

Sandra De Iaco
Monica Palma
Donato Posa *Editors*

Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques

ECOST-DATA, Bari, Italy, July 3-4, 2025

OPEN ACCESS

 Springer

Springer Proceedings in Mathematics & Statistics

Volume 531

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Sandra De Iaco • Monica Palma • Donato Posa
Editors

Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques

ECOST-DATA, Bari, Italy, July 3-4, 2025

Editors

Sandra De Iaco 
University of Salento
Lecce, Italy

Monica Palma
University of Salento
Lecce, Italy

Donato Posa
University of Salento
Lecce, Italy



Workshop on Exploration of Spatio-Temporal Environmental Conditions ecost-data ecost-data2025
ECoST-DATA Bari, Italy 2025 07 03 2025 07 04 <https://sites.google.com/unisalento.it/ecost-dataproject/workshop>

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-032-17525-0 ISBN 978-3-032-17526-7 (eBook)
<https://doi.org/10.1007/978-3-032-17526-7>

Mathematics Subject Classification: 60, 62

This work was supported by University of Salento.

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this book or parts of it.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

This work is subject to copyright. All commercial rights are reserved by the author(s), whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Regarding these commercial rights a non-exclusive license has been granted to the publisher.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

This volume collects the scientific contributions presented at the Workshop “Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques (ECoST-DATA)”, held at the University of Bari on July 3-4, 2025, as Satellite event of DUST 2025. It is a key result of the ECoST-DATA Project, funded by European Union – NextGenerationEU with the Cascade Open Calls published by ALMA MATER STUDIORUM – University of Bologna, inside the Project GRINS within the PNRR – Mission 4, Component 2, Investment 1.3 “Partnership extended to Universities, Research Centers, Firms and research projects funding”, D.D. 341 of 15/03/2022, CUP: J33C22002910001.

The contents are dedicated to advanced tools and techniques for analyzing and predicting environmental variables with a spatial or spatio-temporal structure. In particular, they include theoretical reviews on spatio-temporal covariance modelling as well as innovative approaches for assessing environmental quality and its effects on climate change by integrating georeferenced data from multiple sources.

Among the primary objectives of this collection, it is worth mentioning: transfer advanced analytic skills for the modelling, estimation, and prediction of spatial and spatio-temporal environmental variables; provide practical guidance on integrating heterogeneous georeferenced data while accounting for measurement error and spatial misalignment; support risk assessment by equipping practitioners with methods to quantify exposure and uncertainty relevant to human health; and inform sustainability planning through models that identify spatially and temporally targeted mitigation or adaptation strategies.

As a result, the contributions are intended for researchers and graduate students in spatial statistics, environmental science, and geosciences; data scientists and practitioners working on environmental monitoring, climate impacts, and public-health risk assessment; policy analysts interested in data-driven sustainability planning.

Regarding the outcomes, readers will gain both the theoretical grounding and the practical toolset necessary to build, validate, and deploy spatio-temporal models for environmental variables, produce reliable forecasts, and translate model outputs into actionable insights for health and sustainability interventions.

The volume is organized in three parts.

In the first one, *Random fields and Covariance modelling*, Posa clarifies the relationship between covariance functions and positive definite functions, emphasizing that Crum's theorem is the necessary generalization of Bochner's theorem. De Iaco discusses on some families of spatio-temporal complex covariance functions for vectorial data. It is also described how these functions can be generalized to the spatio-temporal setting, as well as how some other complex-valued covariance families can be constructed as positive mixtures of infinite components or as a convolution of the real part. Hristopulos presents the hybrid spectral approach for generating physically inspired, non-separable covariance kernels from stochastic ordinary differential equations. Pere et al. propose the stationary subspace analysis to decompose a random field into independent stationary and non-stationary subspaces.

In Part II, *Environmental control and integration*, Stein focuses on the use of modern deep learning methods for the analysis and management of large data volumes. His study explores the use of satellite imagery and time series to analyze air quality and monitor environmental changes. A further challenging approach is presented in Gómez-Hernández et al. to identify atmospheric dust sources, using the ensemble smoother with multiple data assimilation methodology. Capasso discusses the optimization of pollution control within an economic growth system, presenting a spatially structured model. This model analyzes an optimal control problem designed to maximize intertemporal utility while minimizing pollution, the disutility of taxation, and intervention costs. Additionally, the study by Dutillo and Di Battista explores the functional structure and variation of Italian carbon emissions from electricity using functional data analysis. Daily and monthly profiles are constructed using Fourier basis functions and analyzed at national and electricity market zone levels.

