





# Warped multifidelity Gaussian processes for data fusion of skewed environmental data

Pietro Colombo<sup>1</sup> , Claire Miller<sup>1</sup>, Xiaochen Yang<sup>1</sup>, Ruth O'Donnell<sup>1</sup>  
and Paolo Maranzano<sup>2</sup> 

<sup>1</sup>School of Mathematics and Statistics, University of Glasgow, 132 University PI, Glasgow G12 6TA, UK

<sup>2</sup>Department Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1 - 20126, Milan, Italy

*Address for correspondence:* Pietro Colombo, School of Mathematics and Statistics, University of Glasgow, 132 University PI, Glasgow G12 8TA, UK. Email: [pietro.colombo@glasgow.ac.uk](mailto:pietro.colombo@glasgow.ac.uk)

## Abstract

Understanding the dynamics of climate variables is critical for sectors like energy and environmental monitoring. This study addresses the pressing need for accurate mapping of environmental variables in national or regional monitoring networks, a challenge exacerbated by skewed data and large gaps. While this may not be immediately apparent, managing skewness across multiple data sources introduces additional complexities, as conventional transformation methods often fail to effectively normalize the data or preserve inter-dataset relationships. Furthermore, the literature highlights that interpolation uncertainty is closely linked to the interpolation distance, making the handling of large gaps particularly problematic. To tackle these challenges, we propose a novel data fusion approach: the warped multifidelity Gaussian process. This method predicts time-series data from multiple sources with varying reliability and resolution, while effectively addressing skewness and demonstrating partial independence from interpolation distance. Through extensive simulation experiments, we explore both the strengths and limitations of the method. Additionally, as a case study, we apply warped multifidelity Gaussian process (WMFGP) to wind speed data from the Agenzia regionale per la protezione ambientale (ARPA) Lombardia network, a regional environmental agency in Italy. Our results demonstrate the efficacy of WMFGP in filling large gaps in wind speed data, providing more accurate predictions that are essential for air quality forecasting, network maintenance.

**Keywords:** ARPA lombardia, data fusion, multifidelity Gaussian process, skew data

## 1 Introduction

Gaining insight into the changes in environmental variables like wind speed is crucial, especially in fields such as the energy sector, where it influences decisions about wind farm development, and in environmental monitoring to assess air quality in specific regions. These tasks range from mitigating atmospheric and acoustic pollution to overseeing water quality and electromagnetic fields, reflecting the broader mission of environmental stewardship and protection, see [Maranzano \(2022\)](#). Data fusion algorithms are becoming highly successful in supporting these challenges, as they intuitively combine multiple data sources to obtain more informative data and hence potentially more cost-effective results. Our study extends the class of data fusion algorithms based on Gaussian processes, referred to as multifidelity, specifically by incorporating a nonparametric warping function for response variables of interest. This approach allows us to effectively merge data having various reliability (fidelity) from different monitoring stations, even when dealing with skewed data distributions, a common occurrence in environmental variables such as wind speed, air pollution, and precipitation. More specifically, we propose an extension of the

Received: May 20, 2024. Accepted: January 7, 2025

© The Royal Statistical Society 2025.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

autoregressive multifidelity Gaussian process (MFGP) as proposed by [Le Gratiel and Cannamela \(2015\)](#) and used the nonparametric warping of [Agou et al. \(2022\)](#) to handle skewed data. Our extension, called *warped multifidelity Gaussian process* (WMFGP), maintains the same assumptions of the standard MFGP method, where it is assumed a highly reliable source of information, labelled high-fidelity (HF) is related to a source of information, namely low-fidelity (LF), regarded as less reliable, connected together by a so-called autoregressive representation of fidelity levels through a regressive parameter  $\rho$ . This implies that the effectiveness of data fusion of the LF with HF data is directly influenced by the linear correlation between the two data sources and more precisely a higher correlation leads to more effective recovery of HF data.

We apply our model to wind speed data provided by one of the regional environmental protection agencies (in Italian, *Agenzia Regionale di Protezione dell'Ambiente*) active in Italy, namely, ARPA Lombardia, which operates in the Northern part of the country. The Lombardy region, nestled in the Po River valley and surrounded by the Alps, faces unique challenges in air quality management. In fact, its geographical configuration limits air circulation, contributing to the elevated pollutant concentrations ([Maranzano, 2022](#)). Furthermore, Lombardy ranks among Europe's most industrialized regions (see [Eurostat Database, 2024](#)), hosting mechanical, electrical, metallurgic, textile, and chemical industries, along with numerous animal farms ([L. Colombo et al., 2023](#)), which recent studies have linked to significant impacts on air quality ([Fassò et al., 2023](#)). There are multiple studies where it is demonstrated that wind speed plays a critical role in determining air pollutant dispersion, such as those presented in [Carta et al. \(2008\)](#), [Yu et al. \(2004\)](#), [Erdem and Shi \(2011\)](#), [McWilliams et al. \(1979\)](#), and [Raffaelli et al. \(2020\)](#). Hence, an accurate mapping of wind speed across the region is essential for understanding and managing air quality effectively.

ARPA Lombardia manages a broad weather and air quality monitoring network, which sometimes exhibit data gaps, not only in air pollutant concentration but also in wind speed data. Filling these gaps is vital for comprehensively understanding wind patterns and their implications for air quality. For example, ARPA could schedule the maintenance of the monitoring stations more effectively, if accurate methods for filling data gaps were available. Moreover, data gaps might occur in the presence of particular meteorological events. The absence of a week's worth of data might result in a lack of understanding of the physical generation process.

Interpolating a week's worth of data presents substantive challenges for standard methods such as Gaussian Process regression, where uncertainty increases linearly with the interpolation distance. However, our exploration reveals that multifidelity models maintain a partial independence from this interpolation distance, a detail further elaborated in section 4.3. Although multifidelity models address the issue of distance effectively, they fall short when handling skewed data. Standard normalization methods often fail with multiple data sources: applying a single transformation across two independent datasets can be effective for only one dataset (effective in 38% of cases in our study). Conversely, using two separate transformations to normalize each dataset independently can distort the inter-dataset relationships, leading to biased estimates. To exemplify the implications of such distortions, we provided an illustrative example in the [online supplementary material](#), see [Appendix C](#). Our novel model addresses these issues by managing large gaps without substantially distorting the relationships between datasets, while effectively normalizing two independent data sources. Experimental results have demonstrated superior performance compared with both standard interpolation methods and methods designed to handle data skewness.

The remainder of the paper is organized as follows: introduction to the methodology, including a historical overview of multifidelity methods and the challenges associated with skewness. Two experiments are described to demonstrate WMFGP's efficacy in handling skewness in a data fusion context. The practical application is described and the results presented.

## 2 Methods

### 2.1 Historical overview

In the literature, the word *fidelity* is used in place of the word quality or reliability ([Le Gratiel and Cannamela, 2015](#); [Perdikaris et al., 2017, 2016](#)). The first appearance of a multifidelity method in the form of a Gaussian process (GP) can be traced back to 2000, when Kennedy and O'Hagan introduced the idea of creating a model to operate at different fidelity levels in their work ([Kennedy and O'Hagan, 2000](#)). The intuition was that such a model could leverage the qualities of different

datasets. Indeed, HF data are usually scarce in time and space and very expensive to collect. While, LF are typically cheaper and more abundant in time and space, but unreliable by definition. More recently, the framework elaborated by Kennedy and O'Hagan is more commonly referred to as multifidelity. The approach is highly flexible as it can manage data of different reliability and resolutions. For example, most of the institutional ground monitoring networks, such as the one managed by the Italian ARPAs, suffer from a poor spatial coverage and the presence of missing sequences representing an optimal case study for multifidelity methods. Our application deals with the this latter problem, using multifidelity methods to recover missing sequences.

Research on multifidelity models has investigated multiple aspects. For example, the paper by [Le Gratiet and Cannamela \(2015\)](#) focused on the estimation procedure, recasting the equations of the original inefficient Bayesian approach, proposing a sequential design and enabling a much faster inference. Some researchers instead modelled the relationship between the HF and the LF data, as seen in [Perdikaris et al. \(2017\)](#), where the authors propose a multifidelity model (NARGP) to handle nonlinear relationships between fidelity levels, since the original model assumed a linear relationship. Instead, in a not peer reviewed work Raissi and Karniadakis combined the multifidelity Gaussian process with a neural network to create a robust model for discontinuous relationships. Cutajar's model ([Cutajar et al., 2019](#)) not only addresses nonlinear relationships between datasets but also overcomes the potential overfitting issues associated with NARGP model. Of particular relevance is the multifidelity model described in [Lu and Shafto \(2021\)](#), where the data-fusion is performed on the language of kernels, capturing both the uncertainty propagation between fidelity levels and potential nonstationarity in latent GP. An excellent overview of different multifidelity methods can be found in [Costabal et al. \(2019\)](#).

The autoregressive multifidelity model, as described in [Kennedy and O'Hagan \(2000\)](#), [Le Gratiet and Cannamela \(2015\)](#), and [Perdikaris et al. \(2017\)](#), assumes the existence of different Gaussian processes modelling the various fidelity levels. Therefore, like the standard Gaussian process, the MFGP is based on normality assumptions, specifically in the error term. However, environmental data such as wind speed, relative humidity, temperature, and precipitation often follow skewed distributions, which renders the normality assumption inappropriate ([Aslam, 2021](#)). Addressing nonnormal properties using a Gaussian process has historically been a challenge. One intuitive approach might involve using a multivariate skew-normal distribution class. However, as discussed by [Genton and Zhang \(2012\)](#), such distribution classes suffer from identifiability issues in inference. Many attempts have been made to mitigate these issues. For example, [Alodat and Shakhathreh \(2020\)](#) proposed a Gaussian process with skewed errors based on a closed skew normal distribution. However, the model proved to be heavily parameterized, making inference computationally challenging. A recent alternative has been formalized in [Khaledi et al. \(2023\)](#), which proposes a parsimonious version of the closed skew normal distribution for use within a Gaussian process framework. This offers an intriguing approach to test in real-case scenarios. A completely different approach from the one of Khaledi or Alodat and Shakhathreh for addressing nonnormal properties, particularly skewness, involves transformations. The idea is to transform the response variable in such a way that it follows a normal distribution. These transformations can be either parametric or nonparametric. An example of a parametric transformation is the Box-Cox transformation. However, parametric transformations like the Box-Cox rely on parameters that are data-dependent. In a multifidelity context, where multiple data sources need to be transformed, the optimal parameters required for appropriate normalization of each data source differ. Applying different transformations might alter the relationship between the HF and LF data. Within the class of parametric transformations for Gaussian processes there is also the warped Gaussian process ([Snelson et al., 2003](#)), in which a nonlinear mapping of the response variable is directly incorporated into the likelihood function such that the GP parameters are learned jointly with the parameters of the normalizing function. This classic approach is very well suited for multiple applications. However, sometimes the transformations learned in this way are not capable of finding effective normalization. A nonparametric nonlinear transformation of the response variable for GP is proposed in [Agou et al. \(2022\)](#). The data-driven transformation learned in such a way, also referred to as warping, is much more flexible than parametric approaches since it can effectively normalize any data source as long as enough data are provided. More importantly, since the proposed method's transformation is based on the quantiles of a CDF, the normalized values preserves the same ordering in normalized space (latent space). This property is also

referred as *quantile invariance*, and ensure that the relationships between different datasets does not change in the latent space.

