

RESEARCH ARTICLE OPEN ACCESS

Enhancing Identifiable Variational Autoencoder in the Presence of Missing Values and Auxiliary Covariates for Spatio-Temporal Ozone Modeling

Mika Sipilä¹  | Sandra De Iaco²  | Claudia Cappello²  | Klaus Nordhausen³ | Monica Palma² | Sara Taskinen¹

¹Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland | ²DES-Sect. of Mathematics and Statistics, University of Salento, Lecce, Italy | ³Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Correspondence: Sandra De Iaco (sandra.deiaco@unisalento.it)

Received: 13 November 2025 | **Revised:** 25 June 2026 | **Accepted:** 13 July 2026

Keywords: blind source separation | deep learning methods | multivariate spatio-temporal process | spatio-temporal forecasting

ABSTRACT

The formation of ground-level ozone follows complex nonlinear photochemical processes that depend on multiple environmental factors and have strong spatio-temporal structures. Environmental data used to study these dynamics usually originate from multiple sources, including in situ monitoring stations and satellite observations. While in situ data provide more accurate measurements, they often suffer from missing values and limited coverage, making it beneficial to incorporate additional satellite-based covariates. To address these challenges for spatio-temporal interpolation and forecasting aims, novel identifiable variational autoencoder (iVAE)-based methods are introduced that explicitly integrate satellite-derived predictors and handle missing values within a nonlinear blind source separation framework. The proposed extensions preserve the identifiability guarantees of the iVAE framework under the missing at random assumption. Performance is benchmarked against established statistical and deep learning approaches by using daily average ozone concentrations in Northern Italy: in interpolation, the proposed method achieves an appreciable reduction of the estimation errors and in forecasting it shows 10 times less training time with respect to the best competing method. The results establish nonlinear blind source separation as a promising approach for spatio-temporal prediction of environmental data.

1 | Introduction

Ground-level ozone (O_3), that is, the ozone concentration found in the lowest layer of the Earth's atmosphere, the troposphere, is a secondary air pollutant that poses significant threats to human health, ecosystems, and agricultural productivity (Emberson 2020; Nuvolone et al. 2018). Unlike stratospheric ozone, which protects the Earth from harmful ultraviolet radiation, tropospheric ozone is a harmful pollutant formed through complex photochemical reactions that occur when primary pollutants, namely nitrogen oxides (NO and NO_2) and volatile organic compounds (VOCs) react with sunlight and stagnant air (Zhang

et al. 2019). Exposure to increased ozone concentrations has been associated with respiratory problems, cardiovascular diseases, and increased non-accidental mortality (Nuvolone et al. 2018). Beyond human health impacts, ozone pollution damages vegetation, reduces crop yields, and contributes to ecosystem degradation (Emberson 2020). Therefore, accurate modeling and prediction of ozone levels are crucial for public health and environmental protection.

The formation of ground-level O_3 follows complex nonlinear photochemical processes that depend on multiple environmental factors and have strong spatio-temporal dependencies (Donzelli

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Environmetrics* published by John Wiley & Sons Ltd.

and Suarez-Varela 2024). While the primary precursors are nitrogen oxides, emitted mainly by vehicular traffic and industrial sources, and VOCs, from both anthropogenic and biogenic sources, many other variables also influence this environmental process, such as relative humidity, wind direction and speed, air temperature and particulate matter ($PM_{2.5}$, PM_{10}). All such environmental variables are characterized by spatial and temporal dependence. For example, in Europe, temperature and solar radiation are higher during the summer as compared to winter, and most pollutants have strong spatial dependence due to substantial variation in terms of traffic and industry activities, as well as between urban and rural areas. Therefore, when modeling and predicting O_3 concentrations, it is important to consider these auxiliary covariates and their spatio-temporal correlation structures. This makes the O_3 modeling task highly complex and often computationally challenging as the relationship among the different variables, as well as the dependencies in both space and time, have to be considered.

In addition to the previously discussed complexities, the modeling of environmental phenomena of O_3 concentration often involves data from various different sources. The two primary sources are ground-based in situ measurements from monitoring stations and satellite-derived estimates. Satellite-based data usually have high spatial and temporal resolutions, making them useful for analyzing spatial and temporal variations on a global scale, although they provide less accurate measurements compared to in situ observations (Donzelli and Suarez-Varela 2024). Satellite observations are measured on regular grid in space allowing wide range of methods, including, for example, convolutional neural network (CNN) based methods (Raitoharju 2022), to be used for forecasting or spatial downscaling purposes. However, modeling performed solely on the basis of satellite data is suboptimal due to the low accuracy of these kind of measurements. On the other hand, in situ observations from air quality monitoring networks provide accurate local measurements, serving as the “ground truth” for model validation. Nevertheless, these measurements are spatially sparse and unevenly distributed, which prevents the use of traditional grid-based modeling approaches. For such irregular data, modeling methods must explicitly account for the spatial relationships between stations, including adjacency, distance, and directional dependencies, since the natural neighborhood structure of regular grids is absent.

Many of the recent studies have focused on predicting O_3 concentrations by combining in situ and satellite data. This is achieved by using various statistical and machine learning methods carried out to predict O_3 concentrations by taking into account other air pollutants, meteorological variables and land-use variables (Li et al. 2020; Wang et al. 2022; Xu et al. 2022). Machine learning algorithms, such as neural networks, random forests and eXtreme Gradient Boosting (XGBoost) have been widely utilized for O_3 prediction thanks to their capability to capture nonlinear dependencies (Li et al. 2020; Pan et al. 2023; Zhou et al. 2023). While these methods have obtained good predictive performance, the spatio-temporal dependencies between the observations have been often neglected. In addition, forecasting the study variable for future time points is not possible with these approaches, as the predictors are available only for current and historical time points.

To account for spatial and spatio-temporal dependencies, kriging has been applied to spatial O_3 prediction (Kethireddy et al. 2014) as well as regression kriging, which incorporates additional covariates through a preliminary regression step, has been used for the spatio-temporal prediction of NO_2 (Van Zoest et al. 2020) and O_3 concentrations (Michael et al. 2019). While these geostatistical approaches explicitly model spatial and spatio-temporal dependence and can incorporate auxiliary covariates, they rely on assumptions of linearity and stationarity that may be inadequate for capturing the complex nonlinear photochemical processes governing O_3 formation.

To model these nonlinear relationships among variables and complex dependence structures, a variety of modern approaches, including deep learning methods, have been considered for O_3 prediction. In the temporal context, O_3 prediction lately been addressed using long short-term memory (LSTM) networks (Guo et al. 2025), hybrid support vector regression (SVR)–LSTM models (Cheng et al. 2021), and functional data analysis approaches (Shah et al. 2024). Recent developments, such as transformers (Vaswani et al. 2017), CNNs and graph neural networks (GNN) (Scarselli et al. 2009) have also produced many new methods for multivariate forecasting, increasing the predictive performance. For example, Chen et al. (2022) utilized CNN-transformer hybrid approach for station-wise forecasting of O_3 . Similarly, Pak et al. (2018) considered CNN-LSTM hybrid model for temporal forecasting, whereas Guo et al. (2023) applied the CNN combined with the LSTM and the gated recurrent unit (GRU) to predict the change of monthly average and extreme atmospheric temperatures. Although these methods capture nonlinear relationships effectively, they treat observations at different stations independently, either ignoring spatial dependence entirely or modeling it only implicitly. For broader surveys of deep learning approaches to air pollutant modeling and forecasting, see Guo et al. (2024), He and Guo (2024), He et al. (2025); Zhang et al. (2022).

GNN based approaches, in particular, can be used for univariate or multivariate spatio-temporal forecasting. GNNs can effectively model the spatial or spatio-temporal dependence structure of irregular observations through adjacency matrices. Iskandaryan et al. (2023), among others, employed various GNN based methods on a multivariate air pollution prediction task. While these approaches do account for the irregular spatial structure, they are typically computationally expensive, require complete data and do not naturally integrate heterogeneous data sources such as satellite covariates alongside in situ measurements. Sipilä et al. (2026a) extended two deep learning architectures (the super-resolution deep residual network and the encoder-decoder-based UNet) with a lightweight temporal module for the spatial downscaling of tropospheric ozone.

Taken together, the existing literature reveals a persistent gap; indeed, frameworks that simultaneously (i) model nonlinear cross-variable relationships and complex spatio-temporal dependencies in spatially irregular settings, (ii) accommodate the large proportion of missing values characteristic of in situ monitoring data, (iii) integrate satellite-derived auxiliary covariates as additional predictors, and (iv) support both spatial interpolation and temporal forecasting within a unified modeling procedure, remain scarce.

A promising candidate for addressing several of these challenges is the nonlinear blind source separation (BSS) framework for multivariate spatio-temporal prediction introduced in Sipilä et al. (2025), Sipilä et al. (2026b). In this approach, cross-variable dependencies are modeled through statistically independent latent components extracted from the observed multivariate data using an identifiable variational autoencoder (iVAE), thereby avoiding the need for explicit parametric modeling of potentially complex cross-variable relationships. The spatio-temporal dependencies of each latent component are then modeled individually using radial basis functions (RBFs) constructed from the observation locations and times, providing a flexible representation of spatio-temporal structure that naturally accommodates the irregular spatial distribution of monitoring stations. However, these existing methods are not designed to incorporate external auxiliary covariates, nor do they provide a principled mechanism for training in the presence of the high rates of missing observations that are typical of air quality monitoring data.

To address this gap, two novel extensions of the iVAE framework (Sipilä et al. 2025, 2026b) are proposed. The first incorporates satellite-derived auxiliary covariates directly into the latent variable model to improve the interpolation of O_3 concentrations across Northern Italy. The second introduces a probabilistic treatment of missing values through a masked-loss training approach, enabling reliable learning of latent components despite the substantial proportion of missing observations in monitoring station data. The resulting models are applied to both spatial interpolation and temporal forecasting tasks, and their performance is evaluated against a range of established competing methods, including geostatistical, machine learning, and deep learning approaches.

The paper is organized as follows. Section 2 recalls the main theoretical aspects of BSS-based modeling methods. The proposed iVAE extensions are then presented in the case of missing values in Section 3.1, as well as when covariates are available in Section 3.2. Then a thorough description of the case study on O_3 concentration follows in Section 4, where the interpolation and forecasting results are compared with the outcomes obtained from various competing statistical approaches. Finally, concluding remarks and discussion are provided at the end of the paper in Section 5.

2 | Multivariate Spatio-Temporal Modeling With BSS

Multivariate spatio-temporal data present significant modeling challenges due to complex dependencies across spatial locations, temporal sequences, and multiple variables simultaneously. BSS offers an approach to simplify the modeling complexity by addressing the cross-variable dependencies by finding independent latent components from the observed data that can be modeled and predicted univariately.

2.1 | Multivariate Spatio-Temporal Processes

In modeling multivariate spatio-temporal phenomena, it is crucial to account for spatio-temporal dependence as well as for the

cross-variable relationships. In such a setting, the multivariate observations $\mathbf{y} := \mathbf{y}(s, t) = (y_1(s, t), \dots, y_S(s, t))^T \in \mathcal{Y} \subset \mathbb{R}^S$ are observed in spatial locations $s \in S \subset \mathbb{R}^D$, where S is the spatial domain and D is the spatial dimension (usually $D = 2$), and at time points $t \in \mathcal{T} \subset \mathbb{R}$, where $\mathcal{T}$ is the temporal domain. The whole spatio-temporal domain can be denoted as $S \times \mathcal{T}$.

The fundamental challenge in modeling this kind of processes lies in capturing the complex dependency structures that exist between variables, across space, and through time. The multivariate spatio-temporal process is usually characterized by the spatio-temporal trend function $\boldsymbol{\mu}(s, t) = (\mu_1(s, t), \dots, \mu_S(s, t))^T$ and spatio-temporal covariance function $C(\mathbf{y}(s, t), \mathbf{y}(s', t'))$ which is an $(S \times S)$ functional depending on the spatio-temporal locations (s, t) and (s', t') . The spatio-temporal process is then the following:

$$\mathbf{y}(s, t) = \boldsymbol{\mu}(s, t) + \boldsymbol{\omega}(s, t), \quad (1)$$

where $\boldsymbol{\omega}(s, t)$ follows a zero-mean distribution with the covariance structure $C(\mathbf{y}(s, t), \mathbf{y}(s', t'))$.

The covariance model is often simplified by assuming either stationarity, which implies that the covariance function depends solely on spatial and temporal distances within the spatio-temporal domain, or separability, indicating that distinct spatial and temporal covariance models can represent the entire spatio-temporal correlation structure. When stationarity in space and time is assumed, the covariance function simplifies to

$$C(\mathbf{y}(s, t), \mathbf{y}(s', t')) = C(\|s - s'\|, |t - t'|), \quad (2)$$

and in case of separability, the covariance can be written as a product of spatial and temporal covariances:

$$C(\mathbf{y}(s, t), \mathbf{y}(s', t')) = C(s, s') \odot C(t, t'), \quad (3)$$

where $\odot$ represents the Hadamard product. Both assumptions, stationarity and separability, can simplify the modeling considerably, but they are not realistic in many real life applications. For example in the context of air pollution modeling, the nonstationarity may occur in space because of variation across geographical regions due to factors such as topography, land use, or emission sources, and in time because of seasonal behavior, changes in air quality control policies and long-term trends. In addition, the nonseparability reflects the common presence of interaction in space and time, that is for different spatial lags, the temporal profile of covariance function are not proportional and similarly for the spatial profile, when different temporal lags are considered. Hence, it is important to account for nonseparability and nonstationarity in spatio-temporal modeling procedures, although this represents a highly complex and computationally challenging task (Porcu et al. 2021; Salvana and Genton 2020).