The volume concludes with Part III, *Environmental Analysis*, where Jalilian et al. apply a flexible generalized additive model framework for surface ozone prediction, using spatio-temporally varying coefficients to correct biases between satellite and in situ data. Gomes et al. investigate the geostatistical characterization of climate extremes dynamics, focusing on drought. A Random Forest-based short-term alert system is presented for predicting drought severity classes at a monthly scale, demonstrating strong predictive performance for local areas and offering a flexible tool for managing drought-related risks. For environmental quality assessment, the work by Masoumi et al. addresses a gap in biodiversity modelling in the Apulia region by comparing the performance of various machine learning models (Random Forests, Support Vector Machines, Extreme Gradient Boosting, Decision Trees,

and Linear Regression) for assessing the effects of environmental variables on the multilevel biodiversity index. Furthermore, in Cappello et al., spatio-temporal blind source separation model is reviewed and employed to model and predict ground-level ozone concentration in a complex multivariate spatio-temporal framework.

Lecce, Italy
September 2025

Sandra De Iaco
Donato Posa
Monica Palma

Acknowledgments This research was supported by the European Union-NextGenerationEU with the Cascade Open Calls published by ALMA MATER STUDIORUM - University of Bologna, inside the Project GRINS funded by PNRR - Mission 4, Component 2, Investment 1.3 “Partnership extended to Universities, Research Centers, Firms and research projects funding”, D.D. 341 of 15/03/2022, “ECoST-DATA, Exploring Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques”, CUP: J33C22002910001.

Contents

Part I Random Fields and Covariance Modelling

Covariance Functions	3
Donato Posa	
1 Introduction	3
2 <i>CFs</i> , <i>PD</i> and <i>SPD</i> Functions	4
3 Bochner's Theorem and the Corresponding Generalization	7
4 Conclusions	11
Bibliography	11
Spatiotemporal Complex Covariance Functions for Vectorial Data	15
Sandra De Iaco	
1 Introduction	15
2 Theory of Spatio-temporal Complex Random Fields	16
3 Some Complex Families from Space to Space-Time	19
4 Conclusions	26
Bibliography	26
Non-separable Covariance Kernels for Spatiotemporal Gaussian Processes Based on the Hybrid Spectral Method	29
Dionissios T. Hristopulos	
1 Introduction	29
2 Spatial Covariance Kernels from SPDEs	30
3 Space-Time Covariance Kernels from SPDEs?	34
4 On the Hybrid Spectral Approach	35
5 Conclusions	38
Bibliography	38
Stationary Subspace Analysis for Spatiotemporal Data	41
Jaakko Pere, Sandra De Iaco, and Klaus Nordhausen	
1 Introduction	41
2 Model Formulation	42
3 The stSSA Method for the Mean	44

4	Simulation Study	45
5	Conclusions	50
	Bibliography	51

Part II Environmental Control and Integration

Space Turns to Time: The Advent of Time Series of Remote Sensing Images	55
--	-----------

Alfred Stein

1	Air Quality	55
2	Time Series of Images	57
3	Integrating the Objects	59
4	Ongoing Research	59
5	Looking Forward	60
6	Concluding Remarks	61
	Bibliography	61

Ensemble Smoother with Multiple Data Assimilation for Atmospheric Dust Source Identification: A Generic Framework Approach	63
---	-----------

J. Jaime Gómez-Hernández, Valeria Todaro, Marco D’Oria, and Maria Giovanna Tanda

1	Introduction	64
2	Theoretical Background	65
3	Methodology for Atmospheric Dust Source Identification	66
4	Applications and Advantages	68
5	Future Developments	69
6	Conclusions	69
	Bibliography	70

Optimizing Pollution Control for an Economic Growth System	73
---	-----------

Vincenzo Capasso

1	Introduction	73
2	A Spatially Structured Model	75
3	An Optimal Control Problem	77
	Bibliography	79

Exploring Functional Structure and Variation of Italian Carbon Emissions	81
---	-----------

