Multifidelity methods for data fusion of environmental data

Pietro Colombo

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy in Statistics

January 2026

Multifidelity methods for data fusion of environmental data

Pietro Colombo

School of Mathematics and Statistics

University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy in Statistics

January 2026

Multifidelity models are a type of data fusion approach that use information from multiple data sources arranged in a hierarchy. This structure is especially useful in environmental modelling, where data sources naturally differ in terms of reliability, resolution, and how often they are available. Although Gaussian processes—the foundation of multifidelity models—are commonly used in environmental science, their application in multifidelity settings has been quite limited. This thesis explores the use of multifidelity modeling for combining environmental data from different sources, with a focus on wind speed as a representative example. It examines when multifidelity methods are useful and compares them with standard methods used in the industry. As part of this work, the multifidelity model is extended to handle skewed data by introducing a new method called the Warped Multifidelity Gaussian Process (WMFGP). In addition, a scalable framework for modeling both spatial and temporal data is developed. This framework leveraged the use of Vecchia approximation to make complex multifidelity models more efficient. Both the WMFGP and the scalable framework prove effective in modeling wind speed data from Lombardy, a region in northern Italy.

Contents

List of Figures

List of Tables

0.1 List of acronyms

Table 1: List of acronyms used in the thesis.

Acronym	**Meaning**
AGNES	Offshore wind-farm case study / project
AR	Autoregressive
AR1	First-order autoregressive multifidelity model
ARPA	Agenzia Regionale per la Protezione dell'Ambiente
AWS	Automatic Weather Station
BCMF	Box–Cox Multifidelity model
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
C3S	Copernicus Climate Change Service
CASG	Covariance-Adjusted Skew-Gaussian
CDF	Cumulative Distribution Function
CO	Colico–Via La Madoneta station/dataset
CRPS	Continuous Ranked Probability Score
CSN	Closed Skew-Normal
DA	Data Assimilation
DDWGP	Data-Driven Warped Gaussian Process
ERA5	ECMWF Reanalysis v5

Acronym	Meaning
GCOS	Global Climate Observing System
GCM	General Circulation Model
GIS	Geographic Information System
GLM	Generalized Linear Model
GP	Gaussian Process
GPHF	Gaussian Process fitted to high-fidelity data
GPLF	Gaussian Process fitted to low-fidelity data
GPR	Gaussian Process Regression
GPRSE	Gaussian Process Regression with Skewed Errors
GPSCF	Gaussian Process with Specified Covariance Function
GWR	Geographically Weighted Regression
HF	High Fidelity
ILPE	Inner-Level Pattern Extraction
INLA-SPDE	Integrated Nested Laplace Approximation with Stochastic Partial Differential
KS	Kolmogorov–Smirnov
LF	Low Fidelity
LIDAR	Light Detection and Ranging
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo
MF	Multifidelity
MFcp	Multifidelity comparison constant mean polynomial ρ function
MFGP	Multifidelity Gaussian Process
ML	Missing sequence Length
MLE	Maximum Likelihood Estimation
MOS	Model Output Statistics
NARGP	Nonlinear Autoregressive Gaussian Process
NDA	Non-Disclosure Agreement
NEE	Normalised Estimation Error
NES	New Experimental Setting

Acronym	Meaning
NN	Nearest Neighbour
PCSN	Parsimonious Closed Skew-Normal
PDF	Probability Density Function
PP	Perfect Prognosis
PSD	Positive Semidefinite
PWGP	Parametric Warped Gaussian Process
QGBRT	Quantile Gradient Boosted Regression Tree
RBF	Radial Basis Function
RMSE	Root Mean Squared Error
RMSED	Root Mean Squared Error of the Distribution
RQ	Research Question
RSS	Residual Sum of Squares
SAEM	Stochastic Approximation Expectation–Maximisation
SAN	San Siro Alpe Recascia station/dataset
SAR	Synthetic Aperture Radar
SCR	Spatially Clustered Regression
SD	Standard Deviation
SI	Simple Imputation
SPDE	Stochastic Partial Differential Equation
STL	Seasonal and Trend decomposition using Loess
SUN	Unified Skew-Normal
TC	Total Computation Cost
TS2	Second-experiment observed time-series
UQ	Uncertainty Quantification
UTM	Universal Transverse Mercator
WGS	World Geodetic System
WGP	Warped Gaussian Process
WMF	Warped Multifidelity
WMFGP	Warped Multifidelity Gaussian Process