If the nonstationarity is explained by the trend function $\boldsymbol{\mu}$ in (1), then it is common in Geostatistics to estimate and model the covariance matrix on the stationary residuals. After predicting the spatio-temporal residuals usually on the basis of some valid models for the covariance matrix (Bevilacqua et al. 2015; De Iaco et al. 2003, 2005; Genton and Kleiber 2015; Gneiting et al. 2010), the estimates of the variable of interest can be obtained by adding the trend to the predicted residuals. In this

context, some techniques useful to support the modeling stages through latent variables identification are available in the literature (Cappello et al. 2022; Muehlmann et al. 2023), and various covariance models, characterized by different types of separability, can be used (Allard et al. 2022; De Iaco and Posa 2013; De Iaco et al. 2025, 2013). Note also that the trend function can be supposed to depend on the (spatial or temporal) coordinates or on some covariates. Indeed, in many applications, some auxiliary information or covariates are available from the phenomenon of interest. Let $\mathbf{x} := \mathbf{x}(s, t) = (x_1(s, t), \dots, x_M(s, t))^T$ denote an M -dimensional covariate vector that is additionally observed at spatio-temporal location (s, t) . If such covariates are available more densely than the variables of interest $\mathbf{y}$, the covariates can be used as predictors to determine $\mathbf{y}$ to new spatio-temporal locations. Then, $\mathbf{y}$ can be modeled as function of s, t and $\mathbf{x}$, that is, $\mathbf{y}(s, t, \mathbf{x})$, and the generative spatio-temporal process of $\mathbf{y}$ can be written as.

$$\mathbf{y}(s, t, \mathbf{x}) = \boldsymbol{\mu}'(\mathbf{x}) + \boldsymbol{\omega}'(s, t), \quad (4)$$

where $\boldsymbol{\omega}'(s, t)$ follows a zero-mean distribution with a covariance structure $\mathbf{C}(\boldsymbol{\omega}'(s, t), \boldsymbol{\omega}'(s', t'))$.

Alternatively, it can be assumed that the trend function $\boldsymbol{\mu}$ also includes the spatio-temporal location, and that the residual part $\boldsymbol{\omega}$ includes the covariates $\mathbf{x}$. This is a more relaxed assumption, allowing more general nonstationarity, that is,

$$\mathbf{y}(s, t, \mathbf{x}) = \boldsymbol{\mu}^*(s, t, \mathbf{x}) + \boldsymbol{\omega}^*(s, t, \mathbf{x}). \quad (5)$$

This last formulation, however, increases the complexity of the model, as it requires the $(S \times S)$ covariance function $\mathbf{C}$ of $\boldsymbol{\omega}^*$, which may be nonstationary, and depends on the parameters s, t and $\mathbf{x}$, to be modeled.

2.2 | Spatio-Temporal BSS

In the BSS framework (Comon and Jutten 2010), it is assumed that the observed data $\mathbf{y}$ are generated from independent latent components $\mathbf{z} := \mathbf{z}(s, t) = (z_1(s, t), \dots, z_P(s, t))^T$ through some mixing function $\mathbf{f} : \mathbb{R}^P \rightarrow \mathbb{R}^S$, namely.

$$\mathbf{y} = \mathbf{f}(\mathbf{z}). \quad (6)$$

BSS aims to estimate a function $\mathbf{q}$ which recovers the latent components $\mathbf{z}$ up to some identifiability class, usually up to permutation, signed scale and location shift. Once the latent components are recovered, they can be modeled and predicted univariately. The predicted latent components can be back-transformed to the observation space by using the inverse of $\mathbf{q}$, which is typically estimated simultaneously.

BSS approach simplifies the spatio-temporal modeling considerably, as instead of the $(P \times P)$ matrix-valued covariance function, it is sufficient to estimate P univariate covariance functions. The cross-variable dependencies are accounted for exclusively by the mixing process of the estimated latent components.

In spatio-temporal context, the spatial and temporal dependencies have to be accounted for when estimating the latent

components. Spatio-temporal BSS (ST-BSS) accounts for spatio-temporal dependencies by jointly diagonalizing multiple spatial and temporal scatter matrices in order to identify the latent components (Muehlmann et al. 2023). ST-BSS, however, assumes stationary latent components and that the mixing process is linear, i.e., $\mathbf{f}(\mathbf{z}) = \mathbf{A}\mathbf{z}$, where $\mathbf{A}$ is a $(P \times P)$ mixing matrix. In addition, ST-BSS cannot be directly applied to predict latent components at new spatio-temporal locations; it must be combined with a separate univariate spatio-temporal method for prediction purposes.

Recently, iVAE-based methods have been developed for nonlinear and nonstationary ST-BSS (Sipilä et al. 2025, 2026b). Spatio-temporal iVAE (Sipilä et al. 2025) uses spatio-temporal locations of the observations as auxiliary data to achieve identifiability and estimates the latent components from the original data. Similarly, auxiliary data have been used to achieve identifiability in nonlinear temporal and spatial BSS, for example, in Hyvärinen and Morioka (2016, 2017), Hyvärinen et al. (2019), Khemakhem et al. (2020), Sipilä et al. (2024). For identifiability results of spatio-temporal iVAE methods, see Sipilä et al. (2026b).

Spatio-temporal iVAE assumes nonstationary latent components and a nonlinear injective mixing function $\mathbf{f}$, making the modeling framework more flexible. The model simultaneously learns an encoder–decoder pair for latent component extraction, along with an auxiliary function that estimates the spatio-temporal trend and variance based on the observation's location. Owing to the auxiliary function, the model can directly be applied to predict data at new spatio-temporal locations.

In Sipilä et al. (2026), iVAE has been further extended to a spatio-temporal framework by incorporating the estimation of autoregressive (AR) coefficient functions for each latent component through the auxiliary function. The AR iVAE (iVAEar) assumes a temporal AR structure for the latent components and utilizes previous time points for modeling and prediction.

Both iVAE and iVAEar account for spatio-temporal dependencies as well as cross-variable relationships through the latent components and the mixing function. By assuming the BSS model (6), they can estimate the multivariate spatio-temporal process defined in (1), while allowing for nonstationary trends and covariance structures in both the latent components and the observed data.

However, in previous studies, covariates $\mathbf{x}$, as presented in (5), have not been incorporated into the model. In the following section, the iVAE and iVAEar models are introduced in detail, showing how covariates can be integrated into the same framework.

3 | Identifiable Variational Autoencoders

Following Sipilä et al. (2026b), $\mathbf{u}(s, t) = (u_1(s, t), \dots, u_K(s, t))^T$ is considered to denote the spatio-temporal location dependent auxiliary variable vector. iVAE and iVAEar methods consist of three main components, namely an encoder $\mathbf{g}(\mathbf{y}, \mathbf{u})$, a decoder $\mathbf{h}(\mathbf{z})$ and an auxiliary function $\mathbf{w}(\mathbf{u})$ that are parametrized by neural networks so that vectors $\boldsymbol{\theta}_g, \boldsymbol{\theta}_h$ and $\boldsymbol{\theta}_w$ contain the parameters for encoder, decoder and auxiliary function, respectively.

The encoder aims to estimate the unmixing function by mapping the observation-auxiliary data pair $(\mathbf{y}, \mathbf{u})$ into the mean vector $\boldsymbol{\mu}_{z|y,u} \in \mathbb{R}^P$ and variance vector $\boldsymbol{\sigma}_{z|y,u} \in \mathbb{R}^P$. Through the reparametrization trick (Kingma and Welling 2013), these parameters are used to obtain the latent representation $\mathbf{z}$. The decoder then estimates the mixing function by aiming to reconstruct the original input $\mathbf{y}$ from the latent representation $\mathbf{z}$. The final output of the decoder is denoted by $\mathbf{y}'$.

The auxiliary function maps the auxiliary data $\mathbf{u}$ into distributional parameters $\boldsymbol{\mu}_{z|u}$ and $\boldsymbol{\sigma}_{z|u}$, effectively learning to predict the latent component statistics based solely on auxiliary information. In iVAE method, the auxiliary function additionally outputs the AR coefficients for each latent component, and calculates the distribution mean $\boldsymbol{\mu}$ by assuming an AR process with AR order R over the previous latent components. The AR order R is hyperparameter to be decided by the user. In Sipilä et al. (2026) a larger R , such as $R = 3$, has been proposed as a safer option for latent component estimation. On the other hand for prediction tasks, the parameter R can be selected based on prediction accuracy on validation data.

A key insight of iVAE is that the auxiliary function $\mathbf{w}(\mathbf{u})$ is trained to approximate the encoder's output $\mathbf{g}(\mathbf{y}, \mathbf{u})$. This is achieved through the evidence lower bound (ELBO) optimization:

$$\text{ELBO} = \mathbb{E}_{q_{\theta_g}(\mathbf{z}|\mathbf{y}, \mathbf{u})} \left[\log p_{\theta_h}(\mathbf{y}|\mathbf{z}) + \log p_{\theta_w}(\mathbf{z}|\mathbf{u}) - \log q_{\theta_g}(\mathbf{z}|\mathbf{y}, \mathbf{u}) \right], \quad (7)$$

where $p_{\theta_h}(\mathbf{y}|\mathbf{z})$ controls the reconstruction accuracy and the Kullback–Leibler (KL) divergence $\log p_{\theta_w}(\mathbf{z}|\mathbf{u}) - \log q_{\theta_g}(\mathbf{z}|\mathbf{y}, \mathbf{u})$ forces the distribution given by the auxiliary function to be as close as possible to the distribution of the encoder. Each of the distributions, p_{θ_h} , p_{θ_w} and q_{θ_g} , are assumed to be Gaussian. Specifically, the distributions are set as $p_{\theta_w} = N(\mathbf{z}|\boldsymbol{\mu}_{z|u}, \text{diag}(\boldsymbol{\sigma}_{z|u}))$, $q_{\theta_g} = N(\mathbf{z}|\boldsymbol{\mu}_{z|y,u}, \text{diag}(\boldsymbol{\sigma}_{z|y,u}))$ and $p_{\theta_h} = N(\mathbf{y}|\mathbf{y}', \beta \mathbf{I})$, where $\beta > 0$ is a constant that defines the assumed observational noise and controls the weight of the reconstruction error in the loss function. In non-noisy BSS setting (6), where no observational noise is assumed, a common choice for the weight parameter is $\beta = 0.01$. The loss function (7) is minimized with respect to the parameters θ_h , θ_w and θ_g of the encoder, the decoder and the auxiliary function.

Because the auxiliary function learns to mimic the encoder's behavior during training, it can be used independently for prediction tasks. Once trained, $\mathbf{w}(\mathbf{u})$ can estimate latent component statistics at new locations where only auxiliary data is available, without requiring observations $\mathbf{y}$. This makes iVAE particularly valuable for spatial and temporal prediction applications.

In this paper, radial basis function (RBF) based iVAE methods (Sipilä et al. 2025, 2026b) have been carried out for spatio-temporal prediction, as these methods can predict smooth trend, variance and AR coefficient functions, enhancing the ability to predict to new spatio-temporal locations. In the RBF approach for iVAE, $\mathbf{u}$ is constructed by using radial basis functions to capture rich spatio-temporal dependencies beyond simple coordinate information. This method transforms spatial and temporal locations into RBF representations that can better encode complex spatial and temporal patterns.

The RBFs are constructed separately for spatial and temporal domains as functions of the distance between the spatial or temporal coordinate of the observation and a spatial or temporal node point. The RBFs are defined at multiple spatial and temporal resolution levels H_{k_1} , $k_1 = 1, \dots, K_S$ and G_{k_2} , $k_2 = 1, \dots, K_T$, where K_S and K_T denote the numbers of unique spatial and temporal resolution levels, respectively. The resolution levels control the density of the node points and the scale of the RBFs; at low resolution levels, the node points are sparse and the scale is large, allowing the functions to capture broad, large-scale dependencies, while at high resolution levels, node points are placed densely over the spatial and temporal domains and the scale small, enabling the representation of fine-scale, local variations in the spatio-temporal process. Prior the construction of the RBFs, spatial coordinates are normalized to range $[0, 1]$. The normalized spatial coordinates are denoted as $\mathbf{s}^*$. The auxiliary variables are then constructed using these spatial and temporal RBFs, defined respectively as.

$$v^S(\mathbf{s}^*; \mathbf{o}_i^S, H_{k_1}) = v\left(\frac{\|\mathbf{s}^* - \mathbf{o}_i^S\|}{\zeta_{H_{k_1}}}\right), \quad v^T(t; \mathbf{o}_j^T, G_{k_2}) = v\left(\frac{|t - \mathbf{o}_j^T|}{\zeta_{G_{k_2}}}\right), \quad (8)$$

where $\{\mathbf{o}_i^S\}$ and $\{\mathbf{o}_j^T\}$ denote the sets of spatial and temporal node points, $\zeta_{H_{k_1}}$ and $\zeta_{G_{k_2}}$ are scale parameters defined by the spatial and temporal resolution levels H_{k_1} and G_{k_2} controlling the spread of each basis function, and $v(\cdot)$ is a chosen kernel function. The scale parameters $\zeta_{H_{k_1}}$ and $\zeta_{G_{k_2}}$ are calculated as $\zeta_{H_{k_1}} = \frac{5}{2H_{k_1}}$ and $\zeta_{G_{k_2}} = \frac{T}{G_{k_2} + 1}$, where T is the number of time points, following Chen et al. (2024); Nag et al. (2023) and Nychka et al. (2015).

The temporal node points for a specific resolution level G_{k_2} are constructed by evenly selecting G_{k_2} points that divide the temporal domain into $G_{k_2} + 1$ equally sized segments, i.e., the node points are $\left\{ \frac{T}{G_{k_2} + 1}, \frac{2T}{G_{k_2} + 1}, \dots, \frac{G_{k_2}T}{G_{k_2} + 1} \right\}$. The spatial node points are constructed by similarly dividing both spatial dimensions into equal segments, forming the following grid of node points: $\left\{ (\mathbf{s}_1^*, \mathbf{s}_2^*) : \mathbf{s}_1^*, \mathbf{s}_2^* \in \left(\frac{1}{H_{k_1} + 1}, \frac{2}{H_{k_1} + 1}, \dots, \frac{H_{k_1}}{H_{k_1} + 1} \right) \right\}$. This means that the number of formed spatial RBFs for the resolution level H_{k_1} is $(H_{k_1})^2$. Following W. Chen et al. (2024) and Nag et al. (2023), Gaussian and Wendland kernels (Wendland 1995), v_G and v_W , are used for temporal and spatial RBFs, respectively. Specifically, Gaussian kernel is defined as $v_G(d) = e^{-1/2d^2}$ and the Wendland kernel that is used in this paper is defined as follows:

$$v_W(d) = \begin{cases} (1 - d)^6 (35d^2 + 18d + 3)/3, & d \in [0, 1] \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The sets $\mathbf{H} = \{H_1, \dots, H_{K_S}\}$ and $\mathbf{G} = \{G_1, \dots, G_{K_T}\}$ specify the spatial and temporal bases, respectively, where each H_k (or G_k) indicates the number of spatial (or temporal) node points used at resolution level k . Together, these define the full collection of spatial and temporal radial basis functions. For more details on the construction of the RBFs, see Sipilä et al. (2025).

