TI 2026-093/III
Tinbergen Institute Discussion Paper

Modeling Spatial Pollution Dynamics under Irregular Sampling, Missing Values, and Outliers

Henry de Vries [1]
Yicong Lin [2]
Andre Lucas [3]

1 Vrije Universiteit Amsterdam

2 Vrije Universiteit Amsterdam and Tinbergen Institute

3 Vrije Universiteit Amsterdam and Tinbergen Institute

Modeling Spatial Pollution Dynamics under Irregular Sampling, Missing Values, and Outliers

Henry de Vries*[1], Yicong Lin[1,2], and Andre Lucas[1,2]

[1]Vrije Universiteit Amsterdam

[2]Tinbergen Institute

November, 2026

Abstract

Low-cost air-quality sensor networks enable spatially extensive monitoring of air pollution, but pose substantial challenges for forecasting due to irregular sensor placement, missing observations, and outliers. We propose a functional framework for spatio-temporal modelling of air pollution data collected from low-cost sensor networks. The framework allows for alternative dynamic specifications as well as dimensionality reduction via functional principal component analysis (FPCA). Extensive simulations demonstrate that the new framework is robust against the commonly encountered challenges in low-cost sensor network data. We find that a simple functional autoregressive and a parameter-driven state-space dynamic specification have the strongest predictive performance, both at and outside observed sensor locations. The functional autoregressive approach has the lower computational complexity of the two. The results are confirmed empirically in an application to daily European PM2.5 concentrations from a low-cost sensor network.

Keywords: functional time series, spatio-temporal modelling, irregular sampling, low-cost sensor networks

*Corresponding author: henryjcdevries@outlook.com.

1 Introduction

Air pollution is a major environmental challenge that has profound consequences for human health and ecosystems (Kampa and Castanas, 2008; Lovett et al., 2009; Manisalidis et al., 2020), creating a growing need for accurate monitoring and forecasting of pollutant concentrations. Traditionally, air quality has been monitored using, for instance, ground-based reference networks, satellites, or outputs from chemical transport models, which can involve sparse coverage, limited resolution, or substantial financial and computational costs (Bonas and Castruccio, 2025). In recent years, however, low-cost air-quality sensor networks have increasingly complemented traditional monitoring stations by providing affordable, spatially extensive, and high frequency pollution measurements (Bagkis et al., 2025). Despite their advantages, low-cost sensor networks introduce substantial modelling challenges for three reasons. First, their spatial distribution is highly irregular because sensors are installed by individual users. Second, measurements are frequently missing due to sensor malfunctions, communication failures, or maintenance. Third, low-cost sensors are highly susceptible to measurement errors and outliers compared to regulatory monitoring stations.

These challenges have motivated the development of spatio-temporal forecasting methods tailored to low-cost sensor networks. Recent approaches include machine and deep learning methods for spatio-temporal modelling of air pollution; see, among many others, Iyer et al. (2022) and Agbehadji and Obagbuwa (2024). Although such methods have shown promising predictive performance for spatio-temporal air pollution forecasting, they typically do not explicitly exploit the functional nature of the evolving pollution field. Existing functional forecasting approaches, on the other hand, have primarily focused on densely and regularly observed curves obtained from conventional monitoring stations (Shang, 2017; Shah et al., 2024; Tao et al., 2026). To the best of our knowledge, no functional forecasting framework has been proposed for two-dimensional pollution surfaces observed through irregular low-cost sensor networks.

We therefore develop a functional time series framework for forecasting two-dimensional pollution surfaces using data from irregular low-cost sensor networks. To this end, we represent the latent pollution field as a smooth functional surface that evolves over time. We then model the temporal evolution through a range of dynamic specifications. A straightforward autoregressive model is one example. More advanced frameworks that

better account for measurement noise are also adopted, including observation-driven (Creal et al., 2013; Lin and Lucas, 2025) and parameter-driven models (Durbin and Koopman, 2012). As such, we provide guidance on which type of dynamic specification for the projection coefficients works best in the challenging context of low-cost, noisy sensor network data. We also investigate whether functional principal component analysis (FPCA) can be useful for obtaining a low-dimensional representation of the pollution surface, or whether a less restrictive parameterization provides a better out-of-sample description of the data.

The main contributions of this paper are therefore twofold. First, we propose a functional forecasting framework for two-dimensional pollution surfaces observed through irregular low-cost sensor networks. Second, we investigate the relative performance of alternative observation-driven, parameter-driven, autoregressive, and non-parametric dynamic specifications for this model within a unified framework, both with and without FPCA-based dimension reduction.

The proposed framework is assessed in two complementary settings. First, we conduct an extensive simulation study that mimics the defining characteristics of low-cost sensor networks, including irregular spatial sampling, missing observations, and measurement outliers. Second, we evaluate the methodology using a dataset of daily PM2.5 concentrations measured by a low-cost sensor network across Europe. We find that the functional autoregressive and parameter-driven models achieve the strongest forecasting performance for this type of data, with the former having lower computational complexity. Moreover, the FPCA brings substantial computational gains. Using only a few estimated eigenfunctions, the FPCA yields similar performance, with only a slight decrease in forecasting accuracy compared to an unrestricted B-spline basis using 49 basis functions, across all dynamic specifications considered.

The rest of this paper is organized as follows. In Section 2 we introduce the time-varying surface dynamics, the FPCA dimensionality reduction, and dynamic specifications. Section 3 shows the performance of the considered models in a variety of controlled settings, including the presence of missing observations, irregular sampling, and measurement outliers. Section 4 illustrates the framework in an empirical context. Section 5 concludes. Additional derivations and results are provided in the appendix. The code and data are available at `https://github.com/henrydevries/pollution-modeling`.

2 Methodology

We model an air pollution surface by a location-scale model with time-varying components in Section 2.1, followed by the dynamic specifications of the time-varying components in Section 2.2. Section 2.3 then introduces an FPCA-based dimensionality reduction. To set the stage, vectors $\boldsymbol{y}$ and matrices $\boldsymbol{Y}$ are denoted in bold, whereas scalars y are non-bold. For a square integrable function $f \in L^2(\mathcal{S})$, its $L^2(\mathcal{S})$ norm is denoted by $\|f\|_{L^2} = \sqrt{\langle f, f\rangle_{L^2}}$, where $\langle f, g\rangle_{L^2} = \int_{\mathcal{S}} f(s)g(s)\mathrm{d}s$.

2.1 Location-scale setup

For a sample size $T \in \mathbb{N}$, consider the functional location-scale model

$$y_t(\boldsymbol{s}) = \mu_t(\boldsymbol{s}) + \sigma\, \varepsilon_t(\boldsymbol{s}), \qquad t = 1, \ldots, T, \tag{2.1}$$

where $\boldsymbol{s} = (s_1, s_2)^\top \in \mathcal{S}$ and $\sigma > 0$. In our context, $y_t(\cdot)$ denotes the observed air pollution concentration, while $\mu_t(\cdot) \in L^2(\mathcal{S})$ represents the underlying air pollution concentration surface. Moreover, $\{\varepsilon_t(\cdot),\, t \in \mathbb{Z}\}$ is a mean-zero error process that contains, for instance, measurement errors. Without loss of generality, we assume $\mathcal{S} = [0, 1]^2$ in what follows. As will be seen later, one typically observes $y_t(\boldsymbol{s})$ only at irregularly distributed locations $\boldsymbol{s} \in \mathcal{S}$, which may form a dense or sparse grid over space. Given that we are particularly interested in data obtained from low-cost sensor networks, the spatial error process $\varepsilon_t(\cdot)$ may also be subject to substantial outliers.

In this paper, we are interested in modelling the latent surface dynamics of $\mu_t(\cdot)$. Note that $\mu_t(\cdot)$ is an infinite dimensional random object. A natural starting point is to project it onto a finite-dimensional space spanned by basis functions along both the longitude and latitude dimensions, i.e., $\{\phi_1(\cdot), \ldots, \phi_{K_\ell}(\cdot)\}$ with $K_\ell \in \mathbb{N}$ and $\ell = 1, 2$. For $\boldsymbol{s} = (s_1, s_2)^\top$, this yields the following approximation (Lin and Lucas, 2025):

$$\mu_t(\boldsymbol{s}) \approx \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} \gamma_{i,j,t}\, \phi_i(s_1)\, \phi_j(s_2) = \big(\boldsymbol{\phi}_{K_2}(s_2) \otimes \boldsymbol{\phi}_{K_1}(s_1)\big)^\top \mathrm{vec}(\boldsymbol{\Gamma}_t), \tag{2.2}$$

where $\boldsymbol{\phi}_{K_\ell}(\cdot) = (\phi_1(\cdot), \ldots, \phi_{K_\ell})^\top \in \mathbb{R}^{K_\ell}$, $\boldsymbol{\Gamma}_t = \big(\gamma_{i,j,t},\, 1 \le i \le K_1,\, 1 \le j \le K_2\big)$, $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix, and $\otimes$ denotes the Kronecker product. In this paper, we adopt B-spline basis functions, and other basis functions can also be used as

well.

Defining $\boldsymbol{\gamma}_t = \text{vec}(\boldsymbol{\Gamma}_t) \in \mathbb{R}^K$ with $K = K_1 \cdot K_2$, and omitting the approximation errors in (2.2), the observed function can then compactly be expressed as $y_t(\boldsymbol{s}) = (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{\gamma}_t + \sigma\, \varepsilon_t(\boldsymbol{s})$. To forecast the dynamics of pollution surfaces, we therefore model the latent dynamics of the coefficient vectors $\{\boldsymbol{\gamma}_t\}$. Once a predicted value $\widetilde{\boldsymbol{\gamma}}_{t+1}$ is obtained, one can reconstruct the predicted surface location at time $t+1$ as $\tilde{\mu}_{t+1}(\boldsymbol{s}) = (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \widetilde{\boldsymbol{\gamma}}_{t+1}$. Note that when K is large, the dimension of $\boldsymbol{\gamma}_t$ is also large. This renders the modelling of its dynamics computationally challenging and potentially unstable. Therefore, certain regularization and dimension reduction methods can be useful. We formalize these ideas in the following sections.

Before doing so, for the low-cost sensor application we have in mind, we can further represent the model in a more compact form. For N sensors at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$, we define the response vector $\boldsymbol{y}_t = (y_t(\boldsymbol{s}_1), \ldots, y_t(\boldsymbol{s}_N))^\top \in \mathbb{R}^N$ and, similarly, $\boldsymbol{\varepsilon}_t = (\varepsilon_1(\boldsymbol{s}_1), \ldots, \varepsilon_N(\boldsymbol{s}_N))^\top \in \mathbb{R}^N$. This yields the compact form

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\, \boldsymbol{\gamma}_t + \sigma\, \boldsymbol{\varepsilon}_t, \tag{2.3}$$

where the ith row of $\boldsymbol{\Phi} \in \mathbb{R}^{N \times K}$ is given by $(\boldsymbol{\phi}(s_{2,i}) \otimes \boldsymbol{\phi}(s_{1,i}))^\top$. Since sensor observations may be missing, let $\mathbb{I}_t \subseteq \{1, \ldots, N\}$ denote the index set of *observed* sensors at time t and $|\mathbb{I}_t|$ be the corresponding cardinality. Define $\boldsymbol{\Phi}_{\mathbb{I}_t} \in \mathbb{R}^{|\mathbb{I}_t| \times K}$ as the submatrix of $\boldsymbol{\Phi}$ containing the rows corresponding to $\mathbb{I}_t$, and let $\boldsymbol{y}_{\mathbb{I}_t} \in \mathbb{R}^{|\mathbb{I}_t|}$ denote the corresponding observations. The observation equation (2.3) then becomes $\boldsymbol{y}_{\mathbb{I}_t} = \boldsymbol{\Phi}_{\mathbb{I}_t} \boldsymbol{\gamma}_t + \sigma \boldsymbol{\varepsilon}_{\mathbb{I}_t}$. Thus, missing observations affect only the observed data vector and the corresponding design matrix, while the latent coefficient vector $\boldsymbol{\gamma}_t$ remains unchanged. So in all methods introduced below, missing values are easily accounted for by replacing N, $\boldsymbol{y}_t$, and $\boldsymbol{\Phi}$ by $|\mathbb{I}_t|$, $\boldsymbol{y}_{\mathbb{I}_t}$, and $\boldsymbol{\Phi}_{\mathbb{I}_t}$, respectively.

2.2 Dynamic specifications

To model the temporal dynamics of $\boldsymbol{\gamma}_t$, we consider different dynamic specifications. We begin with an autoregressive model and benchmark it against observation-driven, parameter-driven, and two simple smoothing approaches. For convenience, an estimate, such as a cross-sectional estimate, is denoted by a hat, e.g., $\hat{f}$. Forecasts are denoted by a

tilde, e.g., $\tilde{f}$.

2.2.1 Functional autoregressions

A commonly used specification is the functional autoregressive model of order one, represented as a vector autoregressive model with one lag (VAR(1)) for the projection coefficients $\boldsymbol{\gamma}_t$, i.e.,

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\omega} + \boldsymbol{B}\boldsymbol{\gamma}_t + \boldsymbol{u}_{t+1}, \quad \boldsymbol{u}_{t+1} \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \tag{2.4}$$

where $\boldsymbol{\omega} \in \mathbb{R}^K$ is a constant drift term, $\boldsymbol{B} \in \mathbb{R}^{K \times K}$ the transition matrix, and $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$ a positive definite covariance matrix.

