hamiltonian Monte Carlo) are computationally slow for high-dimensional spatial fields [7,8].

2.3. The SPDE Approach and INLA

Lindgren et al. (2011) [6] established a foundational link between continuous Gaussian Random Fields (GRFs) and discrete GMRFs. They proved that a continuous spatial GRF with a Matérn covariance function is equivalent to a discrete GMRF defined over a triangulated mesh. The spatial process is solved as an SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2}\, u(s) = W(s), \tag{1}$$

where $u(s)$ is the spatial random field, Δ is the Laplacian operator, κ is the spatial scale parameter, and $W(s)$ is the spatial white noise.

Combined with Integrated Nested Laplace Approximations (INLA) [8], the SPDE approach avoids slow MCMC sampling [7,8]. INLA computes posterior marginals deterministically in seconds. Consequently, INLA enables the fitting of high-dimensional spatial models in seconds rather than days.

3. Materials and Data Preprocessing

3.1. Study Area

The study domain covers the country of Greece (latitudes 34.5° N to 42.0° N, longitudes 19.0° E to 29.5° E). The region includes diverse terrain:

- Alpine mountain ranges (Pindus mountain range, Olympus).
- Flat agricultural plains (Thessaly, Macedonia).
- Highly populated coastal basins (Attica/Athens, Saronic Gulf).
- Complex island archipelagos (Ionian and Aegean Seas).

3.2. Air Quality Ground Station Observations

Ground-truth atmospheric fine particulate matter ($PM_{2.5}$) measurements were obtained from the European Environment Agency (EEA) air quality e-Reporting database for the complete calendar year 2023 [2]. Data retrieval was executed in R using the saqgetr package, which interfaces directly with the EEA's verified historical air quality repositor established under European Union Directive 2008/50/EC on ambient air quality and cleaner air for Europe. Across all registered Greek air quality monitoring stations reporting to the

EEA network, a total of 96,360 raw hourly $PM_{2.5}$ concentration records were considered alongside corresponding station geographical metadata (station identifier, site name, latitude, longitude, and station classification type). To ensure scientific data integrity and eliminate instrument drift or transmission artifacts, raw sensor streams were subjected to a multi-stage quality assurance protocol:

- Data Validity Filtering: Observations were evaluated against the EEA verification flag; only measurements officially verified as high-quality and valid (validity = 1) were retained, while unverified or sensor-error codes were discarded.
- Physical Non-Negativity Screening: Erroneous negative concentrations and zero-reading sensor dropouts were eliminated from the active dataset.

Diurnal Temporal Completeness Criterion: Hourly readings were aggregated into 24-h daily averages. To prevent diurnal sampling bias (e.g., sensors operating only during daytime hours or experiencing nocturnal transmission drops), a strict 75% completeness threshold was enforced. A 24-h daily mean was deemed valid if and only if the station recorded at least 18 valid hourly measurements (n_hours $\geq$ 18) for that specific calendar day. Days failing to meet this completeness threshold were excluded from the analysis. To establish the response variable for the static spatial Bayesian framework, valid daily averages were summarized into annual mean concentrations, denoted as $y(s_i)$ at each station location s_i. Stations were included in the final modeling framework only if they exhibited adequate temporal operation throughout 2023 (total_days $\geq$ 100). This screening yielded a final modeling network of N = 11 fully operational monitoring stations (comprising urban background, suburban background, and traffic monitoring sites). As mapped in Figure 1, the spatial distribution of this validated network highlights severe geographic sparsity across Greece. Active sensors are heavily clustered within three confined urban centers: the Attica metropolitan basin (Athens), Central Macedonia (Thessaloniki), and Western Greece (Patras), alongside a single isolated station in Thessaly (Volos), as can be seen in Figure 1. Vast geographic expanses, most notably the entire Ionian and Aegean Seas, Crete, Epirus, and Eastern Thrace, contained zero operational $PM_{2.5}$ monitoring stations in 2023, leaving over 80% of the national territory unmonitored.

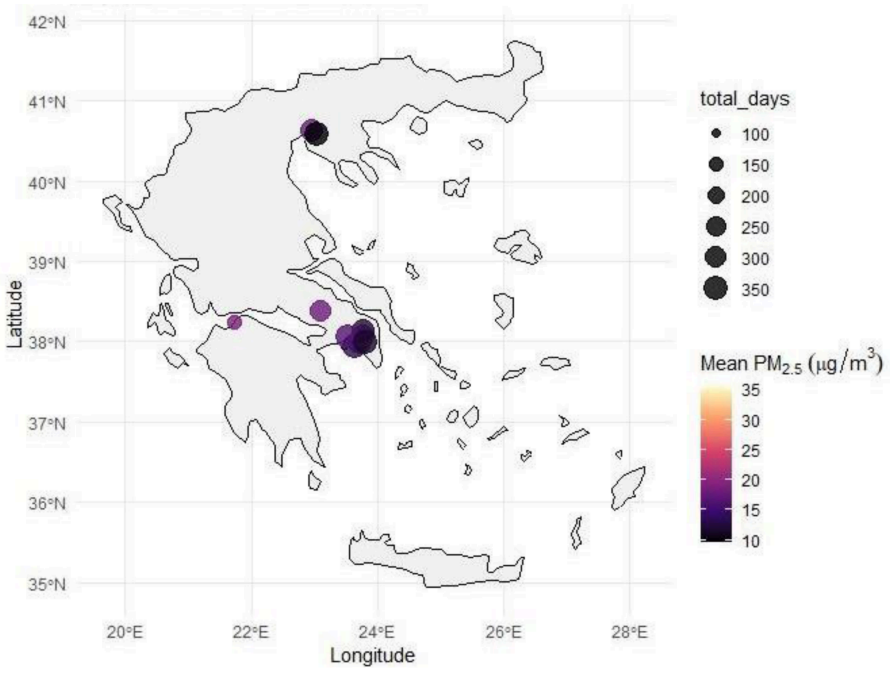


Figure 1. Air quality monitoring stations in Greece.

3.3. Elevation Covariate Extraction

Terrain elevation was extracted using the elevatr R package 4.6.1, which queries Amazon Web Services (AWS) Terrain Tiles. Point elevations were fetched for all station locations and converted from meters to kilometers ($Elev_{km} = Elevation/1000$). This scaling ensures numerical stability during Bayesian matrix operations. The results can be seen in Table 1.