Pierdomenico Dutillo and Tonio Di Battista

1	Introduction	81
2	Data and Methodology	82
3	Explorative Analysis and FPCA Results	85
	Bibliography	90

Part III Environmental Analysis

Ozone Predictions Through a Generalized Additive Model	95
Abdollah Jalilian, Claudia Cappello, Monica Palma, and Sandra De Iaco	
1 Introduction	96
2 Dataset	96
3 Generalized Additive Model Framework	97
4 Model Evaluation and Results	98
5 Conclusions	101
Bibliography	102
 Characterization and Forecasting of Drought: Geostatistical Approaches	105
Miguel Gomes, Ana Castro, and Amilcar Soares	
1 Objectives	105
2 Study Region and Drought Indices	106
3 Spatial Characterization of the Main Temporal Trends	107
4 An Alert System for Short-Term Prediction of Drought Severity	111
5 Conclusions	114
Bibliography	114
 Using Machine Learning Methods to Explore the Effects of Environmental Variables on Biodiversity	117
Iman Masoumi, Amirhossein Najafabadipour, Sabrina Maggio, and Giuseppina Giungato	
1 Introduction	118
2 Methodology and Data	119
3 Environmental Variables and Biodiversity Prediction	125
4 Conclusions	127
Bibliography	128
 Modeling and Prediction of Ground-Level Ozone Concentrations in a Spatiotemporal Multivariate Context	131
Claudia Cappello, Klaus Nordhausen, and Monica Palma	
1 Introduction	131
2 The ST-BSS Based Approach	133
3 Spatiotemporal Modeling of Ozone Concentrations	133
4 Spatiotemporal Ozone Predictions	137
5 Conclusions	138
Bibliography	138

Ozone Predictions Through a Generalized Additive Model



Abdollah Jalilian, Claudia Cappello, Monica Palma, and Sandra De Iaco

Abstract Accurate prediction of surface-level ozone (O_3) concentrations is essential for ensuring efficient air quality control and safeguarding public health. This study proposes a flexible spatiotemporal modeling framework to estimate daily mean O_3 levels across Italy by integrating satellite-derived ozone data with high-resolution environmental predictors and in situ observations. The model relies on a linear regression with dynamic intercepts and slopes to relate in situ O_3 measurements to satellite estimates, accounting for additive and multiplicative biases. The time- and space-varying parameters are modeled within a generalized additive model (GAM) framework to capture complex relationships between ozone and environmental drivers. Model evaluation and cross-validation demonstrate superior explanatory and predictive performance compared to other alternatives. The framework identifies key bias patterns linked to elevation, nitrogen dioxide concentrations, and seasonal variation. Overall, the proposed model provides an accurate, scalable, and interpretable framework for obtaining harmonized databases, facilitating environmental monitoring and informing policy development.

A. Jalilian

Lancaster Ecology and Epidemiology Group, Lancaster University, Lancaster, UK

C. Cappello (✉)

Department of Economic Sciences, University of Salento, Lecce, Italy

e-mail: claudia.cappello@unisalento.it

M. Palma

Department of Economic Sciences, University of Salento, Lecce, Italy

National Centre for HPC, Big Data and Quantum Computing, Bologna, Italy

S. De Iaco

Department of Economic Sciences, University of Salento, Lecce, Italy

National Centre for HPC, Big Data and Quantum Computing, Bologna, Italy

National Biodiversity Future Center, Palermo, Italy

© The Author(s) 2025

S. De Iaco et al. (eds.), *Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques*, Springer Proceedings in Mathematics & Statistics 531, https://doi.org/10.1007/978-3-032-17526-7_9

Keywords Generalized additive model · Surface-level ozone · Harmonized databases

1 Introduction

Accurate estimation and prediction of surface-level exposure to ozone (O_3) are essential for managing air quality and protecting public health and food security, given that ambient O_3 poses serious risks to human health and negatively affects crop quality and yield [24, 34]. Interpolation techniques and predictive modeling approaches in air quality management generally include physics-based photochemical grid models (e.g., CAMx, as discussed in [10]), data-driven machine learning models (as in [8, 19, 25, 28]), statistical models (as in [12, 13, 30]), even referred to spatiotemporal data (as in [4–7, 26]), or based on dimensionality reduction techniques (as in [2, 21, 22, 27]).

In this context, statistical modeling provides an essential framework for deriving interpretable connections between O_3 levels and environmental variables, along with explicit measures of uncertainty [12]. Land use regression (LUR) models, which typically employ a multiple linear regression framework, are commonly used to relate pollutant concentrations to spatially resolved geographic and environmental predictors [29]. To capture complex and potentially nonlinear spatiotemporal patterns, Generalized Additive Models (GAMs), firstly introduced by Hastie [14, 15], have been widely used in air pollution research, allowing for smooth variation across time [20] and space-time [16, 31].