Since the projection coefficients $\boldsymbol{\gamma}_t$ are latent, one needs to replace $\boldsymbol{\gamma}_t$ in (2.4) by some estimates in order to estimate the parameters $\boldsymbol{\omega}$ and $\boldsymbol{B}$. The simplest way to do this is by cross-sectional regression. Due to missing observations, however, there can be problems for specific cross-sections. In particular, $\boldsymbol{\Phi}_{\mathbb{I}_t}^\top \boldsymbol{\Phi}_{\mathbb{I}_t}$ can become nearly singular depending on which observations are missing, where $\boldsymbol{\Phi}_{\mathbb{I}_t}$ is the design matrix for the observed data only as defined below (2.3). In such cases, the cross-sectional (least squares) estimates of $\boldsymbol{\gamma}_t$ become unstable or even infeasible. We therefore propose to regularize the regression by imposing a smoothness penalty on the estimated surface. Specifically, let $\boldsymbol{R}$ denote a positive definite penalty matrix (see Appendix A.1 for the detailed construction). The resulting penalized least squares estimator is then given by

$$\widehat{\boldsymbol{\gamma}}_t = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + l\boldsymbol{R})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}_t \tag{2.5}$$

where $l > 0$ is a small smoothing parameter.

The VAR(1) model in (2.4) is subsequently fitted to the estimated projection coefficients $\widehat{\boldsymbol{\gamma}}_t$ from (2.5). Estimation of the VAR parameters $\boldsymbol{\omega}$ and $\boldsymbol{B}$ can be performed via least squares (LS). For our forecasting analysis for pollution levels, we have no need to estimate $\boldsymbol{\Sigma}$. The recursive structure of the VAR(1) in (2.4) yields straightforward one-step-ahead forecasts of the pollution surface as $\boldsymbol{\Phi}\,\widetilde{\boldsymbol{\gamma}}_{t+1}$ with $\widetilde{\boldsymbol{\gamma}}_{t+1} = \widehat{\boldsymbol{\omega}} + \widehat{\boldsymbol{B}}\,\widehat{\boldsymbol{\gamma}}_t$.

2.2.2 Functional score-driven models

A potential drawback of the two-step functional autoregressive approach from Section 2.2.1 is that $\widehat{\boldsymbol{\gamma}}_t$ is estimated in a first step, and used as input for the VAR(1) in a second stage regression. As a recent alternative, the observation-driven functional generalized autoregressive score (fGAS) framework of Lin and Lucas (2025) integrates these two steps into a single-step estimation approach. It does so by generalizing the score-driven framework of Creal et al. (2013) to functional data. In our current context, the specification boils down to

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\Phi}\boldsymbol{\gamma}_t + \boldsymbol{e}_t, & \boldsymbol{y}_t \mid \boldsymbol{\gamma}_t &\sim p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t; \nu), \\ \boldsymbol{\gamma}_{t+1} &= \boldsymbol{\omega} + \boldsymbol{B}\boldsymbol{\gamma}_t + \boldsymbol{A}\boldsymbol{\mathcal{I}}^{-1}\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}_t, \nu), & \nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}_t, \nu) &= \frac{\partial \log p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}; \nu)}{\partial \boldsymbol{\gamma}}\bigg|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_t}, \end{aligned} \tag{2.6}$$

where $\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}_t)$ denotes the so-called score of the conditional predictive density $p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t; \nu)$ specified below, and ν is a parameter of the density. Moreover, $\boldsymbol{\omega} \in \mathbb{R}^K$ is a drift term, $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ rules the persistence of $\boldsymbol{\gamma}_t$, and $\boldsymbol{A} \in \mathbb{R}^{K\times K}$ determines the impact of the score on the update. The score is scaled by the inverse Fisher information matrix $\boldsymbol{\mathcal{I}} = \mathbb{E}\left[\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}, \nu)\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}, \nu)^\top\right]$ to stabilize the variance of the scaled score, see Creal et al. (2013) for further details. The updates result in expected improvements of the Kullback-Leibler divergence of the model with respect to the unknown data generating process (see Blasques et al., 2015; Creal et al., 2024; De Punder et al., 2024). Unlike the VAR(1) specification, the update in (2.6) is predetermined whenever the data $\boldsymbol{y}_t$ and the current state $\boldsymbol{\gamma}_t$ are given.

To account for the outlying observations in noisy low-cost sensor data, we assume the conditional predictive density is a multivariate Student's t distribution with ν degrees of freedom, i.e.,

$$p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t; \nu) = \frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi\sigma^2)^{\frac{N}{2}}}\left(1 + \frac{\boldsymbol{e}_t(\boldsymbol{\gamma}_t)^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t)}{\nu\sigma^2}\right)^{-\frac{\nu+N}{2}},$$

where $\boldsymbol{e}_t(\boldsymbol{\gamma}_t) = \boldsymbol{y}_t - \boldsymbol{\mu}(\boldsymbol{\gamma}_t)$ and $\boldsymbol{\mu}(\boldsymbol{\gamma}_t) = \boldsymbol{\Phi}\boldsymbol{\gamma}_t$. The corresponding score and Fisher

information matrix are then given by

$$\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}_t, \nu) = \left.\frac{\partial \log p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}; \nu)}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_t} = \frac{\nu + N}{\sigma^2 \nu + \boldsymbol{e}_t(\boldsymbol{\gamma}_t)^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t)} \boldsymbol{\Phi}^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t),$$
$$\boldsymbol{\mathcal{I}} = \mathbb{E}\left[\nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}, \nu) \nabla(\boldsymbol{y}_t, \boldsymbol{\gamma}, \nu)^\top\right] = \frac{\nu + N}{\sigma^2(\nu + N + 2)} \boldsymbol{\Phi}^\top \boldsymbol{\Phi}.$$

To prevent ill-conditioning of the inverse of the Fisher information matrix, we again add a regularization term $l\,\boldsymbol{R}$ for $l > 0$, thus obtaining the following updating equation:

$$\boldsymbol{\gamma}_{t+1} = \boldsymbol{\omega} + \boldsymbol{B}\boldsymbol{\gamma}_t + \boldsymbol{A}\left[\frac{(\nu + N + 2)\sigma^2}{\nu\sigma^2 + \boldsymbol{e}_t(\boldsymbol{\gamma}_t)^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t)} (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + l\,\boldsymbol{R})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t)\right]. \tag{2.7}$$

For estimation, we gather all static parameters of the model in the vector $\boldsymbol{\theta} = (\nu, \boldsymbol{\omega}^\top, \mathrm{vec}(\boldsymbol{A})^\top, \mathrm{vec}(\boldsymbol{B})^\top, \sigma)^\top$. The notation is now changed to $\boldsymbol{\gamma}_t(\boldsymbol{\theta})$ to stress that the time-varying parameters are a function of $\boldsymbol{\theta}$ via the recursion in (2.7). To estimate $\boldsymbol{\theta}$, we note that the log-likelihood can be expressed in an analytic form, such that the maximum likelihood estimator is given by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} T^{-1} \sum_{t=1}^{T} \ell_t(\boldsymbol{\gamma}_t(\boldsymbol{\theta}), \nu),$$

$$\begin{aligned}\ell_t(\boldsymbol{\gamma}_t(\boldsymbol{\theta}), \nu) &= N^{-1} \log p(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t(\boldsymbol{\theta}); \nu) \\ &= \frac{1}{N} \log\left(\frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi\sigma^2)^{\frac{N}{2}}}\right) - \frac{\nu + N}{2N} \log\left(1 + \frac{\boldsymbol{e}_t(\boldsymbol{\gamma}_t(\theta))^\top \boldsymbol{e}_t(\boldsymbol{\gamma}_t(\theta))}{\nu\sigma^2}\right),\end{aligned}$$

where $\boldsymbol{\Theta}$ denotes the parameter space of $\boldsymbol{\theta}$. The one-step-ahead forecasts $\widetilde{\boldsymbol{\gamma}}_{t+1}$ then follow directly from the recursion (2.7).

2.2.3 Functional state-space models

As an alternative to the fGAS approach from Section 2.2.2, we can also adopt a parameter-driven or state-space dynamic specification for $\boldsymbol{\gamma}_t$ to avoid the two-step functional autoregressive approach from Section 2.2.1. For the state-space specification, we assume that the development of the surface over time is determined by a latent process $\{\boldsymbol{\gamma}_t\}$ with an

associated observation sequence $\{\boldsymbol{y}_t\}$ as

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{\Phi}\boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}_t, \qquad & \boldsymbol{\varepsilon}_t &\sim \mathcal{N}(0, \sigma_\varepsilon^2 \boldsymbol{I}_N), \\ \boldsymbol{\gamma}_{t+1} &= \boldsymbol{\omega} + \boldsymbol{B}\boldsymbol{\gamma}_t + \boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim \mathcal{N}(0, \sigma_\eta^2 \boldsymbol{I}_K). \end{aligned} \tag{2.8}$$

This formulation follows the classical setup in Durbin and Koopman (2012), where $\boldsymbol{\omega} \in \mathbb{R}^K$ is a constant drift term and $\boldsymbol{B} \in \mathbb{R}^{K \times K}$ is the state transition matrix. The latent states are now random vectors, evolving according to the state transition equation. The error terms $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed to be serially independent, and independent of each other at all time points.

The parameters $\boldsymbol{\omega}, \boldsymbol{B}, \sigma_\varepsilon^2, \sigma_\eta^2$ are estimated by the Expectation Maximization (EM) algorithm. Due to the closed-form solution in each iteration of the EM algorithm, estimation is computationally efficient, especially compared to the numerical optimization of the log-likelihood function for the fGAS model of Section 2.2.2. Moreover, the recursive structure of (2.8) in combination with the Kalman filter directly yields the one-step-ahead forecast $\widetilde{\boldsymbol{\gamma}}_{t+1} = \widehat{\boldsymbol{\omega}} + \widehat{\boldsymbol{B}}\boldsymbol{\gamma}_{t|t}$, where $\boldsymbol{\gamma}_{t|t} = \mathbb{E}[\boldsymbol{\gamma}_t \mid \mathcal{F}_t]$ denotes the filtered state obtained from the Kalman Filter recursions, and $\mathcal{F}_t = \sigma\big(\boldsymbol{y}_s,\, s \leq t\big)$ represents the natural filtration up to time t.

2.2.4 Smoothing methods

We also include two simple rolling window benchmarks in our analysis. The first one is a rolling average of the estimated cross-sectional coefficients $\widehat{\boldsymbol{\gamma}}_t$ without imposing further parametric assumptions, i.e.,

$$\widetilde{\boldsymbol{\gamma}}_{t+1} = \frac{1}{J} \sum_{j=0}^{J-1} \widehat{\boldsymbol{\gamma}}_{t-j},$$

which uses the average of the J most recent observations of $\widehat{\boldsymbol{\gamma}}_t$ as our prediction. This estimate may be sensitive to the quality of the individual $\widehat{\boldsymbol{\gamma}}_t$ estimates. We therefore also consider an alternative where we switch the order of estimation and pooling (or averaging). Rather than first estimating the projection coefficients for each time point and then averaging them, we now pool the observations from the J most recent periods

and estimate a single coefficient vector $\widetilde{\boldsymbol{\gamma}}_{t+1}$ as

$$\widetilde{\boldsymbol{\gamma}}_{t+1} = (\boldsymbol{\Phi}_t^{*\top}\boldsymbol{\Phi}_t^* + l\,\boldsymbol{R})^{-1}\boldsymbol{\Phi}_t^{*\top}\boldsymbol{y}_t^*,$$

where $\boldsymbol{y}_t^* = (\boldsymbol{y}_{\mathbb{I}_t}^\top, \boldsymbol{y}_{\mathbb{I}_{t-1}}^\top, \ldots, \boldsymbol{y}_{\mathbb{I}_{t-J+1}}^\top)^\top$ is a $\sum_{j=0}^{J-1}|\mathbb{I}_{t-j}|$-dimensional vector, and $\boldsymbol{\Phi}_t^* = (\boldsymbol{\Phi}_{\mathbb{I}_t}^\top, \ldots, \boldsymbol{\Phi}_{\mathbb{I}_{t-J+1}}^\top)^\top$ a $(\sum_{j=0}^{J-1}|\mathbb{I}_{t-j}|) \times K$-dimensional matrix. The regularization parameter $l > 0$ again controls the degree of smoothness imposed on this pooled estimate.

2.3 Dimensionality reduction by FPCA

All of the dynamic specifications from Section 2.2 can operate on an arbitrary basis. So far, we concentrate on the projection of the location function $\mu_t(\cdot)$ onto the finite-dimensional tensor B-spline basis in (2.2). This can create challenges if the dimensionality K of the tensor basis grows large, with subsequent potential problems for forecasting the coefficient vectors $\widetilde{\boldsymbol{\gamma}}_t$. Therefore, next to the tensor B-spline basis, we also consider a functional principal component analysis (FPCA) approach to offer a more parsimonious representation of the data (see also Ramsay and Silverman, 2005, Chapter 8).

Let $X_t(\cdot) = \mu_t(\cdot) - \bar{\mu}(\cdot)$ denote the mean-centered location function. FPCA represents X_t in terms of the eigenfunctions ψ_ℓ, $\ell \geq 1$, of its covariance operator $\mathcal{C}_X$:

$$\mathcal{C}_X\,\psi_\ell = \int_{\mathcal{S}} c_X(\boldsymbol{s},\cdot)\psi_\ell(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s} = \lambda_\ell\,\psi_\ell, \quad \ell = 1, 2, , \ldots, \tag{2.9}$$

where $c_X : \mathcal{S}^2 \to \mathbb{R}$ is the covariance kernel associated with $\mathcal{C}_X$, and $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are the corresponding eigenvalues. The mean-centered location function X_t admits the following Karhunen–Loève decomposition:

$$X_t(\boldsymbol{s}) = \sum_{\ell=1}^{\infty} \xi_{t,\ell}\,\psi_\ell(\boldsymbol{s}), \qquad \xi_{t,\ell} = \langle X_t, \psi_\ell\rangle_{L^2},$$

with $\mathrm{Var}(\xi_{t,\ell}) = \lambda_\ell$ and $\mathrm{Cov}(\xi_{t,\ell}, \xi_{t,k}) = 0$ for $k \neq \ell$.