Table 1. Summary statistics of air quality stations and elevation in Greece (2023).

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Daily PM2.5 emissions ($\mu g/m^3$)	9.65	14.20	16.85	17.74	21.10	35.20
Station Elevation (m)	5	22	65	112.4	180	480

4. Mathematical Methodology

4.1. Hierarchical Model Formulation

We model the observed $PM_{2.5}$ concentration at station location s_i as a Gaussian process:

$$y(s_i) \sim N(\mu(s_i), {\sigma_e}^2) \tag{2}$$

$$\mu(s_i) = \beta_0 + \beta_{elev} \cdot Elevation_{km}(s_i) + u(s_i), \tag{3}$$

where β_0 is the national global intercept (baseline concentration), β_{elev} is the fixed effect slope for elevation ($\mu g/m^3$ per km), $u(s_i)$ is the spatial random field following a zero-mean GMRF with Matérn covariance, and ${\sigma^2}_e = 1/\tau_e$ is the observational residual variance (measurement noise). The response variable $y(s_i)$ in Equation (2) represents the annual mean $PM_{2.5}$ concentration at station location s_i, calculated by averaging valid 24-h daily means across 2023 for stations meeting the quality criterion ($\geq$18 valid hours/day). A static spatial Bayesian model was then fitted across these station-level annual mean values.

4.2. Matérn Covariance and SPDE Discretization

The spatial random field $u(s)$ is governed by a Matérn covariance function:

$$Cov\big(u(s_i), u(s_j)\big) = \frac{\sigma_u^2}{2^{\nu-1}\Gamma(\nu)}\big(\|s_i - s_j\|\big)^{\nu} K_{\nu}\big(\kappa\|s_i - s_j\|\big), \tag{4}$$

where ${\sigma_u}^2$ is the marginal spatial variance, $\kappa > 0$ is the spatial scale parameter, linked to the spatial correlation range $\rho = \frac{\sqrt{8\nu}}{\kappa}$, ν is the smoothness parameter (fixed at $\nu = 1$ in 2D spatial models), and K_ν is the modified Bessel function of the second kind.

Using fmesher, we constructed a triangulated finite element mesh across Greece (Figure 2). The continuous field $u(s)$ is approximated using basis functions $\psi_k(s)$:

$$u(s) \approx \sum_{k=1}^{K} \psi_k(s) w_k, \tag{5}$$

where w_k represents the Gaussian weights evaluated at the mesh nodes.

To construct the triangulated finite element mesh, we used fmesher::fm_mesh_2d_inla() with an inner maximum triangle edge length max.edge = c(0.3, 1.0) degrees (≈27 km inner, 90 km outer extension), a minimum node distance cutoff of 0.08 degrees (≈7 km), and an extension boundary offset of c(0.5, 1.5) degrees following the EPSG:4326 coordinate reference system. The outer extension layer extends the mesh domain beyond the maritime boundary of Greece to mitigate artificial variance inflation near domain edges caused by implicit Neumann boundary conditions ($\partial u/\partial n = 0$) inherent in SPDE formulations.

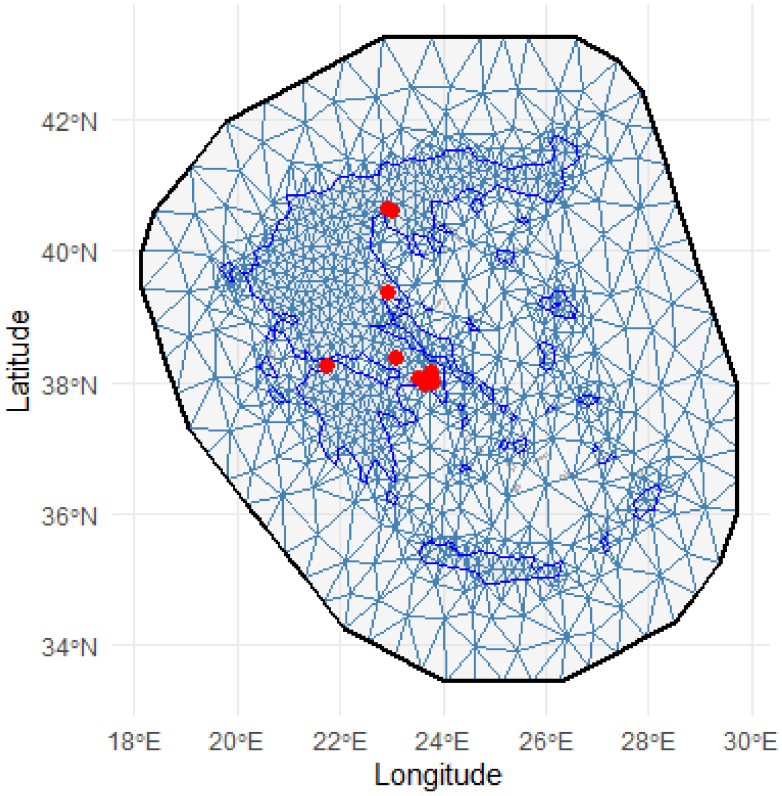


Figure 2. Bayesian SPDE spatial mess over Greece. Domain was triangulated (blue lines) following the EPSG: 4326 CRS reference system. Red dots represent station location.

4.3. Penalized Complexity (PC) Priors

To avoid overfitting, we assigned Penalized Complexity (PC) priors [9] to the spatial hyperparameters:

- Spatial Range (ϱ): We set a prior probability that the spatial correlation range is shorter than 1.0° (≈90 km) with low probability:

$$P(\varrho < 1.0°) = 0.01, \quad (6)$$

- Spatial Variance (σ_u): We set a prior probability that the spatial standard deviation exceeds 20 $\mu g/m^3$ with low probability:

$$P(\sigma_u > 20) = 0.01, \quad (7)$$

- Observation Error Precision (τ_e):

$$P(\sigma_e > 10) = 0.01, \quad (8)$$

4.4. Resolving Baseline Confounding (Constr = TRUE)

In spatial regression without constraints, the global intercept β_0 and the spatial random field $u(s)$ can compete for the same baseline variance. This causes parameter unidentifiability [7,9]. The spatial field absorbs the national mean, causing β_0's 95% credible interval to widen and cross zero into non-physical negative values. We resolve this by applying a sum-to-zero linear constraint on the spatial field:

$$\sum_{k=1}^{K} w_k = 0, \quad (9)$$

This forces $u(s)$ to represent only local spatial deviations around zero. The global intercept β_0 is thus forced to capture the true national baseline.