3.1 | iVAE With Missing Values

In many practical scenarios, observational data are incomplete: sensors may fail, measurement campaigns may be irregular, or

subsets of variables may not be recorded at all spatio-temporal locations. Standard iVAE assumes that the observation vector $\mathbf{y}$ is fully observed, which restricts its applicability when missingness is present. A common way to address missingness is to discard any incomplete samples, but this strategy results in substantial information loss when missing values are frequent. To avoid this, a masking-based extension of the iVAE training objective is proposed.

Let $\mathbf{m} = (m_1, \dots, m_p)^\top$ be a binary mask, where $m_i = 1$ if observation y_i is available and $m_i = 0$ otherwise, and let $\mathbf{y}_{\text{obs}} = \{y_i : m_i = 1\}$ denote the vector that contains all the observed values. Based on the data log-likelihood, the following lower bound is derived:

$$\log p(\mathbf{y}_{\text{obs}}|\mathbf{u}) \geq \mathbb{E}_{q_z(\mathbf{z}|\mathbf{y}_{\text{obs}}, \mathbf{u})} [\log p_y(\mathbf{y}_{\text{obs}}|\mathbf{z}) + \log p_z(\mathbf{z}|\mathbf{u}) - \log q_z(\mathbf{z}|\mathbf{y}_{\text{obs}}, \mathbf{u})], \quad (10)$$

where the distributions p_y , p_z and q_z are the true distributions of the data generating process. Similarly as in regular iVAE loss formulation, these distributions are estimated by encoder, decoder and auxiliary function that are modeled as neural networks with parameters θ_g , θ_h and θ_w , respectively. However, as the length of the vector $\mathbf{y}_{\text{obs}}$ varies based on the number of zeros in the vector $\mathbf{m}$, and the algorithms in practice require constant length, the vector $\mathbf{y}^* \in \mathbb{R}^S$ is formed as follows:

$$y_i^* = \begin{cases} y_i, & m_i = 1, \\ 0, & m_i = 0, \end{cases} \quad (11)$$

where the missing values are replaced with zeros to avoid computational problems. Then, by utilizing the masking variable $\mathbf{m}$ to write $p_{\theta_h}(\mathbf{y}_{\text{obs}}|\mathbf{z}) = \sum_{i=1}^P p_{\theta_h}(y_i^*|\mathbf{z})$ and $q_{\theta_g}(\mathbf{z}|\mathbf{y}_{\text{obs}}, \mathbf{u}) = q_{\theta_g}(\mathbf{z}|\mathbf{y}^*, \mathbf{u}, \mathbf{m})$, the following masked ELBO is obtained:

$$\text{ELBO}_{\text{masked}} = \mathbb{E}_{q_{\theta_g}(\mathbf{z}|\mathbf{y}^*, \mathbf{u})} \left[\sum_{i=1}^P m_i \log p_{\theta_h}(y_i^*|\mathbf{z}) + \log p_{\theta_w}(\mathbf{z}|\mathbf{u}) - \log q_{\theta_g}(\mathbf{z}|\mathbf{y}^*, \mathbf{u}, \mathbf{m}) \right]. \quad (12)$$

In this formulation, the missing entries are excluded from the reconstruction term and in KL divergence, the indices of the missing values are included as conditioning variable in q_{θ_g} . This modification ensures that the unavailable entries are omitted from the loss calculation, while all available information is fully utilized.

In practice, the masked loss (12) can be estimated with iVAE by modifying the encoder so that it takes an additional input $\mathbf{m}$ containing the indices for the missing values. Then, the encoder can learn to distinguish between a true zero observation and a missing entry whose value was imputed with zero. Without adding $\mathbf{m}$ as an input, this ambiguity would lead to bias in estimation of q_{θ_g} . For iVAEar, the same approach can be used to allow missing values in the observed data.

Beyond its computational convenience, the masked ELBO (12) is a lower bound on the observed-data log likelihood $p(\mathbf{y}_{\text{obs}}|\mathbf{u})$, which is the appropriate objective under the missing at random (MAR) assumption. Specifically, MAR assumes that $p(\mathbf{m}|\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{u}) = p(\mathbf{m}|\mathbf{y}_{\text{obs}}, \mathbf{u})$, where $\mathbf{y}_{\text{mis}}$ denotes the unobserved values. Under this assumption, the joint observed-data likelihood factorizes according to the chain rule as

$\log p(\mathbf{y}_{\text{obs}}, \mathbf{m}|\mathbf{u}) = \log p(\mathbf{y}_{\text{obs}}|\mathbf{u}) + \log p(\mathbf{m}|\mathbf{y}_{\text{obs}}, \mathbf{u})$. Furthermore, assuming that the parameters θ in (12) are distinct from those governing the missingness mechanism, the latter is ignorable in the sense of Rubin (1976). In other words, the probability of missingness may depend on the auxiliary variables (including the covariates introduced in Section 3.2) and on the observed values $\mathbf{y}_{\text{obs}}$, but not on the unobserved values $\mathbf{y}_{\text{mis}}$ themselves. Because the masked ELBO targets the correct observed-data log likelihood $p(\mathbf{y}_{\text{obs}}|\mathbf{u})$ under the MAR assumption, the proposed framework remains consistent with the identifiability guarantees established for the underlying iVAE models (Khemakhem et al. 2020; Sipilä et al. 2026b). Related likelihood-based variational approaches for incomplete data under MAR have been proposed by Nazabal et al. (2020) and Mattei and Frellsen (2019). In Section 4, the iVAE and iVAEar models equipped with the masked-loss approach are referred to as masked iVAE and masked iVAEar, respectively.

3.2 | iVAE With Covariates

The iVAE framework relies on auxiliary variables $\mathbf{u}$ to encode spatio-temporal structure. In many situations, however, additional covariates $\mathbf{x}$ are available that provide information about the latent structure of the data. To take advantage of such information, the iVAE model has been extended to incorporate covariates alongside the spatio-temporal auxiliary variables.

Formally, assume the following generative latent variable model:

$$p(\mathbf{y}, \mathbf{z}|\mathbf{u}, \mathbf{x}) = p_y(\mathbf{y}|\mathbf{z})p_z(\mathbf{z}|\mathbf{u}, \mathbf{x}). \quad (13)$$

From data log-likelihood, the following lower bound can be then derived:

$$\log p(\mathbf{y}|\mathbf{u}, \mathbf{x}) \geq \mathbb{E}_{q_z(\mathbf{z}|\mathbf{y}, \mathbf{u}, \mathbf{x})} [\log p_y(\mathbf{y}|\mathbf{z}) + \log p_z(\mathbf{z}|\mathbf{u}, \mathbf{x}) - \log q_z(\mathbf{z}|\mathbf{y}, \mathbf{u}, \mathbf{x})]. \quad (14)$$

To estimate the true distributions q_z , p_y and p_z , the encoder and the auxiliary function are modified so that vector $\mathbf{x}$ of the covariates is added to their input. Then, the encoder $\mathbf{g}(\mathbf{y}, \mathbf{u}, \mathbf{x})$ has access also to the covariates when encoding the observations $\mathbf{y}$, and similarly, the auxiliary function estimates the parameters of the distribution p_z based on both spatio-temporal locations $\mathbf{u}$ and the covariates $\mathbf{x}$. The following ELBO is then maximized with respect to the parameters θ_g , θ_h and θ_w to estimate the assumed generative model:

$$\text{ELBO}_{\text{cov}} = \mathbb{E}_{q_{\theta_g}(\mathbf{z}|\mathbf{y}, \mathbf{u}, \mathbf{x})} [\log p_{\theta_h}(\mathbf{y}|\mathbf{z}) + \log p_{\theta_w}(\mathbf{z}|\mathbf{u}, \mathbf{x}) - \log q_{\theta_g}(\mathbf{z}|\mathbf{y}, \mathbf{u}, \mathbf{x})]. \quad (15)$$

In practice, the covariates $\mathbf{x}(s, t)$ are required to be available for all observed data point $\mathbf{y}(s, t)$ to fit the model. If the objective is to predict the observed variables or the latent components to new, unobserved spatio-temporal locations, the covariates $\mathbf{x}$ should be available also for the new spatio-temporal locations of interest.

Naturally, the masking approach can be combined with the auxiliary variables approach when missing values are present in $\mathbf{y}$ but not in $\mathbf{x}$.

4 | Case Study

The primary objective of this case study is to assess the spatio-temporal interpolation and forecasting of in situ O_3 concentrations across Northern Italy. In particular, this section demonstrates the application of the iVAE methodology developed earlier in the paper and evaluates its performance relative to several established O_3 prediction methods. Through this comparison, we examine the extent to which incorporating satellite-derived covariates and accommodating incomplete observations through the proposed masking strategy can improve predictive performance.

The interpolation and forecasting performance of the competing methods are evaluated separately. For interpolation, the available satellite-derived covariates are incorporated to improve predictive accuracy. In addition, because the in situ observations contain a substantial number of missing values, the analysis must accommodate incomplete observations. Accordingly, the interpolation study employs both iVAE extensions introduced earlier in the paper, the incorporation of satellite covariates and the masking-based treatment of incomplete data, and compares their performance with several state-of-the-art interpolation methods.

As the satellite variables are not available at the future time points, the forecasting is conducted solely based on the in situ variables. Because most of the forecasting methods require complete data, the best performing interpolation method is used to impute missing values in the in situ data prior the forecasting step. The standard iVAE and iVAE methods, along with multiple state-of-the-art prediction methods, are applied for the complete O_3 data. These methods are compared against the proposed masked iVAE and iVAE methods, which in contrast do not use the imputed complete data, but the original incomplete one instead.

Based on these objectives, the case study has been developed through the following steps:

1. exploratory analysis of the available data sets,
2. interpolation of O_3 in situ data to overcome the issue related to the missing values,
3. prediction of O_3 in situ data for future time points not used in the modeling procedures.

A schematic overview of the interpolation/missing value imputation with the proposed iVAE method, is presented in Figure 2 and an analysis of missingness patterns is also provided in the next Section.

Hyperparameter selection is a crucial step for all modeling approaches as it controls model complexity and the balance between fitting training data and generalization to unseen spatio-temporal locations and future time points. To ensure reliable model evaluation, the data have been divided into training, validation, and test sets. The validation set has been used exclusively for hyperparameter selection, while the test set has been reserved for final performance assessment. All hyperparameters have been selected to minimize the Mean Absolute Error (MAE)

on the validation sets. In interpolation stage, all validation data have been used in validation, while the validation in forecasting stage have been performed based on ten-step ahead accuracy. The data preparation is explained in more detail later in this section.

The analysis in the case study has been performed using R 4.4.1 (R Core Team 2024). The code for reproducing all the results of the case study and the implementations for the methods are provided in GitHub repositories.^{1,2,3} All data used for the case study are publicly available as described in Section 4.1.

4.1 | The Analyzed Data

The available data consist of satellite and in situ observations recorded for meteorological and air pollution variables over Northern Italy, from 2018 to 2023. The number of unique time points is 2189, meaning that total number of multivariate observations is 337,106. In situ data, which can be considered the most accurate and reliable measurements in comparison with satellite data, refer to three very hazardous air pollutants, among which O_3 represents the target variable to be estimated. On the other hand, the satellite data consist of two subsets, namely air pollution and meteorological data recorded on grid; these data are adopted as auxiliary variables to enhance the predictions of O_3 in situ measurements.

4.1.1 | In Situ Dataset

The in situ dataset contains daily mean concentrations of O_3 , NO_2 , and PM_{10} (all in $\mu g/m^3$) measured at 154 meteorological stations across Northern Italy from 2018 to 2023. This dataset provides daily aggregated data derived from the hourly measurements available on the Air Quality Portal of the European Environmental Agency (EEA) (Fusta Moro and Fassò 2025).

The spatial locations of the EEA monitoring stations are illustrated in Figure 1, where also the spatial and temporal behaviors of daily O_3 values are shown. Very high daily values have been observed during the summer months, especially in the Northern part of the study area, extending from East to West. This area is the so-called Po Valley, which is also known as the Pianura Padana; it is a large and fertile geographical feature in Northern Italy that extends from the Western Alps to the Adriatic Sea, encompassing a broad, flat plain surrounding the Po River. In the Po Valley, O_3 levels often exceed health and environmental limits, particularly in summer. This is due to elevated levels of precursor pollutants such as nitrogen oxides originating from traffic and industrial activities, coupled with strong solar radiation and the distinctive geographic basin effect that retains pollutants.

From the time plots shown in Figure 1c,d it is also evident that a seasonal behavior characterizes the temporal evolution of O_3 levels during the year. Very high values are seen during the summer (in July and August the values exceed 100–110 $\mu g/m^3$ on average at almost all the monitoring stations), which decrease significantly in the winter months (in winter, O_3 daily concentrations range from 15 to 40 $\mu g/m^3$ for most monitoring stations).