In practice $\mu_t(\cdot)$ is latent and we only observe $y_t(\cdot)$. Moreover, $y_t(\cdot)$ is typically not observed as a continuous function, but only at a finite set of discrete points. To accommodate this, we proceed in three steps (see, e.g., Ramsay and Silverman, 2005, Chapter 8): (i) estimate the mean function $\bar{\mu}(\cdot)$, (ii) construct the denoised and demeaned latent functions $y_t(\cdot) - \bar{\mu}(\cdot)$ and the covariance kernel c_X, and (iii) solve the eigenproblem.

First, as in (2.2), the mean function $\bar{\mu}(\cdot)$ can be approximated by a projection on a tensor B-spline basis: $\bar{\mu}(\boldsymbol{s}) \approx (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{\upsilon}$. We estimate the projection coefficient vector $\boldsymbol{\upsilon}$ by smoothness-penalized least squares as in (2.5), but now using all the data rather than cross-section t only. The estimated demeaned surface is then given by $y_t(\boldsymbol{s}) - \hat{\bar{\mu}}(\boldsymbol{s}) = X_t(\boldsymbol{s}) + \hat{\varepsilon}_t(\boldsymbol{s})$ where $\hat{\varepsilon}_t(\boldsymbol{s}) = \varepsilon_t(\boldsymbol{s}) + \bar{\mu}(\boldsymbol{s}) - \hat{\bar{\mu}}(\boldsymbol{s})$.

Next, to estimate the covariance kernel c_X, we again project $X_t(\cdot)$ on the same tensor B-spline basis as $X_t(\boldsymbol{s}) \approx (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{\kappa}_t$, where the coefficients $\boldsymbol{\kappa}_t$ are estimated by cross-sectional penalized regressions as in (2.5). We denote the resulting estimates by $\hat{\boldsymbol{\kappa}}_t$. Defining the time series averages $\bar{\hat{\boldsymbol{\kappa}}} = T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\kappa}}_t$, and $\bar{X} = T^{-1} \sum_{t=1}^T X_t$, the covariance kernel $\hat{c}_X(\cdot, \cdot)$ of $X_t(\cdot)$ is then estimated as

$$\hat{c}_X(\boldsymbol{a}, \boldsymbol{b}) = \frac{1}{T} \sum_{t=1}^T X_t(\boldsymbol{b}) X_t(\boldsymbol{a}) - \bar{X}(\boldsymbol{b}) \bar{X}(\boldsymbol{a}) = \big(\boldsymbol{\phi}(b_2) \otimes \boldsymbol{\phi}(b_1)\big)^\top \widehat{\boldsymbol{C}}_T \big(\boldsymbol{\phi}(a_2) \otimes \boldsymbol{\phi}(a_1)\big).$$

for some $\boldsymbol{a} = (a_1, a_2)^\top$ and $\boldsymbol{b} = (b_1, b_2)^\top$ in $\mathcal{S}$, where $\widehat{\boldsymbol{C}}_T = T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\kappa}}_t \hat{\boldsymbol{\kappa}}_t^\top - \bar{\hat{\boldsymbol{\kappa}}} \bar{\hat{\boldsymbol{\kappa}}}^\top$.

As a final step, we compute the projections of the eigenfunctions $\psi_\ell(\cdot)$ on the same set of B-spline basis functions, i.e., $\psi_\ell(\boldsymbol{s}) \approx (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{q}_\ell$, under the normalization $\|\psi_\ell\|_{L^2} = 1$, i.e., $\|(\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{q}_\ell\|_{L^2} = \boldsymbol{q}_\ell^\top \boldsymbol{W} \boldsymbol{q}_\ell = 1$ for the symmetric matrix $\boldsymbol{W} = \int_{\mathcal{S}} \boldsymbol{\phi}(s_2)\boldsymbol{\phi}(s_2)^\top \otimes \boldsymbol{\phi}(s_1)\boldsymbol{\phi}(s_1)^\top \mathrm{d}\boldsymbol{s}$. Substituting this expression for $\psi_\ell(\boldsymbol{s})$ into (2.9) yields the eigenproblem

$$\int_{\mathcal{S}} \hat{c}_X(\boldsymbol{s}, \boldsymbol{a})(\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{q}_\ell \, \mathrm{d}\boldsymbol{s} = \lambda (\boldsymbol{\phi}(a_2) \otimes \boldsymbol{\phi}(a_1))^\top \boldsymbol{q}_\ell \iff \widehat{\boldsymbol{C}}_T \boldsymbol{W} \boldsymbol{q}_\ell = \lambda \boldsymbol{q}_\ell,$$

subject to $\boldsymbol{q}_\ell^\top \boldsymbol{W} \boldsymbol{q}_\ell = 1$. Defining $\boldsymbol{v}_\ell = \boldsymbol{W}^{1/2} \boldsymbol{q}_\ell$, the problem reduces to

$$\widehat{\boldsymbol{C}}_T \boldsymbol{W} \boldsymbol{q}_\ell = \lambda \boldsymbol{q}_\ell \iff \boldsymbol{W}^{1/2} \widehat{\boldsymbol{C}}_T \boldsymbol{W} \boldsymbol{q}_\ell = \lambda \boldsymbol{W}^{1/2} \boldsymbol{q}_\ell \iff \boldsymbol{W}^{1/2} \widehat{\boldsymbol{C}}_T \boldsymbol{W}^{1/2} \boldsymbol{v}_\ell = \lambda \boldsymbol{v}_\ell,$$

subject to $\boldsymbol{q}_\ell^\top \boldsymbol{W} \boldsymbol{q}_\ell = \boldsymbol{v}_\ell^\top \boldsymbol{v}_\ell = 1$, which is a standard PCA applied to the transformed covariance matrix $\boldsymbol{W}^{1/2} \widehat{\boldsymbol{C}}_T \boldsymbol{W}^{1/2}$. This gives the estimates $\widehat{\boldsymbol{q}}_\ell$ of $\boldsymbol{q}_\ell$. Using the above expressions for $\psi_\ell(\boldsymbol{s})$ and truncating the Karhunen–Loève representation to the first $r < K$ components, we obtain the parsimonious representation

$$\Big(\mu_t(\boldsymbol{s}_1), \mu_t(\boldsymbol{s}_2), \ldots, \mu_t(\boldsymbol{s}_N)\Big)^\top \approx \hat{\bar{\boldsymbol{\mu}}} + \widehat{\boldsymbol{\Phi}} \, \boldsymbol{\xi}_t, \qquad \hat{\bar{\boldsymbol{\mu}}} = \Big(\hat{\bar{\mu}}(\boldsymbol{s}_1), \hat{\bar{\mu}}(\boldsymbol{s}_2), \ldots, \hat{\bar{\mu}}(\boldsymbol{s}_N)\Big)^\top,$$

with $\widehat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}\widehat{\boldsymbol{Q}}_r \in \mathbb{R}^{N\times r}$, $\widehat{\boldsymbol{Q}}_r = (\widehat{\boldsymbol{q}}_1, \ldots, \widehat{\boldsymbol{q}}_r) \in \mathbb{R}^{K\times r}$, and coefficient vector $\boldsymbol{\xi}_t = (\xi_{t,1}, \ldots, \xi_{t,r})^\top \in \mathbb{R}^r$ for $r < K$.[1] The pollution surface representation becomes

$$\boldsymbol{y}_t = \hat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\Phi}}\,\boldsymbol{\xi}_t + \sigma\,\boldsymbol{\varepsilon}_t \tag{2.10}$$

for the much lower dimensional $\boldsymbol{\xi}_t$ compared to $\boldsymbol{\gamma}_t$, such that we can again apply any of the dynamic specifications from Section 2.2 to model the dynamics of $\boldsymbol{\xi}_t$. Both the B-spline and FPCA basis thus fit the same modelling framework. Note that under the new representation (2.10), we can work directly with the new basis coefficient matrix $\widehat{\boldsymbol{\Phi}}$ and model the dynamics of $\boldsymbol{\xi}_t$ using the dynamic specifications in Section 2.2, rather than estimating the elements $\xi_{t,\ell}$ using FPCA.

3 Simulation study

We compare the performance of the proposed models in Section 2.2 under irregular sampling, missing values, and heavy-tailed innovations. We consider both settings with and without FPCA dimensionality reduction. The simulation settings are designed to closely resemble the features of the empirical application. Specifically, consider again the location-scale model $y_t(\boldsymbol{s}) = \mu_t(\boldsymbol{s}) + \sigma\varepsilon_t(\boldsymbol{s})$. We assume that $\varepsilon_t(\boldsymbol{s}) \sim t(\nu)$ independent across $t \in \mathbb{Z}^+$, with $\nu \in \{5, 500\}$, such that we consider settings with and without severe measurement errors. The covariance kernel is given by $\mathrm{Cov}(\varepsilon_t(\boldsymbol{s}), \varepsilon_t(\boldsymbol{t})) = \nu(\nu-2)^{-1}\exp(-\|\boldsymbol{s}-\boldsymbol{t}\|/0.1)$ for $\boldsymbol{s}, \boldsymbol{t} \in [0,1]^2$, making the innovations independent over time, but spatially correlated. We set $\sigma = 1$ and generate the location function under a functional factor form

$$\begin{aligned} \mu_t(\boldsymbol{s}) &= \sum_{\ell=1}^{r_0} \xi_{t,\ell}\psi_\ell(\boldsymbol{s}), \\ \xi_{t+1,\ell} &= c_\ell + a_\ell \xi_{t,\ell} + u_{t+1,\ell}, \quad u_{t+1,\ell} \sim \mathcal{N}(0, \sigma_u^2), \end{aligned} \tag{3.1}$$

where $r_0 \in \mathbb{N}$ denotes the true number of factors, and where the functions $\psi_\ell(\boldsymbol{s})$ are given by the first $r_0 = 5$ estimated eigenfunctions from the empirical application in order to have an empirically congruent simulation set-up. The functions $\psi_\ell(\boldsymbol{s})$ are kept fixed throughout

[1] Rather than estimating $\widehat{\boldsymbol{Q}}_r$ via PCA, one can also directly (possibly using the PCA estimate as a starting value) estimate $\boldsymbol{Q}_r$ jointly with the other parameters under the restriction $\boldsymbol{Q}_r^\top \boldsymbol{W} \boldsymbol{Q}_r = \boldsymbol{I}_r$, thus avoiding a two-step approach at the cost of a higher computation time.

the simulations. The factor scores $\xi_{t,\ell}$ evolve independently (across ℓ) as AR(1) processes, with autoregressive parameters a_ℓ linearly spaced between 0.1 and 0.8, drift parameters $c_\ell = 0.01$ and innovation standard deviations $\sigma_u = 0.3$ for all ℓ. The factor scores are initialized at their unconditional means, $\xi_{1,\ell} = c_\ell/(1 - a_\ell)$.

To model missing values, we define the selection variable $Z_{j,t} \in \{0, 1\}$ that follows a binary Markov Chain in t, where $Z_{j,t} = 0$ indicates a missing observation at location j and time t, and $Z_{j,t} = 1$ indicates a non-missing observation. We fix the probability remaining non-missing $\mathbb{P}(Z_{j,t} = 1 \mid Z_{j,t-1} = 1) = 0.9$ and vary the probability of remaining missing as $\mathbb{P}(Z_{j,t} = 0 \mid Z_{j,t-1} = 0) \in \{0.6, 0.9\}$, corresponding to approximately 20% and 50% missing observations, respectively. We also consider a complete-data benchmark $\mathbb{P}(Z_{j,t} = 0) = 0$.

To assess the consequences of irregular sampling, we compare the uniform sampling $\boldsymbol{s}_j = (s_{j,1}, s_{j,2})^\top \sim \mathrm{U}[0, 1]^2$ scheme with an irregular sampling scheme $s_{j,\ell} \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(\alpha = 5, \beta = 2)$, for $j \in \{1, \ldots, N\}$ and $\ell = 1, 2$, with $N \in \{100, 200\}$. The Beta(5, 2) distribution is skewed in such a way that all points tend to be clustered near the upper-right corner of the unit square, in line with the sampling behaviour of the sensors in the empirical application.

All methods use $K_1 = K_2 = 7$ cubic B-spline basis functions with 5 knots positioned at $s = 0, 0.25, 0.5, 0.75, 1$. All window sizes are fixed to $J = 20$, and the number of functional factors is set equal to the number of retained FPCA components, $r = r_0 = 5$. All the regularization parameters are fixed at $l = 0.01$ for numerical simplicity. Furthermore, forecast accuracy is evaluated using out-of-sample simulated median-absolute-errors (MedAE). Specifically, we generate $T + T_{\text{out}}$ discrete paths of $\{y_t(\cdot)\}$, where $T \in \{500, 1500\}$ denotes the in-sample and $T_{\text{out}} = 1000$ the out-of-sample time series length. For each of the 500 Monte Carlo iterations the MedAE is calculated as $\text{med}\left\{|\mu_t(\boldsymbol{s}_j) - \tilde{\mu}_t(\boldsymbol{s}_j)| : t = T + 1, \ldots, T + T_{\text{out}},\ j = 1, \ldots, N\right\}$. To assess the performance beyond the observed locations, we also evaluate accuracy on a fixed 10×10 grid:

$$\mathbb{X} = \left\{0.05 + \frac{i-1}{9}(0.95 - 0.05) : i = 1, 2, \ldots, 10\right\}^2 \tag{3.2}$$

with corresponding MedAE computed as $\text{med}\left\{|\mu_t(\boldsymbol{x}) - \tilde{\mu}_t(\boldsymbol{x})| :\ t = T+1, \ldots, T+T_{\text{out}},\ \boldsymbol{x} \in \mathbb{X}\right\}$. The fixed grid is only used for evaluation and remains unobserved during the forecasting process. This dual evaluation allows us to compare surface reconstruction

and forecasting performance at observed as well as unobserved locations. All results are presented in Table 1.