## 2.2 The autoregressive multifidelity Gaussian process

We will first describe the autoregressive multifidelity model, as our method builds upon an extension of this framework. Consider a set of high fidelity scalar responses denoted as  $\mathbf{y}_H = [y(x_{H_1}), y(x_{H_2}), \dots, y(x_{H_{N_H}})]$ . It is typically assumed that  $x_H \in \mathbb{R}^d$ , referred to as the input space. The subscript  $H$  emphasizes that these observations are sourced from HF data, implying that if they pertain to an environmental factor like wind speed, the measured wind speed closely approximates the actual. The set of all locations where HF data are collected is denoted as  $\mathbf{x}_H$ .

However, it is important to note that the observations  $\mathbf{y}_H$  often have limited spatial and temporal coverage. Constructing an accurate model becomes especially challenging when the sample size of collected HF samples  $N_H$  is small. To address this, multifidelity models incorporate observations that are correlated with the variable of interest and are available near the spatial locations of  $\mathbf{y}_H$ . We denote these observations as  $\mathbf{y}_L$ . The observations labelled as LF are measured with a lower degree of reliability but are observed at locations  $\mathbf{x}_L$ , with a sample size  $N_L \gg N_H$ . We also define  $N = N_L + N_H$ , being the total number of observations.

Moreover, we assume that  $\mathbf{x}_H \subset \mathbf{x}_L$ , meaning we have a nested sampling design in which each HF observation  $y(x_{H_i})$  has a corresponding LF observation  $y(x_{L_i})$ , but not vice versa.

Such a situation can be modelled using Gaussian Process Regression (GPR); see [Williams and Rasmussen \(2006\)](#) for more details. For a finite collection  $u_L(x_{L_1}), u_L(x_{L_2}), \dots, u_L(x_{L_{N_L}})$ , the joint distribution is Gaussian with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , i.e.

$$\mathbf{u}_L \equiv \begin{bmatrix} u_L(x_{L_1}) \\ u_L(x_{L_2}) \\ \vdots \\ u_L(x_{L_{N_L}}) \end{bmatrix} \sim \mathcal{N}_{N_L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = [m(x_{L_1}) \quad m(x_{L_2}) \quad \dots \quad m(x_{L_{N_L}})]^\top,$$

and the  $(i, j)$ th element of  $\boldsymbol{\Sigma}$  is given by a generic covariance function  $C(x_{L_i}, x_{L_j}, \boldsymbol{\theta})$ .

In most situations, we only have noisy observations of  $u_L(x_L)$ , i.e.

$$y_L(x_L) = u_L(x_L) + \epsilon_L(x_L), \quad \epsilon_L(x_L) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_L^2). \quad (1)$$

Here,  $y_L(x_L)$  represents the observed value at the input  $x_L \in \mathbb{R}^d$ , while  $\epsilon_L(x_L)$  denotes the measurement error, which is assumed to be i.i.d. In this equation,  $u_L(x_L)$  represents the true unobserved phenomenon of the LF data, and  $\boldsymbol{\theta}_L$  is the vector of parameters characterizing the covariance function of the LF data.

The multifidelity model, which jointly models both HF and LF data, is best described by the following recursive equation:

$$y_H(x_L) = \rho u_L(x_L) + \delta(x_L) + \epsilon_\delta(x_L). \quad (2)$$

This equation establishes a hierarchical relationship between an HF function  $y_H(\cdot)$  and a function of the LF data  $u_L(\cdot)$  at each location where low-quality information is available, denoted as  $\mathbf{x}_L$ .

In other words, the function that approximates the HF data is constructed using three components:

1. A  $\rho$ -scaled version of  $u_L(\cdot)$ , where  $\rho$  is a parameter to be estimated.

2. The Gaussian process  $\delta(\cdot)$ , which is a latent process aiming to model the discrepancies between  $y_L$  and  $y_H$ . This process is referred to as an *independent* Gaussian process, meaning it has its own covariance function and parameters  $\theta_\delta$ , independent of  $u_L(\cdot)$ .
3. The measurement error of the discrepancies, expressed by  $\epsilon_\delta(x_L) \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\delta^2)$ . This noise might be related to the fact that both HF and LF data are measured with some error that could affect the derivation of discrepancies.

Equation 2 intuitively defines also the name of the model as the  $\rho$  is autoregressive parameter, given that it establishes an autoregressive relationship between different fidelity levels. In other words, the term autoregressive does not refer to any ordering in time or space but to the degree of fidelity. The model can be generalized for multiple fidelity levels. From a practical point of view, the full model can be thought of as a standard Gaussian process where the vector of observations is:

$$\mathbf{y} = \begin{bmatrix} y_L \\ y_H \end{bmatrix}, \tag{3}$$

while the covariance matrix is defined as follows:

$$\mathbf{K} = \begin{bmatrix} C_{LL}(\mathbf{x}_L, \mathbf{x}_L; \boldsymbol{\theta}_L) + \sigma_L^2 \mathbf{I} & C_{LH}(\mathbf{x}_L, \mathbf{x}_H; \boldsymbol{\theta}_L, \rho) \\ C_{HL}(\mathbf{x}_H, \mathbf{x}_L; \boldsymbol{\theta}_L) & C_{HH}(\mathbf{x}_H, \mathbf{x}_H; \boldsymbol{\theta}_L, \boldsymbol{\theta}_\delta, \rho) + \sigma_H^2 \mathbf{I} \end{bmatrix}. \tag{4}$$

The  $\mathbf{K}$  matrix depends on multiple covariance functions defined by the positions of the set of locations<sup>1</sup>  $\mathbf{x}_L$  and  $\mathbf{x}_H$ . Moreover, the term  $C_{HH}(\mathbf{x}_H, \mathbf{x}_H; \boldsymbol{\theta}_L, \boldsymbol{\theta}_\delta, \rho) + \sigma_H^2 \mathbf{I}$ , which defines the correlation between the HF data, is designed using the parameters of the discrepancy process  $\theta_\delta$  and the LF process  $\theta_L$ . More precisely:

$$C_{HH}(\mathbf{x}_H, \mathbf{x}_H; \boldsymbol{\theta}_L, \boldsymbol{\theta}_\delta, \rho) = \rho^2 C(\mathbf{x}_H, \mathbf{x}_H; \boldsymbol{\theta}_L) + C(\mathbf{x}_H, \mathbf{x}_H; \boldsymbol{\theta}_\delta). \tag{5}$$

The model for two fidelity levels has seven parameters, doubling the standard Gaussian process parameters, the signal variance (here inside the vector  $\boldsymbol{\theta}$ ), the nuggets  $\sigma_H^2$  and  $\sigma_L^2$ , and decay parameter of the covariance function (inside the vector  $\boldsymbol{\theta}$  as well) plus the autoregressive parameter  $\rho$ . Assuming a zero mean function a common assumption in multifidelity modelling the estimation can be derived by minimizing the negative log-likelihood:

$$\mathcal{NLM}(\theta_1, \theta_2, \rho) = \frac{1}{2} \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} + \frac{1}{2} \log |\mathbf{K}| + \frac{N}{2} \log (2\pi), \tag{6}$$

and for new input locations  $\mathbf{x}^*$ , we can derive the joint distribution:

$$\begin{bmatrix} \mathbf{u}_H^* \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left( \mathbf{0}, \begin{bmatrix} k_{HH}(\mathbf{x}^*, \mathbf{x}^*; \boldsymbol{\theta}_L, \boldsymbol{\theta}_\delta, \rho) & \mathbf{q}^T \\ \mathbf{q} & \mathbf{K} \end{bmatrix} \right), \tag{7}$$

and then, derive the predictions equation for new input locations of the conditional mean and variance:

$$\mathbf{u}_H^* = \mathbf{q}^T \mathbf{K}^{-1} \mathbf{y} \tag{8}$$

$$V(\mathbf{u}_H^*) = C_{HH}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{q}^T \mathbf{K}^{-1} \mathbf{q} \tag{9}$$

with  $\mathbf{q}$  being equal to  $\mathbf{q}^T = [C_{HL}(\mathbf{x}^*, \mathbf{x}_L; \theta_1, \rho) \quad C_{HH}(\mathbf{x}^*, \mathbf{x}_H; \theta_1, \theta_2, \rho)]$ .

<sup>1</sup> Depending of  $d$ -dimension,  $\mathbf{x}$  set can be either a vector or a matrix. For example, if we have time has only dimension than the  $\mathbf{x}$  are vectors.

### 2.3 Warped multifidelity Gaussian process

The warped multifidelity approach involves a nonlinear and nonparametric transformation of the nonnormal data  $\mathbf{y}_H$  and  $\mathbf{y}_L$  into a latent space where they are normalized. Subsequently, a multifidelity model is run in this latent space. We refer to the normalized data as  $\mathbf{z}_H$  and  $\mathbf{z}_L$ . The two nonlinear mappings,  $g_L: \mathbf{y}_L \rightarrow \mathbf{z}_L$  and  $g_H: \mathbf{y}_H \rightarrow \mathbf{z}_H$ , are designed such that  $z(x_L) = g_L[y(x_L)]$  and  $z(x_H) = g_H[y(x_H)]$  exhibit both marginal and joint normal distributions. The warping transformations,  $g_L(\cdot)$  and  $g_H(\cdot)$ , are obtained by first computing  $CDF_H$  and its marginal  $F_y(y_H)$  for the HF data, and  $CDF_L$  and its marginal  $F_y(y_L)$  for the LF data. Then, normal scores are obtained by inverting these cumulative distribution functions (CDFs). More precisely, with  $\phi$  being a normal CDF we have:

$$\mathbf{z}_L = \phi_L(Fy(\mathbf{y}_L))^{-1} = g_L(\mathbf{y}_L) \quad (10)$$

$$\mathbf{z}_H = \phi_H(Fy(\mathbf{y}_H))^{-1} = g_H(\mathbf{y}_H). \quad (11)$$

This means the procedure needs to estimate the CDF from the data. The estimation of the CDF, in turn, is obtained by integrating an antecedent kernel density; more details of the specific kernel density estimation are provided in [Pavlidis et al. \(2022\)](#). Once the multifidelity model is trained using the latent normalized observations  $\mathbf{z}_L$  and  $\mathbf{z}_H$ , it becomes necessary to back-transform the estimated quantities [the means and variances of the multifidelity model, as shown in [equations 8 and 9](#)] into the original space. Therefore, we need an inverse warping transform, referred to as  $\tilde{g}(\cdot) = g(\cdot)^{-1}$ . In other words,  $\tilde{g}(\cdot): \mathbf{z} \rightarrow \mathbf{y}$ . This latter inverse transformation is also a monotonic mapping and it is obtained by means of a lookup table, see [Agou et al. \(2022\)](#), since an explicit version of the inverse function is not available.