This study proposes an additive regression model with spatiotemporal varying coefficients to relate daily mean in situ O_3 concentrations to satellite-derived O_3 estimates across Italy. Similarly to LUR models, this approach incorporates a comprehensive set of environmental variables within an additive framework to capture variations in both intercept and slope coefficients in space and time. It is worth noting that the proposed additive structure improves the flexibility of the model, allowing it to capture the potentially nonlinear effects of the covariates within the linear relationship. After describing the dataset used in the case study (Sect. 2), the modeling framework has been introduced in Sect. 3. The model's goodness of fit and its spatial and temporal predictive capabilities have been evaluated through an extensive cross-validation procedure (Sect. 4). The article concludes with the main findings and implications (Sect. 5).

2 Dataset

In this study, ground-based O_3 concentrations measured at Italian air quality monitoring stations have been used as the primary response variable. Explanatory variables have comprised satellite-derived O_3 estimates and environmental predictors representing atmospheric composition, meteorological conditions, topography,

land use, and anthropogenic activity. Data spanning a 10-year period, from January 1, 2013 to December 30, 2023, have been included in the analysis.

2.1 Ozone Monitoring Data (Response Variable)

Hourly O_3 concentrations have been obtained from the European Air Quality Portal maintained by the European Environment Agency (EEA) [11]. In particular, data for 402 Italian air quality monitoring stations have been considered and then aggregated to calculate daily mean O_3 concentrations for the 10-year period. Mean daily O_3 values have exhibited spatiotemporal variability and a right-skewed distribution, characterized by frequent moderate and infrequent high concentrations. To mitigate skewness and approximate normality, a square-root transformation has been applied, consistently with prior studies [9, 18] and supported by Box–Cox analysis [1]. Monthly boxplots of transformed O_3 concentrations have revealed a distinct seasonal cycle, peaking in summer due to meteorological factors like sunlight and temperature [33].

2.2 Covariates

A comprehensive set of spatiotemporal covariates from satellite as well as static covariates has been employed in the analysis:

1. *CAMS Air Quality Data*: daily mean concentrations of O_3 , NO_2 , $PM_{2.5}$, and PM_{10} from the Copernicus Atmosphere Monitoring Service (CAMS) European reanalysis data at a 0.1° resolution (approximately 10 km) [3]. The square-root-transformed CAMS O_3 concentration has shown a strong linear relationship with the in situ data (Pearson correlation coefficient of 0.9139). Note that the empirical variogram of the differences between CAMS and station values has indicated no spatial structure in the discrepancies.
2. *ERA5-Land Climate Variables*: daily means of meteorological variables, including air temperature at 2m, surface net solar radiation, relative humidity, and boundary layer height, from the ERA5-Land dataset [23].
3. *Spatial Environmental/Anthropogenic Variables*: elevation, population density (proxy for anthropogenic emissions), road density (indicator of traffic-related emissions), and land cover/use (influencing biogenic emissions and chemistry).

3 Generalized Additive Model Framework

Let $Y(\mathbf{u}, t)$ be the stochastic process associated with the square root of the mean daily O_3 concentration at location $\mathbf{u} \in \mathbb{R}^2$ and day $t \in \mathbb{Z}$ and $z(c_0(\mathbf{u}), t)$ be the

corresponding CAMS O_3 concentration at the day t and the satellite grid cell $c_0(\mathbf{u})$ that contains the location $\mathbf{u}$. The spatiotemporal process $Y(\mathbf{u}, t)$ has been modeled as

$$Y(\mathbf{u}, t) = \mu(\mathbf{u}, t) + \alpha(\mathbf{u}, t)z(c_0(\mathbf{u}), t) + \epsilon(\mathbf{u}, t), \quad (1)$$

where:

- $\mu(\mathbf{u}, t)$ is a spatially and temporally varying intercept that accounts for systematic differences between satellite-derived and in situ O_3 observations,
- $\alpha(\mathbf{u}, t)$ represents a dynamic slope, enabling the relationship between satellite and in situ O_3 measurements to vary in space and time,
- $\epsilon(\mathbf{u}, t)$ corresponds to an unstructured random error component, assumed to be independent and identically distributed (i.i.d.) following a Student's t -distribution with $\nu > 2$ degrees of freedom.