Chapter 1

Introduction

1.1 Overview

New advances in technology have made it increasingly easier to collect environmental data from multiple sources. *Data fusion*, which involves merging data to obtain information that is more consistent, informative, and accurate than the original raw data—often uncertain or imprecise—offers a comprehensive route to insight. However, factors such as reliability, resolution, and sparsity (due to missingness processes) make such integration challenging. Many studies have examined data fusion as a means of integrating and combining diverse data sources, considering factors such as spatial support, environmental applications, and source reliability. Yet, providing a unified overview of this complex field remains challenging because each research branch is typically developed around a specific application problem. However, a general broad view might be the following: *Data assimilation* couples dynamical models with observations (e.g., Kalman or ensemble Kalman filters), providing strong physical consistency but at a high computational cost. *Statistical downscaling* methods map coarse to fine scale information (as in Perfect Prognosis or Model Output Statistics approaches), offering effective bias correction yet often struggling to capture sub-grid variability and extremes. *General data-fusion* techniques—such as multi-source regression, co-kriging, or manifold learning—are flexible but typically rely on dense spatial coverage or well-paired datasets. Fi-

nally, multi-fidelity data sources in statistics refer to collections of datasets that observe the same underlying system or phenomenon at differing levels of accuracy, resolution, cost, and reliability. Multi-fidelity modelling treats these different data sources as hierarchically related observations of a common latent process, explicitly balancing the scarcity of high-fidelity (high quality, highly costly, high resolution) data against the abundance of low-fidelity (low quality, cheap, low resolution) data while providing calibrated uncertainty quantification.

The following examples further clarify how these methodological strands have been used across different application domains. For instance, Gotway and Young (2002) primarily focused on integrating different spatial supports with an application to agricultural science, where datasets are measured on grids with different cell sizes (spatial support). Such integration of different supports is generally useful for improved inference.

Climate modelling, by contrast, often fuses satellite data—low resolution but broad spatial coverage—with in-situ measurements—limited coverage but high resolution (Fernández-Godino et al., 2019). The broad class of methods used here is known as *statistical downscaling* (Maraun, 2016; Maraun et al., 2010). These techniques include a "bias correction" step that improves the quality of low-resolution data using high-resolution observations. In this context, fusion focuses less on the support and more on the values of the phenomenon. In its simplest form, downscaling maps large-scale (low-resolution) predictors to the expected value of a small-scale predictand. A useful classification of statistical downscaling approaches, based on the nature of the predictors, is provided by Rummukainen (1997): *Perfect Prognosis* (PP: linear models, GLMs, nonlinear regression, quantile and multisite regression; only observational data) and *Model Output Statistics* (MOS: delta change, quantile mapping, additive scaling; relationships between model output and observed data at specific locations).

Some papers (e.g., Xu et al., 2021) propose methods for data sparsity in fusion, while others model complex input-output relationships using *manifolds*—topological spaces locally Euclidean but globally non-Euclidean. Manifold approaches can capture nonlinear relations (Lin et al., 2019) and reveal low-dimensional structure in high-

dimensional environmental data (Calandra et al., 2016).

A separate but related line of work is *data assimilation* (DA), which blends a time-evolving dynamical model with observations to update the model state (e.g., Kalman filter, Ensemble Kalman Filter, 3D/4D-Var). Unlike statistical downscaling, DA leverages model dynamics explicitly and is primarily temporal in nature. We mention DA here to situate our approach within the broader landscape; the methods later developed in this thesis are not DA and do not require a prognostic numerical model.

In environmental statistics, more recent contributions include Wilkie et al. (2019), who developed Bayesian hierarchical models with spatially varying coefficients to combine data of different spatial-temporal characteristics (applied to log(chl-a) in Lake Balaton), and Villejo et al. (2023), who used an INLA-SPDE algorithm to link environmental pollutant levels to health outcomes.