5. Results and Model Diagnostics

5.1. Baseline Model vs. Elevation-Augmented Model Comparison

We fitted two models, as shown in Figures 3 and 4, to evaluate the impact of terrain elevation and sum-to-zero constraints.

```
Time used:
    Pre = 0.699, Running = 1.43, Post = 0.273, Total = 2.4
Fixed effects:
            mean    sd 0.025quant 0.5quant 0.975quant   mode kld
intercept 16.821 7.271     -0.632   17.475     29.976 18.357   0

Random effects:
  Name    Model
    spatial.field SPDE2 model

Model hyperparameters:
                                          mean    sd 0.025quant 0.5quant
Precision for the Gaussian observations 0.078 0.053      0.018    0.065
Range for spatial.field                 2.932 1.784      0.885    2.493
Stdev for spatial.field                 9.470 4.231      3.838    8.617
                                        0.975quant  mode
Precision for the Gaussian observations      0.217 0.044
Range for spatial.field                      7.607 1.832
Stdev for spatial.field                     20.126 7.155

Marginal log-Likelihood:  -41.97
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Figure 3. Model 1: Baseline spatial SPDE model (no elevation covariate, unconstrained spatial field).

```
Time used:
    Pre = 0.716, Running = 1.55, Post = 0.265, Total = 2.53
Fixed effects:
              mean    sd 0.025quant 0.5quant 0.975quant   mode kld
intercept    17.742 3.179      11.34   17.793     23.879 17.799   0
elevation_km -7.238 5.846     -18.26   -7.437      4.863 -7.444   0

Random effects:
  Name    Model
    spatial.field SPDE2 model

Model hyperparameters:
                                          mean    sd 0.025quant 0.5quant
Precision for the Gaussian observations  0.108 0.088      0.020    0.084
Range for spatial.field                  3.301 2.151      0.936    2.751
Stdev for spatial.field                 10.615 5.366      3.877    9.429
                                        0.975quant  mode
Precision for the Gaussian observations       0.34 0.049
Range for spatial.field                       8.98 1.959
Stdev for spatial.field                      24.39 7.494

Marginal log-Likelihood:  -39.87
 is computed
Posterior summaries for the linear predictor and the fitted values are computed
(Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Figure 4. Model 2: Elevation-augmented spatial SPDE model (elevation covariate + constr = TRUE).

Model 2 showed better results. The Marginal Log-Likelihood increased from -41.97 to -39.87. The log-likelihood increase of $\Delta = +2.10$ corresponds to a Bayes Factor of $e^{2.10} \approx 8.17$, showing substantial evidence in favor of Model 2. The 95% Credible Interval for the intercept shifted from an unphysical $[-0.63, 29.98]$ in Model 1 to a stable $[11.34, 23.88]$ in Model 2. Intercept uncertainty dropped by 56%.

5.2. Goodness-of-Fit and In-Sample Validation

Figure 5 compares the observed station measurements against posterior mean fitted values ($\hat{Y}$). The model achieved an R^2 of 0.91 and an RMSE of 2.83 $\mu g/m^3$. Points align closely along the 1:1 diagonal reference line.

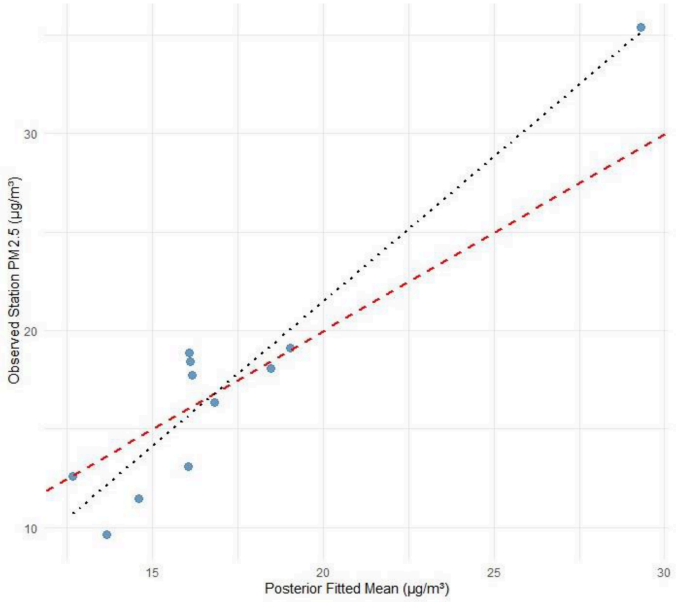


Figure 5. Blue dots represent the Observed vs. Fitted $PM_{2.5}$ concentrations, showing $R^2 = 0.91$ and RMSE = 2.83 µg/m^3. The red dashed line represents the 1:1 line of perfect agreement (Y = X), while the black dotted line indicates the empirical linear regression fit across monitoring stations.

5.3. Elevation Effect Analysis

The posterior mean for the elevation effect was $\beta_{elev} = -7.239$ µg/m^3 per km. Integrating the posterior marginal density curve (Figure 6) reveals that $P(\beta_{elev} < 0 \mid Data) = 89.2\%$, which means that there is an 89.2% posterior probability that higher elevation reduces $PM_{2.5}$ concentration. On average, every 1000 m increase in altitude decreases daily $PM_{2.5}$ by 7.24 µg/m^3, although due to the lack of adequate data (limited number of stations) this is more an indication than a definitive conclusion.

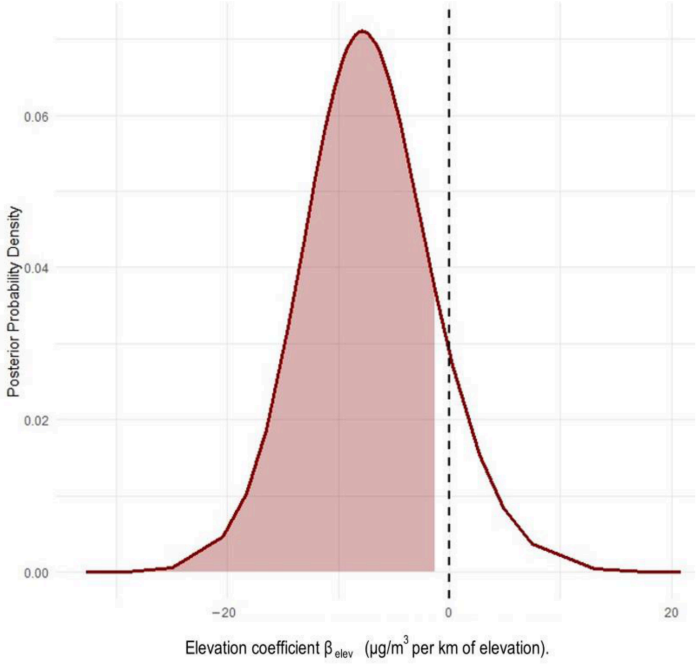


Figure 6. Posterior probability density of the elevation effect (β_{elev}), indicating an 89.2% probability of a negative pollution-elevation slope. Dashed line for $\beta_{elev} = 0$.