The lookup table consists of two columns: the first column contains the query points, forming a dense grid of values ranging from  $[y_{\min} - b, y_{\max} + b]$ , while the second column holds the corresponding probability levels  $p_i$ , which expresses the probability of observing  $y_i$  ordered observation. This dense grid allows us to create a detailed mapping over the entire range of the data.

In other words, some of the probability levels  $p_i$  are obtained directly from the actual data points by estimating their cumulative distribution function values. These CDF values are then transformed into normal scores using the inverse of the standard normal distribution function ( $\Phi^{-1}$ ), resulting in  $z = \{\Phi^{-1}(p_i)\}_{i=1}^N$ . However, relying solely on the sample data limits the connection between  $y$  and  $p$  to the discrete data points. To overcome this limitation, we perform linear interpolation on the dense grid of query points. In this interpolation, the actual data (either  $\mathbf{y}_H$  or  $\mathbf{y}_L$ ) serve as the independent variable, and the estimated CDF values ( $p_i$ ) act as the dependent variable. This process establishes a continuous relationship between the data and their estimated CDF values across the dense grid.

After running the multifidelity models in the latent (normal score) space, the predictions are expressed as normal scores. These predicted normal scores, denoted as  $\hat{\mathbf{z}}$ , are then transformed back into estimated probability levels ( $\hat{p}_i$ ) by applying the standard normal cumulative distribution function ( $\Phi$ ). For each estimated  $\hat{p}_i$ , we locate the nearest probability levels  $p_i$  in the second column of the lookup table and retrieve their corresponding indices. Using these indices, we extract the final predictions on the original data scale from the first column of the lookup table. A concise description of the method is provided in [Algorithm 1](#).

## 3 Data

### 3.1 Data for the first simulation experiment

The data for the first simulation experiment are the wind-speed reanalysis downloaded from the Copernicus Climate Data Storage (CCDS), see reference ERA5 ([Hersbach et al., 2023](#)). The CCDS offers a wide range of open access climate data, with in-depth description of the data and of the production process. The single-level<sup>2</sup> dataset offers a temporal coverage from 1940, with an hourly temporal resolution and spatial resolution of  $0.25^\circ \times 0.25^\circ$ , with data updated with a latency of 5 days. We chose offshore wind speed data from a site designated for a future wind farm. While the

<sup>2</sup> See ERA5 website, for more information about single-level dataset.

**Algorithm 1** Warped Multifidelity Gaussian Process, description of the steps involved in the derivation of WMFGP estimates of the mean  $\mathbf{u}_H^*$ .

---

**Input :**  $D = [y_H, y_L]$

**Output :** WMFGP interpolation of the mean  $\mathbf{u}_H^*$

for  $j$  in  $D$  for

- a. Perform kernel density-based estimation and derivation of the bandwidth  $b$
- b. Compute the kernel-based estimate of the CDF to derive  $p_i$  probability levels based on the sample values  $y$  of the time-series, the chosen kernel and estimated  $b$
- c. Compute the normal scores by inverting  $\{\Phi^{-1}(p)\}_{i=1}^N = \mathbf{z}$
- d. Interpolate between the estimated CDF estimates ( $p_i$ ) at the data points (either  $y_H$  or  $y_L$ ) to obtain a dense grid (4000 points) of  $p_i$ , lets call it  $\mathbf{p}_d$ , having the following range ( $[y_{\min} - b, y_{\max} + b]$ )
- e. Generate a lookup table to link the actual data with their probability levels. The table contains the dense grid  $\mathbf{p}_d$  of probability levels, and the corresponding values in the range ( $[y_{\min} - b, y_{\max} + b]$ )

end

f. Run the MFGP on the normal scores  $\mathbf{z}_L$  and  $\mathbf{z}_H$  to obtain  $\hat{\mathbf{z}}$

g. Use the lookup table to backtransform  $\hat{\mathbf{z}}$  MFGP estimations, into the original scale  $\mathbf{u}_H^*$

h. Return  $\mathbf{u}_H^*$

---

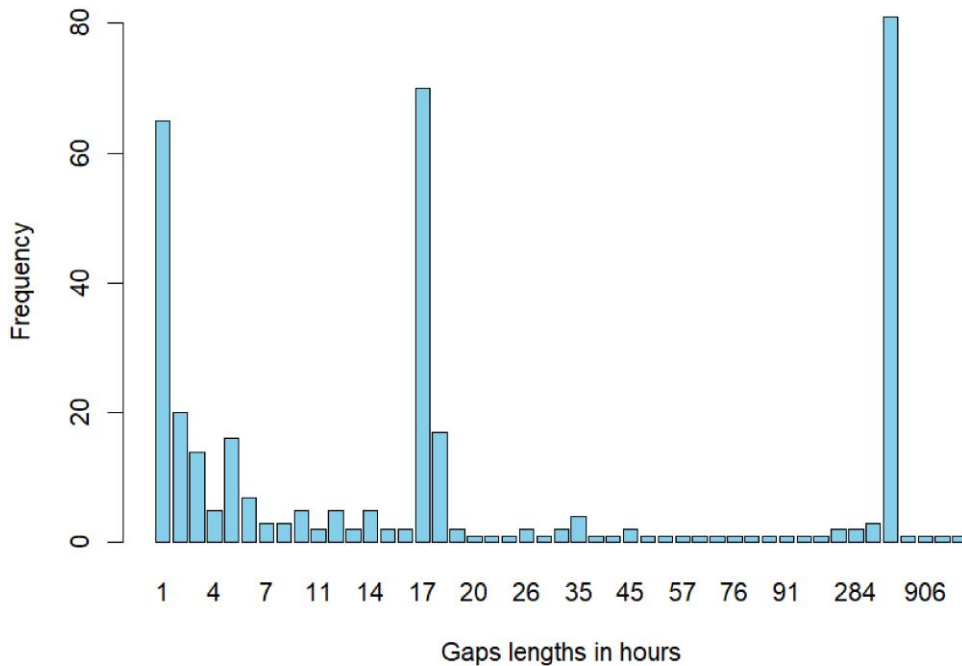
exact location is not critical for our current project, we made this decision to uphold the principle of replicating wind speed measurements in areas of potential significance. The vectorized wind speed components  $u$  and  $v$  are obtained at an altitude of 100 m, and the wind speed is reconstructed using the Pythagorean theorem, i.e.  $wind\ speed = \sqrt{u^2 + v^2}$ . A seasonal and trend decomposition using LOESS (Cleveland et al., 1990) is applied to extract the structural components (denoted as  $T$  in equation 12) of the downloaded time-series. These components serve as the basis for the simulation discussed in section 4.1. Please note that the precise location of the downloaded data aligns with the construction site of the Agnes wind farm (AGNES, n.d).

### 3.2 Data for the application and second simulation experiment

According to Italian and European legislation, the monitoring and protection of the environment (e.g. quantification of air and water pollution levels, measurement of meteorological phenomena, checks and inspections on the environmental impacts of enterprises, and scientific support to national and local institutions) is entrusted to the Regional Environmental Protection Agencies (ARPAs) that are members of the National System for Environmental Protection (SNPA).<sup>3</sup> Each ARPA is in charge of gathering air quality information within the area of responsibility. They also collect data on various environmental variables such as wind speed, and humidity. Data are publicly available either using the agency's data portal (ARPA Data Portal, n.d), the Regione Lombardia open data portal (Lombardy Region Weather Station, n.d) or by means of specialized open source software, such as *ARPALData* (Maranzano and Algieri, 2024) and *EAAq* (Tassan Mazzocco and Maranzano, 2023). Here the *ARPALData* package was used to retrieve the weather measurements used in both the simulation design and the application. Measurements span from 1 January to 31 December 2022.

The Lombardy network comprises more than 120 ground sites monitoring wind speed and direction, some of them active since the 1990s and others activated only in the last few years. Therefore, a subset of 94 monitoring stations for wind speed activated before 1 January 2015 and scattered throughout the region is investigated in this paper. Figure 5, which is the result of the clustering experiment described in section 5, displays the illustration of the station position and territory characteristics. By default, weather stations collect data at 10-min intervals, however, in both the simulation experiments and the application, we considered the hourly average wind speed as we are interested in recovering general patterns. Figure 1 illustrates the frequency and duration in hours of missing data sequences in the Lombardy region throughout 2022. The

<sup>3</sup> It is an Italian acronym of Sistema Nazionale di Protezione Ambientale



**Figure 1.** Frequency and duration of missing data sequences for all monitoring stations of Lombardy in the 2022. For a map of stations position, see [Figure 5](#).

histograms reveal numerous gaps with duration of  $<15$  h, which can be addressed using standard interpolation methods. Gaps ranging from 15 to 192 h (between one day and more than one week) are of particular interest in this application, as they are difficult to fill using simpler interpolation techniques. Longer gaps exceeding 192 h are not considered, as they likely indicate structural malfunctions rather than structural missing data and it would be unwise to address these only using statistical methodologies.

#### 4 Simulation studies

In this section, we introduce two experimental designs to evaluate the performance of the warped multifidelity Gaussian process in processing skewed data. These experiments illuminate the inherent challenges and effective strategies for reconstructing missing sequences inside an HF time-series. The first experiment assesses how well a series<sup>4</sup> of models performs when faced with random patterns of missing data in HF time-series. This scenario is standard in multifidelity modelling, where HF data are randomly sampled with equal probability. In this case, our goal is to understand how different models handle the task of recovering missing skewed HF information when HF data is sparsely sampled. We refer to this experimental design as *randomized missingness* since the sampling frequency of HF data is very low (10%) and occurs randomly. The second experiment addresses the issue of structural missingness, which involves the presence of long sequences of missing data in a time-series. In this latest experiment, following [Maranzano et al. \(2023\)](#), rather than simulating random locations and temporal patterns, we employ actual series from the ARPA Lombardia network in a Monte Carlo experiment, where the real locations are randomly sampled. The map of the monitoring stations and of the geography of Lombardy, shown in [Figure 5](#), evidences that in northern part of the region there is prevalence of mountains, with a higher spatial concentration of monitoring stations, while in the southern part, where the plains and cities prevail, the stations are more sparsely scattered. Due to the absence of physical obstacles, the correlations of the time-series recorded in the southern region will be higher, potentially

<sup>4</sup> See section [4.1](#) for the list of the models.



bringing a more efficacious recovery of the missing sequence. In both experiments, we employ simulations mimicking real-world data scenarios to analyse the model's performance, focusing on the challenges posed by skewed data. Our data fusion methodology, is applied to the data relying on a single assumption of linear correlation<sup>5</sup> between datasets. Among other benefits, our methodology is able to impute long missing sequences (up to a week of data with hourly resolutions). Moreover, it can manage different varying seasonal, sub-seasonal, and cyclic components automatically, without ad hoc study for each fused time-series.