A distinct fusion scheme known as *multifidelity* (MF; Fernández-Godino et al., 2019) has been developed under different principles and is widely used in engineering. MF modelling assumes both high-fidelity (HF) and low-fidelity (LF) data. MF surrogates harness multiple sources, ordered by reliability, to exploit the strengths of each. Although not necessarily more accurate than carefully crafted mono-fidelity models in every situation, MF surrogates are often *far more cost-effective*: HF samples are expensive and sparse, whereas LF samples are plentiful and cheap but biased or noisy. A further advantage is principled *uncertainty quantification* (UQ), since MF models are commonly cast in a Gaussian process (GP) framework with well-defined predictive variances (see Section 1.3.1). The hierarchical structure of MF—different fidelities as layers of accuracy—aligns well with real environmental data ecosystems. Recent research activity is vibrant (Ding et al., 2025; Sella et al., 2025), with growing use in spatio-temporal environmental statistics (Christelis et al., 2023; Giannoukou et al., 2025; Lee et al., 2024).

Problem statement. Environmental data fusion is impeded by three recurrent issues: (i) incomplete data (gaps and missingness); (ii) skewed distributions that challenge Gaussian assumptions; and (iii) under-coverage in space–time. This thesis

investigates multifidelity Gaussian process (MFGP) models as a pragmatic and flexible framework for addressing these challenges simultaneously, with a focus on methods that remain applicable under realistic environmental monitoring constraints.

The remainder of this chapter introduces the methodological foundations and empirical context of the thesis. Sections 1.3.1 and 1.3.2 outline the Gaussian process regression framework, followed by Section 1.3.3 and Section 1.3.4, which introduce the multifidelity modelling paradigm and its GP formulation. Section 1.3.6 describes estimation and prediction in linear MFGP models. The motivating real-world applications that drive the methodological development are then presented in Section 1.4. Section 1.5 details the datasets used throughout the thesis. Finally, Sections 1.6–1.8 explain the focus on wind-speed data, summarize the original contributions of the thesis, and provide a roadmap for the chapters that follow.

1.2 Why study multifidelity models: motivations and research questions

This section elaborates on the motivations underlying the problem statement in Section 1.1 and formulates the research questions that guide the remainder of the thesis. Although PP/MOS bias correction can bridge observational and high-resolution data, they typically cannot capture sub-grid variability or complex terrain effects. They may approximate long-term means but often miss local variability and tails behaviour. Several studies (Maraun, 2016; Maraun and Widmann, 2018; Maraun et al., 2010) show that even quantile mapping—designed to match variances and tails behaviour—can introduce artefacts in downscaling, discouraging its use for highly localized applications (e.g., the application later discussed AGNES or ARPA Lombardia are highly localized.). Other fusion approaches may require dense networks, e.g., Winstral et al. (2017) for wind-field downscaling (∼200 stations), infeasible in settings like AGNES with at most three HF points. In contrast, MFGP models handle data scarcity while retaining accuracy without large HF datasets. The GP foundation enables UQ, often as important as the mean predictions themselves (e.g., Global Climate Observing System (GCOS) emphasises

trusted uncertainty; probabilistic forecasting for wind is widely used in research and industry (Browell and Gilbert, 2020)). Moreover, MFGP for spatio-temporal statistics remains relatively under-explored, despite its potential.

Research Questions (RQs).

RQ1: *Can MFGP reduce dependence on high-fidelity samples while achieving comparable or better predictive accuracy than mono-fidelity baselines under realistic noise and sparsity? To which extent is this feasible?*

RQ2: *How can we robustly accommodate skewed environmental variables (e.g., wind speed) in an MF setting without breaking cross-fidelity structure or miscalibrating uncertainty?*

RQ3: *Can MFGP scale to large spatio-temporal domains (fields over many sites/times) while remaining numerically stable, and can it flexibly fuse fidelities across space (e.g., via spatially varying coupling)?*

RQ1 is tackled through systematic comparisons across regimes; RQ2 with a development of the warped MFGP that preserves LF-HF coupling; and RQ3 with the development of a scalable approximations (Vecchia) and spatially varying integration.

1.3 The Gaussian process and the multifidelity framework

1.3.1 Gaussian process framework

A Gaussian process is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions (Rasmussen, 2004).