To enhance interpretability, $\mu(\mathbf{u}, t)$ and $\alpha(\mathbf{u}, t)$ have been related to p environmental covariates (x_j) using an additive structure, forming the full model (M_3):

$$\mu(\mathbf{u}, t) = \beta_0 + \sum_{j=1}^p f_j(x_j(c_j(\mathbf{u}), t)) + s(d(t)) + \tau_{w(t)} \quad (2)$$

$$\alpha(\mathbf{u}, t) = \alpha_0 + \sum_{j=1}^p \tilde{f}_j(x_j(c_j(\mathbf{u}), t)) + \tilde{s}(d(t)) + \tilde{\tau}_{w(t)}. \quad (3)$$

In this Generalized Additive Model (GAM) structure:

- f_j and $\tilde{f}_j$ are smooth functions, implemented using thin plate regression splines, as in [32], capturing potentially nonlinear effects of covariates,
- $s(d(t))$ and $\tilde{s}(d(t))$ are penalized cyclic cubic regression splines, as in [32], based on the day of the year $d(t)$, capturing seasonal variation,
- $\tau_{w(t)}$ and $\tilde{\tau}_{w(t)}$ are Gaussian i.i.d. random effects for the day of the week $w(t)$, accounting for potential weekly patterns like the “weekend ozone effect.”

Parameter estimation is achieved efficiently using the restricted maximum likelihood approach via the `mgcv` package in R [32].

4 Model Evaluation and Results

Model M_3 , with both intercept and slope modeled as additive nonlinear functions with seasonal and weekly components, has been compared with simpler nested models. In particular, the following alternative models have been considered:

- M_0 , with constant intercept $\mu(\mathbf{u}, t) = \beta_0$ and slope $\alpha(\mathbf{u}, t) = \alpha_0$

- M_1 , with linear varying intercept $\mu(\mathbf{u}, t) = \beta_0 + \sum_{j=1}^P \beta_j x_j(\mathbf{u}, t)$ and constant slope $\alpha(\mathbf{u}, t) = \alpha_0$
- M_2 , with seasonal and weekly components $\mu(\mathbf{u}, t) = \beta_0 + \sum_{j=1}^P f_j(x_j(\mathbf{u}, t)) + s(d(t)) + \tau_{w(t)}$ and constant slope $\alpha(\mathbf{u}, t) = \alpha_0$

The inclusion of additive nonlinear structures and dynamic slope coefficients in M_3 has led to a consistent and statistically significant improvement in all goodness of fit metrics (R_{adj}^2 , %Dev, AIC, and BIC) compared to M_0 , M_1 , and M_2 , as confirmed by partial F -statistics ($p < 0.001$ for all comparisons).

The full model M_3 has achieved an adjusted coefficient of determination (R_{adj}^2) of 0.867 and explained 72.5% of the null deviance. Residual diagnostics have supported the assumption of the Student's t -distribution ($\hat{v} = 3.6$) and showed no clear structure against fitted values, supporting homoscedasticity.

4.1 Model Interpretability of Dynamic Coefficients

The dynamic coefficient framework, proposed in Eq. (1), has successfully quantified how specific environmental variables influence the additive bias ($\mu(\mathbf{u}, t)$) and the multiplicative bias ($\alpha(\mathbf{u}, t)$).

- *Elevation* has shown a substantial positive effect on $\mu(\mathbf{u}, t)$, associated with a nonlinear upward shift (up to $6 \mu\text{g}/\text{m}^3$ on the original scale) in satellite estimates relative to in situ observations at higher elevations. In contrast, the influence of elevation on multiplicative bias $\alpha(\mathbf{u}, t)$ was negligible.
- *Nitrogen Dioxide (NO_2)* has shown a positive effect on $\mu(\mathbf{u}, t)$ (upward bias) but a negative effect on $\alpha(\mathbf{u}, t)$ (slight scaling contraction). This implies that high NO_2 conditions are linked to an upward bias in satellite estimates alongside a reduction in the strength of their association with ground measurements.
- *Seasonal Variation (Day of Year)* has affected both $\mu(\mathbf{u}, t)$ and $\alpha(\mathbf{u}, t)$, indicating that seasonal factors influence both the baseline offset and the scaling relationship between the two measurement sources.

4.2 Predictive Performance

Predictive performance has been assessed using three cross-validation (CV) schemes: General k -fold, Spatial CV (testing unobserved locations using k -medoid clustering), and Temporal CV (forecasting using only past data).