#### 4.1 First simulation design: randomized missingness

This experiment evaluates the performance of various models in recovering randomized missing data patterns of HF time-series under varying skewness conditions. For the simulation, we used the wind speed data described in section 3.1, which are abundant and highly skewed. By using this real-world data in our simulations, we ensure the generation of realistic time-series to effectively test our models.

In this context, we generated two wind speed time-series—one with limited reliability but a higher number of data points in time and the other with greater reliability but fewer data points. Although our simulation mimics wind speed data, the experiment can be generalized to any skewed data sources characterized by a randomized data missingness structure. The experiment involves the comparison of five models across different skewness levels. The selected models include a standard Gaussian process (GP), a warped Gaussian process (WGP), a classic multifidelity (MFGP) model, a multifidelity model integrated with a simple Box-Cox class transformation (BCMF), and finally, a warped multifidelity (WMFGP) model. We conducted tests in the following dimensions:

1. Randomized missingness: each experiment simulates a different missing data pattern, with a total of 200 replications.
2. Level of skewness: time-series with varying degrees of skewness were generated, where skewness represents a parameter describing the deviation from normality.

The time-series were generated following this structure:

$$\begin{aligned} y_L &= \mathbf{T} + \mathbf{w}_L, \\ y_H &= \mathbf{T} + \mathbf{w}_H. \end{aligned} \quad (12)$$

In this context,  $\mathbf{T}$  represents the true time series we aim to recover. This time series is generated by summing the trend and seasonal components from the ERA5 reanalysis data, as outlined in section 3.1. The seasonal and trend components are extracted using an STL decomposition (Cleveland et al., 1990). Consequently, a new time series is produced by adding a fixed deterministic component ( $\mathbf{T}$ ) to a randomly generated error. The variables  $\mathbf{w}_L$  and  $\mathbf{w}_H$  represent random errors drawn from skewed distributions. Detailed plots of these simulated random errors are provided in the online supplementary material, specifically in Figures A1 and A2. Initially, we generated  $\mathbf{w}_H$  from a closed-skew normal distribution parametrized as follows:  $\text{CSN}(\mu, \sigma_1, \gamma, \nu, \delta)$ , where  $\mu$  represents the location parameter,  $\sigma_1$  is the scale parameter,  $\gamma$  is the skewness parameter,  $\nu$  is the shape parameter, and  $\delta$  is the truncation parameter. We also generated data from another distribution commonly used for wind speed data, allowing for more extreme skewness scenarios: the Weibull distribution with shape and scale parameters. We simulated two scenarios for both distributions, labelled as *high-skewness* and *low skewness*. The parameters for both distributions for these HF and LF errors ( $\mathbf{w}_L$  and  $\mathbf{w}_H$ ) are detailed in Table 1 and for the Weibull Table 2.

As we simulated missingness in the HF data, we conducted model comparisons on the test set using the following performance metrics: Mean Absolute Error (MAE), Bias, and Variance (Var).<sup>6</sup> For consistency, we maintained the HF sample size at a fixed 10% of the LF data. It is important to highlight that in employing the multifidelity approach, the ratio  $\frac{N_H}{N_L}$  plays a pivotal role

<sup>5</sup> In our application, the correlation is always positive.

<sup>6</sup> For the purposes of this paper, we focused only on the MAE and its variability.

**Table 1.** Parameters of the CSN distribution for the different skewness scenario

Data	Low skewness	High skewness
	$w_H, w_L$	$w_H, w_L$
$\mu$	-0.25, -0.5	-0.25, -0.5
$\sigma_1$	0.04, 0.8	0.8, 2.4
$\gamma$	4, 4	50, 50
$\nu$	2	2
$\delta$	3	3

**Table 2.** Summary of statistical measures and generating parameters of the errors generated from the Weibull distribution, for different skewness scenario

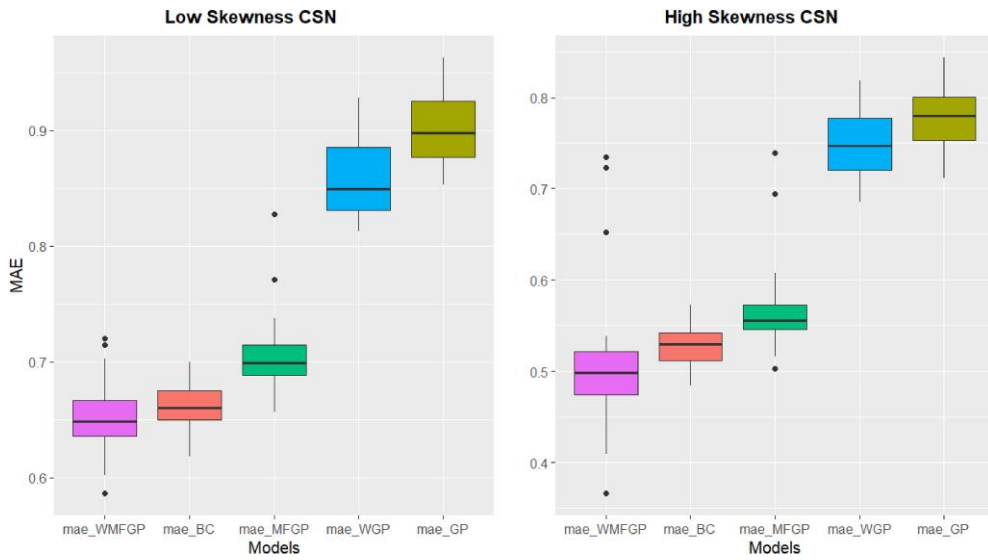
	Error	Scale	Shape	Mean	SD	Skewness
High skewness	$w_L$	2	0.8	1.3	2.8	2.8
	$w_H$	0.5	0.8	0	0.72	2.9
Low skewness	$w_L$	2	2.3	1.18	0.82	0.46
	$w_H$	0.5	2	0	0.23	0.66

as it validates the utility of multifidelity models. With a high ratio the probability of the multifidelity models to not yield superior outcomes compared with standard monofidelity models is increased. Specifically, our analysis reveals that multifidelity proves advantageous up to a ratio of 30%–35% in this particular data configuration.

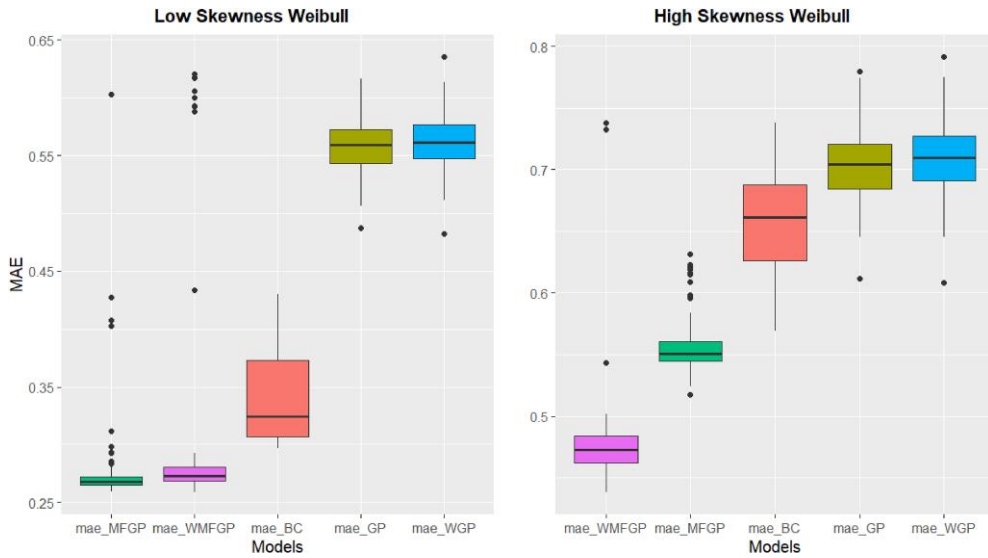
In all four scenarios, the WMFGP model emerged as the top-performing model, as illustrated in the bottom panels of Figures 2 and 3. It is noteworthy that, in general, the performance of the WMFGP model remained stable independently of the skewness level, while simpler multifidelity models face challenges. Occasionally, all multifidelity models (WMF, MF, BCMF) exhibited anomalous performances attributable to numerical instability, as evidenced by certain outlier data points in the boxplot. However, in synthesis, the WMFGP consistently outperforms the other models. The marginal improvement seen with the small skewness Weibull distribution under the standard multifidelity model is minor. The WMFGP resulted to be the best methods in 3 out of 4 scenarios, which more than expected since the method is thought to provide an advantage only in high-skewness scenarios. The Box–Cox multifidelity model generally was a good method, when the error was generated from a closed skew-normal, while when the error was generated from Weibull, where overall the skewness condition were more extreme, it was not able to return good performance. The latter confirmed the expectations: the Box–Cox method exhibits high sensitivity to data; when dealing with multiple datasets, the parametric transformation required to normalize each data source might vary significantly. This can inevitably affect the relationships between datasets in the latent space. Clearly, the more different the datasets are, the more likely it is that they will require different transformations to be normalized.

#### 4.2 Second simulation design: structural missingness

Lombardia data described in section 3.2 have been used to create the second simulation experiment with the aim of proposing multifidelity methods for dealing with structural missingness. Structural missingness can be seen as the presence of long-missing sequences in a time-series. The dataset includes long data gaps that, as explained in the introduction hinder a comprehensive evaluation of the wind-speed. This issue is of particular concern since standard interpolation methods such as Gaussian process regression have very poor predictive power if, for example, the



**Figure 2.** Boxplots of the 200 simulation experiment. On the x-axis the models, while on y-axis the MAE. Random errors generated from a closed skew normal distribution. Each box-plot refers to Mean Absolute Error in the test set for each model. Note that WMFGP has the lowest performance in both scenarios.



**Figure 3.** Boxplots from 200 simulation experiments are presented here. The models are listed on the x-axis, and the Mean Absolute Error (MAE) is displayed on the y-axis. Random errors in these simulations were generated from a Weibull distribution. Each boxplot represents the MAE on the test set for each model. In scenarios with high skewness, the Warped Multifidelity Gaussian Process (WMFGP) demonstrated superior performance. The performance of BCMF, highlighted in red, deteriorated as the underlying data became noisier and exhibited higher variance.

missing sequence is longer than the so-called range parameter or *length scale*<sup>7</sup> that controls the rate at which the covariance between two data points decreases as the distance between them increases. We focused on hourly resolutions, as we are concerned with recovering general patterns, we

<sup>7</sup> The decay parameter inside the kernel function.