In a Gaussian process, the mean function $m(x)$ defines the expected value of the process at any given input x. It encapsulates our prior beliefs about the underlying function behaviour. The covariance function $k(x, x')$, also known as the kernel function, governs the degree of similarity between the function values at

two different input points x and x'. It quantifies the covariance (or correlation) between these points and plays a crucial role in modelling the smoothness and variability of the process. Gaussian processes (GPs) are characterised by several key assumptions:

1. **Gaussian Distribution**: The joint distribution of any finite number of function values is Gaussian. This implies that any subset of function values follows a multivariate Gaussian distribution.

2. **Mean and Covariance Functions**: A GP is fully specified by its mean function $m(x)$ and covariance function (or kernel) $k(x, x')$. The mean function represents the expected value of the function at any point x, while the covariance function captures the pairwise relationships between function values at different input points x and x'.

3. **Stationarity**: The covariance function typically assumes stationarity, meaning that the covariance between two points x and x' depends only on their separation $|x - x'|$, not on their absolute locations. This assumption implies that the behaviour of the process remains consistent across the input space.

4. **Noisy Observations**: Gaussian processes can handle noisy observations. It is assumed that observations are obtained by adding Gaussian noise to the true function values. This noise is usually assumed to be independent and identically distributed (i.i.d.) normal error.

5. **Flexibility**: Gaussian processes provide a flexible framework for modelling functions without imposing specific parametric forms. This flexibility allows GPs to capture complex and nonlinear relationships in the data.

These assumptions make Gaussian processes a versatile and widely applicable tool for various tasks such as regression, classification, and optimisation, especially in scenarios where uncertainty quantification and flexibility are paramount.

1.3.2 Gaussian process regression

A Gaussian process regression model is a flexible regression technique that models many relations between input x and output y with few decisions regarding the underpinning covariance function. Assuming that $y(x)$ is the observed scalar response of a set of random variable Y indexed at some input location x, we can write down a Gaussian process regression model in the following form:

$$y(x) = f(x) + r, \tag{1.1}$$

where f is some smooth function with a Multivariate Gaussian distribution defined by a covariance matrix $\mathbf{K}$, and a mean vector $\boldsymbol{\mu}$, while $r \sim N(0, \sigma_r)$ is a noise component. The degree of smoothness of the GP function that relates the input and output is determined by the covariance/kernel function. There are many classes of covariance functions. A general class is the *Matérn* , which generalizes also the exponential class. The *Matérn* takes the following form:

$$k_\nu(d) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(d\sqrt{\frac{2\nu}{l}} \right)^\nu K_\nu \left(d\sqrt{\frac{2\nu}{l}} \right). \tag{1.2}$$

K_ν is a Bessel function, ν is a parameter that controls the smoothness, l is the so-called length scale parameter that determines the speed of decrease of correlation for increasing distance inputs, and d is the Euclidean distance for chosen inputs. The *Matérn* reduces to the *Squared Exponential* , when $\nu \to \infty$:

$$k_\infty(d) = \exp\left\{ -\frac{d^2}{2l} \right\}. \tag{1.3}$$

Sometimes, the covariance functions are parametrised by additional parameters such as the *nugget* g and the signal variance σ_s. The first is also called *noise variance* and it modulates the micro-variation, the roughness at small scales. Since the GP can be considered a smoothing technique, what the nuggets do it is to break the perfect smoothing of the process introducing a disturbance. Finally, σ_s controls the amplitude of the harmonics of the smoothed signal. If multiple inputs (x is a

matrix, not a vector) are used also the length scale parameters might be multiple. In other words, the "direction matters" (anisotropy), and the correlation extent become dimension-dependent.

The inference for a Gaussian process regression is usually performed by means of the maximum likelihood estimations, where the following represents the log-likelihood function to optimise:

$$\log p(\mathbf{y} \mid \mathbf{X}) = -\frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}\log\det \mathbf{K} - \frac{n}{2}\log(2\pi). \tag{1.4}$$

In the latter likelihood, the $\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}$ the quadratic term represents the fit to the data, $\log\det \mathbf{K}$ a penalty term and quantity $\frac{n}{2}\log(2\pi)$ a normalising constant. The optimised parameters are those of the chosen covariance function, which can be set adapt to any level of smoothness automatically as more data are fed into the model. This leads to the interpretation of a GP as a distribution over functions, which means that the model does not have predefined "shape", it adapts naturally to any shape as more data are added to the model. For this latter reasons, the GP model is often referred as a non-parametric regression technique.