In particular, in the general CV each observation has been randomly assigned to one of k mutually exclusive groups, yielding a partition of the dataset into subsets $\mathcal{P}_1, \dots, \mathcal{P}_k$. For each fold $j = 1, \dots, k$, the model has been fitted using all observations except those contained in $\mathcal{P}_j$, after which predictions are generated

for the excluded subset. This general k -fold cross-validation procedure provides an integrated evaluation of the model's predictive capability and its generalization performance across data partitions.

In the spatial CV approach, the 402 air quality monitoring stations have been grouped into k spatial clusters, $\mathcal{C}_1, \dots, \mathcal{C}_k$, using the k -medoids algorithm [17]. These clusters define spatially coherent partitions of the dataset into subsets $\mathcal{P}_1, \dots, \mathcal{P}_k$, where each subset $\mathcal{P}_j$ contains all observations from the stations belonging to cluster $\mathcal{C}_j$ across the full 4016-day study period. For each fold $j = 1, \dots, k$, the observations in $\mathcal{P}_j$ are omitted, the model has been trained on data from the remaining $k - 1$ subsets, and O_3 concentrations are subsequently predicted for the held-out subset. This spatial cross-validation framework is designed to evaluate the model's capability to predict O_3 concentrations at unobserved locations and to assess its spatial generalization performance.

Finally, in the temporal CV the full study period comprising 4016 days has been divided into $k + 1$ consecutive time intervals of equal length, denoted as $\mathcal{I}_0, \mathcal{I}_1, \dots, \mathcal{I}_k$. The dataset has been accordingly partitioned into temporal subsets $\mathcal{P}_0, \dots, \mathcal{P}_k$, where each subset $\mathcal{P}_j$ includes all observations from every monitoring station corresponding to days within interval $\mathcal{I}_j$. For each fold $j = 1, \dots, k$, the model is trained using data from the immediately preceding subset $\mathcal{P}_{j-1}$, and predictions of O_3 concentrations have been generated for the subsequent subset $\mathcal{P}_j$. This temporal cross-validation approach emulates realistic forecasting conditions by restricting model training to past data, thereby preventing temporal information leakage and enabling a rigorous assessment of the model's predictive accuracy over time.

Model predictive performance for each cross-validation strategy has been quantified using the following accuracy metrics [31]: the Root Mean Square Prediction Error (RMSPE), the Mean Absolute Prediction Error (MAPE), and the cross-validated coefficient of determination (CVR^2), defined as follows:

$$\begin{aligned} \text{RMSPE} &= \frac{1}{k} \sum_{j=1}^k \sqrt{\frac{1}{\#(\mathcal{P}_j)} \sum_{(\mathbf{u}, t) \in \mathcal{P}_j} \left(Y^2(\mathbf{u}, t) - \widehat{Y}^2(\mathbf{u}, t) \right)^2}, \\ \text{MAPE} &= \frac{1}{k} \sum_{j=1}^k \left(\frac{1}{\#(\mathcal{P}_j)} \sum_{(\mathbf{u}, t) \in \mathcal{P}_j} \left| Y^2(\mathbf{u}, t) - \widehat{Y}^2(\mathbf{u}, t) \right| \right), \\ \text{CVR}^2 &= \frac{1}{k} \sum_{j=1}^k \left(1 - \frac{\sum_{(\mathbf{u}, t) \in \mathcal{P}_j} \left(Y(\mathbf{u}, t) - \widehat{Y}(\mathbf{u}, t) \right)^2}{\sum_{(\mathbf{u}, t) \in \mathcal{P}_j} \left(Y(\mathbf{u}, t) - \frac{1}{\#(\mathcal{P}_j)} \sum_{(\mathbf{u}, t) \in \mathcal{P}_j} Y(\mathbf{u}, t) \right)^2} \right), \end{aligned}$$

where $Y^2(\mathbf{u}, t)$ represents the observed O_3 concentration at site $\mathbf{u}$ and day t , and $\widehat{Y}^2(\mathbf{u}, t)$ denotes the predicted one. The notation $\#(\mathcal{P}_j)$ corresponds to the number of data points included in subset $\mathcal{P}_j$. Note that since RMSPE squares the prediction

Table 1 Cross-validation performance metrics (in bold the best predictive performance)

Model	Spatial CV			Temporal CV			General CV		
	RMSPE	MAPE	CVR ²	RMSPE	MAPE	CVR ²	RMSPE	MAPE	CVR ²
M ₀	12.024	6.572	0.738	11.862	6.225	0.814	11.880	6.133	0.818
M ₁	11.438	6.268	0.763	11.124	5.922	0.837	11.046	5.792	0.842
M ₂	11.446	6.391	0.762	10.871	5.813	0.844	10.683	5.615	0.853
M ₃	11.382	6.306	0.763	10.744	5.714	0.857	10.505	5.472	0.867

errors, it emphasizes larger discrepancies, whereas MAPE expresses the mean absolute deviation and is generally easier to interpret.