**Table 3.** The table contains medians of MAE performances and their standard deviations for each simulated missing sequence

	GP	MFGP	WMFGP	Surrogate (abs(LF-HF))	Simple Imputation
<b>ML:24</b> (SD)	0.79 (0.71)	0.55 (0.45)	0.46 (0.51)	0.75 (1.05)	1.48 (1.36)
<b>ML:48</b> (SD)	0.77 (0.69)	0.57 (0.52)	0.49 (0.54)	0.75 (1.05)	1.62 (1.18)
<b>ML:72</b> (SD)	0.79 (0.59)	0.52 (0.48)	0.47 (0.49)	0.74 (0.89)	1.44 (1.16)
<b>ML:96</b> (SD)	0.80 (0.60)	0.54 (0.44)	0.48 (0.45)	0.77 (0.85)	1.43 (1.00)
<b>ML:196</b> (SD)	0.84 (0.40)	0.51 (0.34)	0.45 (0.34)	0.71 (0.65)	1.71 (0.89)

Note. ML stands for missing sequence length.

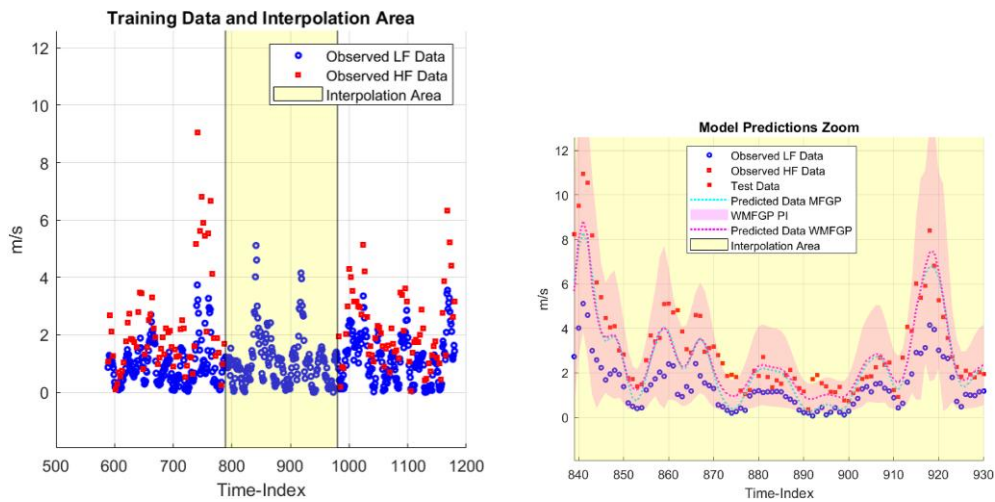
selected a sub-sample of 72 out of 94 stations from the network, to make sure that no real missing values were present in the experiment, with data referring to 2022. Then, we induced missing sequences of the following lengths in hours:

$$\text{missing lengths} = [24, 48, 72, 96, 192] \quad (13)$$

where 24 stands for 24 hr of missing values. Note that by testing many different missing sequences, we show that the method is relatively independent to the gap length, see section 4.4 for a theoretical explanation. We tested the capacity to recover the missing sequences using five methods: the GP, Surrogate, MFGP, WMFGP, and a simple imputation (SI). In this experiment, we excluded BCMF, since, as shown in the previous experiment, the method was never better than a standard MFGP method. However, we included an SI method as a benchmark for comparison. The simple imputation is a fast implementation method that assumes the values of the missing sequence to be an average of the closest observations, a moving average  $k$ -nearest neighbour. The wind speed data used in this experiment are generally skewed; therefore, our expectations are that the WMFGP will outperform MFGP. The surrogate performance represents the discrepancy we were to obtain if we replaced the missing sequence with the data observed in the nearby monitoring station, that we labelled as LF data source. In simple terms, surrogate is just LF--HF. The surrogate performance is a benchmark; we want the discrepancy between the estimates and the HF to be at least smaller than the differences between HF and LF data. Intuitively, once the surrogate results are available, we can measure the percentage error reduction of each method when compared with the original discrepancy between LF and HF. Our strategy relies upon the principle that monitoring stations that are in the same neighbourhood often present correlated information but different structural missingness. In this experiment, we assume the HF data to be those with a missing sequence, while the LF data come from a nearby highly correlated station. Table 3 summarizes the medians<sup>8</sup> of Mean Absolute Error (MAE) performance along with their respective standard deviations (SD) obtained from a series of 500 replication experiments. These experiments were conducted to assess the performance of five different imputation methods when dealing with simulated missing data sequences of varying lengths. Each row of the table corresponds to a specific missing data (ML) sequence length, denoted as ML:24, ML:48, ML:72, ML:96, and ML:196. The columns represent the different imputation methods used: GP, MFGP, WMFGP, Surrogate and Simple Imputation. The values in the table show the medians of MAE scores for each combination of imputation method and missing data sequence length, the median have been chosen to limit the impact of numerical errors, so the importance of anomalous performance is reduced. Additionally, the standard deviations (SD) are provided in parentheses, giving an indication of the variability in the MAE scores across the 500 replications.

The trend for MFGP and WMFGP in the table is that the MAE tends to increase as the discrepancy between LF and HF (Surrogate LF) data increases. This suggests that the missing sequence lengths do not play a role in the performance of the methods. This is an interesting result as in previous interpolation studies, based on simple Gaussian process, the gap length was associated

<sup>8</sup> The medians are chosen to limit the impact generated by the numerical instability of the algorithm.



**Figure 4.** In the top panel, the training data for high-fidelity (HF) simulations are depicted in red, while low-fidelity (LF) simulations are shown in blue. The simulated test area is highlighted in yellow. This bottom illustration showcases the successful recovery of 96 HF wind speed data points, emphasizing the robust performance of multifidelity models. Notably, across various time frames, the WMFGP—represented by the dashed magenta line—closely matches the simulated missing data (depicted as red squares) better than the MFGP, indicated by the dashed blue line. The vertical axis represents wind speed in metres per second (m/s), while the horizontal axis denotes the time index in hours from the initial observation.

with increased uncertainty (P. Colombo and Fassò, 2022; Fassò et al., 2020). Considering the overall performance, WMFGP appears to be a strong candidate for imputing missing data, as it maintains relatively low MAE values across different missing sequence lengths. This result is generally expected as the wind speed data from the monitoring stations of Lombardy are generally quite skewed (from 0.3 to 2). The second best method is MFGP, confirming the using multiple data-sources is beneficial. The simple imputation method is not a good choice since such high resolution of the data makes it difficult to understand what the smaller seasonal and cyclic components are?. Figure 4 presents a simulated HF wind speed data reconstruction. In the top panel, we display the training data, which consists of LF data points shown as a blue hole dots, the target HF data represented by red square, and the interpolation area highlighted in yellow. Moving to the bottom panel, we showcase the recovery of HF information in the yellow area using two different approaches: MFGP, represented by the light blue dashed line, and WMFGP, shown as the magenta dashed line. The red square dots on this panel represent the test data to recover. Notably, for this specific scenario, both the MFGP and WMFGP methods demonstrate their proficiency in approximating the missing information effectively, with a relatively better approximation performed by WMFGP. The plot also incorporates the 95% prediction interval of the MFGP model, indicated by the shaded magenta area. Interestingly, only one data point out of the 96 in the test area falls outside the prediction interval. The average discrepancy between the surrogate LF time-series and HF time-series ranged between 0.85 and 1.26 m/s in terms of wind speed. Meanwhile, the prediction of the WMFGP reduces the discrepancy to 0.45–0.65 m/s. These results constitute almost a 50% error reduction.

### 4.3 Uncertainty discussion

Discussing the uncertainty is important for informed decision-making, risk management, and resource allocation. In the structural missingness simulation study, we calculated both the prediction intervals and their corresponding coverage probability.<sup>9</sup> This measure is typically expressed as a

<sup>9</sup> The coverage probability serves as a measure of the prediction interval’s accuracy. In essence, it indicates the proportion of times, within a repeated or hypothetical series of experiments or predictions, that the prediction interval successfully encompasses the true value of the variable of interest.

percentage or a probability value. During 500 simulation experiments, we determined that the coverage probability of the prediction interval, established with a 95% confidence level for the MFGP, was 91%. Similarly, for the WMFGP, it was found to be 90%. This suggests that, on the whole, our prediction interval computations exhibit a high degree of accuracy. The coverage probability remains satisfactory, given the inherent complexities and possible challenges, see section 5.

#### 4.4 Relationship between interpolation distance and uncertainty

Previous research, particularly studies using Gaussian processes as discussed in [Fassò et al. \(2020\)](#) and [P. Colombo and Fassò \(2022\)](#), have shown a strong connection between uncertainty and interpolation distance. This relationship is attributed to the limitations of Gaussian processes, where the interpolation efficacy is constrained by the decay parameter of the covariance function. When there are large gaps, the decay parameter may not fully encompass the prediction areas, leading to increased uncertainty as the interpolation distance grows.

However, interpolation conducted through multifidelity models is less influenced by these decay parameters. These models leverage LF data to approximate the position of the data points to be interpolated and then incorporate the previously learned discrepancies between LF and HF data. Since the discrepancies between LF and HF data are typically constant across the interpolation space, the prediction uncertainty in multifidelity models remains largely independent of the gap size, provided the assumption of constant discrepancies holds.

This latter point is also confirmed by our simulation experiments where the performances of the multifidelity Gaussian processes MFGP and WMFGP have shown similar performances (see [Table 3](#)) independent of the gap sizes ranging from 24 up to 196 h.

An exception to this independence occurs if the relationship between LF and HF data within the interpolation area changes. For instance, if the linear correlation between HF and LF data varies multiple times within the interpolation area, these changes could introduce discrepancies dependent on the sudden shifts in correlation, thereby making the interpolation distance a relevant factor in predicting interpolation uncertainty. However, this scenario is highly unlikely, and it has not been encountered in our application.