1.3.3 Multifidelity problem

The origins of MF modelling trace to Kennedy and O'Hagan (2000), who introduced autoregressive[1] MF models to predict outputs of complex computer codes when fast approximations are available. This foundational work established MF's core idea: balance cost and accuracy by combining a small set of expensive, accurate evaluations with abundant, cheaper approximations.

The term "multifidelity" became common with Forrester and Keane (2009): a large set of inexpensive LF data is coupled with a smaller HF set to improve accuracy relative to using HF alone. Formally, suppose $f : X \to Y$ is the target, with costly f_H and cheaper f_L approximations. We can view $f_L = h_{\text{trans}} \circ f_{\text{exact}}$ with bias induced by h_{trans}, while $f_H \approx f_{\text{exact}}$.

[1] Here, 'autoregressive' refers to the linkage across fidelity levels, not to a temporal or spatial AR process.

1.3.4 Multifidelity GP regression

Among MF models (see Costabal et al., 2019 for a review), the simplest and most widely used is the *autoregressive* MF scheme (Kennedy and O'Hagan, 2000), later recast in recursive and nonlinear forms (Le Gratiet and Garnier, 2014; Perdikaris et al., 2017). Consider datasets $D_1, \ldots, D_S$ at increasing fidelity (index s), with sizes $n_1 > \cdots > n_S$. The AR structure is

$$f_s(x) = \rho_{s-1}\, f_{s-1}(x) + \delta_s(x), \tag{1.5}$$

with coupling ρ_{s-1} and discrepancy δ_s. For two levels, D_L (LF) and D_H (HF), we drop s and write ρ for the coupling.

A common assumption places independent GP priors on f_L and δ:

$$f_L(x) \sim \mathcal{GP}\big(m_L(x),\, k_L(x, x')\big), \tag{1.6}$$

$$\delta(x) \sim \mathcal{GP}\big(m_\delta(x),\, k_\delta(x, x')\big), \tag{1.7}$$

and we observe noisy versions

$$y_L(x) = f_L(x) + \varepsilon_L, \quad \varepsilon_L \sim \mathcal{N}(0, \sigma_L^2), \tag{1.8}$$

$$y_H(x) = \rho\, f_L(x) + \delta(x) + \varepsilon_H, \quad \varepsilon_H \sim \mathcal{N}(0, \sigma_H^2). \tag{1.9}$$

Then $f_H(x) = \rho f_L(x) + \delta(x)$ is the latent HF process. The coupling ρ can be a scalar or function $\rho(x)$. As a scalar, it captures both correlation and rescaling of LF information; as $\rho(x)$, it permits spatially varying integration but requires regularization to avoid identifiability conflicts with δ.

An empirical ρ can be written as the best linear mean-square coefficient (given aligned inputs):

$$\rho(x) \;=\; \frac{\operatorname{cov}\big(f_H(x), f_L(x)\big)}{\operatorname{var}\big(f_L(x)\big)}\,, \tag{1.10}$$

this empirical ρ assumes truly paired HF-LF observations—same location, time, and measurement support. Without such alignment, the estimate can be biased.

1.3.5 Relation to Bayesian Melding.

A related framework for combining multiple sources of information is Bayesian Melding (Poole and Raftery, 2000), which provides a principled approach for integrating deterministic model outputs with observational data. In Bayesian Melding, prior distributions are specified both on model inputs and outputs, and these are combined to produce a coherent posterior distribution that reconciles discrepancies between simulated and observed quantities.

While both Bayesian Melding and multifidelity Gaussian process (MFGP) models aim to integrate heterogeneous sources of information, they differ fundamentally in their formulation and objectives. Bayesian Melding is typically designed for settings where a deterministic or mechanistic model is available, and the goal is to calibrate or adjust this model using observed data. In contrast, MFGP models are fully data-driven and treat multiple datasets as noisy observations of a latent process at different levels of fidelity, without requiring an explicit underlying physical simulator.

Moreover, MFGP explicitly models the cross-correlation structure between fidelity levels through parameters such as ρ, enabling flexible data-driven learning of inter-source relationships. This makes MFGP particularly well suited for environmental monitoring contexts where multiple observational sources (e.g., satellite and in-situ data) are available, but no single deterministic model fully captures the underlying process.