To assess how the varying intercept term (2) and the dynamic slope term (3) contribute to predictive performance, the full model M₃ has been evaluated against three simplified alternatives (M₀–M₂) in predicting O₃ concentrations, $Y^2(\mathbf{u}, t)$. A summary of the results has been provided in Table 1, which reports RMSPE, MAPE, and CVR² values under the spatial, temporal, and general cross-validation frameworks. Under both temporal and general CV, the full model M₃ exhibit superior predictive performance, achieving the lowest RMSPE and MAPE, and the highest CVR². Predicting O₃ concentrations at unobserved spatial locations (Spatial CV) has been confirmed to be more challenging, yielding higher errors compared to temporal CV. Moreover, M₃ achieved the lowest RMSPE (11.382) in Spatial CV, demonstrating superior overall accuracy by reducing the magnitude of larger prediction errors when generalizing to novel locations.

5 Conclusions

This research successfully established a flexible and interpretable modeling framework for surface-level O₃ prediction by employing a Generalized Additive Model with spatiotemporally varying intercept and dynamic slope coefficients. The novelty lies in combining the established GAM methodology with dynamic coefficient modeling, enabling a clear and unified mechanism to understand and adjust for both additive and multiplicative biases in satellite-based O₃ measurements relative to in situ observations. The results from goodness-of-fit metrics and cross-validation strongly demonstrated that incorporating these dynamic components substantially improved model fit and predictive accuracy compared to models assuming constant coefficients. Moreover, by revealing how environmental factors drive the relationship between in situ and satellite observations, the model contributes to obtaining harmonized data, supporting environmental monitoring and evidence-based policy development.