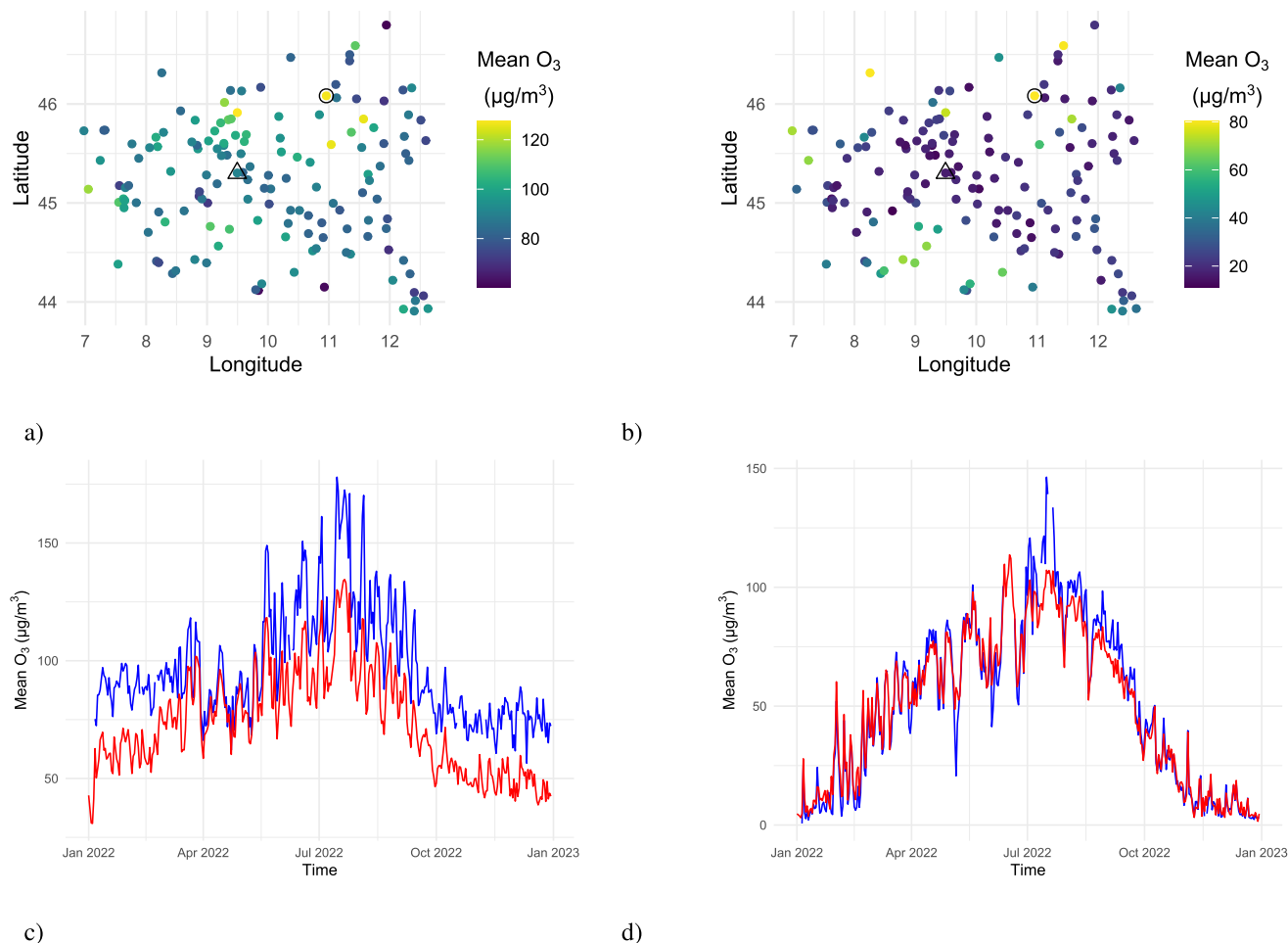


FIGURE 1 | At the top, the spatial plots of the in situ O_3 average values recorded over Northern Italy during summer 2022 (a) and winter 2022 (b). At the bottom, the time plots of the O_3 daily values recorded in situ (blue line) and at the closest satellite grid-nodes (red line) related to two selected EEA monitoring stations indicated with a circle (c) and a triangle (d) in the spatial plots. Note that the color scales are different for the summer and winter average spatial plots, presented in the top row.

In situ pollutant data are characterized by many missing values: 5.6%, 11.0% and 26.9% are the proportions of missing values for O_3 , NO_2 and PM_{10} , respectively. Among the multivariate observations, 34.5% are incomplete, meaning that at least one variable is missing. The missing values occur relatively evenly and randomly across the whole temporal domain. Over spatial domain, the number of missing values is unevenly distributed between the stations. O_3 has been measured at all monitored stations even if with some missing values for a few time points; on the other hand NO_2 and PM_{10} have been completely skipped at, respectively, 4 and 30 stations out of 154. The missing values of O_3 are illustrated in the left hand side of Figure 2, and the spatial and temporal missingness patterns of the in situ variables are illustrated in Figure 3.

4.1.2 | Satellite Datasets

As mentioned above, in the present case study satellite data concerning air pollutants and meteorological variables, have also been used. Air pollution data have been downloaded from the Copernicus Atmosphere Monitoring Service⁴ (AQ-CAMS) and refer to hourly values of O_3 , CO, NH_3 , NO, NO_2 , $PM_{2.5}$, PM_{10} ,

SO_2 and VOC recorded from 2018 to 2023 on a regular ($0.1^\circ \times 0.1^\circ$) grid covering the Northern Italy. For the purposes of this study, the hourly data were aggregated to daily averages, and the resulting daily mean values were used in the subsequent analyses. The total number of 130×130 gridded observations is 2189, meaning that the satellite derived variables are available for all the time points in which the in situ variables are observed.

In terms of spatio-temporal density and number of missing values, the satellite air pollution data are high quality, but the measurements are less accurate due to the fact that they are measured/approximated by satellites. Hence, the satellite data do not give the optimal estimates for the variables at a given location. As an example, in Figure 1c,d the daily O_3 values measured during 2022 at selected EEA monitoring stations have been compared to the daily O_3 satellite data recorded at the closest grid node: the discrepancy between the two different time-series is evident from the plots, especially in the areas (Figure 1c), where the spatial variance is high.

As regards satellite meteorological variables, the hourly values concerning relative humidity (RH, expressed in percentage),

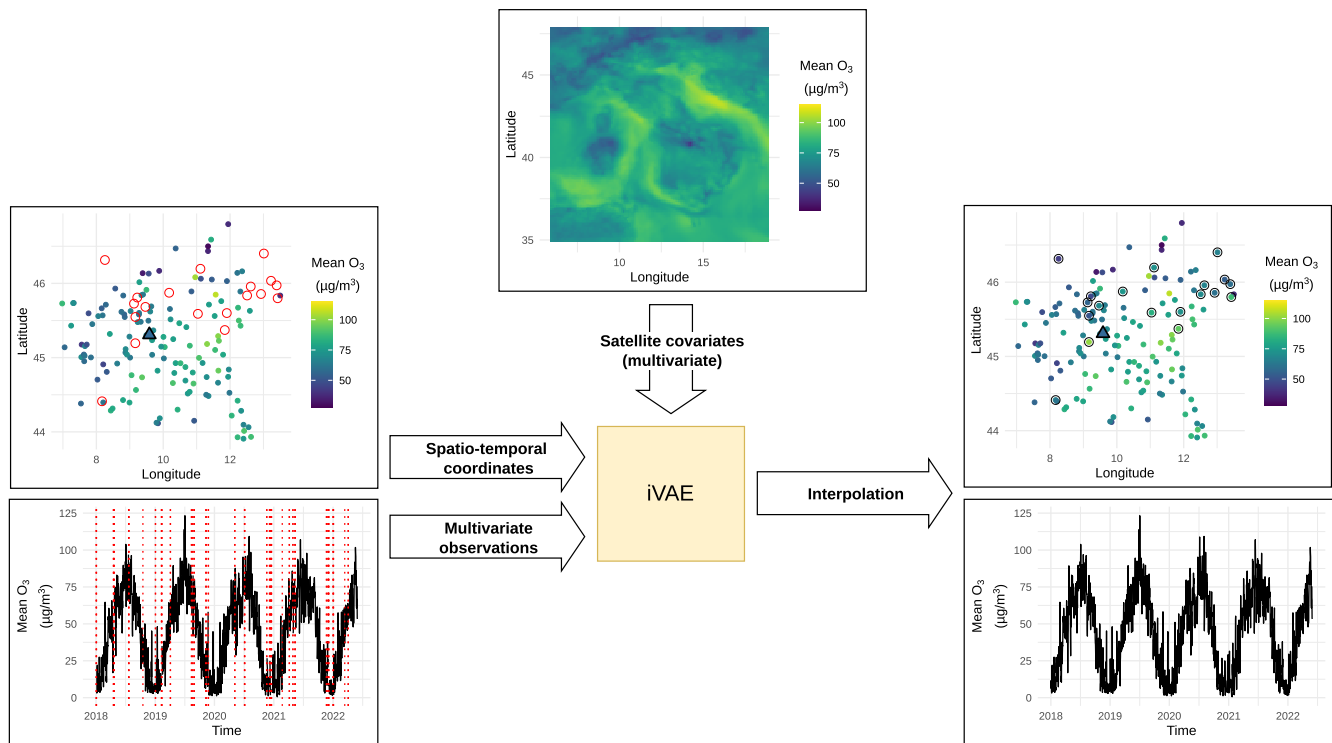


FIGURE 2 | Schematic overview of the interpolation process using iVAE with additional satellite-derived covariates. The iVAE model is trained on incomplete multivariate air pollution data (O_3 , NO_2 , and PM_{10}), which is irregularly and sparsely observed in space, using spatio-temporal coordinates and satellite-derived covariates as auxiliary information. Missing values are then interpolated using the trained model. On the left, the original O_3 data are shown spatially on May 25, 2022, and temporally at the location indicated by the black triangle. In the spatial plot, empty red circles denote O_3 missing values. In the temporal plot, time points marked with dotted red lines correspond to missing observations. On the right, the dataset after imputation is presented. At the top, the multivariate satellite covariates are displayed, including satellite-derived ozone concentrations for May 25, 2022. The satellite covariates are densely observed and complete.

surface solar radiation (SSR, in W/m^2), temperature at 2 m ($T2m$, in $^{\circ}C$), wind speed (WS, in m/s) and low and high vegetation indices (lvi , hvi , in m^2/m^2) have been downloaded from ERA5Land dataset (<https://eeadmz1-cws-wp-air02-dev.azurewebsites.net/download-data/>) and refer to the same grid as AQ-CAMS, for the period 2018–2023. Note that, also in this case, the extracted data have been converted from hourly to daily averages; moreover, these data have no missing values on ground, but all observations from the sea area are missing. The final number of total time points is 2189, similarly as for the in situ air pollution dataset.

In the present case study, the available satellite observations have been utilized as covariates when predicting the in situ measured variables to new unsampled spatio-temporal locations.

4.1.3 | Analysis of Missingness Patterns

The proposed masked iVAE and iVAEar methods assume the missingness mechanism to be either MAR or missing completely at random (MCAR), implying that the probability of missingness cannot depend on the unobserved values themselves or is completely uniform across the entire dataset, that is for the spatial and temporal profiles. This assumption could be violated, for example, if sensors fail more frequently during periods of high pollution. To assess the plausibility of the MAR assumption and rule out obvious evidence of missing not at random

(MNAR) behavior, Spearman correlations have been computed between the binary missingness indicators of the in situ measurements and the corresponding complete satellite-derived variables, which serve as proxies for the unobserved pollution levels.

The results presented in Table 1 indicate that the satellite-derived variables serve as suitable proxies for the in situ measurements, as evidenced by their strong correlations with the corresponding in situ observations. Despite this, all correlations between the missingness indicators and the satellite proxies are below 0.1 in absolute value. This provides strong evidence that sensor failures are not systematically associated with the underlying pollution levels, suggesting that an MNAR mechanism is unlikely.

Furthermore, the average spatial and temporal missingness patterns shown in Figure 3 indicate that the proportion of missing values varies across monitoring stations and, temporally, increases at the beginning of 2022 for all variables. These patterns suggest that the missingness mechanism is more consistent with MAR than with MCAR, as the probability of missingness appears to depend on the spatial and temporal location within the study area.

4.1.4 | Data Preparation

For the aim of the present case study, as previously specified in Section 4, steps (2) and (3), the available in situ air pollution

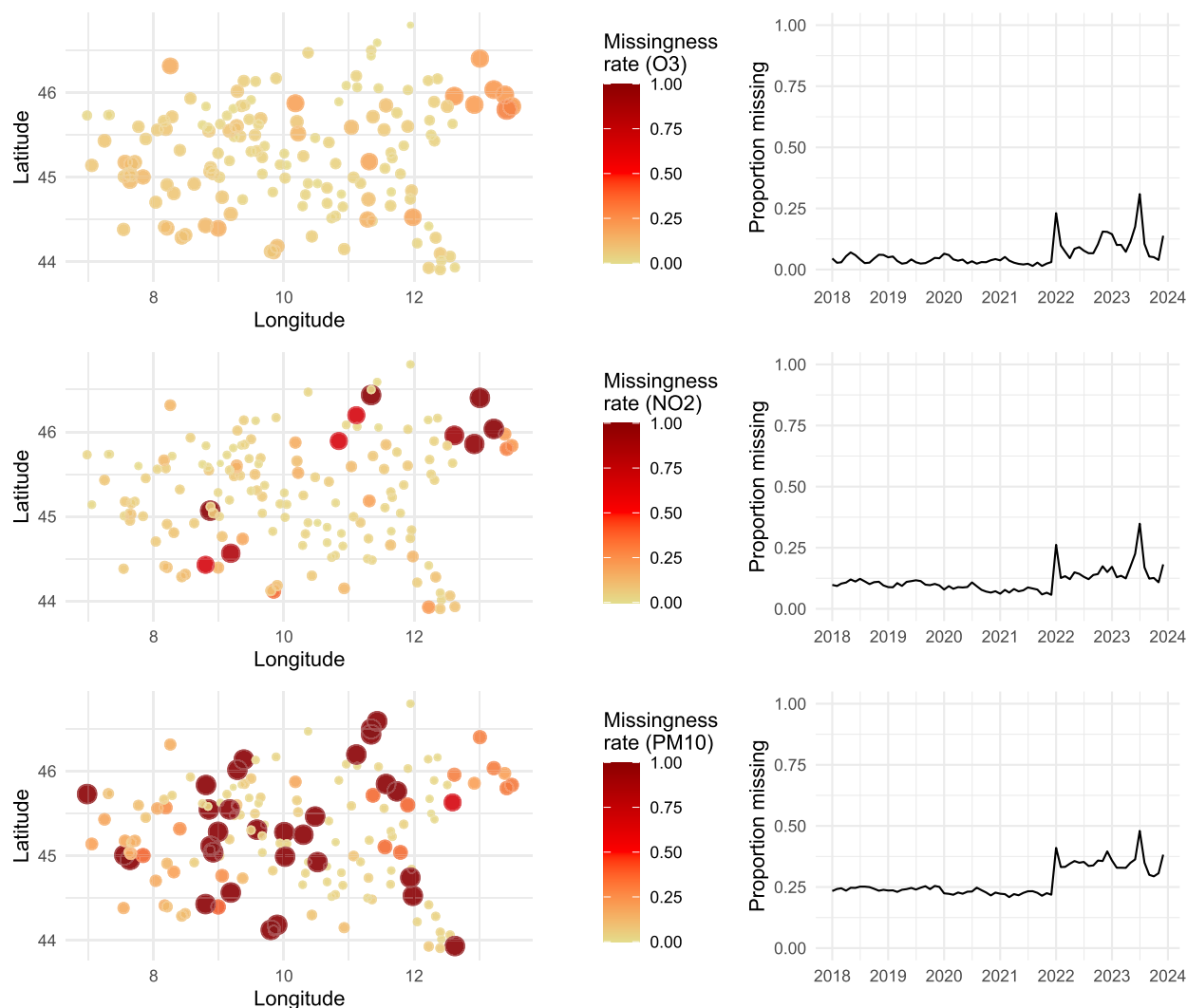


FIGURE 3 | The average spatial (first column) and temporal (second column) missingness patterns, calculated over all time points and all spatial stations, respectively, for the three in situ variables, O_3 , NO_2 and PM_{10} . The dimension of the symbol is proportional to the value of the missingness rate.