For these reasons, the MFGP framework is adopted in this thesis as a flexible and scalable alternative to model-based data integration approaches such as Bayesian Melding.

1.3.6 Prediction and estimation in linear MF models

MFGPs can be viewed as hierarchical multi-output GPs. Let $\boldsymbol{\theta}$ collect hyperparameters (signal variances for f_L and δ, length-scales, noise variances σ_L^2, σ_H^2, and ρ). For observations $\mathbf{y} = [\mathbf{y}_L, \mathbf{y}_H]^\top$, the negative log marginal likelihood is

$$\mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{2}\mathbf{y}^\top \mathbf{K}_y^{-1}\mathbf{y} + \tfrac{1}{2}\log|\mathbf{K}_y| + \tfrac{n_L+n_H}{2}\log(2\pi),$$

with block covariance

$$\mathbf{K}_y = \begin{bmatrix} \mathbf{K}_{LL} + g_L \mathbf{I}_{n_L} & \mathbf{K}_{LH} \\ \mathbf{K}_{HL} & \mathbf{K}_{HH} + g_H \mathbf{I}_{n_H} \end{bmatrix}, \quad \mathbf{K}_{LL} = [k_L(x_i^{(L)}, x_j^{(L)})],$$

$$\mathbf{K}_{HL} = \rho\,[k_L(x_i^{(H)}, x_j^{(L)})], \quad \mathbf{K}_{LH} = \mathbf{K}_{HL}^\top, \quad \mathbf{K}_{HH} = \rho^2[k_L(x_i^{(H)}, x_j^{(H)})] + [k_\delta(x_i^{(H)}, x_j^{(H)})].$$

Prediction at new inputs $\mathbf{X}_*$ follows standard GP conditioning:

$$\boldsymbol{\mu}_H(\mathbf{X}_*) = \mathbf{K}_{*y}\mathbf{K}_y^{-1}\mathbf{y}, \qquad \boldsymbol{\Sigma}_H(\mathbf{X}_*) = \mathbf{K}_{**}^H - \mathbf{K}_{*y}\mathbf{K}_y^{-1}\mathbf{K}_{y*}.$$

1.3.7 Limitations of the MFGP framework

The challenges discussed in Section 1.2—such as missing data, skewness, and limited spatio-temporal coverage—are intrinsic to environmental datasets. However, it is important to emphasise that the methodological developments in this thesis are carried out *conditional on adopting a Gaussian process (GP)-based multifidelity framework*. This modelling choice introduces additional, model-specific limitations that must be acknowledged.

A first limitation concerns the role of the coupling parameter ρ, which governs the contribution of the low-fidelity (LF) process to the high-fidelity (HF) signal. In the standard autoregressive formulation,

$$f_H(x) = \rho f_L(x) + \delta(x), \tag{1.11}$$

the effectiveness of the multifidelity structure relies on a sufficiently strong correlation between LF and HF data. In the limiting case where $\rho \to 0$, the LF component vanishes and the model reduces to a single-fidelity Gaussian process defined by the discrepancy term $\delta(x)$. In such scenarios, the multifidelity framework provides little practical advantage, highlighting its dependence on the existence of meaningful cross-fidelity relationships.

A second, more fundamental issue is that of identifiability. When flexible specifications are adopted for the discrepancy process $\delta(x)$, the coupling parameter $\rho(x)$,

and the mean functions, the model may admit multiple parameter configurations that explain the data equally well. In particular, variability in the HF observations can be ambiguously attributed either to the scaled LF component $\rho f_L(x)$ or to the discrepancy term $\delta(x)$. This lack of identifiability can lead to overfitting, unstable parameter estimation, and reduced interpretability of the model components.

These issues are exacerbated when ρ is allowed to vary over the input space, as increased flexibility in $\rho(x)$ may conflict with the role of $\delta(x)$ in capturing residual structure. For this reason, practical implementations of MFGP models often require regularisation, constraints, or parsimonious modelling choices to ensure a meaningful decomposition across fidelity levels.

The considerations above motivate several of the modelling decisions adopted in this thesis, including the emphasis on controlled flexibility, the use of structured transformations, and the evaluation of regimes in which multifidelity approaches offer clear advantages over mono-fidelity alternatives.