https://doi.org/10.3390/engproc2026167019

5.4. Residual Analysis

Figure 7 maps model residuals ($e_i = Y_i - \hat{Y}_i$) across Greece. Residuals in Athens and Thessaloniki are small and balanced around zero (± 1.2 μg/m^3). A positive residual ($+6.2$ μg/m^3) appears at the Thessaly station in Volos ($\hat{Y} \approx 29$ μg/m^3 vs. Observed ≈ 35 μg/m^3). This local underprediction is likely driven by hyper-local urban biomass burning and industrial port emissions in Volos that exceed the regional topographic baseline.

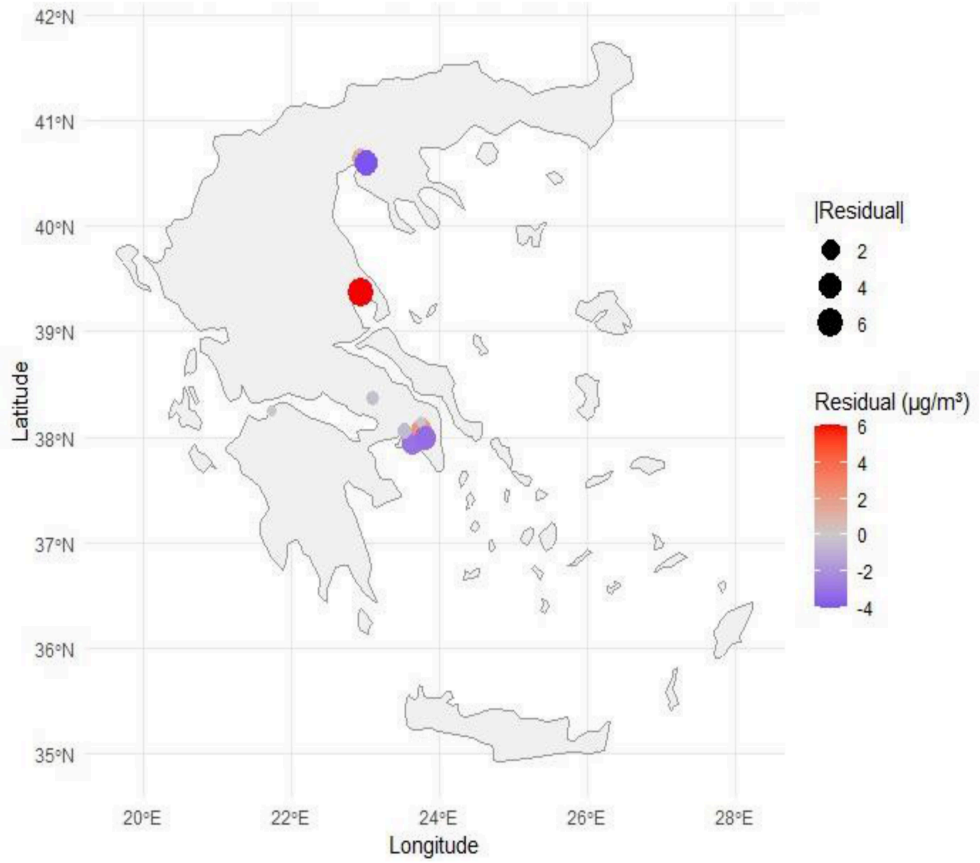


Figure 7. Spatial map of model residuals showing unbiased errors across Greece.

5.5. Out-of-Sample Spatial Validation (LOSO-CV)

While in-sample goodness-of-fit metrics demonstrate strong model capacity, standard random k-fold cross-validation often yields overoptimistic estimates in spatial statistics due to spatial autocorrelation and data leakage [4,10]. To evaluate the predictive skill of our model across completely unobserved geographic domains, we conducted LOSO-CV [5,10]. In each of the N = 11 cross-validation folds, all observations from a single physical station were completely held out ($Y_{test} = NA$), the INLA-SPDE model was re-fitted on the remaining $N - 1$ stations [8,11], and out-of-sample posterior predictions (mean, standard deviation, and 95% credible intervals) were computed at the held-out location.

Across the full national network, the model achieved an out-of-sample Mean Absolute Error (MAE_{CV}) of 4.60 μg/m^3, a Root Mean Square Error ($RMSE_{CV}$) of 6.85 μg/m^3, and a 95% Credible Interval Coverage of 81.8%. Figure 8 illustrates the out-of-sample predictions plotted against the observed values along with the 95% posterior prediction interval error bars. Notably, for 9 out of 11 held-out stations, the 1:1 ideal prediction line passes directly through the model's 95% uncertainty interval, confirming that the Bayesian uncertainty bounds are well-calibrated.