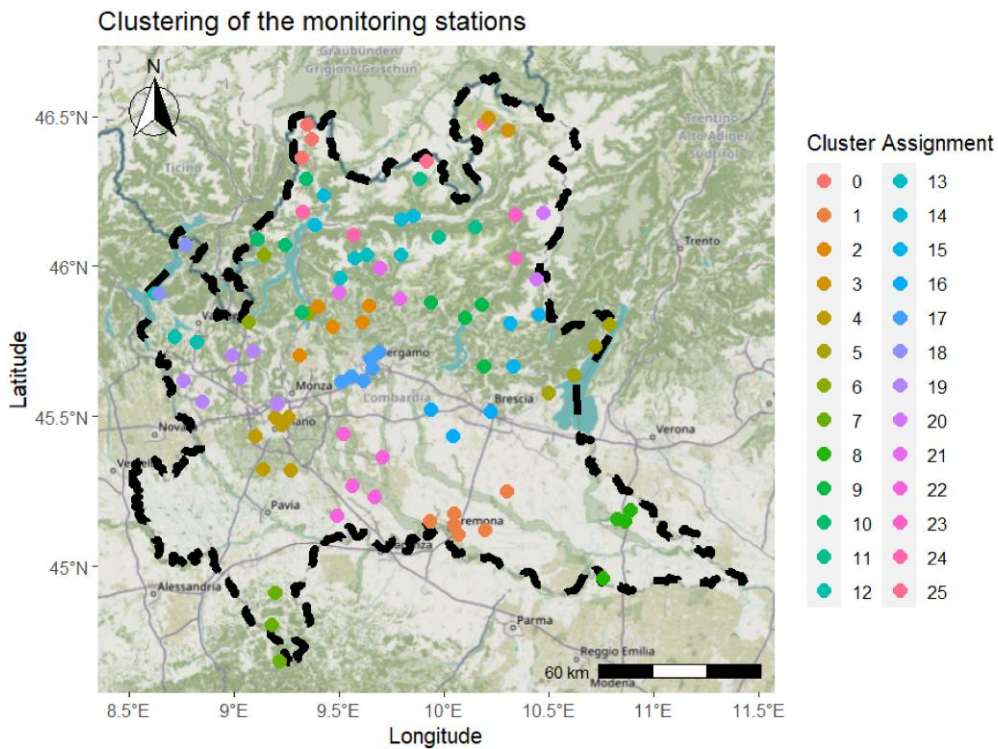
## 5 Application

In the previous sections, we stated that the ARPA monitoring networks often contain numerous missing values with different gap lengths. Therefore, we utilize the WMFGP and MFGP methods to recover these gaps in the sequences. Our approach leverages the natural temporal correlation among time-series observed at nearby locations, resulting in a highly computational efficiency strategy and of easy implementation. By excluding the spatial dimension from the multifidelity part, we might incur endogeneity issues ([Le Gallo and Fingleton, 2021](#)), as external factors like orography, geography, and land cover can strongly influence the model's performance. To address this concern, we conducted a clustering experiment based on the latitude, longitude, and altitude of the stations of the ARPA Lombardia dataset.

### 5.1 Empirical strategy

This procedure on one hand aims to identify monitoring stations that can serve as surrogate time-series data (LF); on the other to fill the gaps a target HF time-series data. The assumption is that nearby stations are likely to have similar information ([Zhu and Turner, 2022](#)), and therefore, correlated data can be used in a multifidelity context. We follow these steps to discover clusters of stations:

1. **Constrained  $k$ -means:** We utilize constrained  $k$ -means to identify clusters based on spatial information about the monitoring stations, such as latitude, longitude, and altitude. We also set constraints on the maximum and minimum cluster sizes. More details about the specific functioning of this method are available in [Bradley et al. \(2000\)](#). Fixing the number of clusters is crucial because, for our purposes, it makes little sense to have clusters containing, for example, 20 stations, even if they are well defined.



**Figure 5.** Depiction of the monitoring station clustering in two dimension (longitude and latitude).

2. Randomized pairing: We select a station with gaps and pair it with another randomly selected station within the same cluster.
3. The stations with gaps constitute our HF dataset, while the paired stations serve as the LF dataset. We then run multifidelity models using these two datasets.

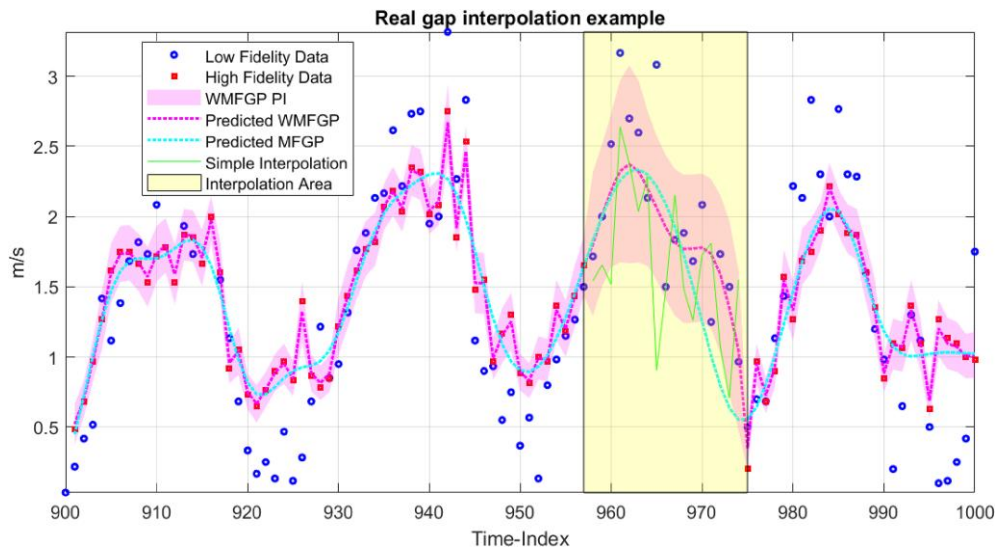
### 5.2 Clustering experiment

The results of the constrained  $k$ -means are shown in Figure 5. The plot is shown in two dimensions for clarity; however, we also used the altitude for clustering purposes. The altitudes play a crucial role in defining the clusters as in the north part of the region, the presence of the mountains hinders the cluster aggregation; in the same way, stations at similar latitudes and longitudes but different altitudes might have very different wind speeds.

We obtained 26 clusters by optimizing standard metrics, such as the elbow plot and average silhouette. This experiment resulted in an average cluster size of 3.6, with the largest cluster containing six elements and the smallest containing three elements. Cluster sizes ranging between 3 and 6 are convenient for two main reasons. Firstly, if the same missing sequence is present in two nearby stations, we can select the data from the third station of the clusters. This scenario, where the same gap appears in three nearby stations, is implausible in our dataset. Secondly, avoiding larger groups reduces the probability of associating stations influenced by different phenomena.

### 5.3 Missing data imputation

Figure 6 illustrates the imputations of 17 observation gaps for the Veddasca Monte Cadrigna station. Based on the simulation experiment, we know that among stations sharing similar altitude, longitude, latitude, and correlation between LF and HF, the MAE of WMFGP and MFGP was 30% lower than the MAE of the Surrogate. More importantly, the WMFGP had an MAE 13%



**Figure 6.** Depiction of the prediction performed by MFGP (light blue dotted line), the WMFGP (magenta dotted line) and the SI (continuous green line) for the Veddasca Monte Cadrigna station. Note that in the interpolation area in yellow, SI returns and unlikely recovery of the HF information, while both the WMFGP and MFGP seem to return a plausible pattern.

lower than the MFGP. The latter is aligned with what is shown in Figure 6, where the seasonality is correctly recovered by both the multifidelity models, but with a higher adaptability of WMFGP (magenta line), when compared with the MFGP (light blue). The green line represents a simple interpolation method, which, in this specific scenario, provides a poor recovery of the HF signal. The data imputation remains independent of the presence of different structural components, in different time-series. Meaning that if each sub-experiment involves time-series having different seasonal and cyclic components, it is not necessary to perform an ad hoc study to recover the right harmonic. In other words, the advantage of MFGP data imputations lies in the possibility of automating the data imputation process by simply identifying a correlated time-series with the one of interest.

## 6 Conclusion

Integrating multiple data sources can effectively leverage the relative abundance of low-quality data with the relative scarcity of HF data. Many data fusion algorithms based on Gaussian processes have been developed to model various linear and nonlinear relationships between different datasets. However, little or no work has been done to account for the presence of nonnormal properties in the datasets to be integrated. This latter aspect is an explicit limitation, considering that GP models assume independent, identical, normal errors, which could lead the model to have poor predictive performance. Moreover, the standard strategies for accounting for deviations from normality, such as parametric transformations (i.e. log transformation), might not be straightforward to implement in a data fusion application where each dataset might require a different transformation, the normalization process to distort the inter-dataset relationships. In addition to these challenges, the presence of big gaps is also been associated to an increase interpolation uncertainty.

In this work, we proposed an extension of the autoregressive multifidelity Gaussian process, a standard method for resolving data fusion problems with Gaussian process, based on a non-parametric warping strategy. This extension helps deal with multifidelity applications where the discrepancy data exhibit skewness, a common situation when at least one of the different data sources presents skewed data. Since the model preserves the inter-dataset relationship, it can be effectively applied to multiple data sources, moreover, the use of multifidelity class algorithm allows for partial independence to the interpolation distance. The algorithm



models the relationships between HF and LF observations in a latent space, where, owing to the warping, the discrepancy data follow a normal distribution, thereby improving the prediction of HF data. We showcase the efficacy of our model in modelling multiple data sources through an extended simulation experiment in which we controlled for the HF sample size and skewness.

In this paper, wind speed data have been imputed, however, the model has the potential to fuse and fill data coming from any skewed time-series, such as time-series regarding pollutant concentrations or rain occurrence. The wind speed data of the ARPA Lombardia monitoring network represented our motivating application as, due to both maintenance and adverse weather conditions, gaps of various lengths are often found in the measured time series. These gaps can cause problems when it comes to studying specific weather phenomena or when trying to assess the effect of wind speed on air quality. We, therefore, created a second simulation experiment to test the efficacy of multifidelity models to fill these long missing sequences. Our new method performs better than other tested methods and, particularly, better than the standard MFGP since the wind-speed data of Lombardy presented a considerable skewness. More importantly, our experiment showed how the multifidelity models, see Table 4.2, are partially independent to the interpolation distance offering a simple solution to a long standing problem.

A nice feature of this method is that it can be automatized, while other imputation methods would require an ad hoc study for each time-series. This is an appealing characteristic when dealing with big environmental datasets. One limitation of the method might be related to the number of observations necessary to learn the warping. We observed that, generally, good warping transformations are learned when there are more than 1000 data points for each data source, which might not always be the case in data fusion applications. The simulation study and the application developed at this point are solely based on the temporal dimension (and correlation). This can be seen both as an advantage and a limitation. On one hand, the developed framework requires few computational resources, providing a ready-to-use strategy for different datasets. On the other hand, it does not consider the additional strength that could be gained by incorporating the spatial dimension. A spatial model could be useful in case of wind-farm resource assessment, where measurement taken with Lidar or anemometric technologies could be integrated with reanalysis data or it could be useful to map the wind speed in areas of Lombardy network that are not covered by any monitoring station. However, including the spatial dimension will necessarily increase the algorithm's computational complexity, requiring the implementation of suitable approximation methods for the inversion of the  $\mathbf{K}$  variance-covariance matrix. Implementing a suitable approximation strategy in a multifidelity context is not trivial, as different points have different importance. We have reserved this line of research for a future study.

The strategy that we propose can be easily put into practise by practitioners and governmental agency that deal with missing sequence from a network. It is efficient, as it requires minimal computational resources, and it can be applied to any other datasets having similar characteristics to the ARPA Lombardia case study.

*Conflicts of interest:* There is no conflict of interest to declare.