1.4 Motivating examples

Before introducing the datasets used in this thesis, we first present two real-world motivating examples that exemplify the practical challenges driving our methodological choices. These case studies highlight how environmental data fusion is typically confronted with sparse and heterogeneous observations, systematic biases across sources, incomplete records, and strong spatio-temporal variability. By grounding the discussion in concrete applications, we clarify the types of data environments for which the proposed multifidelity methods are designed, and establish the empirical context in which the subsequent datasets and modelling developments are situated.

Wind farm development (AGNES Project). From wind farm development to accurate mapping of environmental variables, wind resource assessment has long been challenging. The AGNES project (Chief Operating Officer, 2022) involves offshore wind farms (Romagna 1 and 2). Wind speed is critical: too low and farms underperform; uncertainty also propagates to financial risk (AGNES was eligible for 70M€

public funding in June 2021 (Il Fatto Quotidiano, 2022)). Data acquisition costs are a major component for offshore renewables (Medina-Lopez et al., 2021). Synthetic Aperture Radar[1] (SAR) helps with detailed wind-field maps, but uncertainty remains high (discrepancies up to 11% (Zen et al., 2021)). Without accurate resource determination, neither profitability nor infrastructure deterioration can be adequately assessed; weather uncertainty directly affects resource forecasting and maintenance access.

Reanalysis products (e.g., ERA5) suffer from (i) coarse resolution (grid cells larger than farm scales) and (ii) bias relative to in-situ. These motivate advanced fusion to integrate multiple sources and improve accuracy and reliability. Figure 1.1 schematises a typical multi-source setup with LF grids (green), HF points (red), and unobserved HF targets (blue) to be reconstructed via fusion.

ARPA Lombardia monitoring network. The regional network provides valuable environmental variables (wind speed, temperature, humidity) but suffers frequent data gaps and under-coverage. Such gaps, see Figure 1.2, risk missing key meteorological events, impeding causal understanding and leaving areas unmonitored. These limitations highlight the need for spatio-temporal modelling and MF fusion to reach reliable regional-scale inference under sparse, incomplete observations.

1.5 Datasets

The datasets used in this PhD are introduced briefly here; details reappear throughout the chapters of the thesis to follow. Datasets are used entirely (Section 3), partially (Section 4), or as the basis for simulation experiments (Section 2, Section 3, and Section 4).

1.5.1 ERA5 reanalysis

The **ERA5**(Copernicus Climate Change Service (C3S), 2025) data contain hourly wind speed measurements from 1979 onwards. These data combine a meteorological

[1]It is a radar imaging system that produces detailed surface images by synthesizing a large antenna aperture through platform motion.