Table 1 shows that an increase in missing observations yields little to no difference in forecasting performance. All methods are generally robust to missing observations, both at observed locations and on the (unobserved) evaluation grid. Heavy-tailed innovations lead to a modest performance loss.

Across all specifications, predictive performance is very similar at observed locations and unobserved locations, suggesting that the models generalize well across the spatial domain under uniform sampling. In contrast, irregular sampling induced by the Beta distribution leads to deterioration in spatial generalization. In particular, forecasting performance at the observed sensors is similar to that under uniform sampling, while performance on the evaluation grid deteriorates. This is intuitive because the clustering of observations places sensors closer together and can cause spatial overfitting and a weaker generalisation of the surface's fit outside the observed sensor locations. The pattern is particularly clear for the B-spline compared to the FPCA basis. The state-space model appears to be largely robust to this. This is consistent with its correct specification: although irregular sampling reduces the information in the observation equation, the state equation remains correctly specified, allowing the underlying state dynamics to be estimated from the clustered observations and subsequently generalised across the spatial domain.

All specifications show modest improvements when using the FPCA basis rather than the B-spline basis in line with the stucture of the data generating process. Note that FPCA requires only a few estimated eigenfunctions ($r = r_0 = 5$) to achieve the same or better performance than the B-spline approach, compared with $K = 49$ B-spline basis functions. Notably, the FPCA basis also eliminates the spatial overfitting observed for fGAS under beta-distributed sampling. By restricting the model to a lower dimensional factor space, FPCA yields a more stable fit and better generalisation across the surface. The pooled and rolling benchmarks show little to no predictive benefit of FPCA, although estimation is faster. This likely reflects the regularization incorporated in these models, which already enables the richer B-spline basis to generalise well. Overall, the best performing models are the three parametric VAR(1), fGAS, and state-space models. The strong performance of the VAR(1) highlights the effectiveness of the regularized LS projections.

Table 1: Empirical out-of-sample MedAE, both at uniformly observed locations (obs.) and unobserved locations on fixed 10×10 grid (grid). Surface is generated by time-varying factor specification (3.1).

N	missings	ν	samp.	VAR(1)		fGAS		state-space		pooled estimator		rolling average	
				obs.	grid	obs.	grid	obs.	grid	obs.	grid	obs.	grid
						$T = 500$							
B-spline basis ($K = 49$)													
100	0%	500	unif.	0.48	0.48	0.47	0.47	0.52	0.51	0.54	0.53	0.53	0.52
100	50%	500	unif.	0.49	0.49	0.47	0.48	0.55	0.55	0.55	0.54	0.53	0.52
100	20%	5	unif.	0.50	0.50	0.48	0.49	0.56	0.55	0.54	0.54	0.53	0.52
100	50%	5	unif.	0.50	0.50	0.48	0.49	0.59	0.59	0.55	0.55	0.53	0.53
100	0%	5	beta	0.49	0.58	0.47	0.85	0.52	0.56	0.51	0.61	0.50	0.59
100	50%	5	beta	0.50	0.62	0.47	0.82	0.55	0.60	0.51	0.67	0.50	0.61
200	0%	5	unif.	0.49	0.49	0.48	0.47	0.55	0.54	0.54	0.54	0.54	0.53
200	50%	5	beta	0.49	0.61	0.45	0.74	0.52	0.66	0.50	0.61	0.49	0.60
FPCA basis ($r = r_0 = 5$)													
100	0%	500	unif.	0.45	0.44	0.45	0.45	0.44	0.44	0.53	0.52	0.53	0.52
100	50%	500	unif.	0.46	0.45	0.45	0.45	0.45	0.45	0.53	0.52	0.53	0.52
100	20%	5	unif.	0.46	0.46	0.46	0.46	0.46	0.46	0.53	0.52	0.53	0.52
100	50%	5	unif.	0.47	0.46	0.47	0.46	0.47	0.46	0.53	0.53	0.53	0.53
100	0%	5	beta	0.46	0.55	0.47	0.57	0.46	0.56	0.50	0.61	0.50	0.61
100	50%	5	beta	0.46	0.56	0.48	0.56	0.46	0.57	0.50	0.62	0.50	0.64
200	0%	5	unif.	0.45	0.45	0.46	0.45	0.45	0.45	0.53	0.53	0.53	0.53
200	50%	5	beta	0.45	0.57	0.45	0.57	0.45	0.57	0.49	0.61	0.49	0.62
						$T = 1500$							
B-spline basis ($K = 49$)													
100	0%	5	unif.	0.46	0.46	0.47	0.47	0.50	0.50	0.54	0.52	0.53	0.51
100	50%	5	beta	0.44	0.54	0.43	0.66	0.49	0.56	0.49	0.65	0.48	0.60
FPCA basis ($r = r_0 = 5$)													
100	0%	5	unif.	0.45	0.44	0.46	0.45	0.45	0.44	0.53	0.51	0.53	0.51
100	50%	5	beta	0.43	0.53	0.43	0.56	0.43	0.54	0.48	0.62	0.48	0.63

4 Empirical application to PM2.5 forecasting

We now compare the different methods in an empirical example for PM2.5 forecasting. PM2.5 is particulate matter that is 2.5 micrometers or smaller in aerodynamic diameter. Rather than a single pollutant, it is a complex mixture of airborne solid and liquid particles with diverse chemical and physical characteristics (World Health Organization, 2021). Exposures are associated with increased premature mortality, respiratory diseases, cardiovascular diseases, and neurological disorders (Sangkham et al., 2024). The World Health Organization considers PM2.5 one of the greatest environmental health risks.

We evaluate the proposed methods using daily average PM2.5 measurements obtained from the PurpleAir sensor network.[2] PurpleAir is a widely used citizen-science network for real-time air quality monitoring that has been adopted in numerous studies; see Bonas and Castruccio (2025) and the references therein. Our dataset consists of sensors distributed across Europe, approximately along the diagonal from London to Athens, covering 2,188 consecutive days from 2 July 2020 to 30 June 2026. The first 1823 observations (2 July 2020 to 30 June 2025) are used for model estimation, such that we reserve a full year's cycle for out-of-sample evaluation; hence, $T = 1823$ and $T_{\text{out}} = 365$. To ensure positive forecasts, we log-transform the PM2.5 concentrations, treating zero values, which may arise from measurement errors or limited sensor sensitivity, as missing, and transform the forecasts back to the original scale for evaluation. Further details on the data preprocessing are provided in Appendix B.1.

Figure 1 displays the time series for the resulting $N = 200$ sensors. The series exhibit highly erratic peaks for some sensors. Moreover, the empirical transition probabilities of the binary Markov Chain, derived in Appendix A.2, are $\widehat{\mathbb{P}}(Z_{j,t} = 1 \mid Z_{j,t-1} = 1) = 0.892$ and $\widehat{\mathbb{P}}(Z_{j,t} = 0 \mid Z_{j,t-1} = 0) = 0.588$. Thus, recording periods are highly persistent, while missing periods are comparatively short-lived: sensor outages are typically temporary, though they may persist for several consecutive days.

For evaluating model performance, the $N = 200$ sensors are further partitioned into an estimation set $\mathbb{S}^{\text{est}}$ consisting of 125 sensors, and a hold-out evaluation set $\mathbb{S}^{\text{eval}}$ consisting of 75 sensors, to assess the models' ability to generalize to previously unseen spatial locations. To account for the dependence of forecasting performance on the spatial

[2] See https://www.purpleair.com/?srsltid=AYCX9vIX1ElSHTpCdtf1gNpkW3ivQ0MvuTWczQd2_Y-LFfqGmfk-iqjn.

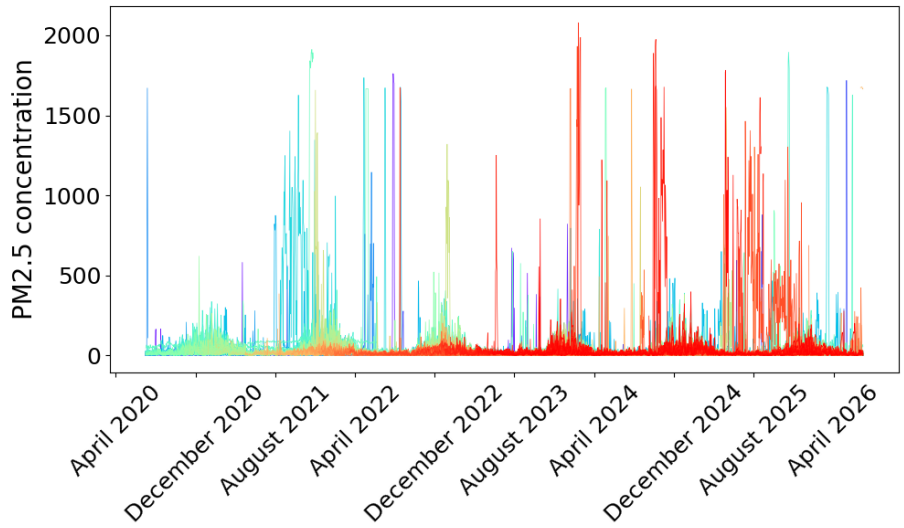


Figure 1: Time series plots of daily PM2.5 concentrations, with each curve representing the data recorded of an individual sensor.

configuration of the sensors, we consider 200 random partitions of the sensors into $\mathbb{S}^{\text{est}}$ and $\mathbb{S}^{\text{eval}}$. To illustrate forecasting performance visually, we highlight one representative partition in Figure 2. This partition is then further used for the FPCA in Section 4.1, and for illustrating forecasting performance in Section 4.2 and Appendix B.4.

4.1 FPCA of PM2.5 surface

Before forecasting, we perform FPCA on the training data of the log-transformed PM2.5 dataset using a small regularization parameter $l = 0.01$ for all the LS regularization parameters; a different but similarly small regularization parameter, such as $l = 0.02$, does not qualitatively change the results. Figure 3 displays two examples of the estimated demeaned PM2.5 concentration surfaces reconstructed using the FPCA basis with $r = 7$ principal components, which explain approximately 95% of the total variance. These examples demonstrate again that the FPCA basis can reconstruct a smooth surface despite the irregular sampling and inherent measurement errors. The cumulative explained variance as a function of r can be found in Appendix B.3.

An advantage of the FPCA dimension reduction is that we can interpret the most dominant principal components, whereas the B-spline coefficients lack a direct interpretation. The estimated eigenfunctions and functional scores for the first three principal components are presented in Figure 4. The complete set of the first seven estimated eigenfunctions with corresponding scores is provided in Appendix B.3. The first principal component

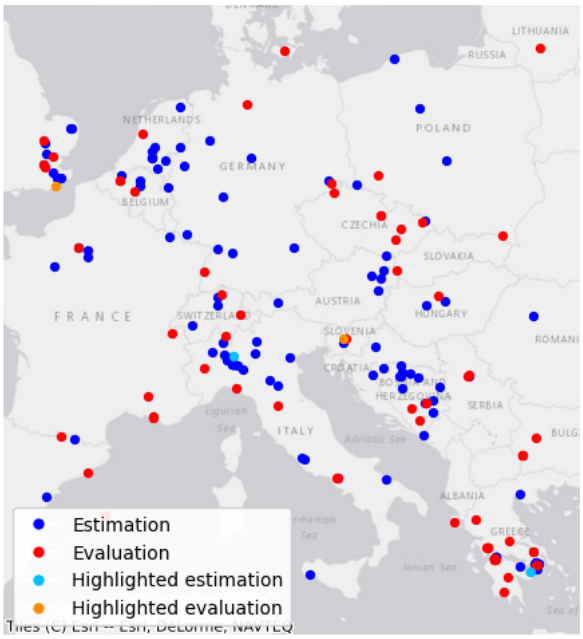
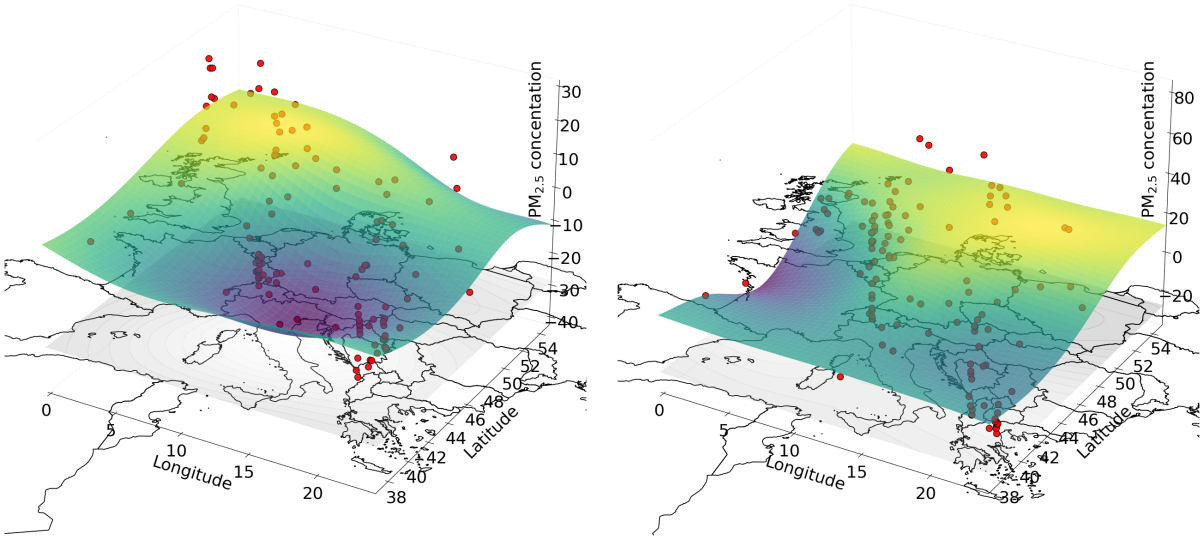


Figure 2: Locations of 200 sensors. All dots represent specific sensor locations. Red dots represent evaluation sensors and blue dots estimation sensors. Milan (Italy), Aegina (Greece), Ljubljana (Slovenia), and Hastings (UK) are highlighted in either light blue or orange.