TABLE 1 | Pairwise Spearman correlations between in situ variables and the respective satellite derived estimates and the Spearman correlations between the binary missingness indicators and the satellite derived estimates.

	O_3	NO_2	PM_{10}
In situ vs. satellite	0.94	0.82	0.81
Missingness vs. satellite	−0.05	−0.076	−0.002

dataset has been split into training, validation, and test data. All the data recorded between 1/1/2018 and 31/5/2022, containing 1612 time points, have been selected as training set and the data recorded between 1/6/2022 and 31/5/2023, containing 364 time points, have been used as validation set, namely for the hyperparameter selection. The rest of the data (213 time points), recorded between 1/6/2023 and 31/12/2023, have been used as test set for final evaluation of the prediction methods. The training, validation and test datasets contain the total number of spatio-temporal observations 248,248, 56,056 and 32,802, respectively. Note that,

although the validation and test datasets span long time periods, validation and evaluation in forecasting stage are performed based on ten-step ahead prediction accuracy; this is because the test set is chosen to cover the beginning of summer—a period of elevated ozone concentrations—while the validation set is aligned to start at a similar time of year, ensuring that the hyperparameter selection reflects conditions representative of the target evaluation period. Moreover, once validation has been performed, also the validation time period (1/6/2022–31/5/2023) has been included to train the forecast methods for the final evaluation.

As the many of the considered prediction methods require complete datasets, the missing values of the in situ training datasets are imputed prior training. In order to determine the best imputation approach, various spatio-temporal interpolation methods are evaluated prior performing the imputation and the forecasting step. For interpolation (step 2), the data from 1/1/2018 to 31/5/2022, corresponding to the training data period for the forecasting step, are further divided into interpolation

training, validation and test datasets. The distinct validation and test datasets are constructed by randomly selecting 5% of the spatio-temporal locations for each set. The resulting training, validation and test datasets contain 223,424, 12,412 and 12,412 spatio-temporal observations, respectively. The validation set has been used for the hyperparameter selection and the test data to assess the various interpolation procedures. The best performing interpolation method has been carried out to impute all missing values in the in situ data set for the prediction methods that require complete data prior training the models.

For both interpolation and forecasting procedures, the performances of the competing methods have been assessed through the error metrics Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) computed between the true values and the estimated ones for the test data set. In order to assess the uncertainty of the MAE and RMSE point estimates for the trained models, 95% bootstrap intervals are calculated based on 1000 bootstrap samples. In interpolation, the bootstrapping is performed over the spatio-temporal observations of the test set, while in forecasting, the bootstrapping is performed over the spatial stations. For each selected station, all time points are preserved to obtain balanced bootstrap intervals over different forecasting horizons.

As the satellite and in situ datasets are observed at different spatial locations, in situ observations have been paired with the satellite ones recorded at the same grid cell. In particular, the pairing has been performed by selecting the spatially closest satellite data for each in situ monitoring station. Moreover, as the meteorological satellite variables were not observed at grid cells corresponding to the sea, 7 meteorological stations on the shore have been dropped out in order to have no missing values in the considered satellite dataset.

4.2 | Interpolation

In interpolation context, the daily O_3 in situ values have been estimated at the spatio-temporal locations where the measurements were missing. For this aim, the satellite data have been considered as additional predictors $\mathbf{x}$, as they are observed densely and are available for all observed points within the in situ dataset. The interpolation procedure using iVAE is illustrated in Figure 2.

4.2.1 | Competing Methods

As statistical baseline methods, linear regression and linear regression combined with either inverse distance weighting (IDW) or spatio-temporal kriging have been considered (De Iaco and Posa 2012). As IDW and kriging do not allow additional predictors by default, they have been combined with linear regression pre-processing step to predict the trend as in (4). Then, IDW and kriging have been applied on the residuals of linear regression, yielding regression IDW and regression kriging. As machine/deep learning alternatives, random forest (Breiman 2001) and deep kriging (Nag et al. 2023) have been carried out on the study data sets.

Linear regression and random forest perform prediction only based on the satellite predictors and neglect the spatio-temporal

structure of the data. Linear regression estimates the target variable as linear combination of the predictors, while random forest allows nonlinear relationship by combining multiple decision trees, where each tree is trained on a random subset of the data and the predictions are made by averaging across all trees.

Regression IDW and regression kriging account for spatio-temporal structure of the data and for the satellite predictors. For both methods, the mean of ozone is modeled by linear regression as $\mu(s, t) = \alpha + \mathbf{u}(s, t)^\top \beta$. Then, the residuals $r(s, t) = x(s, t) - \mu(s, t)$ are predicted to new spatio-temporal location (s', t') in IDW by $\hat{r}(s', t') = \sum_{i=1}^n w_i(s', t') r(s_i, t_i)$ where the weights w_i are inversely proportional to spatio-temporal distances, and in kriging by $\hat{r}(s', t') = \sum_{i=1}^n \lambda_i r(s_i, t_i)$ where the weights λ_i are determined based on the fitted product-sum variogram model. Note that the parameters of the product-sum model have been estimated by using the L-BFGS-B method in the `gstat` R package (Gräler et al. 2016). The final predictions are obtained by $\hat{x}(s', t') = \hat{\mu}(s', t') + \hat{r}(s', t')$.

Deep kriging is a deep learning alternative of ordinary kriging. It models the data by forming several spatial and temporal radial basis functions, similarly as iVAE, and predicts the observed variable with neural network which takes the radial basis functions as input. It allows nonstationary spatio-temporal mean and covariance functions, but it performs the modeling univariately. In addition, it does not take nearby observations explicitly in account in predictions, and instead models the variable only based on the formed radial basis functions. In order to account for additional satellite variables, a further deep neural network (DNN) module has been included, in order to take the covariates as input to the model. The spatio-temporal DNN module, which takes the values of the radial basis functions as input, is combined with the covariate module in final DNN module, which forms the predictions.

For each method considered, all satellite variables have been used as predictors when the models have been trained.

4.2.2 | iVAE Configuration

In iVAE, the in situ variables, O_3 , NO_2 and PM_{10} , have been considered as observed data and all meteorological variables as covariates that are used for predicting the latent components to new spatio-temporal locations. The number of latent components is set to $P = S = 3$, matching the number of observed pollutants. This ensures that no information is discarded during the blind source separation, as $P < S$ would risk losing relevant information for prediction. Choices of $P > S$ are excluded on theoretical grounds, as the identifiability results of Khemakhem et al. (2020), Sipilä et al. (2026) require $P \leq S$; additional components beyond S would be unidentified and contribute noise rather than signal. The appropriateness of this choice is confirmed empirically through validation, described next in hyperparameter selection. The number of hidden layers in encoder and decoder has been set to 3 with 128 hidden units in each layer, following previous simulation studies on latent component estimation (Sipilä et al. 2025, 2026b). The remaining hyperparameters have been selected through validation to optimize interpolation performance.

4.2.3 | Hyperparameter Selection

All hyperparameters have been selected to minimize MAE on the validation set, containing hold-out spatio-temporal locations. For deep learning methods, hyperparameters have been selected through progressive search strategies, gradually increasing model complexity to balance modeling capacity with computational feasibility, while for statistical methods, standard model selection procedures and grid search over reasonable parameter ranges have been employed.

4.2.3.1 | Regression Kriging. For regression kriging, the empirical variogram has been calculated using temporal lags from 0 to 20 and spatial cutoff parameter of 100 km. Based on the empirical variogram, a spatio-temporal product-sum variogram model (De Iaco et al. 2001) has been fitted by *gstat* R package. The fitted variogram model consists of a spatial spherical component with partial sill 56.06 and range 10257.36 (meters), and a temporal spherical component with partial sill 1.90 and range 21.11 (days). The nugget effects are 0 for spatial part and 5.21 for temporal part. The product-sum interaction parameter is 0.059. The best performing spatio-temporal anisotropy parameter, determined on the basis of validation MAE, is 2500, out of the available options of 1250, 2500, and 5000. The anisotropy of 2500 means that 2.5 km is considered to be equal to one day in time when selecting the nearby observations. A set of 200 nearest observations has been used in prediction step.

4.2.3.2 | Random Forest. The random forest model has been trained using *randomForest* R package (Liaw and Wiener 2002). The number of trees has been set to 50 to balance predictive performance with computational feasibility. Other hyperparameters have been kept at their default values.

4.2.3.3 | Deep Kriging. For deep kriging the number of epochs, spatial and temporal radial basis functions, number of hidden layers and units in each DNN module and type of spatial kernel (Gaussian versus Wendland) have been identified through hyperparameter selection. The spatial and temporal radial basis functions have been selected by progressively adding finer resolution levels (from coarse regional scales to fine local scales), balancing computational complexity with the capacity to capture multi-scale spatio-temporal structure. For epochs, options 20, 40, 60 and 80 have been considered. For the DNN modules, the number of hidden layers and hidden units have been gradually increased to find a setting with sufficient modeling capacity: the spatio-temporal module has been increased from one layer with 128 hidden units up to three layers with 1024, 512 and 256 hidden units, and the covariate module from one layer with 128 hidden units up to three layers with 512, 256 and 128 units, respectively.

Based on the validation MAE, the following best performing parameters have been fixed: 20 epochs, spatial basis of $\mathbf{H} = \{2, 9, 17, 37, 60\}$, temporal basis of $\mathbf{G} = \{9, 17, 37, 73, 211, 555\}$, Wendland spatial kernel, two layers with 256 and 128 hidden units in the covariate module, three layers with 1024, 512 and 256 hidden units in the spatio-temporal module, and one layer with 128 hidden units in the final module. Based on the selected resolution levels $\mathbf{H}$ and $\mathbf{G}$, the radial basis functions have been formed as described in Section 3. All models have been trained with a batch size of 64.

4.2.3.4 | iVAE. For iVAE, the number of epochs, spatial and temporal basis functions, spatial kernel type (Gaussian versus Wendland), and number of layers and hidden units in the auxiliary function have been selected using the same progressive search strategy as for deep kriging, gradually increasing model complexity to find a setting with sufficient modeling capacity. Based on the validation MAE, the best-performing parameters have been: 160 epochs, spatial basis of $\mathbf{H} = \{2, 9, 17, 37, 60\}$, temporal basis of $\mathbf{G} = \{9, 17, 37, 73, 211, 777\}$, Wendland spatial kernel, and three hidden layers with 1024, 512 and 256 units, respectively, in the auxiliary function.

While the selected resolution levels and DNN module structure have been similar to the selected hyperparameters of deep kriging model, the best performing number of training epochs has been high compared to deep kriging model. The difference is most likely because iVAE performs multivariate modeling through latent component extraction, requiring more training time compared to deep kriging.

Finally, to validate the choice of latent dimension, the models have been additionally fitted with $P = S - 1 = 2$ and compared against $P = S = 3$ on the validation set. The choice $P = S$ yields better validation performance, confirming that reducing the latent dimension risks discarding information relevant for prediction.

Table 2 summarizes the selected hyperparameters for all forecasting methods.

4.3 | Interpolation Results

Table 3 reports the MAE and RMSE point estimates and their 95% bootstrap intervals related to the O_3 interpolated data for the test set in Northern Italy obtained using iVAE with covariates and the competing methods discussed above. As a further very simple baseline, satellite-derived O_3 levels have been considered as predictions for spatio-temporal locations of the test set. In this case the predictions have been obtained by pairing each in situ observation with the nearest satellite-derived value. This method, denoted as direct satellite predictions in Table 3, exhibits the worst performance, in terms of both MAE and RMSE, with respect to the other selected interpolation methods.

Standard linear regression yields modest improvements over the direct satellite-derived estimates, while random forest gains slight further improvement on prediction performance. Both methods, however, are considerably outperformed by all methods that incorporate spatio-temporal structure.

Regression IDW and regression kriging achieve a substantial improvement over random forest and linear regression, regression kriging being the superior method of the two. Deep kriging, while allowing for nonstationary dependence structures, performs better than regression IDW but slightly worse than regression kriging, likely due to not explicitly leveraging nearby observations during prediction.

The best performance is achieved by the proposed iVAE method, which outperforms all competitors. iVAE performs the modeling

TABLE 2 | Best performing hyperparameters for interpolation methods, selected based on validation MAE for hold-out spatial locations.