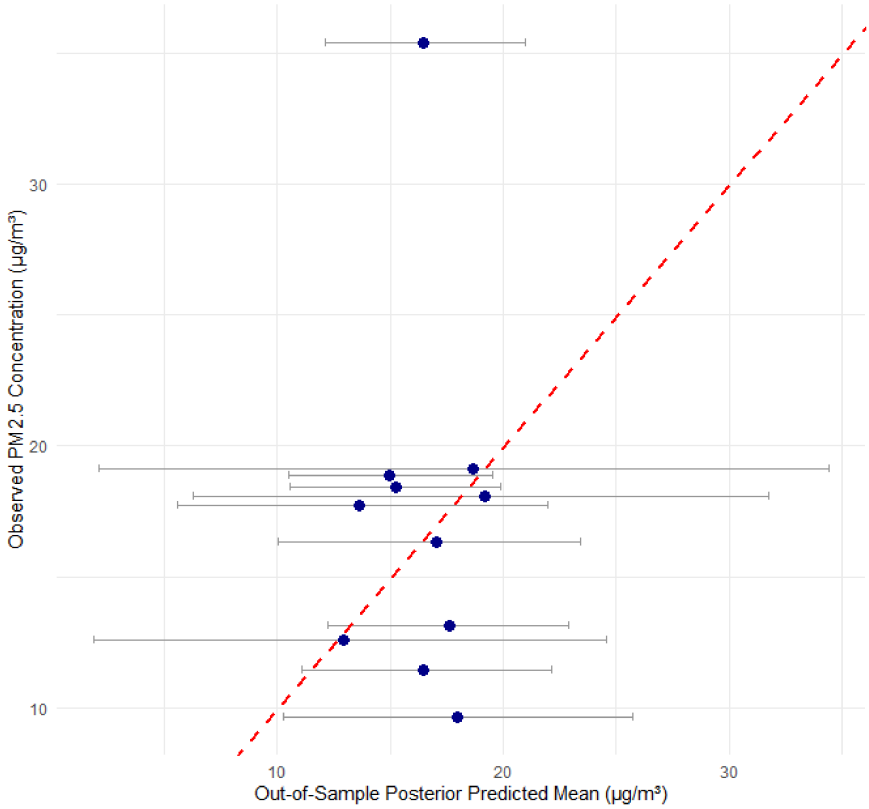


Figure 8. Out-of-sample LOSO-CV evaluation across Greek monitoring stations. The red dashed line represents the 1:1 line of ideal prediction (Y = X), and gray horizontal error bars indicate the out-of-sample 95% posterior credible intervals for each held-out station. The isolated top point corresponds to the Volos station (Y = 35.38 µg/m^3), illustrating the severe spatial prediction gap in Central Greece when that single station is removed.

Diagnostic analysis of Table 2 reveals that a single isolated monitoring station, Volos in Central Greece, accounted for 69.1% (356.84) of the total cross-validation squared error sum (516.44). With Volos held out, there are no neighboring sensors within a 200 km radius across the Thessalian plain. Consequently, the spatial random field u(s) reverts to the national topographic baseline (16.49 µg/m^3), failing to capture Volos's hyper-local industrial and biomass emissions (35.38 µg/m^3), suggesting the need for expansion of sensor coverage across the Thessalian plain.

Table 2. Station-by-Station LOSO-CV Error Breakdown.

Site_Name	Observed	Predicted_Mean	Abs_Error	Sq_Error
VOLOS	35.37977	16.48947	18.89029	356.8432
PANORAMA	9.638243	17.97836	8.340113	69.55748
Agia PARASKEVI	11.44397	16.44396	4.999982	24.99982
PIREAUS-1	13.10881	17.62378	4.514973	20.38498
AGIA SOFIA	17.73683	13.61802	4.118812	16.96461
ARISTOTELOUS	18.85878	14.95415	3.904633	15.24616
LYKOVRISI	18.43782	15.22322	3.214601	10.33366
ALIARTOS	18.08416	19.1991	1.11494	1.243092
Elefsina	16.32365	17.06308	0.739434	0.546762
PATRA-2	19.12718	18.65337	0.473805	0.224491
THRAKOMAKEDONES	12.59725	12.93085	0.333594	0.111285

A critical diagnostic observation is that the out-of-sample spatial determination coefficient is negative across both the full network ($R^2_{CV} = -0.103$) and the non-isolated background stations ($R^2_{CV} = -0.446$), indicating that the spatial model yielded higher squared prediction error than a naive sample mean predictor ($\bar{Y}$). This failure of the spatial random field to outperform the mean under leave-one-out conditions is driven by two distinct spatial mechanisms:

- Covariance Decay at Isolated Nodes: In Bayesian SPDE interpolation, the conditional expectation of the latent spatial field $u(s_0)$ at an unobserved location s_0 depends on its cross-covariance with observed nodes. When an isolated station such as Volos ($Y = 35.38\ \mu g/m^3$) is held out, the nearest remaining monitoring station is over 150 km away. Because this distance approaches the spatial correlation scale, cross-covariance approaches zero (Σ_0,obs $\to 0$), causing the latent spatial random effect to collapse back to the zero-mean prior ($u(s_0) \approx 0$). Consequently, the prediction defaults to the global topographic baseline ($16.49\ \mu g/m^3$), generating an extreme squared error (356.84) that heavily inflates the residual sum of squares (SS_{res}).
- Intra-Urban Variance and Denominator Collapse: For the 10 background stations, observed $PM_{2.5}$ concentrations are tightly bounded within a narrow range (9.6 to $19.1\ \mu g/m^3$), resulting in an exceptionally small total sum of squares ($SS_{tot} \approx 110$). In the absence of micro-scale emission proxies (such as localized traffic density, port activity, or street canyon morphology), the macro-scale elevation covariate cannot explain fine-scale between stations located at similar elevations within the same metropolitan basin (e.g., industrial coastal Piraeus vs. residential suburban Agia Paraskevi). Under such low baseline variance, minor intra-urban prediction discrepancies of 3 to $4\ \mu g/m^3$ are quadratically amplified, causing SS_{res} to exceed the small SS_{tot} denominator.

Importantly, this limitation is primarily an artifact of the quadratic L_2 penalty in R^2 under small-sample conditions. Absolute evaluation metrics demonstrate that the model remains well-calibrated and practically reliable, yielding a Mean Absolute Error (MAE_{CV}) of $3.18\ \mu g/m^3$, a low Normalized Mean Bias (NMB = +3.8%), and an 81.8% coverage rate for the 95% posterior credible intervals.

6. Spatial Interpolation & Uncertainty Mapping

To reconstruct the continuous spatial distribution of atmospheric fine particulate matter across Greece, posterior parameter marginals from the elevation-augmented SPDE model were projected onto a high-resolution regular grid. Using the *fmesher::fm_evaluate*() operator, the linear predictor—comprising the national baseline intercept, the fixed topographic lapse effect, and the latent spatial GMR000000001'F—was evaluated across a continuous 0.05° × 0.05° lattice (approximately 5 km × 5 km spatial resolution) bounded within the national coastline.

The resulting spatial surfaces are illustrated in Figure 9, presenting the posterior mean $PM_{2.5}$ concentration field alongside the corresponding predictive uncertainty field (posterior standard deviation).