(a) Estimated surface on 14 May 2023 **(b)** Estimated surface on 1 February 2026

Figure 3: Estimated demeaned PM2.5 concentration surfaces using FPCA basis with $r = 7$ principal components. Red dots represent observed locations, while the coloured surface represents the fitted surface.

(bottom panel) represents the overall dominant PM2.5 concentration across Europe: the eigenfunction is roughly constant, indicating that it acts as a common factor affecting all regions simultaneously. The corresponding scores in the top panel exhibit a clear seasonal pattern, with negative values in winter and positive values in summer, consistent with

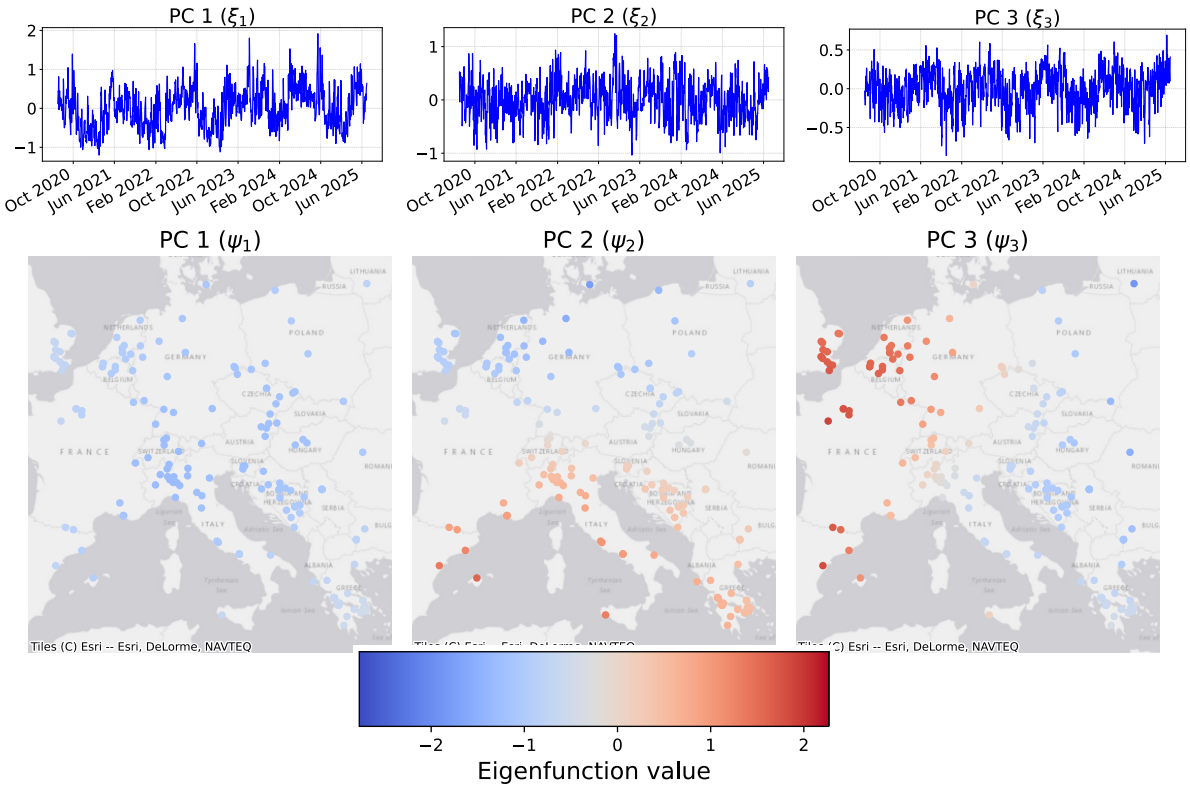


Figure 4: Estimated first three eigenfunctions and corresponding functional scores for log-scaled daily PM2.5 concentrations.

the higher winter concentrations shown in Figure 1. The second principal component captures a north-south contrast, with opposite signs in northern and southern Europe. Its scores show greater variability in winter, indicating that regional differences are more pronounced during this period. The third principal component captures an east-west contrast, with opposite signs in eastern and western Europe, and similarly exhibits greater score variability during winter.

4.2 One-step-ahead forecasting performance for PM2.5

We evaluate the one-step-ahead forecasting performance of all models both with and without FPCA dimensionality reduction, where the FPCA basis is again estimated on the training set. For the FPCA basis, we consider $r \in \{3, 7\}$ principal components. The choice of $r = 3$ provides a highly parsimonious representation and is of particular interest, because the first three principal components admit a clear physical interpretation. The choice $r = 7$ explains 95% of the variance. Further computational details for the empirical implementation can be found in Appendix B.2.

We randomly select 125 sensors to form the estimation set $\mathbb{S}^{\text{est}}$ and 75 sensors to form

Table 2: Medians of out-of-sample median absolute errors for PM2.5 forecasts at observed estimation sensors (obs.) and unobserved evaluation sensors (eval.).

Basisfuncs.	VAR(1)		fGAS		state-space		pooled estimator		rolling average	
	obs.	eval.	obs.	eval.	obs.	eval.	obs.	eval.	obs.	eval.
FPCA ($r = 3$)	4.23	4.88	4.38	5.04	4.24	4.89	4.76	5.31	4.75	5.30
FPCA ($r = 7$)	4.04	4.72	4.24	4.91	4.06	4.74	4.73	5.33	4.73	5.34
B-spline ($K = 49$)	3.91	4.58	4.21	5.22	3.93	4.49	4.60	5.59	4.64	5.24

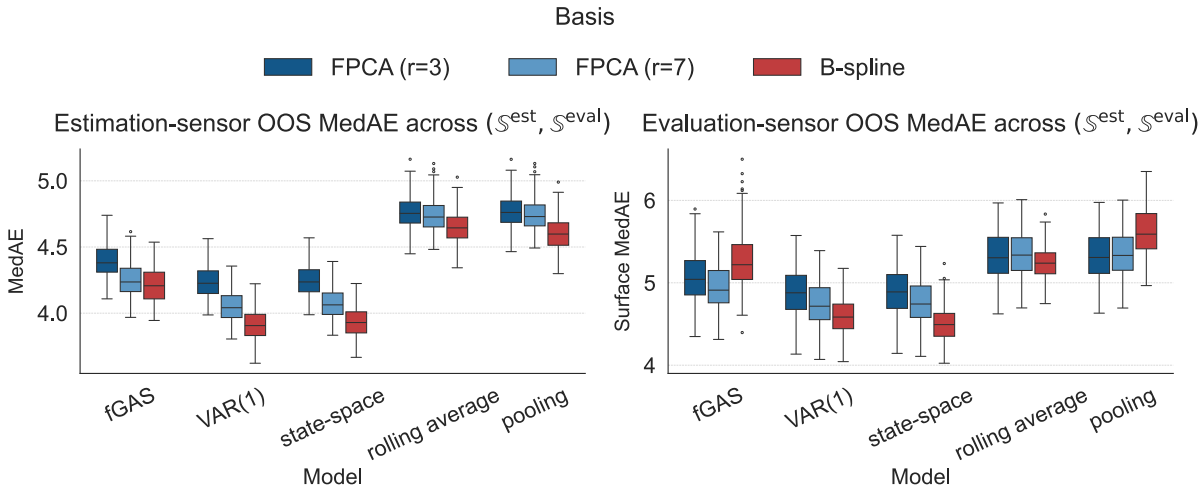


Figure 5: Distributions of the median absolute out-of-sample (OOS) forecasting error (MedAE) across randomly drawn partitions of estimation and evaluation sensors ($\mathbb{S}^{\text{est}}, \mathbb{S}^{\text{eval}}$). Note that the error scales differ between the two sensor groups.

the evaluation set $\mathbb{S}^{\text{eval}}$. We repeat this procedure 200 times, as forecasting performance can depend substantially on the particular partition of sensors into $\mathbb{S}^{\text{est}}$ and $\mathbb{S}^{\text{eval}}$. For each specification, we report the median of the MedAE across the 200 sensor partitions. The MedAE for the observed sensors is evaluated as $\text{med}\left\{|y_t(\boldsymbol{s}_j) - \tilde{\mu}_t(\boldsymbol{s}_j)| : t = T+1, \ldots, T+T_{\text{out}}, j \in \mathbb{I}_t^{\text{est}}\right\}$ where $y_t(\cdot)$ corresponds to the observed PM2.5 concentration. Notice that model performance can only be assessed on observed values. Accordingly, the MedAE for the unobserved evaluation sensors is evaluated as $\text{med}\left\{|y_t(\boldsymbol{s}_j) - \tilde{\mu}_t(\boldsymbol{s}_j)| : t = T+1, \ldots, T+T_{\text{out}}, j \in \mathbb{I}_t^{\text{eval}}\right\}$, where $\mathbb{I}_t^{\text{eval}} = \mathbb{S}^{\text{eval}} \cap \mathbb{I}_t$. The results are presented in Table 2 and the distribution of the daily MedAE is presented in Figure 5.

Table 2 and Figure 5 show that forecasting performance both at observed locations and unobserved locations generally benefits from a richer basis. The main exceptions occur at the evaluation sensors for the B-spline basis combined with the fGAS and pooling models, which perform more poorly, probably due to spatial overfitting. The performance of the VAR(1) and state-space models is similar and yields the best results. Given its

substantially lower complexity, the VAR(1) seems to be the more favourable specification. All parametric models substantially outperform the simple rolling average and pooled estimator.

The MedAE at the evaluation sensors is slightly higher than at the estimation sensors. This is not surprising, given that the former are entirely excluded from model estimation, and forecasting performance thus relies on the spatial generalizability to these unseen locations. Moreover, the variance of the MedAE of the unobserved locations is higher than the MedAE of observed locations, again highlighting that the forecasting ability can be highly dependent on the partition between observed ($\mathbb{S}^{\text{est}}$) and unobserved ($\mathbb{S}^{\text{eval}}$) sensors. We also see that for the empirical data and the VAR(1) and state-space specification, the B-spline bases have a better out-of-sample performance than the FPCA, though the difference is small. With $r = 7$ principal components and 95% of the variance explained, the FPCA basis still does not fully capture the granularity of PM2.5 percolation, but comes close to the unrestricted B-spline basis using 49 basis functions.

In Figure 6, we zoom in on four specific sensors across different regions of Europe over time using the same partition of estimation and evaluation sensors as in Section 4.1. The two selected estimation sensors are located in Milan (Italy) and Aegina (Greece), and the two evaluation sensors in Ljubljana (Slovenia) and Hastings (UK). They are highlighted in light blue and orange, respectively, in Figure 2. These sensors were selected as representative examples of missing observations and outliers. The VAR(1) forecasts with and without dimensionality reduction are shown in Figure 6, and the difference between the two is indeed small. Forecasts for the other specifications are provided in Appendix B.4. We notice that the data are very noisy and have considerable stretches of missing values, such that it is non-trivial to filter out the location dynamics. Still, the VAR(1) fit as obtained from the full European pollution landscape captures the general movements of the sensor measurements quite well irrespective of the chosen basis specification, despite the challenging nature of the data.

5 Conclusion

In this paper, we proposed a functional framework for spatio-temporal modelling of air pollution data collected from low-cost sensor networks. The framework represents

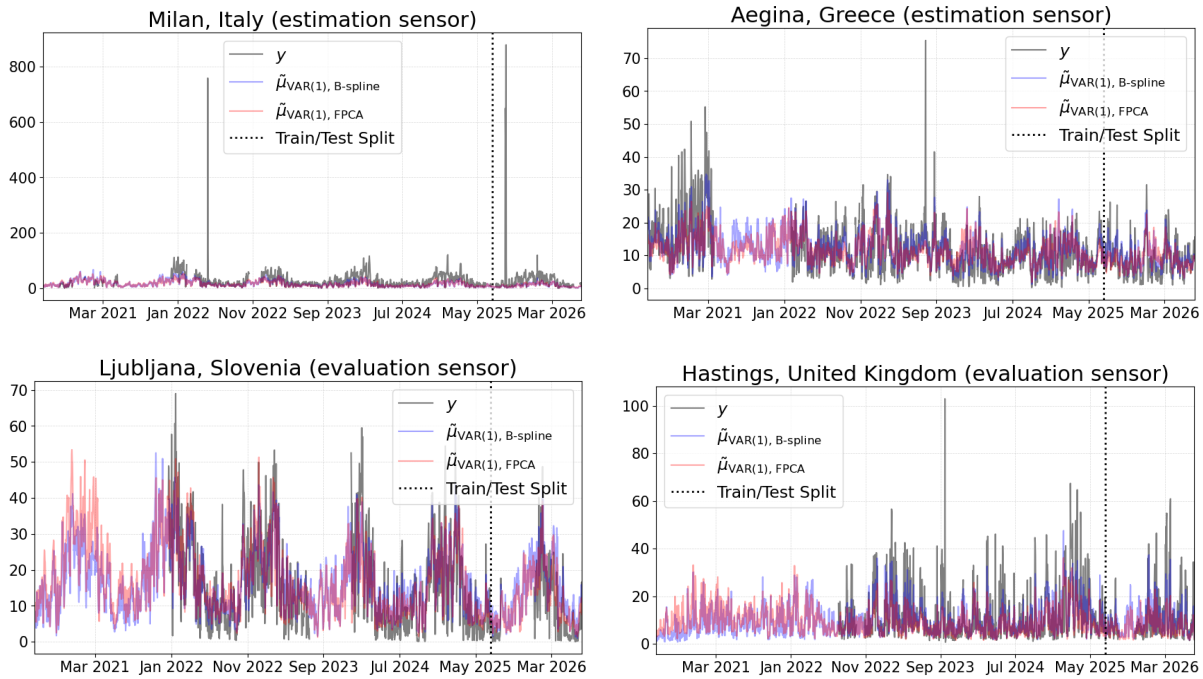


Figure 6: In-sample and out-of-sample predicted PM2.5 concentrations for Milan, Aegina, Ljubljana and Hastings using VAR(1). Both regular B-spline basis and FPCA basis with $r = 7$ principal components are considered. y denotes the observed values, $\tilde{\mu}_{\text{VAR(1), basis}}$ denotes predicted PM2.5 level, and dotted line represents split between in-sample and out-of-sample fit.

observed two-dimensional pollution surfaces using a finite set of basis functions with dynamic projection coefficients. Different dynamic specifications were investigated to see which specification is most promising empirically.