## Funding

The authors acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) through a PhD scholarship under Grant 2021 No: 00863600

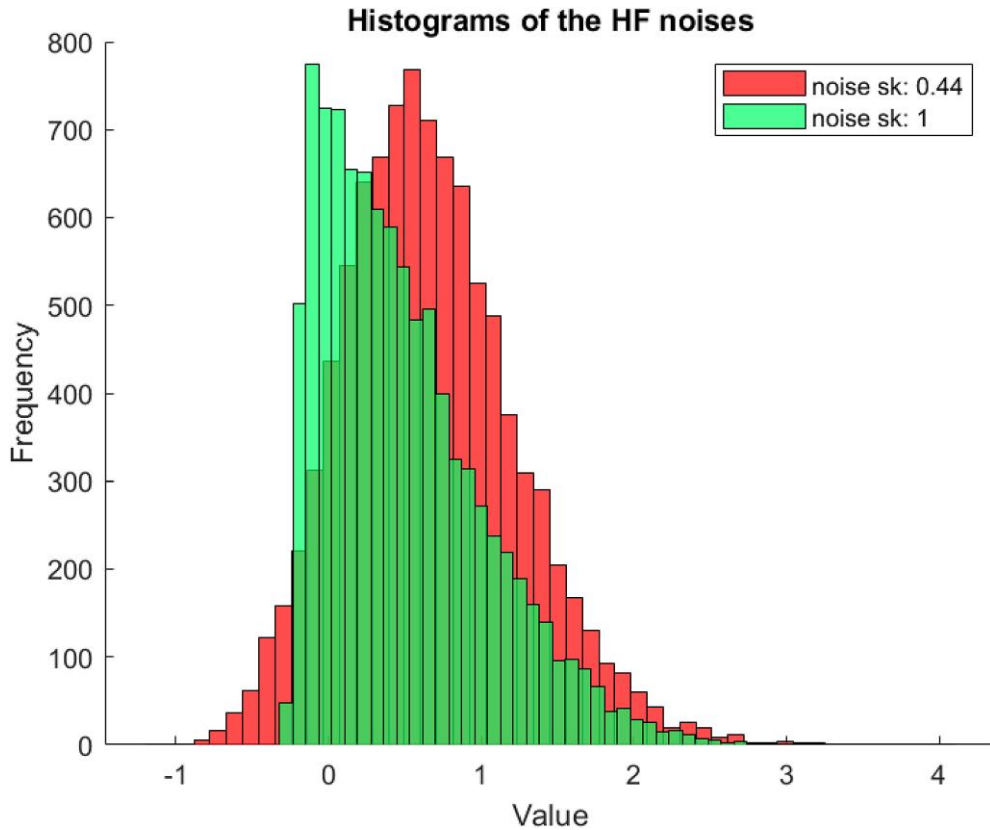
## Data availability

Data used in the paper are public (source Eurostat). Since we use only public data, no Special Permission is need to use copyrighted material from other sources (including the Internet). For reproducibility purposes, all scripts and the data are available at the following GitHub folder [https://github.com/PaoloMaranzano/RC\\_PM\\_RM\\_SCSAR\\_AgroConcentration.git](https://github.com/PaoloMaranzano/RC_PM_RM_SCSAR_AgroConcentration.git). For the WMFGP implementation, refer instead to the following repository: <https://github.com/Pietrostat193/WMFGP>.

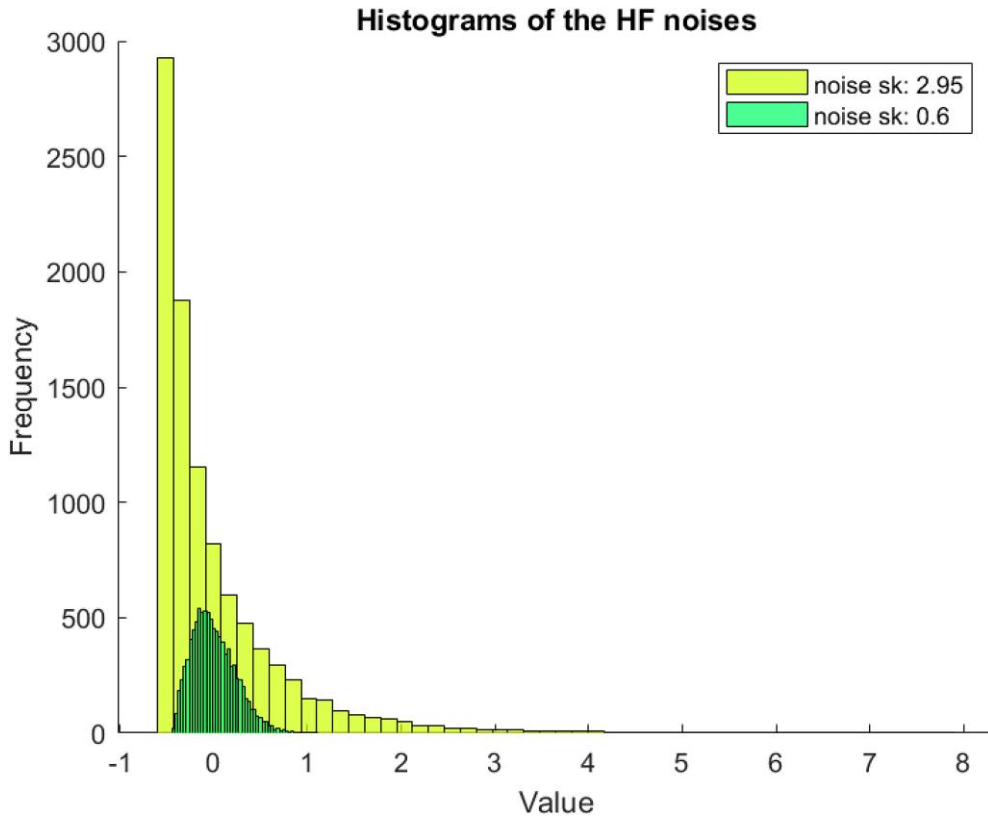
## Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series C*.

### Appendix A: Histograms of the simulated skew noises



**Figure A1.** Histograms of HF data noise for different skewness scenarios. Noise values from a CSN distribution.  $w_H$  depicted in red for low skewness, while  $w_H$  in green for high skewness.



**Figure A2.** Histograms of HF data noise for different skewness scenarios. Noise values from a Weibull distribution. In this case,  $w_H$  depicted in green for low skewness, while  $w_H$  in yellow for high skewness. Note the increased skewness of both noises and the increased variance compared with [Figure A1](#).

### Appendix B: Limitations of the MFGP model for skew data

To understand the limitation of the MFGP, we have to understand the skewness of the discrepancies between two random variables. To determine the skewness of the difference of two random variables  $aX - bY$ , where  $X$  and  $Y$  are positively skewed and positively correlated, we can use properties of skewness and correlation. Let us denote the skewness of  $X$  as  $\text{Skew}(X)$ , the skewness of  $Y$  as  $\text{Skew}(Y)$ , and the correlation between  $X$  and  $Y$  as  $\rho_{XY}$ . The skewness of a linear combination of random variables can be expressed as follows:

$$\text{Skew}(aX - bY) = \frac{E[(aX - bY)^3]}{(E[(aX - bY)^2])^{3/2}}$$

Given that  $X$  and  $Y$  are positively correlated, we have  $\rho_{XY} > 0$ . Now, we can calculate the third central moment  $E[(aX - bY)^3]$  and the second central moment  $E[(aX - bY)^2]$ .

$$\begin{aligned} E[(aX - bY)^3] &= E[a^3X^3 - 3a^2bXY^2 + 3ab^2X^2Y - b^3Y^3] \\ &= a^3E[X^3] - 3a^2bE[XY^2] + 3ab^2E[X^2Y] - b^3E[Y^3] \\ E[(aX - bY)^2] &= E[a^2X^2 - 2abXY + b^2Y^2] \\ &= a^2E[X^2] - 2abE[XY] + b^2E[Y^2] \end{aligned}$$

Given that  $X$  and  $Y$  are positively skewed,  $\text{Skew}(X) > 0$  and  $\text{Skew}(Y) > 0$ , the third central moments  $E[X^3]$  and  $E[Y^3]$  are positive. Using the properties of expectation and correlation, we can determine the signs of the terms involved:

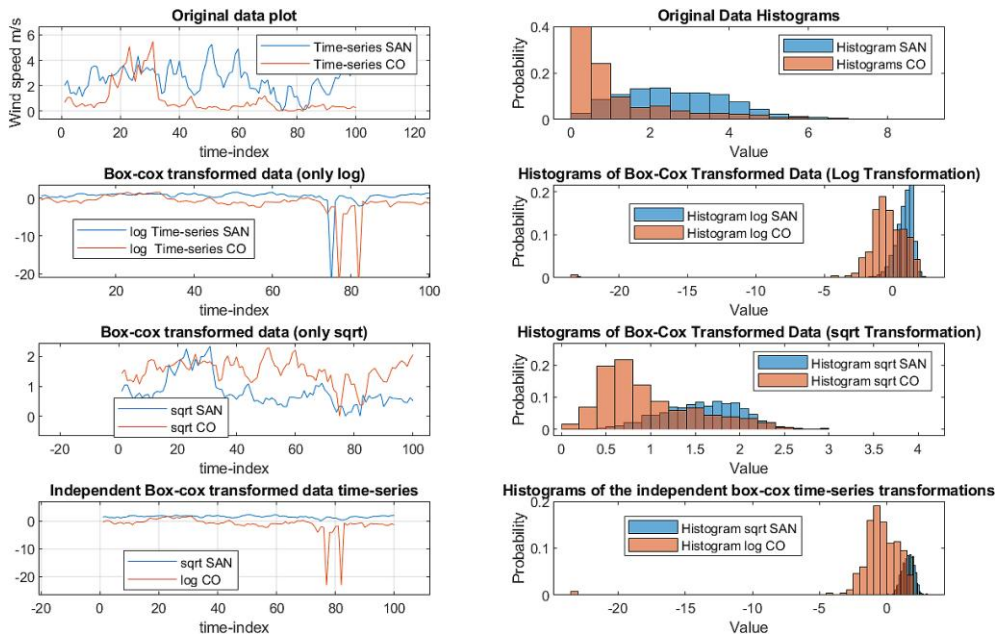
1.  $E[XY^2]$  and  $E[X^2Y]$  will both be positive due to the positive skewness of  $X$  and  $Y$  and the positive correlation between them.
2.  $E[XY]$  will be positive due to the positive correlation between  $X$  and  $Y$ .