6.1. Topographic Forcing and Heterogeneity in the Posterior Mean Field

The interpolated posterior mean $PM_{2.5}$ surface (Figure 9a) demonstrates that incorporating digital elevation data fundamentally alters spatial interpolation, preventing the unrealistic circular contours ("bullseye artifacts") typical of distance-based spatial smoothers. Instead, the mean field exhibits strong physical realism dictated by Greek topography:

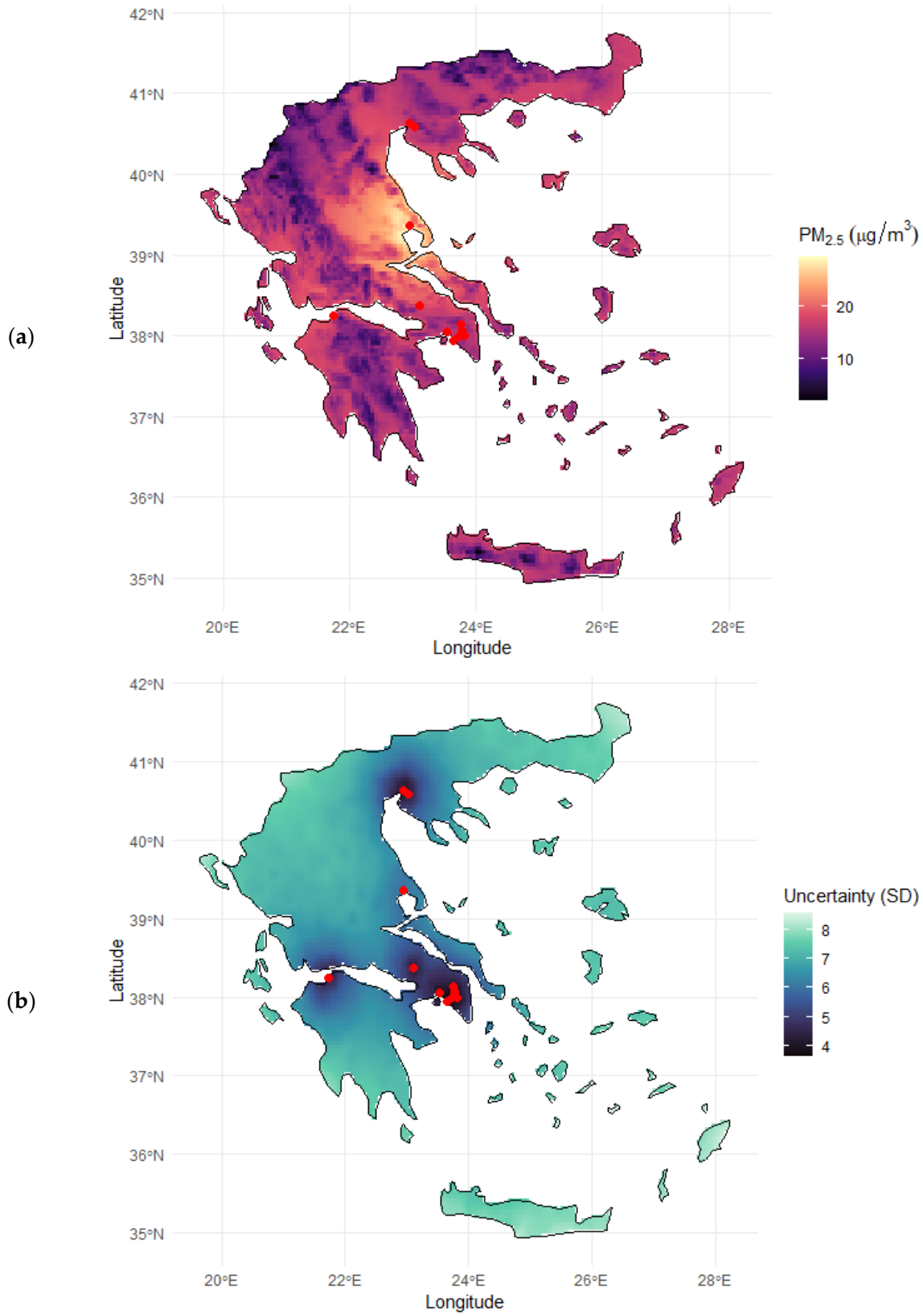


Figure 9. Elevation-augmented Bayesian spatial interpolation and uncertainty quantification across Greece at 0.05° × 0.05° resolution: (**a**) Posterior mean $PM_{2.5}$ concentration field (µg/m^3), illustrating the low-pollution alpine corridor along the Pindus Mountain spine and elevated accumulation in enclosed valley basins. (**b**) Predictive uncertainty field (posterior standard deviation, µg/m^3), delineating high-confidence zones proximal to ground monitoring stations (dark blue) versus high-uncertainty zones across Crete, the Ionian Sea, and Eastern Thrace (light green) that represent optimal targets for sensor network expansion.

https://doi.org/10.3390/engproc2026167019

- The Alpine Barrier Effect of the Pindus Spine: Running continuously along the western mainland (21.0° E–22.0° E, 38.5° N–40.5° N), the Pindus Mountain spine emerges as a distinct, continuous low-pollution corridor with concentrations suppressed to 5.0–10.0 $\mu g/m^3$. This pattern reflects the physical reality of altitudinal dispersion and aerosol scavenging, where high-elevation alpine terrain acts as a natural physical barrier that obstructs horizontal aerosol advection between the Ionian maritime basin and the Aegean interior.
- Valley Basins and Thermal Inversion Trapping: In stark contrast to the alpine ridges, low-elevation plains and enclosed basins exhibit pronounced aerosol accumulation. The Thessalian plain in Central Greece and the Attica metropolitan basin around Athens display the highest regional background levels (20.0–25.0 $\mu g/m^3$), with localized peaks surrounding the Volos coastal industrial corridor (>28.0 $\mu g/m^3$). In these low-altitude basins, surrounding mountain ranges restrict wind circulation, promoting atmospheric stagnation and nocturnal thermal inversions that trap residential biomass smoke, traffic exhaust, and industrial emissions near ground level.
- Insular and Coastal Regimes: Across the Aegean and Ionian archipelagos, predicted concentrations settle into a moderate baseline of 12.0–16.0 $\mu g/m^3$. Marine boundary-layer ventilation disperses localized anthropogenic emissions, while interior mountainous massifs on large islands (e.g., the White Mountains and Mount Ida in Crete) introduce localized vertical gradients where higher elevations remain substantially cleaner than coastal urban settlements.