Simulations demonstrated that the proposed framework is robust to key challenges encountered in low-cost sensor networks, including irregular sampling, missing observations, and outliers. These results are confirmed in an empirical application to a large low-cost European sensor network for PM2.5 air pollution. The results demonstrate strong forecasting performance of the time-varying functional approach, with the functional autoregressive and parameter-driven state-space models performing best irrespective of the basis functions used. Out of these two, the autoregressive model has a substantially lower computational complexity. Across all specifications, forecasting performance is slightly weaker at unobserved locations. Forecasting performance at unobserved locations also exhibits greater variability across sensor configurations, highlighting the sensitivity of a model's spatial generalizability to sensor placement.

Acknowledgement

During the preparation of this work the authors used ChatGPT 6 to improve the readability of the original manuscript. After using this tool, we further reviewed and edited the content as needed and take full responsibility for the content of the article.

References

Agbehadji, I. E. and Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. *Atmosphere*, 15:1352.

Bagkis, E., Hassani, A., Schneider, P., DeSouza, P., Shetty, S., Kassandros, T., Salamalikis, V., Castell, N., Karatzas, K., Ahlawat, A., and Khan, J. (2025). Evolving trends in application of low-cost air quality sensor networks: challenges and future directions. *npj Clim Atmos Sci*, 8(335).

Blasques, F., Koopman, S. J., and Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2):325–343.

Bonas, M. and Castruccio, S. (2025). High-resolution urban air quality monitoring from citizen science data with echo-state transformer networks. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 74(4):905–924.

Creal, D., Koopman, S., and Lucas, A. (2013). General autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Creal, D., Koopman, S. J., Lucas, A., and Zamojski, M. (2024). Observation-driven filtering of time-varying parameters using moment conditions. *Journal of Econometrics*, 238(2):105635.

De Punder, R., Dimitriadis, T., and Lange, R.-J. (2024). Kullback-Leibler-based characterizations of score-driven updates. *arXiv preprint arXiv:2408.02391*.

Durbin, J. and Koopman, S. (2012). *Time Series Analysis by State Space Methods*. Oxford Statistical Science Series. Oxford University Press, Oxford, second edition.

Iyer, S. R., Balashankar, A., Aeberhard, W. H., Bhattacharyya, S., Rusconi, G., Jose, L., Soans, N., Sudarshan, A., Pande, R., and Subramanian, L. (2022). Modeling fine-grained spatio-temporal pollution maps with low-cost sensors. *npj Clim Atmos Sci*, 5(76).

Kampa, M. and Castanas, E. (2008). Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367.

Lin, Y. and Lucas, A. (2025). Functional location-scale models with robust observation-driven dynamics. Technical Report 25-027/III, Tinbergen Institute.

Lovett, G., Tear, T., Evers, D., Findlay, S., Cosby Jr, B., Dunscomb, J., Driscoll, C., and Weathers, K. (2009). Effects of air pollution on ecosystems and biological diversity in the eastern United States. *Annals of the New York Academy of Sciences*, 1162:99 – 135.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., and Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8:14.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, New York.

Sangkham, S., Phairuang, W., Sherchan, S. P., Pansakun, N., Munkong, N., Sarndhong, K., Islam, M. A., and Sakunkoo, P. (2024). An update on adverse health effects from exposure to PM2.5. *Environmental Advances*, 18:100603.

Shah, I., Gul, N., Ali, S., and Houmani, H. (2024). Short-term hourly ozone concentration forecasting using functional data approach. *Econometrics*, 12(2):12.

Shang, H. L. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and Statistics*, 1:184–200.

Tao, Z., Liu, W., Xu, Q., and Wang, P. (2026). Exploiting functional time series prediction for PM2.5 based on multivariate variational mode decomposition and anomaly detection. *Journal of Forecasting*, 45(3):1077–1091.

World Health Organization (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, Geneva. Licence: CC BY-NC-SA 3.0 IGO.

Online Appendix

Modeling Spatial Pollution Dynamics under Irregular Sampling, Missing Values, and Outliers

Henry de Vries[1], Yicong Lin[1,2], and Andre Lucas[1,2]

[1]Vrije Universiteit Amsterdam [2]Tinbergen Institute

Appendix A Technical derivations **Appendix p. 1**

A.1 Derivation of regularization term in surface estimation . . . Appendix p. 1

A.2 Binary Markov chain properties . . . . . Appendix p. 2

Appendix B More empirical details **Appendix p. 3**

B.1 The PM2.5 data . . . . . Appendix p. 3

B.2 Computational details . . . . . Appendix p. 4

B.3 Additional FPCA results on PurpleAir PM2.5 concentration surface . . . . . Appendix p. 5

B.4 Additional sensor-level forecasts . . . . . Appendix p. 8

A Technical derivations

A.1 Derivation of regularization term in surface estimation

To ensure smoothness in the estimation of a certain latent surface μ, we penalize the roughness of the estimated surface. Let $\|\cdot\|_F$ denote the Frobenius norm. We consider the roughness as the integrated squared curvature $\int_{\mathcal{S}} \|\nabla^2_{\boldsymbol{s}} \mu(\boldsymbol{s})\|_F^2 \, \mathrm{d}\boldsymbol{s}$, which is zero for linear functions. Consider again the location function represented as $\mu(\boldsymbol{s}) = (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{\gamma} = \boldsymbol{\phi}(s_1)^\top \boldsymbol{\Gamma} \boldsymbol{\phi}(s_2)$ where $\mathrm{vec}(\boldsymbol{\Gamma}) = \boldsymbol{\gamma}$. The second-order partial derivatives for μ are given by

$$
\begin{aligned}
\mu_{11}(\boldsymbol{s}) &= \frac{\partial^2 \mu(\boldsymbol{s})}{\partial s_1^2} = \boldsymbol{\phi}''(s_1) \boldsymbol{\Gamma} \boldsymbol{\phi}(s_2) = (\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}''(s_1))^\top \boldsymbol{\gamma}, \\
\mu_{22}(\boldsymbol{s}) &= \frac{\partial^2 \mu(\boldsymbol{s})}{\partial s_2^2} = \boldsymbol{\phi}(s_1) \boldsymbol{\Gamma} \boldsymbol{\phi}''(s_2) = (\boldsymbol{\phi}''(s_2) \otimes \boldsymbol{\phi}(s_1))^\top \boldsymbol{\gamma}, \\
\mu_{12}(\boldsymbol{s}) &= \frac{\partial^2 \mu(\boldsymbol{s})}{\partial s_2 \partial s_1} = \boldsymbol{\phi}'(s_1) \boldsymbol{\Gamma} \boldsymbol{\phi}'(s_2) = (\boldsymbol{\phi}'(s_2) \otimes \boldsymbol{\phi}'(s_1))^\top \boldsymbol{\gamma},
\end{aligned}
$$

where $\boldsymbol{\phi}'(\cdot) = (\phi'_1(\cdot), \ldots, \phi'_{K_\ell}(\cdot))$, $\boldsymbol{\phi}''(\cdot) = (\phi''_1(\cdot), \ldots, \phi''_{K_\ell}(\cdot))$, and $\mu_{12} = \mu_{21}$. Consider first $\mu^2_{11}(\boldsymbol{s})$:

$$
\begin{aligned}
\mu^2_{11}(\boldsymbol{s}) &= \left((\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}''(s_1))^\top \boldsymbol{\gamma}\right)^2 = \left((\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}''(s_1))^\top \boldsymbol{\gamma}\right)^\top \left((\boldsymbol{\phi}(s_2) \otimes \boldsymbol{\phi}''(s_1))^\top \boldsymbol{\gamma}\right) \\
&= \boldsymbol{\gamma}^\top (\boldsymbol{\phi}(s_2)\boldsymbol{\phi}(s_2)^\top \otimes \boldsymbol{\phi}''(s_1)\boldsymbol{\phi}''(s_1)^\top) \boldsymbol{\gamma}.
\end{aligned}
$$

Integration over $\mathcal{S}$ in combination with linearity of the Kronecker product yields

$$
\begin{aligned}
\int_{\mathcal{S}} \mu^2_{11}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} &= \boldsymbol{\gamma}^\top \left[\left(\int_0^1 \boldsymbol{\phi}(s_2)\boldsymbol{\phi}(s_2)^\top \mathrm{d}s_2 \right) \otimes \left(\int_0^1 \boldsymbol{\phi}''(s_1)\boldsymbol{\phi}''(s_1)^\top \mathrm{d}s_1 \right) \right] \boldsymbol{\gamma} \\
&= \boldsymbol{\gamma}^\top (\boldsymbol{\Omega}_0 \otimes \boldsymbol{\Omega}_2) \boldsymbol{\gamma}
\end{aligned}
$$

where we define $\boldsymbol{\Omega}_0 = \int_0^1 \boldsymbol{\phi}(z)\boldsymbol{\phi}(z)^\top \mathrm{d}z$, $\boldsymbol{\Omega}_1 = \int_0^1 \boldsymbol{\phi}'(z)\boldsymbol{\phi}'(z)^\top \mathrm{d}z$, and $\boldsymbol{\Omega}_2 = \int_0^1 \boldsymbol{\phi}''(z)\boldsymbol{\phi}''(z)^\top \mathrm{d}z$. By analogous derivations for $\mu^2_{22}(\boldsymbol{s})$ and $\mu^2_{12}(\boldsymbol{s})$,

$$
\begin{aligned}
\int_{\mathcal{S}} \mu^2_{11}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} &= \boldsymbol{\gamma}^\top (\boldsymbol{\Omega}_0 \otimes \boldsymbol{\Omega}_2) \boldsymbol{\gamma}, \\
\int_{\mathcal{S}} \mu^2_{22}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} &= \boldsymbol{\gamma}^\top (\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_0) \boldsymbol{\gamma}, \\
\int_{\mathcal{S}} \mu^2_{12}(\boldsymbol{s}) \mathrm{d}\boldsymbol{s} &= \boldsymbol{\gamma}^\top (\boldsymbol{\Omega}_1 \otimes \boldsymbol{\Omega}_1) \boldsymbol{\gamma}.
\end{aligned}
$$

Combining all the components, the roughness penalty becomes

$$\int_{\mathcal{S}} \left\| \nabla_{\boldsymbol{s}}^2 \mu(\boldsymbol{s}) \right\|_F^2 = \boldsymbol{\gamma}^\top \left[(\boldsymbol{\Omega}_0 \otimes \boldsymbol{\Omega}_2) + 2(\boldsymbol{\Omega}_1 \otimes \boldsymbol{\Omega}_1) + (\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_0) \right] \boldsymbol{\gamma} =: \boldsymbol{\gamma}^\top \boldsymbol{R} \boldsymbol{\gamma}$$

where $\boldsymbol{R}$ denotes the resulting penalty matrix.

A.2 Binary Markov chain properties

Consider the two-state Markov chain of observed state $Z_{j,t} \in \{0, 1\}$ defined in Section 3. Here $Z_{j,t} = 0$ represents an unobserved state at time t at location j, while $Z_{j,t} = 1$ represents an observed state. We again define the probability of an observed state keeping observed by $p = \mathbb{P}(Z_{i,t} = 1 \mid Z_{i,t-1} = 1)$, and an unobserved state keeping unobserved by $q = \mathbb{P}(Z_{i,t} = 0 \mid Z_{i,t-1} = 0)$. We count the transitions then by

$$n_{11} = \sum_{i=1}^{N} \sum_{t=2}^{T} \mathbb{1}\{Z_{i,t-1} = 1, Z_{i,t} = 1\}, \qquad n_{10} = \sum_{i=1}^{N} \sum_{t=2}^{T} \mathbb{1}\{Z_{i,t-1} = 1, Z_{i,t} = 0\},$$
$$n_{01} = \sum_{i=1}^{N} \sum_{t=2}^{T} \mathbb{1}\{Z_{i,t-1} = 0, Z_{i,t} = 1\}, \qquad n_{00} = \sum_{i=1}^{N} \sum_{t=2}^{T} \mathbb{1}\{Z_{i,t-1} = 0, Z_{i,t} = 0\}.$$

The maximum likelihood estimators for p, q are then given by (Norris, 1997, Chapter 1.10):

$$\hat{p} = \frac{n_{11}}{n_{11} + n_{10}}, \quad \hat{q} = \frac{n_{00}}{n_{00} + n_{01}}.$$

B More empirical details

B.1 The PM2.5 data

As the sensors are user-installed, their placement can cause significant measurement errors. PurpleAir therefore assigns a 1- to 5-star rating to the placement of the sensor. We consider only sensors with either a 4- or 5-star rating. Moreover, sensors with observations missing on more than half of the recorded days typically contain data only in the evaluation period, with little training data under our train/test split. Also sensors with a substantial outlier rate contain very small amount of realistic data, and are therefore of little practical relevance. Some examples of these are shown in Figure 7 below. Given these observations, we adopt a pragmatic rule by excluding sensors with missing observations on more than half of the recorded days. Furthermore, we exclude sensors with more than 25 daily PM2.5 concentrations of 500 μg/m^3, or with more than 14 consecutive days of such concentrations.