Method	Hyperparameter	Selected value
Regression kriging	Anisotropy parameter	2500
	Nearest observations	200
Random forest	Number of trees	50
Deep kriging	Epochs	20
	Spatial basis	{2, 9, 17, 37, 60}
	Temporal basis	{9, 17, 37, 73, 211, 555}
	Spatial kernel	Wendland
	Covariate module layers	2 (256, 128 units)
	Spatio-temporal module layers	3 (1024, 512, 256 units)
	Final module layers	1 (128 units)
	Batch size	64
	Epochs	160
	Spatial basis	{2, 9, 17, 37, 60}
	Temporal basis	{9, 17, 37, 73, 211, 777}
	Spatial kernel	Wendland
	Auxiliary function layers	3 (1024, 512, 256 units)
	Encoder and decoder layers	3 (128, 128, 128)
iVAE	Latent dimension	3
	Batch size	64

TABLE 3 | MAEs and RMSEs for various methods on the interpolation task of O₃ in Northern Italy, when the covariates are used as predictors.

Interpolation method	MAE	RMSE
Direct satellite predictions	7.47 (7.35, 7.60)	10.56 (10.37, 10.78)
Linear regression	7.17 (7.04, 7.31)	10.14 (9.91, 10.36)
Regression IDW	3.68 (3.62, 3.74)	5.02 (4.90, 5.19)
Regression kriging	3.46 (3.40, 3.52)	4.74 (4.63, 4.85)
Deep kriging with covariates	3.57 (3.51, 3.63)	4.86 (4.75, 5.00)
iVAE with covariates	3.22 (3.17, 3.28)	4.33 (4.25, 4.42)
Random forest	6.40 (6.28, 6.52)	9.17 (8.94, 9.38)

Note: The values in parentheses correspond to the 95% bootstrap intervals over 1000 bootstrap samples of the test data set.

of O₃, NO₂ and PM₁₀ multivariately simultaneously and captures both spatio-temporal dependencies through radial basis functions and cross-variable dependencies through latent component modeling. These results highlight the importance of incorporating covariates, modeling spatio-temporal dependencies, and performing the modeling multivariately for the most accurate ozone interpolation.

The calculated 95% bootstrap intervals of MAE and RMSE are relatively small for all of the methods. The bootstrap intervals of the best performing method, iVAE, and the second best performing method, regression kriging, are by far non-overlapping, suggesting strong statistical significance between the MAE and RMSE obtained by these two methods.

To assess computational complexity of the different interpolation methods, the training and inference times are reported in Table 4,

expressed in seconds (s). The simplest methods, linear regression and regression IDW, are by far the fastest to fit, while iVAE with covariates requires the longest training time. This is expected, as iVAE simultaneously accounts for spatio-temporal structures, cross-variable dependencies, and satellite-based covariates, making it the most information-intensive method among those compared. Deep kriging has the second longest training time, followed by random forest. Despite the longer training time of iVAE, its inference over the full test set remains practically negligible. Regression kriging, by contrast, has a moderate fitting time but by far the slowest inference—over 30 min—making it impractical for operational use at scale.

4.4 | Forecasting

In forecasting stage, the missing values of O₃, NO₂ and PM₁₀ of the train data set are first estimated by using iVAE method which

TABLE 4 | Computation wall-clock times for fitting the models and for inference over the test sets.

	Method	Fit (s)	Inference (s)
Interpolation	Linear regression	0.1	0.002
	Regression IDW	0.1	27.0
	Regression kriging	80.7	2062.9
	Deep kriging with covariates	1386.7	0.7
	iVAE with covariates	9092.5	3.2
	Random forest	612.5	0.36
Forecasting	ARIMA	30.2	0.006
	Kriging	98.6	15.4
	iVAEar	1873.1	0.30
	iVAEar masked	1851.9	0.30
	iVAE	796.2	0.10
	iVAE masked	814.4	0.14
	HiSTGNN	17725.3	1.0

has been the best performing method in terms of interpolation results, as discussed in the previous section. The imputed data set is used for forecasting all methods besides the masked iVAE variants.

When forecasting to future, the future satellite covariates are not available and hence they cannot be utilized as predictors. Therefore, in forecasting, only methods that perform the modeling based on temporal, spatio-temporal and cross-variable dependencies, are available. For the multivariate methods, there are all iVAE variants and the hierarchical spatio-temporal graph neural network (HiSTGNN) developed by Ma et al. (2023), where the three in situ variables O_3 , NO_2 and PM_{10} , and three most important satellite variables, temperature at 2 m, relative humidity and VOC, can be considered as input variables. The iVAE methods additionally use the spatio-temporal coordinates as auxiliary data, where as the HiSTGNN models the spatio-temporal structure through graphs by constructing an input using all spatio-temporal observations from 20 most recent time points at once. For the univariate methods, namely, ARIMA and kriging, only the in situ observed O_3 is used as an input.

All hyperparameter selection is performed based on the validation data. The final models are then trained by including also the validation data, and evaluated using the test data set.

4.4.1 | Competing Methods

As competing methods, two statistical baselines, ARIMA (Box et al. 2015) and ordinary kriging, have been considered. On the other hand, as modern deep learning alternative, HiSTGNN developed in Ma et al. (2023), has been adopted.

Ordinary kriging has been used similarly as in the interpolation scenario, but without the linear regression pre-processing step as the satellite variables are not available.

ARIMA is a classical time series forecasting model that combines autoregression (AR), integration (I), and moving average (MA) components. The model is commonly denoted as ARIMA (p, d, q), where p represents the AR order, d the degree of differencing needed for stationarity, and q the order of the moving average. ARIMA models linear relationships in univariate time series by expressing the current observation as a linear combination of past observations, past forecast errors, and a stochastic term. The differencing component handles nonstationary data by transforming it into a stationary series. ARIMA does not account for spatial dependence or the dependence between different variables.

For ARIMA and kriging, seasonal trends have been removed to fulfill the stationarity assumption. In particular, the detected seasonal component has been modeled as follows

$$x_i(s, t) = \beta_{0,i} + \beta_{1,i} \cos(2\pi t/365) + \beta_{2,i} \sin(2\pi t/365) + x_{res,i}(s, t),$$

where residuals $x_{res,i}$ have been predicted using ARIMA and kriging.

HiSTGNN is a modern deep learning architecture designed for spatio-temporal forecasting that captures multi-scale spatial and temporal dependencies through a hierarchical graph structure (Ma et al. 2023). The model constructs multiple graph representations at different spatial resolutions, from fine-grained local interactions to coarse-grained global patterns, enabling it to model both short-range and long-range spatial dependencies. The temporal dependencies within each subgraph are modeled using temporal convolution to capture the temporal evolution. HiSTGNN incorporates cross-dependencies between input variables as well as spatio-temporal dependencies. It also allows nonstationarity in the dependence structures.

4.4.2 | iVAE and iVAEar Configurations

In iVAE and iVAEar, the time points relative to seasonal period instead of the absolute time points have been used to form the temporal radial basis functions. This approach keeps the future values of the radial basis functions within the boundaries of the training data, allowing the methods to generalize better to the future. In order to allow difference between seasons, additional one-hot encoded year factor has been added to the spatio-temporal auxiliary data. In encoder and decoder, three hidden layers with 128 units in each, and leaky ReLU activation function have been applied, following previous studies (Sipilä et al. 2025, 2026). The learning rate is set to 0.0001 and the batch size to 64. The most crucial hyperparameters, spatial and temporal bases, the depth and number of units in auxiliary function and the type of the spatial kernel, are selected through hyperparameter selection. The number of latent components is set to $P = S = 6$, following the same reasoning as in the interpolation case (Section 4.2). The appropriateness of this choice is confirmed through validation as described in the hyperparameter selection.

4.4.3 | Hyperparameter Selection

All hyperparameters, that were selected on the validation data, have been evaluated based on ten-step-ahead validation MAE.

4.4.3.1 | ARIMA. The best parameters have been chosen for each station individually on the basis of the corrected AIC (the well known Akaike Information Criterion) computed for AR of order $p = 0, \dots, 5$ and MA of order $q = 0, \dots, 5$. The integration parameter has been determined through the KPSS test (Kwiatkowski et al. 1992) selecting from options 0 and 1. Then, it has been set to 1 for all stations, according to the higher validation accuracy.

4.4.3.2 | Kriging. The spatio-temporal anisotropy parameter has been set to 1250 based on the average ten-step-ahead validation MAE, selected from options 1250, 2500, and 5000. A set of 200 nearest observations has been used in the prediction step.

4.4.3.3 | HiSTGNN. The initial neighborhood graphs have been constructed by selecting all spatial points within 20 km to be connected observations, representing a typical scales of local air quality influence. The 20 most recent observations in time have been considered to predict the 10 time steps ahead. The seasonal time point of the observation has been included as predictor to allow the model to learn seasonal differences. The model has been trained multivariately for the pollutants O_3 , NO_2 and PM_{10} recorded in situ, and the satellite environmental variables which are the most important in terms of influence on air quality conditions, namely relative humidity, temperature and VOC. Based on preliminary experiments, the default values for the model layers have been chosen: the number of convolution and residual channels have been set both to 32, while the number of skip channels and the one for the final convolution channels have been fixed, respectively, to 64 and 128. The number of graph convolution heads has been set to one, the depth of the graph convolutions to two and the number of consecutive graph convolution layers in the model to two. With these parameters, the model already contains approximately 1.8 million parameters and makes it computationally heavy to train, which is the reason for not performing an extensive hyperparameter search. Then, the number of training epochs have been selected by monitoring the validation accuracy to avoid overfitting. The number of epochs for the final model is 40.

4.4.3.4 | iVAE and iVAEar. For iVAE and iVAEar, the hyperparameter search has been similar as in the interpolation step previously discussed. For iVAEar the following best performing parameters have been selected based on 10 steps ahead validation MAE: AR of order $R = 3$, one layer with 64 units in the auxiliary function, spatial basis of (2, 9, 17, 37), temporal basis of (9, 17, 37) and 70 training epochs. For masked iVAEar, the same parameters have been chosen, with except that the AR order is $R = 4$. For both iVAE and masked iVAE methods, the following parameters have been selected: one layer with 128 units in the auxiliary function, the spatial basis of (2, 9, 17, 37), the temporal basis of (9, 17, 37) and 60 training epochs. As in the interpolation case, the choice $P = S = 6$ is confirmed against $P = S - 1 = 5$ on the validation set, with $P = S$ yielding better validation performance for all iVAE and iVAEar variants.

Table 5 summarizes the optimal hyperparameters for all forecasting methods.

TABLE 5 | Best performing hyperparameters for forecasting methods, selected based on validation average MAE for ten-step-ahead predictions.

Method	Hyperparameter	Optimal value
ARIMA	AR order (p)	0–5 (station-specific)
	MA order (q)	0–5 (station-specific)
	Integration order	1
Kriging	Anisotropy parameter	1250
	Nearest observations	200
HiSTGNN	Spatial neighborhood	20 km
	Temporal window	20 time steps
	Convolution/residual channels	32
	Skip channels	64
	Final convolution channels	128
	Graph convolution heads	1
	Graph convolution depth	2
	Consecutive GC layers	2
	Training epochs	40
iVAEar	AR order	3
	Auxiliary function layers	1 (64 units)
	Spatial basis	(2, 9, 17, 37)
	Temporal basis	(9, 17, 37)
	Training epochs	70
Masked iVAEar	Latent dimension	6
	AR order	4
	Auxiliary function layers	1 (64 units)
	Spatial basis	(2, 9, 17, 37)
	Temporal basis	(9, 17, 37)
iVAE/Masked iVAE	Training epochs	70
	Latent dimension	6
	Auxiliary function layers	1 (128 units)
	Spatial basis	(2, 9, 17, 37)
	Temporal basis	(9, 17, 37)
	Training epochs	60
	Latent dimension	6

4.5 | Forecasting Results

Table 6 presents the MAE, the RMSE and their 95% bootstrap intervals of daily average O_3 concentrations over Northern Italy for all evaluated forecasting methods. In particular, MAE and RMSE for one-step-ahead predictions, as well as MAE and RMSE averaged over the first ten forecasting steps, calculated on the in situ test set, have been considered and the most accurate results have been written in boldface. For one-step-ahead forecasting, masked iVAE approach achieves the lowest error, followed by masked iVAEar and HiSTGNN. However, when forecasting further into the future, the best performance is shown by the HiSTGNN and the masked iVAEar, with no significant differences, although the former performs slightly better, closely followed by the latter.

The masked iVAE variants, trained on the original training data with missing values, are performing better than the counterparts that are trained on the imputed dataset, especially in

TABLE 6 | MAEs and RMSEs for the adopted forecasting methods of daily average O₃ concentrations in Northern Italy.

Method	1 step MAE	1 step RMSE	1–10 steps MAE	1–10 steps RMSE
ARIMA	6.33 (5.42, 7.32)	8.35 (7.37, 9.37)	11.51 (10.94, 12.11)	14.63 (13.79, 15.61)
Kriging	6.86 (5.93, 7.78)	8.49 (7.47, 9.52)	11.23 (10.75, 11.69)	14.34 (13.52, 15.25)
iVAEar	6.53 (5.63, 7.55)	8.42 (7.35, 9.40)	11.11 (10.58, 11.57)	14.28 (13.48, 15.13)
iVAEar masked	5.95 (5.18, 6.84)	7.59 (6.63, 8.60)	10.66 (10.10, 11.28)	13.77 (12.74, 14.91)
iVAE	7.94 (6.91, 8.99)	9.91 (8.76, 11.01)	11.38 (10.84, 11.90)	14.78 (13.86, 15.80)
iVAE masked	5.80 (4.88, 6.77)	7.93 (6.62, 9.24)	12.06 (11.51, 12.56)	15.52 (14.69, 16.38)
HiSTGNN	6.12 (5.39, 6.95)	7.81 (6.64, 9.13)	10.58 (10.12, 11.03)	13.70 (12.92, 14.63)

Note: The errors are calculated for one step (day) ahead and the mean of the first 10 steps (days) ahead predictions of the in situ O₃ test data set. Bold values in parentheses correspond to the 95% bootstrap confidence intervals over 1,000 bootstrap samples.

one-step-ahead forecasts. The masked iVAEar yields lower errors also for ten-steps-ahead forecasts, but the masked iVAE ranks the worst in ten-steps-ahead forecasting. Overall, iVAE methods perform worse than iVAEar methods.