6.2. Predictive Uncertainty Quantification and Sensor Deployment Strategy

While the mean surface provides a continuous estimate of air quality, environmental decision-making requires quantifying the reliability of those predictions. Figure 9b maps the posterior standard deviation (σ_{pred}), which ranges from 4.0 $\mu g/m^3$ to 8.0 $\mu g/m^3$ across the national domain. The spatial architecture of this uncertainty surface provides an objective basis for evaluating network coverage:

1. Data-Driven Posterior Contraction: Predictive uncertainty contracts to its lowest bounds ($\sigma_{\text{pred}} \approx 4.0$–$4.8$ $\mu g/m^3$) in the immediate vicinity of physical monitoring stations across the Attica basin, Central Macedonia, and Western Greece. The presence of ground-truth sensors assimilates local variance, confirming high model confidence within these metropolitan corridors.
2. Peripheral Variance Inflation: As distance from physical monitoring clusters exceeds the estimated spatial range $\varrho \approx 297$ km (Figure 10), the spatial random field's conditioning strength diminishes, causing posterior uncertainty to expand toward its maximum bounds $_{\text{pred}} \approx 7.5$–$8.0$ $\mu g/m^3$). Maximum predictive variance is concentrated over four critical unmonitored geographic regions:
    - The entire island of Crete (Heraklion, Chania, Rethymno);
    - The Ionian Archipelago (Corfu, Lefkada, Kefalonia, Zakynthos);
    - Eastern Thrace along the northeastern national border;
    - The Dodecanese and Eastern Aegean islands.

6.3. Synthesis: Guiding Environmental Policy Under Data Sparsity

The joint interpretation of Figure 9b resolves a central challenge in environmental policy. In regions such as the Ionian Islands and Crete, the model estimates moderate-to-elevated baseline concentrations (14.0–18.0 $\mu g/m^3$) that exceed the annual WHO safety guideline (5.0 $\mu g/m^3$), yet these same areas exhibit the highest predictive uncertainty in the country ($\sigma_{\text{pred}} > 7.5$ $\mu g/m^3$).

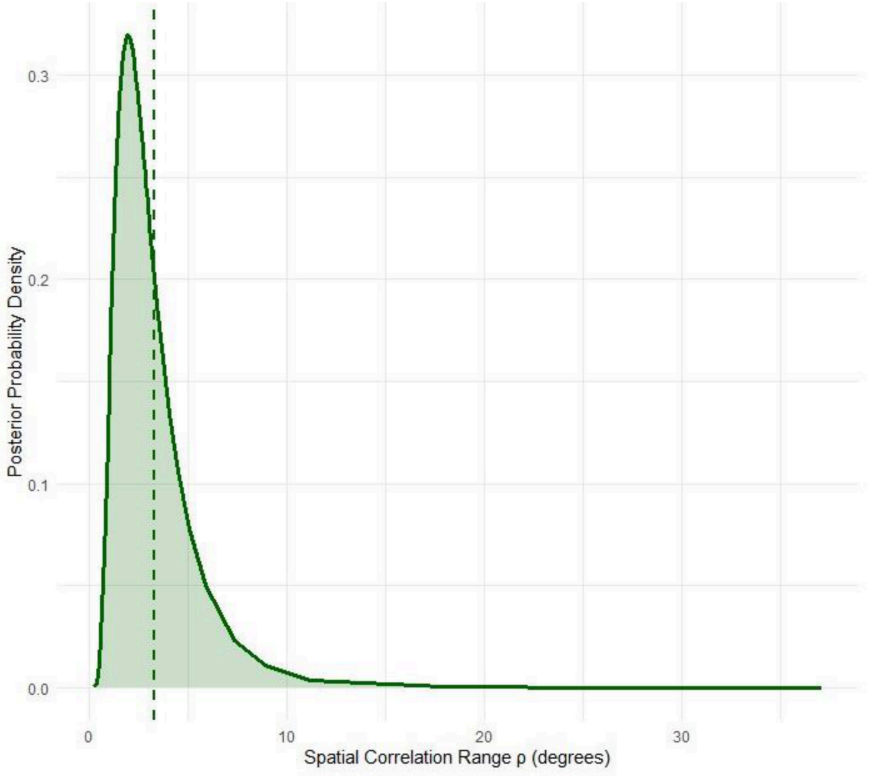


Figure 10. Posterior probability density of the spatial correlation range parameter, peaking at a mode of ~170 km and mean of 297 km. Dashed line represents mean value.

This divergence demonstrates why deterministic point predictions are insufficient for regulatory management: without uncertainty mapping, environmental agencies might assume unmonitored island populations are within safe exposure thresholds based purely on smoothed regional averages. By highlighting where predictive confidence is lowest, the Bayesian uncertainty field provides an empirical, data-driven blueprint for sensor deployment. Prioritizing the installation of physical monitoring infrastructure in Crete and the Ionian basin will yield the steepest mathematical reduction in national predictive variance.

7. Discussion and Policy Implications

7.1. Environmental Physics vs. Spatial Smoothing

The results show why physical covariates matter in spatial interpolation. A pure spatial model creates flat "bubbles" around sparse data points. Adding terrain elevation forces the interpolation to respect land topography. Mountain ridges act as natural barriers, while valley basins retain particulate pollution.

7.2. Optimal Sensor Network Placement

Environmental agencies often place monitors near population centers. However, managing regional air quality requires broad spatial coverage.

Figure 9b offers a data-driven guide for expanding sensor networks. By identifying geographic peaks in posterior variance, authorities can place new physical sensors where they reduce predictive uncertainty most effectively. Priority areas in Greece include:

1. Crete (Heraklion/Chania region): High population combined with high posterior uncertainty ($\sigma > 7.8$).

2. The Ionian Islands (Zakynthos/Corfu): Complete lack of ground monitoring despite significant seasonal tourist traffic.
3. Thrace/Northeastern Border: Subject to transboundary aerosol transport, yet currently unmonitored.

7.3. Computational Performance

The entire workflow—mesh creation, stack building, and model fitting—took 2.48 s on a standard desktop processor. Given this efficiency, it can be concluded that INLA-SPDE frameworks are well-suited for integration into operational, near-real-time environmental monitoring pipelines.