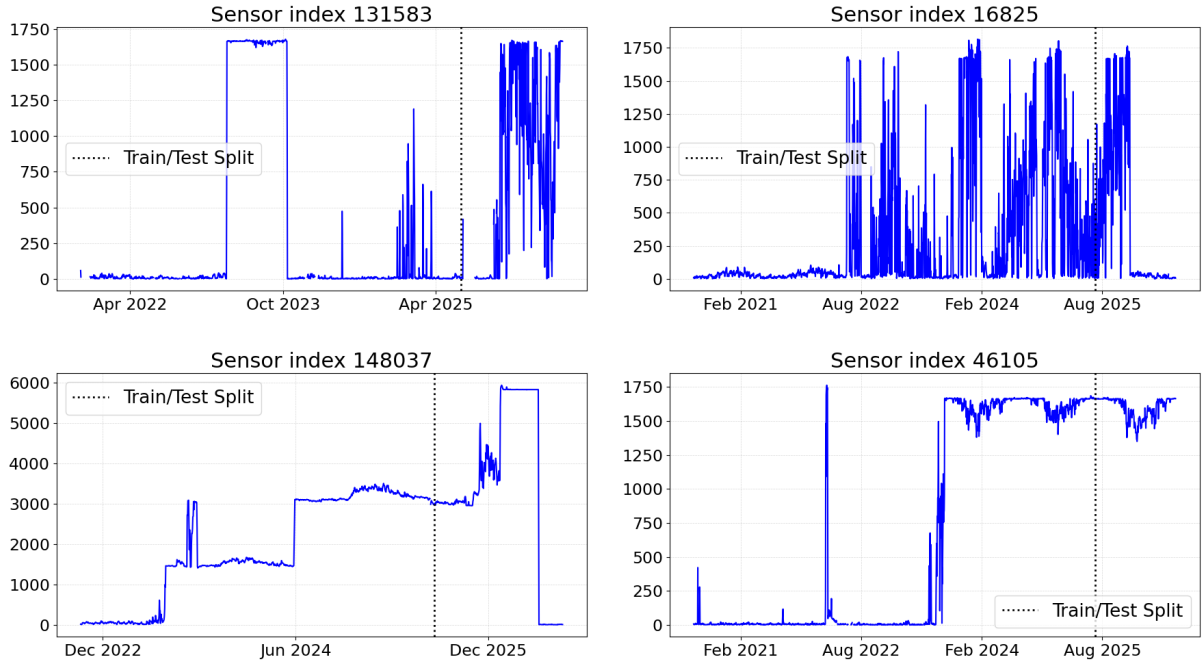


Figure 7: Examples of disregarded PurpleAir PM.25 data. Sensor index corresponds to internal PurpleAir sensor indices.

The PurpleAir sensor indices that are used in our analysis are summarized in Table 3.

Table 3: Sensor indices included in the empirical analysis.

1006	1030	1566	1672	2542	2546	2556	3064
3324	3988	4033	4127	4331	4406	4619	5078
5816	5892	5942	6142	6464	6540	6598	6694
6698	7024	7682	9240	9942	10116	10602	11178
13027	13029	14143	14371	14857	16445	16833	16835
16849	16897	16959	16961	16971	16973	16981	16993
17615	18523	20783	20787	21947	22897	22905	22913
23335	23759	23783	23837	23999	24233	26187	26387
26677	27127	27531	27559	28891	29423	30167	31569
32145	32277	35325	36491	36511	36533	36735	37599
37717	38805	42031	42095	43753	43835	43891	44187
44385	44433	44515	47165	47283	49607	50903	51975
52367	52757	54919	55445	55513	56167	56507	57949
60443	60929	77025	77083	79679	79739	83327	83339
83983	86175	86913	87233	87351	87475	89247	89581
90209	91697	91941	93379	94351	95347	95635	95847
96479	97051	97623	98487	98829	99633	99663	99697
101237	101481	101597	101603	101609	101611	101651	102562
103244	104076	111696	112714	112746	113232	113330	115919
118937	120429	121199	121331	121513	121525	121529	121539
121781	125895	126107	126373	126385	126489	127587	127645
133001	133658	133974	134054	134192	134310	138408	138424
138432	141814	144624	146844	147040	148269	148683	150828
156537	156577	156737	161467	164859	169575	171189	171777
171945	171999	174553	174679	174707	175649	176857	178233

B.2 Computational details

The initial state of the state-space and fGAS model is set to $\widetilde{\gamma}_0 = \boldsymbol{g}_1 = (\boldsymbol{\Phi}_T^{*\top}\boldsymbol{\Phi}_T^* + 0.1\boldsymbol{R})^{-1}\boldsymbol{\Phi}_T^{*\top}\boldsymbol{y}_T^*$ where $\boldsymbol{y}_T^* = (\boldsymbol{y}_{\mathbb{I}_1^{\text{est}}}^\top, \ldots, \boldsymbol{y}_{\mathbb{I}_T^{\text{est}}}^\top)^\top$, $\boldsymbol{\Phi}_T^* = (\boldsymbol{\Phi}_{\mathbb{I}_1^{\text{est}}}^\top \ldots, \boldsymbol{\Phi}_{\mathbb{I}_T^{\text{est}}}^\top)^\top$, $\mathbb{I}_t^{\text{est}} = \mathbb{S}^{\text{est}} \cap \mathbb{I}_t$. The first prediction thus starts at the regularized mean coefficient vector of the observed time series.

The initial variance for the state-space model in the empirical analysis is set to $\boldsymbol{P}_1 = \boldsymbol{I}_K$ (or $\boldsymbol{P}_1 = \boldsymbol{I}_r$ in case of the FPCA basis). In the fGAS model in combination with the B-spline basis, the regularization parameter for the scaled score is fixed at $l = 0.01$. When incorporating an FPCA basis, no regularization is needed, and we set $l = 0$. Finally, all window sizes are set to $J = 20$ and the LS regularization parameters l are selected via cross-validation using a grid search based on in-sample forecasting performance. Specifically, for a given l, the parameters of the VAR(1) are estimated on the first $\lfloor 0.7 \cdot T \rfloor$ days of the training data. The optimal l is then selected as the value that minimizes the forecasting

error over the remaining 30% of the training period. For the smoothers, we simply pick l the forecasting error over the entire training period.

B.3 Additional FPCA results on PurpleAir PM2.5 concentration surface

Figure 8 reports the corresponding cumulative explained variances. Because $r = 7$ principal components explain 95% of the variance, the estimated eigenfunctions of the first 7 principal components are presented in Figure 9 with the corresponding functional scores in Figure 10.

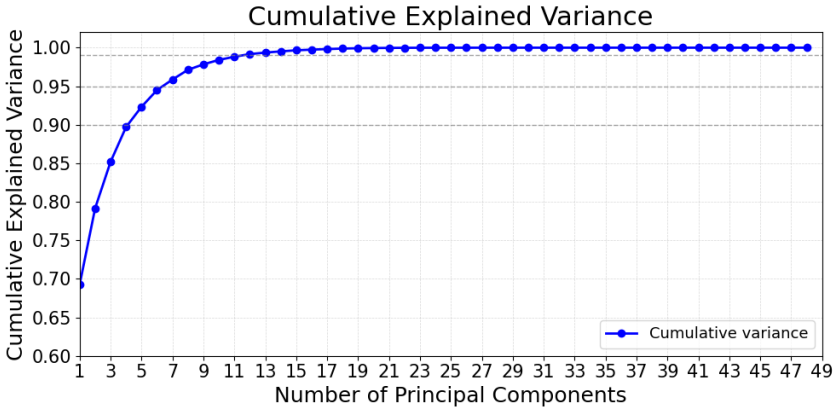


Figure 8: Cumulative explained variance for each selected amount of principal components of FPCA. Explained variance is computed as $\sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{K} \lambda_j \times 100$.

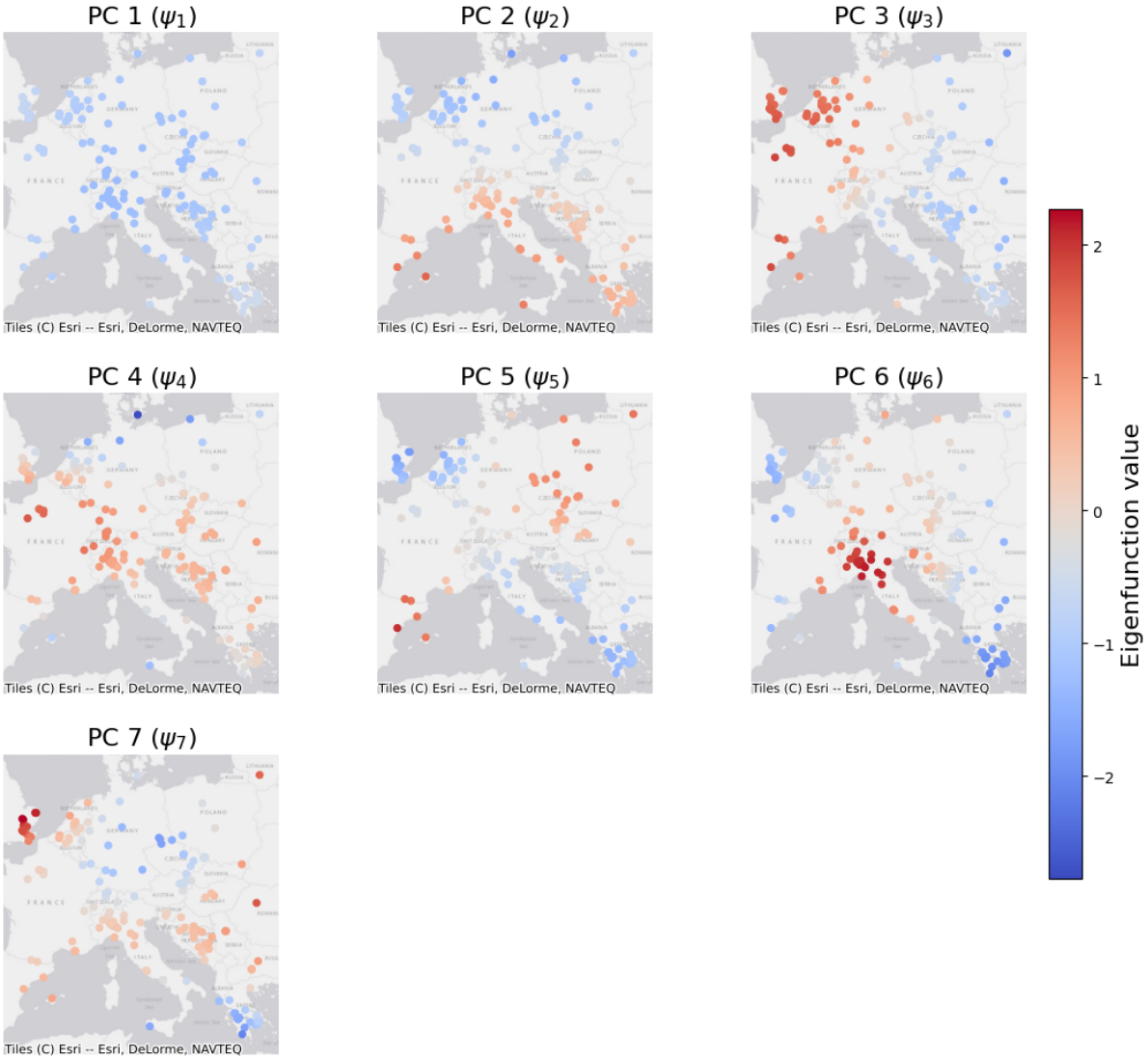


Figure 9: First 7 estimated eigenfunctions across the spatial domain for daily PM2.5 concentrations.

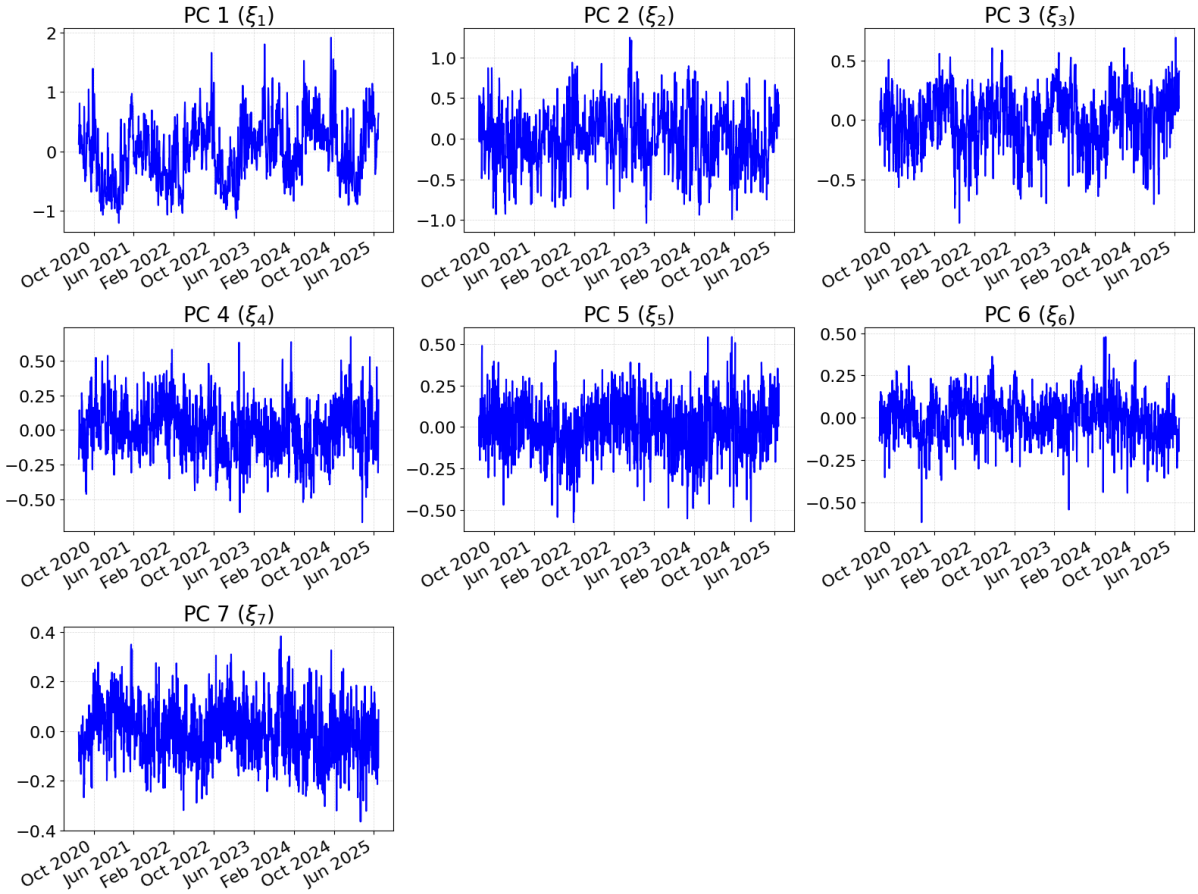


Figure 10: Estimated first 7 functional scores over time for daily PM2.5 concentrations.