Bibliography

1. Box, G.E., Cox, D.R.: An analysis of transformations. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **26**(2), 211–243 (1964)
2. Cabral Pinto, F., Manchuk, J.G., Deutsch, C.V.: Decomposition of multivariate spatial data into latent factors. *Comput. Geosci.* **153**, 104773 (2021)
3. Copernicus Atmosphere Monitoring Service: CAMS European air quality reanalyses (2021). Accessed on 10 Apr 2025
4. De Iaco, S., Myers, D.E., Posa, D.: Space-time variograms and a functional form for total air pollution measurements. *Comput. Stat. Data Anal.* **41**(2), 311–328 (2002)
5. De Iaco, S., Myers, D.E., Posa, D.: The linear coregionalization model and the product-sum space-time variogram. *Math. Geol.* **35**(1), 25–38 (2003)
6. De Iaco, S., Myers, D.E., Palma, M., Posa, D.: Using simultaneous diagonalization to identify a space-time linear coregionalization model. *Math. Geosci.* **45**(1), 69–86 (2013)
7. De Iaco, S., Palma, M., Posa, D.: Choosing suitable linear coregionalization models for spatio-temporal data. *Stochastic Environ. Res. Risk Assess.* **33**(7), 1419–1434 (2019)
8. De Iaco, S., Hristopulos, D.T., Lin, G.: Special issue: geostatistics and machine learning. *Math. Geosci.* **54**(3), 459–465 (2022)
9. Dou, Y., D. Le, N., V. Zidek, J.: Modeling hourly ozone concentration fields. *Ann. Appl. Stat.* **4**(3), 1183–1213 (2010)
10. Emery, C.A., Baker, K.R., Wilson, G.M., Yarwood, G.: Comprehensive air quality model with extensions, v7. 20: formulation and evaluation for ozone and particulate matter over the US. *Geosci. Model Dev. Discuss.* **2024**, 1–48 (2024)
11. European Environment Agency: European air quality portal. <https://www.eea.europa.eu/themes/air/air-quality-index> (2025). Accessed 10 Apr 2025
12. Gelfand, A.E., Fuentes, M., Hoeting, J.A., Smith, R.L.: *Handbook of Environmental and Ecological Statistics*. CRC Press, Boca Raton (2019)
13. Genton, M., Kleiber, W.: Cross-covariance functions for multivariate geostatistics. *Stat. Sci.* **30**(2), 147–163 (2015)
14. Hastie, T., Tibshirani, R.: Generalized additive models. *Stat. Sci.* **1**(3), 297–310 (1986)
15. Hastie, T.J., Tibshirani, R.J.: *Generalized Additive Models*. Chapman and Hall, London (1990)
16. Jalilian, A., Cappello, C., Palma, M., De Iaco, S.: Generalized Additive Model With Dynamic Coefficients for Spatiotemporal Ozone Predictions. *Environmetrics* **37**(3), (2026)
17. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken (1990)
18. Lu, X., Gelfand, A.E., Holland, D.M.: Local real-time forecasting of ozone exposure using temperature data. *Environmetrics* **29**(7), e2509 (2018)
19. Meng, X., Wang, W., Shi, S., Zhu, S., Wang, P., Chen, R., Xiao, Q., Xue, T., Geng, G., Zhang, Q., et al.: Evaluating the spatiotemporal ozone characteristics with high-resolution predictions in mainland China, 2013–2019. *Environ. Pollut.* **299**, 118865 (2022)
20. Morlini, I., Albertson, S., Orlandini, S.: Characterization of annual urban air temperature changes with special reference to the city of Modena: a comparison between regression models and a proposal for a new index to evaluate relationships between environmental variables. *Stochastic Environ. Res. Risk Assess.* **38**(3), 1163–1178 (2024)
21. Muehlmann, C., De Iaco, S., Nordhausen, K.: Blind recovery of sources for multivariate space-time random fields. *Stochastic Environ. Res. Risk Assess.* **37**, 1593–1613 (2023)
22. Muehlmann, C., Cappello, C., De Iaco, S., Nordhausen, K.: Anisotropic local covariance matrices for spatial blind source separation. *AStA Adv. Stat. Anal.* (2025)
23. Muñoz Sabater, J.: ERA5-Land hourly data from 1950 to present (2019). Accessed on 10 Apr 2025
24. Nuvolone, D., Petri, D., Voller, F.: The effects of ozone on human health. *Environ. Sci. Pollut. Res.* **25**(9), 8074–8088 (2018)

25. Pan, Q., Harrou, F., Sun, Y.: A comparison of machine learning methods for ozone pollution prediction. *J. Big Data* **10**(1), 63 (2023)
26. Rouhani, S., Wackernagel, H.: Multivariate geostatistical approach to space-time data analysis. *Water Resour. Res.* **26**(4), 585–591 (1990)
27. Sipilä, M., Cappello, C., De Iaco, S., Nordhausen, K., Taskinen, S.: Modelling multivariate spatio-temporal data with identifiable variational autoencoders. *Neural Netw.* **181**, 106774 (2025)
28. Sipilä, M., Cappello, C., De Iaco, S., Nordhausen, K., Palma, M., Taskinen, S.: Enhancing identifiable variational autoencoder in the presence of missing values and auxiliary covariates for spatio-temporal ozone modeling (2025). Submitted
29. van Nunen, E., Vermeulen, R., Tsai, M.Y., Probst-Hensch, N., Ineichen, A., Davey, M., Imboden, M., Ducret-Stich, R., Naccarati, A., Raffaele, D., et al.: Land use regression models for ultrafine particles in six European areas. *Environ. Sci. Technol.* **51**(6), 3336–3345 (2017)
30. Wackernagel, H.: *Multivariate Geostatistics: an Introduction with Applications*. Springer, Berlin (2003)
31. Wang, J., Cohan, D.S., Xu, H.: Spatiotemporal ozone pollution LUR models: suitable statistical algorithms and time scales for a megacity scale. *Atmos. Environ.* **237**, 117671 (2020)
32. Wood, S.N.: *Generalized Additive Models: an Introduction with R*. Chapman and Hall/CRC, Boca Raton (2017)
33. Yan, Y., Pozzer, A., Ojha, N., Lin, J., Lelieveld, J.: Analysis of European ozone trends in the period 1995–2014. *Atmos. Chem. Phys.* **18**(8), 5589–5605 (2018)
34. Zhang, J., Wei, Y., Fang, Z.: Ozone pollution: a major health hazard worldwide. *Front. Immunol.* **10**, 2518 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