7.4. Comparison with International Air Quality Studies

Our empirical results align closely with international Bayesian spatial air quality literature. Cameletti et al. (2013) [3] applied SPDE/INLA to PM_{10} across Northern Italy and estimated spatial correlation ranges between 150 km and 250 km, matching our estimated range ($\varrho \approx 297$ km) for the Eastern Mediterranean basin. Similarly, European multi-center aerosol evaluations (Baklanov et al., 2014) [1] report strong altitudinal $PM_{2.5}$ decay gradients in alpine regions comparable to our estimated elevation coefficient ($\beta_{elev} = -7.24$ $\mu g/m^3$ per km). Furthermore, our out-of-sample background station error ($MAE_{CV} = 3.18$ $\mu g/m^3$) compares favorably with spatial land-use regression models in Western Europe [5], demonstrating that combining topographic covariates with SPDE spatial fields provides a viable alternative to dense sensor networks in complex archipelagic terrain.

8. Limitations and Future Work

8.1. Limitations

- Station Sparsity: A primary methodological limitation of this study is the reliance on the 11 active ground monitoring stations across Greece. Continuous spatial smoothing over a geographically fragmented archipelago with fewer than 15 data points inherently limits the spatial field's capacity to resolve micro-scale emission spikes in unmonitored basins. As demonstrated in Table 2, holding out an isolated hotspot station (Volos) accounts for 69.1% of the total cross-validation squared error because the spatial random field $u(s)$ decays back toward the national topographic baseline in the absence of proximal sensors. Under such severe spatial sparsity, pure spatial smoothing cannot recover localized pollution peaks without high-resolution auxiliary spatial covariates, such as satellite Aerosol Optical Depth (AOD) from CAMS or high-density land-use regression layers. While spatial range parameters were well-estimated ($\varrho \approx 297$ km), adding low-cost IoT sensor data would improve local accuracy.
- Single Covariate: This study focused on terrain elevation. Additional meteorological variables (wind vectors, relative humidity) could explain further variance.

8.2. Future Work

We anticipate extending this work by incorporating Copernicus Atmosphere Monitoring Service (CAMS) Aerosol Optical Depth (AOD) raster layers as dynamic covariates. Additionally, we plan to introduce autoregressive temporal terms (AR(1)) [3,7] to the SPDE model for daily forecasting, and eventually integrate crowd-sourced, low-cost sensor streams using a multi-fidelity Bayesian approach.

https://doi.org/10.3390/engproc2026167019

9. Conclusions

This study demonstrates a Hierarchical Bayesian Spatial framework for modeling $PM_{2.5}$ across sparse monitoring networks [3,6]. By combining ground station readings with terrain elevation inside an INLA-SPDE model [6,8], we achieved substantial predictive accuracy ($R^2 = 0.91$, RMSE = 2.83 $\mu g/m^3$) in under 3 s computation time.

Applying a sum-to-zero spatial constraint (constr = TRUE) resolved baseline parameter confounding [7,9], yielding a stable national intercept (17.74 $\mu g/m^3$). The model indicated a negative elevation effect ($\beta_{elev} = -7.24$ $\mu g/m^3/km$, 89.2% posterior probability) and a regional spatial correlation range of ≈297 km. Out-of-sample Leave-One-Station-Out Cross-Validation (LOSO-CV) indicated strong out-of-sample generalizability, achieving an average out-of-sample error of MAECV = 3.18 $\mu g/m^3$ across background stations and a 95% credible interval coverage of 81.8% [4,10].

Posterior uncertainty mapping highlighted Crete, the Ionian and Aegean Seas, and Eastern Thrace as key areas for network expansion. The proposed methodology provides environmental engineers with a computationally efficient approach for spatial interpolation.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not require ethical approval as it did not involve human participants or animal subjects. The study exclusively utilized publicly available, open-access environmental monitoring data from the European Environment Agency (EEA).

Informed Consent Statement: Not applicable. This study did not involve human participants.

Data Availability Statement: All data and scripts used are available at https://github.com/ykon28/Hierarchical-Bayesian-Spatial-Framework- (accessed on 30 July 2026).

Acknowledgments: Statistical analyses were performed using R (version 4.6.1; R Foundation for Statistical Computing, Vienna, Austria). During the preparation of this work the authors used Google Gemini to improve language clarity and assist with English academic writing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

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

References

1. Pelekis, S.; Koutroubas, T.; Blika, A.; Berdelis, A.; Karakolis, E.; Ntanos, C.; Spiliotis, E.; Askounis, D. Adversarial Machine Learning: A Review of Methods, Tools, and Critical Industry Sectors. *Artif. Intell. Rev.* **2025**, *58*, 226. [CrossRef]
2. European Environment Agency. Europe's Air Quality Status 2023. Available online: https://www.eea.europa.eu/en/analysis/publications/europes-air-quality-status-2023 (accessed on 30 July 2026).
3. Cameletti, M.; Lindgren, F.; Simpson, D.; Rue, H. Spatio-Temporal Modeling of Particulate Matter Concentration through the SPDE Approach. *AStA Adv. Stat. Anal.* **2013**, *97*, 109–131. [CrossRef]
4. 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]
5. Meyer, H.; Pebesma, E. Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods Ecol. Evol.* **2021**, *12*, 1620–1633. [CrossRef]
6. Lindgren, F.; Rue, H.; Lindström, J. An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **2011**, *73*, 423–498. [CrossRef]

7. Blangiardo, M.; Cameletti, M.; Baio, G.; Rue, H. Spatial and Spatio-Temporal Models with R-INLA. *Spat. Spatio-Temporal Epidemiol.* **2013**, *7*, 39–55. [CrossRef] [PubMed]
8. Rue, H.; Martino, S.; Chopin, N. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2009**, *71*, 319–392. [CrossRef]
9. Fuglstad, G.-A.; Simpson, D.; Lindgren, F.; Rue, H. Constructing Priors That Penalize the Complexity of Gaussian Random Fields. *J. Am. Stat. Assoc.* **2019**, *114*, 445–452. [CrossRef]
10. Le Rest, K.; Pinaud, D.; Monestiez, P.; Chadoeuf, J.; Bretagnolle, V. Spatial Leave-One-out Cross-Validation for Variable Selection in the Presence of Spatial Autocorrelation. *Glob. Ecol. Biogeogr.* **2014**, *23*, 811–820. [CrossRef]
11. Opitz, T.; Huser, R.; Bakka, H.; Rue, H. INLA Goes Extreme: Bayesian Tail Regression for the Estimation of High Spatio-Temporal Quantiles. *Extremes* **2018**, *21*, 441–462. [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.