Kriging performs decently in both one-step-ahead and ten-steps-ahead prediction. Although its one-step-ahead performance is second worst, it achieves nearly as low prediction errors as iVAEar. In ten-step-ahead forecasts, kriging achieves similar prediction errors as iVAEar, being the fourth best method. ARIMA performs relatively well in one-step-ahead forecasting, achieving the fourth best performance, but is the second worst in ten-steps-ahead forecasting.

The 95% bootstrap confidence intervals for the forecasting metrics are considerably wider than those for interpolation, and overlap notably across all methods. This indicates that the differences in forecasting performance between methods are statistically less conclusive than in the interpolation case. The larger uncertainty is consistent with the nature of the forecasting task: predictions are made beyond the observed time window, so there is no information from the future side of the interval, and the spatio-temporal distance to adjacent observations becomes large. Together, these factors increase prediction uncertainty and reduce the ability to distinguish between methods based on a single test period.

The training and inference times for the forecasting methods are reported in Table 4. Based on the methods, ARIMA and kriging are the fastest to fit, while the iVAE variants require moderate training times. The autoregressive iVAEar variants take approximately twice as long to train as the non-autoregressive iVAE variants, reflecting the additional complexity of modeling temporal dependencies. Most notably, HiSTGNN requires almost ten times longer training than iVAEar, representing a substantial computational overhead while providing similar forecasting accuracy. For inference, kriging is the slowest by a considerable margin, while all other methods complete inference in at most 1 s. Among the neural network approaches, iVAE variants perform the inference the fastest, followed by iVAEar approaches, and the HiSTGNN is the slowest.

Further comparison between two best performing methods, masked iVAEar and HiSTGNN, has been conducted by comparing their spatial prediction error maps, that are provided in

Figures 4 and 5 for the prediction steps 1–10. Based on the figures, HiSTGNN has slightly higher prediction errors in Milano area (around coordinates $(X, Y) = (550, 5050)$), especially at steps 2 and 3. In contrast, masked iVAEar has higher errors in some of the locations that are at the edge of the study area, for example, the most south-east location (see steps 5 and 6). Both of the methods have the three similar large errors in steps 6 and 7, which are individual observations with very high ozone concentration.

When comparing the masked iVAEar and HiSTGNN, it should be noted that HiSTGNN has been trained particularly by minimizing the ten-steps-ahead prediction error, while masked iVAEar is trained by maximizing the lower bound of the data log-likelihood. As result, iVAEar can more robustly make predictions to any forecasting horizon once trained. Furthermore, HiSTGNN is significantly more complex model, requiring substantially greater computational resources and longer training times when compared to iVAE methods.

The superior one-step-ahead prediction performance of masked iVAE variants compared against their non-masked counterparts, and their competitive performance in longer forecasting horizons, is promising. These methods require fewer computational steps and avoid missing values imputation, which may otherwise lead to noise in the imputed training data. This suggests that learning from the original partially observed data can be beneficial performance wise, while being computationally more efficient and simpler approach.

5 | Conclusions and Discussion

To address the challenges of fully multivariate spatio-temporal modeling, the existing iVAE framework was extended to incorporate observed covariates directly into the estimation and prediction of the latent component. In addition, a masking strategy was introduced to explicitly handle missing values, resulting in the masked iVAE approach. These extensions enabled the incorporation of satellite-based covariates together with spatio-temporal coordinates for interpolation, while simultaneously allowing latent components to be estimated even at locations with incomplete observations.

The proposed iVAE extensions were evaluated against several state-of-the-art methods for interpolation and forecasting of in situ observed O₃ concentrations in Northern Italy. In the

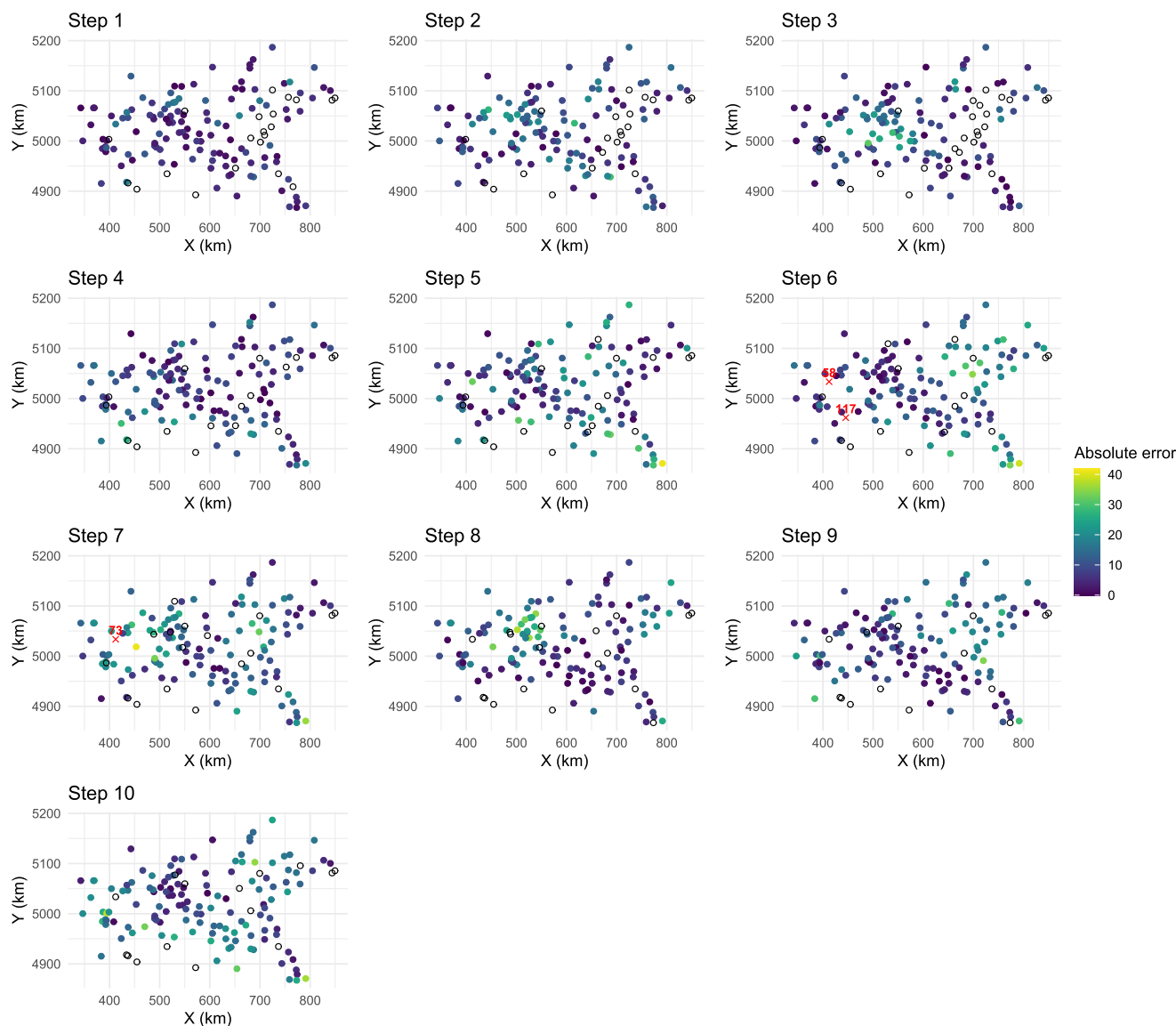


FIGURE 4 | The absolute prediction error maps of in situ measured daily O_3 concentrations for prediction steps 1–10 for masked iVAEar approach. The three large errors in time steps 6 and 7 are marked with red crosses and their respective values to keep the color scale clear for majority of the data. The predictions are available also for the spatio-temporal points marked with empty circle, but these points are missing in the test data set.

interpolation task, the masked iVAE with satellite covariates achieved the strongest predictive performance, clearly outperforming competing methods. In the forecasting task, masked iVAEar yielded the best one-step-ahead prediction accuracy and remained competitive at longer forecast horizons when compared to the modern and computationally intensive deep learning model HiSTGNN. Notably, masked iVAEar attained its strong performance without requiring prior imputation of missing values and with substantially shorter training times than HiSTGNN.

Although the present study focuses on ozone prediction, the proposed framework is not limited to this specific application. The methodology is applicable to a broad class of multivariate spatio-temporal prediction problems where latent factor structures and partially observed covariates arise. For instance, similar modeling challenges occur in environmental monitoring tasks such as water quality assessment, hydrological prediction, and other ecological or atmospheric applications. A

systematic evaluation across different pollutants, geographical regions, and related monitoring domains would provide further insight into the generalizability and practical utility of the proposed approach.

Despite its advantages, the proposed framework has certain limitations. In particular, the iVAEar specification assumes a strict autoregressive temporal structure, which implies a separable spatio-temporal covariance structure. This assumption may restrict flexibility in settings where space–time interactions are more complex. Future research could extend the model to non-separable spatio-temporal covariance structures.

Another promising direction is to further enhance the forecasting pipeline through improved integration of satellite-based information. One possible extension is to first predict full satellite grids using CNN architectures such as U-Net combined with temporal models (e.g., LSTM), and subsequently use these

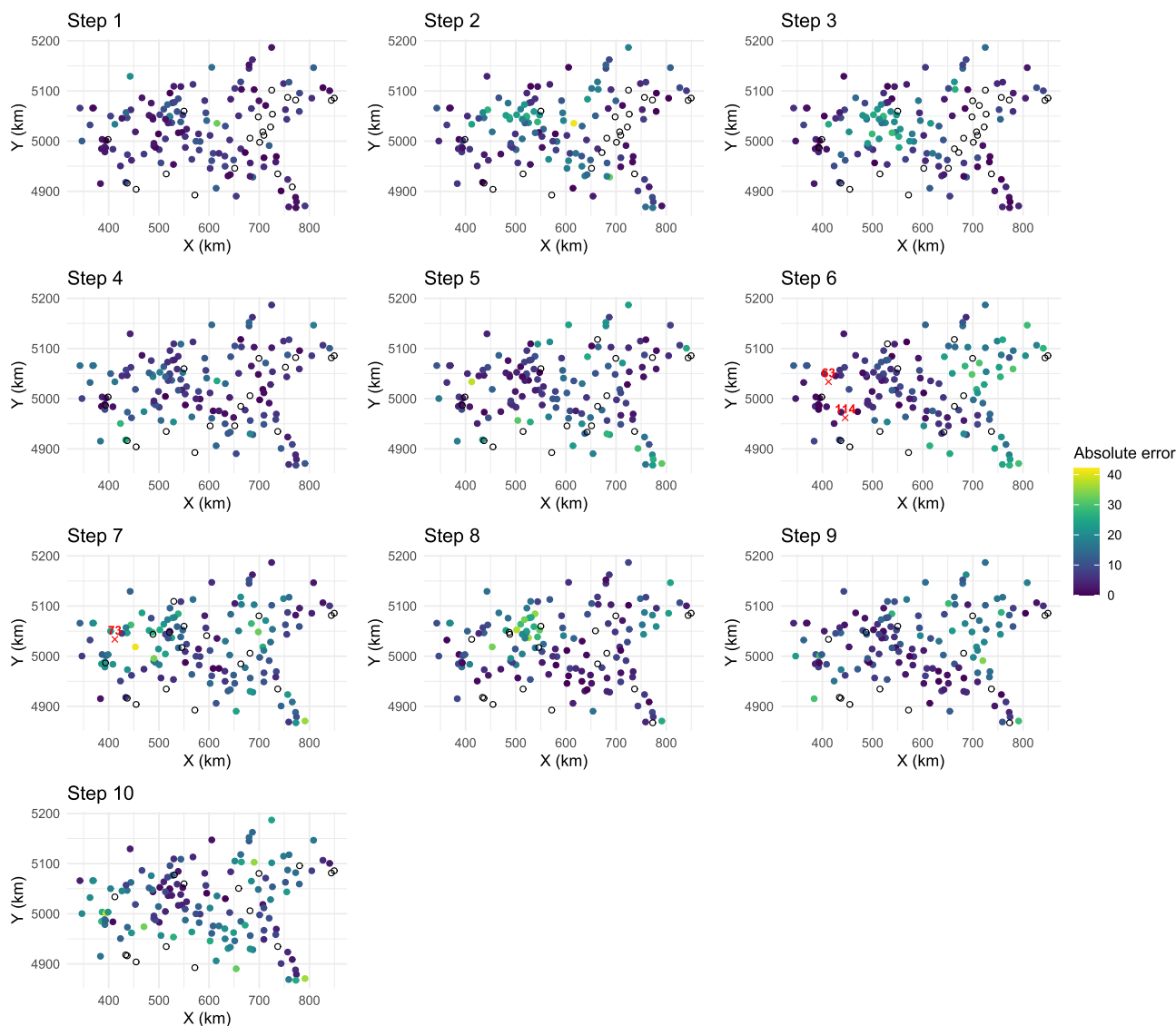


FIGURE 5 | The absolute prediction error maps of in situ measured daily O_3 concentrations for prediction steps 1–10 for HiSTGNN approach. The three large errors in time steps 6 and 7 are marked with red crosses and their respective values to keep the color scale clear for majority of the data. The predictions are available also for the spatio-temporal points marked with empty circle, but these points are missing in the test data set.

predicted high-resolution covariates within the iVAE forecasting framework. Such an approach could leverage high-resolution remote sensing information more effectively and may further improve prediction accuracy, particularly in regions with sparse ground-based monitoring infrastructure.

Finally, the nonlinear BSS model adopted in this work is assumed to be non-noisy, which leads to a Gaussian reconstruction distribution with an infinitesimally small variance (set to 0.01 in practice). While this is a standard assumption in the BSS literature and provides a principled motivation for the Gaussian reconstruction term in the ELBO, environmental phenomena such as ozone and PM_{10} pollution are characterized by episodic extreme events—for example, summer ozone peaks or pollution episodes driven by stagnant meteorological conditions—that can induce skewed or heavy-tailed error distributions. Since the reconstruction term models the distribution of observational noise rather than the marginal distribution of the variables directly, the

expected impact on model performance is modest; however, in the presence of genuinely heavy-tailed measurement noise, the Gaussian assumption may lead to less accurate decoder estimation at extreme values. A natural future extension would therefore be to consider a noisy nonlinear BSS variant, in which observation noise is explicitly modeled through a heavier-tailed reconstruction distribution, such as a Student-t, skew-normal or skew-t likelihood, and to study the robustness of iVAE estimation under such conditions.