B.4 Additional sensor-level forecasts

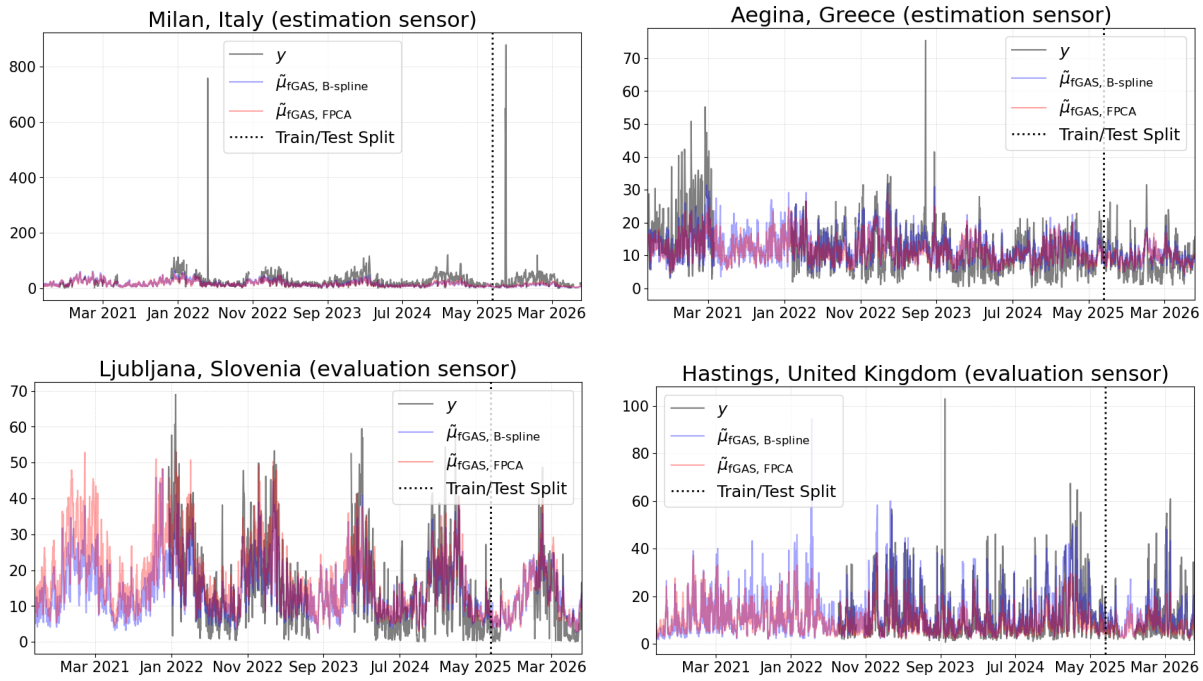


Figure 11: In-sample and out-of-sample predicted PM2.5 concentrations for Milan, Aegina, Ljubljana and Hastings using fGAS. Both regular B-spline basis and FPCA basis with $r = 7$ principal components are considered. y denotes the observed values, $\tilde{\mu}_{\text{fGAS, basis}}$ denotes predicted PM2.5 level, and dotted line represents split between in-sample and out-of-sample fit.

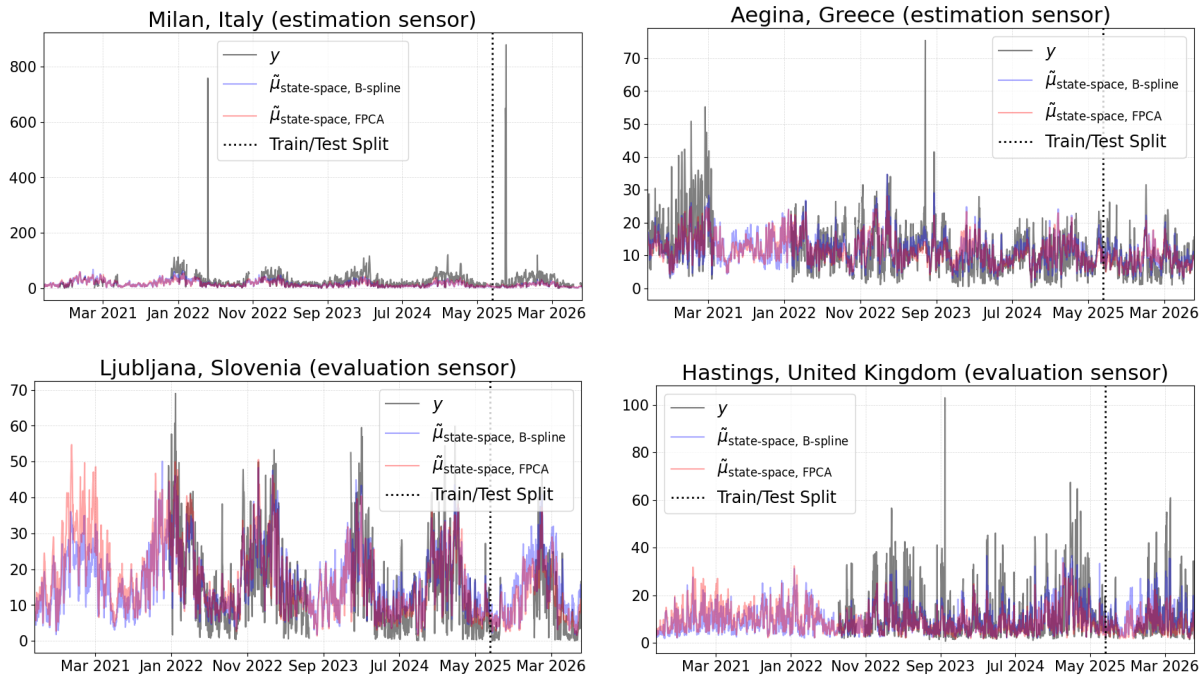


Figure 12: In-sample and out-of-sample predicted PM2.5 concentrations for Milan, Aegina, Ljubljana, and Hastings using state-space. Both regular B-spline basis and FPCA basis with $r = 7$ principal components are considered. y denotes the observed values, $\tilde{\mu}_{\text{state-space, basis}}$ denotes predicted PM2.5 level, and dotted line represents split between in-sample and out-of-sample fit.

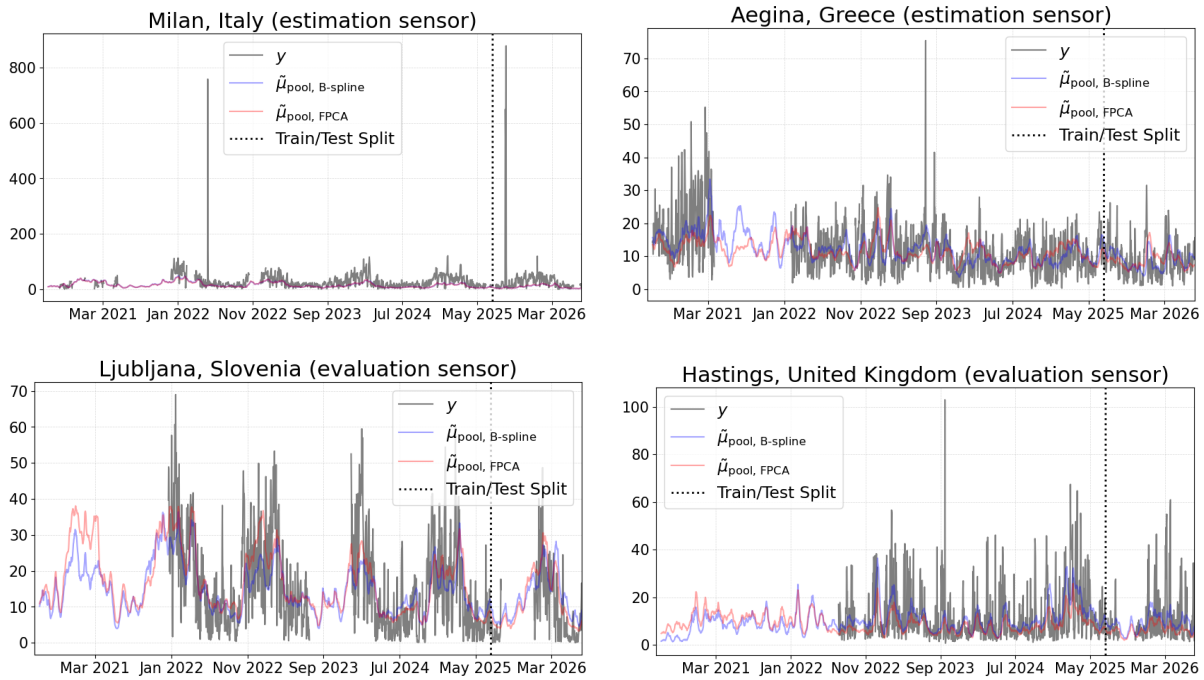


Figure 13: In-sample and out-of-sample predicted PM2.5 concentrations for Milan, Aegina, Ljubljana and Hastings using the pooled estimator. Both regular B-spline basis and FPCA basis with $r = 7$ principal components are considered. y denotes the observed values, $\tilde{\mu}_{\text{pool, basis}}$ denotes predicted PM2.5 level, and dotted line represents split between in-sample and out-of-sample fit.

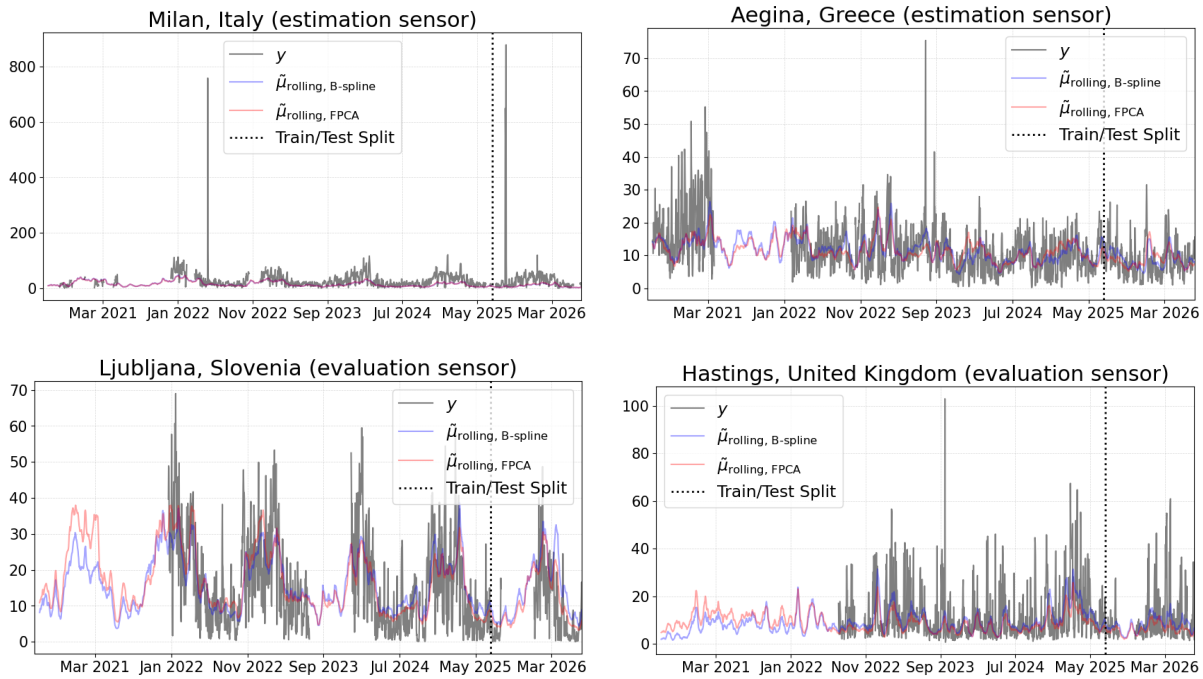


Figure 14: In-sample and out-of-sample predicted PM2.5 concentrations for Milan, Aegina, Ljubljana and Hastings using the rolling average. Both regular B-spline basis and FPCA basis with $r = 7$ principal components are considered. y denotes the observed values, $\tilde{\mu}_{\text{rolling, basis}}$ denotes predicted PM2.5 level, and dotted line represents split between in-sample and out-of-sample fit.

References in Appendix

Norris, J. R. (1997). *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.