Therefore, all terms in the numerator of  $\text{Skew}(aX - bY)$  are positive, and all terms in the denominator are positive as well. This implies that  $\text{Skew}(aX - bY) > 0$ . Thus, if  $X$  and  $Y$  are positively skewed and positively correlated, the difference  $aX - bY$  will also be positively skewed. MFGP depends on two independent Gaussian processes: the discrepancy process  $\delta(\cdot)$  and the LF process  $u_L(\cdot)$ . Both of these processes come with the standard Gaussian process assumptions, including the assumption that both  $\epsilon_L$  and  $\epsilon_\delta$  are independent and identically distributed normal errors. However, these assumptions might not hold in the presence of skewed data. More precisely, suppose that  $\mathbf{x}$  is a generic vector of input locations. If both  $u_H(\mathbf{x})$  and  $u_L(\mathbf{x})$  are skewed, since  $\rho$  is positive, their discrepancies will also be skewed. Rearranging the terms of [equation 2](#), we obtain:

$$\delta(\mathbf{x}) + \epsilon_\delta(\mathbf{x}) = u_H(\mathbf{x}) - \rho u_L(\mathbf{x}). \quad (\text{B1})$$

This leads the first term of [equation B1](#), which is standard Gaussian process, to model the skewed data resulting from  $u_H(\mathbf{x}) - \rho u_L(\mathbf{x})$ , where  $\rho$  is disregarded as it is a positive constant. Since the normality assumption on  $\epsilon_\delta$ , we might incur in an imprecise HF process estimation, due to the inappropriate use of the discrepancy process.

## Appendix C: Failure of Box–Cox transformation in normalization two data-sources jointly



**Figure C1.** The figure displays time-series plots and corresponding histograms for San Siro Alpe Rescascia (SAN) and CO stations under various transformations. In the top-left panels, the original time-series and histograms are shown. The second row presents the time-series and histograms for the log-transformed data of both SAN and CO. The third row shows the time-series and histograms after applying a square root transformation to SAN and CO. Finally, the bottom row features the time-series and histogram of the square root-transformed SAN data alongside the log-transformed CO data.

As an illustrative example, consider the wind-speed monitoring stations San Siro Alpe Rescascia (SAN) and Colico—Via La Madoneta (CO), taken from the dataset described in section 3.2. These two data sources exhibit a linear correlation of 0.53 and considerable skewness, as shown in the top-right panel of Figure C1. To normalize the data according to the Box–Cox criterion, we should apply the log-transformation for CO and the square root transformation for SAN. However, if we were to apply the log-transformation to both data sources, it would only symmetrize the CO data, as shown in the second panel of the second row of Figure C1. Similarly, applying the square root transformation would mainly symmetrize the SAN data.

If we were to use independent transformations, this might reduce the skewness satisfactorily for each dataset. However, it would disrupt the inter-dataset relationships, as shown in the first panel of the fourth row of Figure C1. Here, the inter-dataset relationship is defined as the quantile ordering between the different datasets. For example, if at time  $t$ , a wind speed measurement from SAN is greater than one from CO, this should still hold after transformation. The inter-dataset relationship can be partially captured using the linear correlation coefficient: if it remains constant before and after the transformation, the relationship is preserved. Our algorithm as explained in section 2.3 minimize the distortion introduced by potential transformation while effectively normalizing two data-sources.

## References

- AGNES info, Agnes company website. [https://www.thewindpower.net/windfarm\\_en\\_16755\\_agnes-1.php](https://www.thewindpower.net/windfarm_en_16755_agnes-1.php). Date accessed January 16, 2025.
- Agou V. D., Pavlides A., & Hristopoulos D. T. (2022). Spatial modeling of precipitation based on data-driven warping of Gaussian processes. *Entropy*, 24(3), 321. <https://doi.org/10.3390/e24030321>
- Alodat M. T., & Shakhatreh M. K. (2020). Gaussian process regression with skewed errors. *Journal of Computational and Applied Mathematics*, 370, Article 112665. <https://doi.org/10.1016/j.cam.2019.112665>
- ARPA Data Portal. <https://www.arpalombardia.it/dati-e-indicatori/>. Date accessed January 16, 2025.
- Aslam M. (2021). A study on skewness and kurtosis estimators of wind speed distribution under indeterminacy. *Theoretical and Applied Climatology*, 143(3-4), 1227–1234. <https://doi.org/10.1007/s00704-020-03509-5>
- Bradley P. S., Bennett K. P., & Demiriz A. (2000). Constrained k-means clustering (Technical Report MSR-TR-2000-65). Microsoft Research, Redmond, WA.
- Carta J. A., Ramirez P., & Bueno C. (2008). A joint probability density function of wind speed and direction for wind energy analysis. *Energy Conversion and Management*, 49(6), 1309–1320. <https://doi.org/10.1016/j.enconman.2008.01.010>
- Cleveland R. B., Cleveland W. S., McRae J. E., & Terpenning I. (1990). Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, 6(1), 3–73.
- Colombo L., Marongiu A., Malvestiti G., Fossati G., Angelino E., Lazzarini M., Gurrieri G. L., Pillon S., & Lanzani G. G. (2023). Assessing the impacts and feasibility of emissions reduction scenarios in the Po Valley. *Frontiers in Environmental Science*, 11, Article 1240816. <https://doi.org/10.3389/fenvs.2023.1240816>.
- Colombo P., & Fassò A. (2022). Quantifying the interpolation uncertainty of radiosonde humidity profiles. *Measurement Science & Technology*, 33(7), Article 074001. <https://doi.org/10.1088/1361-6501/ac5bff>
- Costabal F. S., Perdikaris P., Kuhl E., & Hurtado D. E. (2019). Multi-fidelity classification using gaussian processes: Accelerating the prediction of large-scale computational models. *Computer Methods in Applied Mechanics and Engineering*, 357, Article 112602. <https://doi.org/10.1016/j.cma.2019.112602>
- Cutajar K., Pullin M., Damianou A., Lawrence N., & Gonzalez J. (2019). Deep gaussian processes for multi-fidelity modeling third Bayesian deep learning workshop. In *Advances in Neural Information Processing Systems*, NeurIPS 2018, Montreal.
- Erdem E., & Shi J. (2011). Arma based approaches for forecasting the tuple of wind speed and direction. *Applied Energy*, 88(4), 1405–1414. <https://doi.org/10.1016/j.apenergy.2010.10.031>
- Eurostat Database. (2024). <https://ec.europa.eu/eurostat/web/main/data/database>. Date accessed February 27.
- Fassò A., Rodeschini J., Moro A. F., Shaboviq Q., Maranzano P., Cameletti M., Finazzi F., Golini N., Ignaccolo R., & Otto P. (2023). Agrimonia: A dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data*, 10(1), Article 143. <https://doi.org/10.1038/s41597-023-02034-0>
- Fassò A., Sommer M., & von Rohden C. (2020). Interpolation uncertainty of atmospheric temperature profiles. *Atmospheric Measurement Techniques*, 13(12), 6445–6458. <https://doi.org/10.5194/amt-13-6445-2020>
- Genton M. G., & Zhang H. (2012). Identifiability problems in some non-gaussian spatial random fields. *Chilean Journal of Statistics*, 3(2), 171–179.

- Hersbach H., Bell B., Berrisford P., Biavati G., Horányi A., Muñoz Sabater J., Nicolas J., Peubey C., Radu R., & Rozum I. (2023). ERA5 hourly data on single levels from 1940 to present. doi:10.24381/cds.adbb2d47.
- Kennedy M. C., & O'Hagan A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1), 1–13. <https://doi.org/10.1093/biomet/87.1.1>
- Khaledi M. J., Zareifard H., & Boojari H. (2023). A spatial skew-Gaussian process with a specified covariance function. *Statistics & Probability Letters*, 192, Article 109681. <https://doi.org/10.1016/j.spl.2022.109681>
- Le Gallo J., & Fingleton B. (2021). Endogeneity of spatial model. In: M. Fischer, and P. Nijkamp (Eds.), *Handbook of regional science*. Springer.
- Le Gratiet L., & Cannamela C. (2015). Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, 57(3), 418–427. <https://doi.org/10.1080/00401706.2014.928233>
- Lombardy Region Weather Station. <https://www.dati.lombardia.it/Ambiente/Mappa-Stazioni-Meteorologiche/8ux9-ue3c>.
- Lu C-K., & Shafto P. (2021). Conditional deep Gaussian processes: Multi-fidelity kernel learning. *Entropy*, 23(11), Article 1545. <https://doi.org/10.3390/e23111545>
- Maranzano P. (2022). Air quality in Lombardy, Italy: An overview of the environmental monitoring system of ARPA lombardia. *Earth*, 3(1), 172–203. <https://doi.org/10.3390/earth3010013>
- Maranzano P., & Algieri A. (2024). ARPALData: An R package for retrieving and analyzing air quality and weather data from ARPA lombardia (Italy). *Environmental and Ecological Statistics*, 31(2), 187–218. <https://doi.org/10.1007/s10651-024-00599-6>
- Maranzano P., Otto P., & Fassò A. (2023). Adaptive LASSO estimation for functional hidden dynamic geostatistical model. *Stochastic Environmental Research and Risk Assessment: Research Journal*, 37(9), 3615–3637. <https://doi.org/10.1007/s00477-023-02466-5>
- McWilliams B., Newmann M. M., & Sprevak D. (1979). The probability distribution of wind velocity and direction. *Wind Engineering*, 3, 269–273. <https://www.jstor.org/stable/43749150>
- Pavlidis A., Agou V. D., & Hristopulos D. T. (2022). Non-parametric kernel-based estimation and simulation of precipitation amount. *Journal of Hydrology*, 612, Article 127988. <https://doi.org/10.1016/j.jhydrol.2022.127988>
- Perdikaris P., Raissi M., Damianou A., Lawrence N. D., & Karniadakis G. E. (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198), 20160751. <https://doi.org/10.1098/rspa.2016.0751>
- Perdikaris P., Venturi D., & Karniadakis G. E. (2016). Multifidelity information fusion algorithms for high-dimensional systems and massive data sets. *SIAM Journal on Scientific Computing*, 38(4), B521–B538. <https://doi.org/10.1137/15M1055164>
- Raffaelli K., Deserti M., Stortini M., Amorati R., Vasconi M., & Giovannini G. (2020). Improving air quality in the Po Valley, Italy: Some results by the LIFE-IP-PREPAIR project. *Atmosphere*, 11(4), Article 429. <https://doi.org/10.3390/atmos11040429>
- Snelson E., Ghahramani Z., & Rasmussen C. (2003). *Advances in neural information processing systems*. Vol. 16. MIT press. [https://proceedings.neurips.cc/paper\\_files/paper/2003/file/6b5754d737784b51ec5075c0dc437bf0-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2003/file/6b5754d737784b51ec5075c0dc437bf0-Paper.pdf).
- Tassan Mazzocco A., Maranzano P., & Borgoni R. (2023). EEAAq: Handle air quality data from the European environment agency data portal. R package version 0.0.3. <https://cran.r-project.org/web/packages/EEAAq/index.html>.
- Williams C. K. I., & Rasmussen C. E. (2006). *Gaussian processes for machine learning*. (Vol. 2). MIT Press.
- Yu K. N., Cheung Y. P., Cheung T., & Henry R. C. (2004). Identifying the impact of large urban airports on local air quality by nonparametric regression. *Atmospheric Environment*, 38(27), 4501–4507. <https://doi.org/10.1016/j.atmosenv.2004.05.034>
- Zhu A-X., & Turner M. (2022). How is the third law of geography different? *Annals of GIS*, 28(1), 57–67. <https://doi.org/10.1080/19475683.2022.2026467>