Acknowledgments

Open access publishing facilitated by Universita del Salento, as part of the Wiley - CRUI-CARE agreement.

Funding

This work was partially supported by the European Union-Next GenerationEU 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. The work of M.S. was supported by the Vilho, Yrjö and Kalle Väisälä foundation. K.N. and S.T. were supported by the Research Council of Finland (363261, 356484). We acknowledge computational resources from CSC—IT Center for Science, Finland.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in GRINS AQLIM: The GRINS datasets on Air Quality and CLimate at doi:<https://doi.org/10.5281/zenodo.17103414>, reference number zenodo.17103414.

Endnotes

¹<https://github.com/mikasp/O3-prediction>.

²<https://github.com/mikasp/NonlinearBSS>.

³<https://github.com/mikasp/ECOST>.

⁴<https://ads.atmosphere.copernicus.eu/datasets/cams-europe-air-quality-reanalyses>.

References

- Allard, D., L. Clarotto, and X. Emery. 2022. “Fully Nonseparable Gneiting Covariance Functions for Multivariate Space–Time Data.” *Spatial Statistics* 52: 100706.
- Bevilacqua, M., A. S. Hering, and E. Porcu. 2015. “On the Existence of Multivariate Covariance Models: Comment on the Paper by Genton and Kleiber.” *Statistical Science* 30, no. 2: 167–169.
- Box, G. E. P., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. 5th ed. Wiley.
- Breiman, L. 2001. “Random Forests.” *Machine Learning* 45: 5–32.
- Cappello, C., S. De Iaco, and M. Palma. 2022. “Computational Advances for Spatio-Temporal Multivariate Environmental Models.” *Computational Statistics* 37, no. 2: 651–670.
- Chen, W., Y. Li, B. J. Reich, and Y. Sun. 2024. “Deepkriging: Spatially Dependent Deep Neural Networks for Spatial Prediction.” *Statistica Sinica* 34: 291–311.
- Chen, Y., X. Chen, A. Xu, Q. Sun, and X. Peng. 2022. “A Hybrid CNN-Transformer Model for Ozone Concentration Prediction.” *Air Quality, Atmosphere & Health* 15, no. 9: 1533–1546.
- Cheng, Y., Q. Zhu, Y. Peng, X.-F. Huang, and L.-Y. He. 2021. “Multiple Strategies for a Novel Hybrid Forecasting Algorithm of Ozone Based on Data-Driven Models.” *Journal of Cleaner Production* 326: 129451.
- Comon, P., and C. Jutten. 2010. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press.
- De Iaco, S., C. Cappello, M. Palma, and K. Nordhausen. 2025. “A Multivariate Approach for Modeling Spatio-Temporal Agrometeorological Variables.” *Environmetrics* 36: e2891.
- De Iaco, S., D. Myers, and D. Posa. 2003. “The Linear Coregionalization Model and the Product-Sum Space-Time Variogram.” *Mathematical Geology* 35: 25–38.
- De Iaco, S., D. E. Myers, and D. Posa. 2001. “Space-Time Analysis Using a General Product-Sum Model.” *Statistics & Probability Letters* 52, no. 1: 21–28.
- De Iaco, S., M. Palma, and D. Posa. 2005. “Modeling and Prediction of Multivariate Space-Time Random Elds.” *Computational Statistics & Data Analysis* 48, no. 3: 525–547.
- De Iaco, S., and D. Posa. 2012. “Predicting Spatio-Temporal Random Elds: Some Computational Aspects.” *Computers & Geosciences* 41: 12–24.
- De Iaco, S., and D. Posa. 2013. “Positive and Negative Non-Separability for Space–Time Covariance Models.” *Journal of Statistical Planning and Inference* 143, no. 2: 378–391.
- De Iaco, S., D. Posa, and D. E. Myers. 2013. “Characteristics of Some Classes of Space-Time Covariance Functions.” *Journal of Statistical Planning and Inference* 143: 2002–2015.
- Donzelli, G., and M. M. Suarez-Varela. 2024. “Tropospheric Ozone: A Critical Review of the Literature on Emissions, Exposure, and Health Effects.” *Atmosphere* 15, no. 7: 779.
- Emberson, L. 2020. “Effects of Ozone on Agriculture, Forests and Grasslands.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 378, no. 2183: 20190327.
- Fusta Moro, A., and A. Fassò. 2025. “AQLIM_GRINS Dataset - the Italian Daily Dataset on Air Quality (v.1.0.0).” [data set] (Tech. Rep.). Zenodo.
- Genton, M., and W. Kleiber. 2015. “Cross-Covariance Functions for Multivariate Geostatistics.” *Statistical Science* 30, no. 2: 147–163.
- Gneiting, T., W. Kleiber, and M. Schlather. 2010. “Matérn Cross-Covariance Functions for Multivariate Random Elds.” *Journal of the American Statistical Association* 105, no. 491: 1167–1177.
- Gräler, B., E. Pebesma, and G. Heuvelink. 2016. “Spatio-Temporal Interpolation Using Gstat.” *R Journal* 8: 204–218.
- Guo, Q., Z. He, and Z. Wang. 2023. “Prediction of Monthly Average and Extreme Atmospheric Temperatures in Zhengzhou Based on Artificial Neural Network and Deep Learning Models.” *Frontiers in Forests and Global Change* 6: 1249300.
- Guo, Q., Z. He, and Z. Wang. 2024. “Monthly Climate Prediction Using Deep Convolutional Neural Network and Long Short-Term Memory.” *Scientific Reports* 14, no. 1: 17748.
- Guo, Q., Z. He, and Z. Wang. 2025. “Assessing the Effectiveness of Long Short-Term Memory and Artificial Neural Network in Predicting Daily Ozone Concentrations in Liaohe City.” *Scientific Reports* 15, no. 1: 6798.
- He, Z., and Q. Guo. 2024. “Comparative Analysis of Multiple Deep Learning Models for Forecasting Monthly Ambient PM_{2.5} Concentrations: A Case Study in Dezhou City, China.” *Atmosphere* 15, no. 12: 1432.
- He, Z., Q. Guo, Z. Wang, and X. Li. 2025. “A Hybrid Wavelet-Based Deep Learning Model for Accurate Prediction of Daily Surface PM_{2.5} Concentrations in Guangzhou City.” *Toxics* 13, no. 4: 254.
- Hyvärinen, A., and H. Morioka. 2016. “Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA.” *Advances in Neural Information Processing Systems*, 29.
- Hyvärinen, A., and H. Morioka. 2017. “Nonlinear ICA of Temporally Dependent Stationary Sources.” *Artificial Intelligence and Statistics*. (pp. 460–469).
- Hyvärinen, A., H. Sasaki, and R. Turner. 2019. “Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning.” *The 22nd International Conference on Artificial Intelligence and Statistics*. (pp. 859–868).
- Iskandaryan, D., F. Ramos, and S. Trilles. 2023. “Graph Neural Network for Air Quality Prediction: A Case Study in Madrid.” *IEEE Access* 11: 2729–2742.

- Kethireddy, S. R., P. B. Tchounwou, H. A. Ahmad, A. Yerramilli, and J. H. Young. 2014. "Geospatial Interpolation and Mapping of Tropospheric Ozone Pollution Using Geostatistics." *International Journal of Environmental Research and Public Health* 11, no. 1: 983–1000.
- Khemakhem, I., D. Kingma, R. Monti, and A. Hyvärinen. 2020. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework." In *International Conference on Artificial Intelligence and Statistics* (pp. 2207–2217).
- Kingma, D. P., and M. Welling. 2013. "Auto-Encoding Variational Bayes." *arXiv Preprint arXiv:1312.6114*.
- Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin. 1992. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are We That Economic Time Series Have a Unit Root?" *Journal of Econometrics* 54, no. 1–3: 159–178.
- Li, R., L. Cui, F. Hongbo, J. Li, Y. Zhao, and J. Chen. 2020. "Satellite-Based Estimation of Full-Coverage Ozone (o3) Concentration and Health Effect Assessment Across Hainan Island." *Journal of Cleaner Production* 244: 118773.
- Liaw, A., and M. Wiener. 2002. "Classification and Regression by random-Forest." *R News* 2, no. 3: 18–22.
- Ma, M., P. Xie, F. Teng, et al. 2023. "HiSTGNN: Hierarchical Spatio-Temporal Graph Neural Network for Weather Forecasting." *Information Sciences* 648: 119580.
- Mattei, P.-A., and J. Frellsen. 2019. "Miwae: Deep Generative Modelling and Imputation of Incomplete Data Sets." In *International Conference on Machine Learning* (pp. 4413–4423).
- Michael, R., C. R. O'Lenick, A. Monaghan, et al. 2019. "Application of Geostatistical Approaches to Predict the Spatio-Temporal Distribution of Summer Ozone in Houston, Texas." *Journal of Exposure Science & Environmental Epidemiology* 29, no. 6: 806–820.
- Muehlmann, C., S. De Iaco, and K. Nordhausen. 2023. "Blind Recovery of Sources for Multivariate Space-Time Random Elds." *Stochastic Environmental Research and Risk Assessment* 37: 1593–1613.
- Nag, P., Y. Sun, and B. J. Reich. 2023. "Spatio-Temporal Deepkriging for Interpolation and Probabilistic Forecasting." *Spatial Statistics* 57: 100773.
- Nazabal, A., P. M. Olmos, Z. Ghahramani, and I. Valera. 2020. "Handling Incomplete Heterogeneous Data Using VAEs." *Pattern Recognition* 107: 107501.
- Nuvolone, D., D. Petri, and F. Voller. 2018. "The Effects of Ozone on Human Health." *Environmental Science and Pollution Research* 25: 8074–8088.
- Nychka, D., S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. 2015. "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets." *Journal of Computational and Graphical Statistics* 24, no. 2: 579–599.
- Pak, U., C. Kim, U. Ryu, K. Sok, and S. Pak. 2018. "A Hybrid Model Based on Convolutional Neural Networks and Long Short-Term Memory for Ozone Concentration Prediction." *Air Quality, Atmosphere & Health* 11: 883–895.
- Pan, Q., F. Harrou, and Y. Sun. 2023. "A Comparison of Machine Learning Methods for Ozone Pollution Prediction." *Journal of Big Data* 10, no. 1: 63.
- Porcu, E., R. Furrer, and D. Nychka. 2021. "30 Years of Space–Time Covariance Functions." *WIREs Computational Statistics* 13, no. 2: e1512.
- R Core Team. 2024. "R: A Language and Environment for Statistical Computing [Computer Software Manual]." Vienna, Austria.
- Raitoharju, J. 2022. "Convolutional Neural Networks." In *Deep Learning for Robot Perception and Cognition*, 35–69. Elsevier.
- Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63, no. 3: 581–592.
- Salvana, M. L. O., and M. G. Genton. 2020. "Nonstationary Cross-Covariance Functions for Multivariate Spatio-Temporal Random Elds." *Spatial Statistics* 37: 100411.
- Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2009. "The Graph Neural Network Model." *IEEE Transactions on Neural Networks* 20, no. 1: 61–80.
- Shah, I., N. Gul, S. Ali, and H. Houmani. 2024. "Short-Term Hourly Ozone Concentration Forecasting Using Functional Data Approach." *Econometrics* 12, no. 2: 12.
- Sipilä, M., C. Cappello, S. De Iaco, K. Nordhausen, and S. Taskinen. 2025. "Modelling Multivariate Spatio-Temporal Data With Identifiable Variational Autoencoders." *Neural Networks* 181: 106774.
- Sipilä, M., S. Maggio, S. De Iaco, K. Nordhausen, M. Palma, and S. Taskinen. 2026a. "Time-Aware UNet and Super-Resolution Deep Residual Networks for Spatial Downscaling." *Environmetrics*: e70134. <https://doi.org/10.1002/env.70134>.
- Sipilä, M., K. Nordhausen, and S. Taskinen. 2024. "Nonlinear Blind Source Separation Exploiting Spatial Nonstationarity." *Information Sciences* 665: 120365.
- Sipilä, M., K. Nordhausen, and S. Taskinen. 2026b. "Identifiable Autoregressive Variational Autoencoders for Nonlinear and Nonstationary Spatio-Temporal Blind Source Separation." In *Machine Learning and Knowledge Discovery in Databases. Research Track*, 362–380. Springer Nature Switzerland.
- Van Zoest, V., F. B. Osei, G. Hoek, and A. Stein. 2020. "Spatio-Temporal Regression Kriging for Modelling Urban NO2 Concentrations." *International Journal of Geographical Information Science* 34, no. 5: 851–865.
- Vaswani, A., N. Shazeer, N. Parmar, et al. 2017. "Attention Is All You Need. Advances in Neural Information Processing Systems." 30.
- Wang, W., X. Liu, J. Bi, and Y. Liu. 2022. "A Machine Learning Model to Estimate Ground-Level Ozone Concentrations in California Using TROPOMI Data and High-Resolution Meteorology." *Environment International* 158: 106917.
- Wendland, H. 1995. "Piecewise Polynomial, Positive Definite and Compactly Supported Radial Functions of Minimal Degree." *Advances in Computational Mathematics* 4: 389–396.
- Xu, S., C. Cui, M. Shan, et al. 2022. "Spatio-Temporal Prediction of Ground-Level Ozone Concentration Based on Bayesian Maximum Entropy by Combining Monitoring and Satellite Data." *Atmosphere* 13, no. 10: 1568.
- Zhang, B., Y. Rong, R. Yong, et al. 2022. "Deep Learning for Air Pollutant Concentration Prediction: A Review." *Atmospheric Environment* 290: 119347.
- Zhang, J., Y. Wei, and Z. Fang. 2019. "Ozone Pollution: A Major Health Hazard Worldwide." *Frontiers in Immunology* 10: 2518.
- Zhou, Z., C. Qiu, and Y. Zhang. 2023. "A Comparative Analysis of Linear Regression, Neural Networks and Random Forest Regression for Predicting Air Ozone Employing Soft Sensor Models." *Scientific Reports* 13, no. 1: 22420.