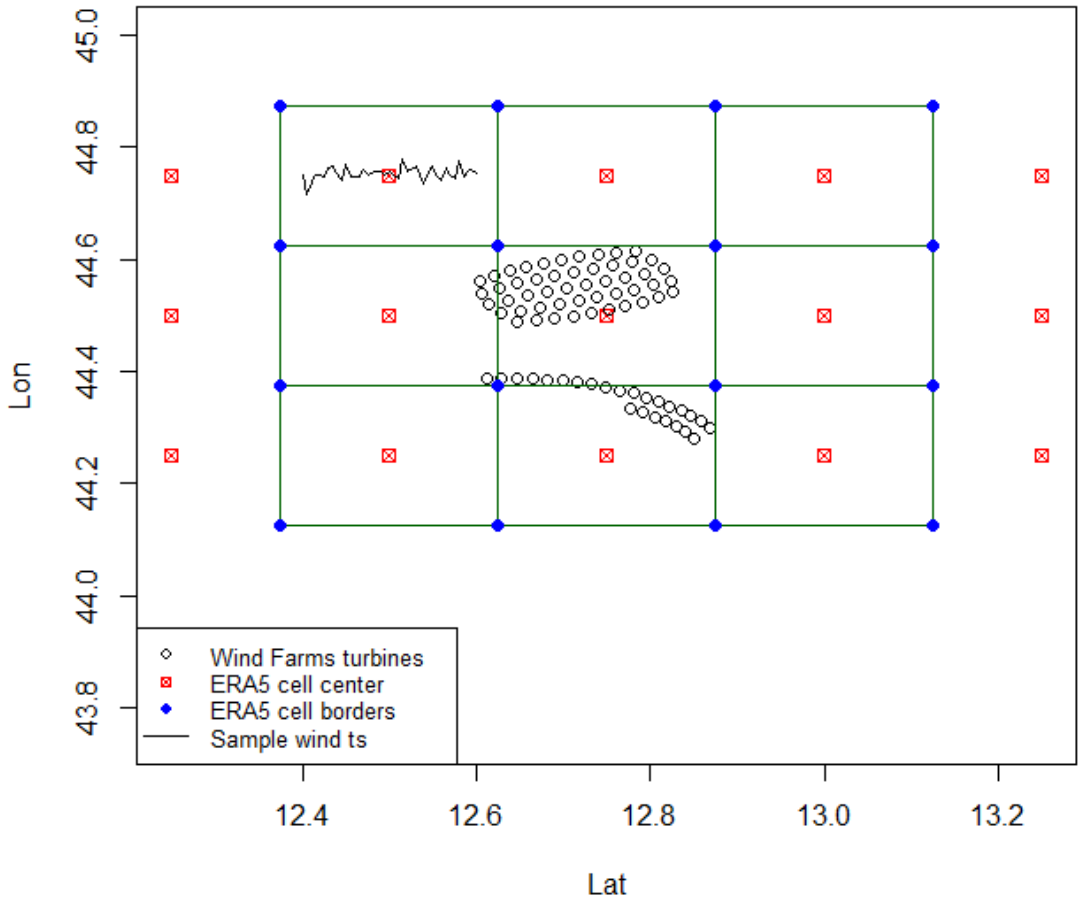


Figure 1.1: Schematic data-fusion scenario. Green squares: low-fidelity, low-resolution wind-field information; red points: high-fidelity, high-resolution measurements; blue points: target HF values at yet-unobserved locations. The task is to reconstruct blue from green and red (units: m/s for wind speed).

model, satellite observational data, and terrestrial measurements to build a coherent weather map. These data have a spatial resolution of $0.25^\circ \times 0.25^\circ$, and their wind speed representation tends not to be precise enough for understanding local variability (Wang et al., 2026). Throughout the thesis, these data are consistently used either as direct sources of LF data (see Section 4) or as the basis for simulation procedures (see Section 2, Section 3, and Section 4).

1.5.2 ARPA Lombardia

The Italian environmental agency, ARPA, plays a pivotal role in collecting comprehensive air quality and environmental data, encompassing vital parameters such

as wind speed, temperature, and humidity. These datasets are made accessible to researchers and policymakers through a variety of platforms, including ARPA's dedicated data portal (ARPA Lombardia, 2025), the Regione Lombardy open portal, and the R package `ARPALData`. For both the simulation and application phases of our study described in Chapters 3 and 4, we procured data spanning from January 1st to December 31st, 2022. This extensive temporal coverage allowed us to capture a robust representation of environmental conditions throughout the year.

The monitoring network deployed across the Lombardy region comprises 94 wind speed monitoring stations strategically positioned to provide comprehensive spatial coverage. These stations record data at 10-minute intervals, furnishing researchers with high-resolution insights into local atmospheric dynamics. However, analysis of the collected data for 2022 revealed a common issue: frequent sequences of missing data. These gaps in the data pose significant challenges for analytical work, particularly when employing interpolation methods to fill in the missing information. Of particular concern are gaps lasting from 15 to 192 hours, which emerged as the most frequent temporal anomalies within the dataset. While statistical techniques can effectively address shorter gaps, those exceeding 192 hours are typically indicative of structural malfunctions or prolonged maintenance periods. Consequently, attempting to employ statistical methods to interpolate data over such extended durations would be inappropriate and could lead to unreliable outcomes. Hence, our analysis in Chapter 3 focuses on developing robust methodologies capable of effectively handling missing data sequences within the bounds of practical and statistical feasibility.

1.5.3 South Lombardy Dataset

The South Lombardy Wind Speed Dataset contains wind speed measurements from the southern region of Lombardy. The primary purpose of this dataset is to evaluate different modelling approaches. While the dataset is large (roughly 26000 rows between HF and LF stations), it is not prohibitively massive. This means that although spatio-temporal studies require specialized approximations to remain computationally feasible, the dataset's size still allows for the testing of various modelling strategies

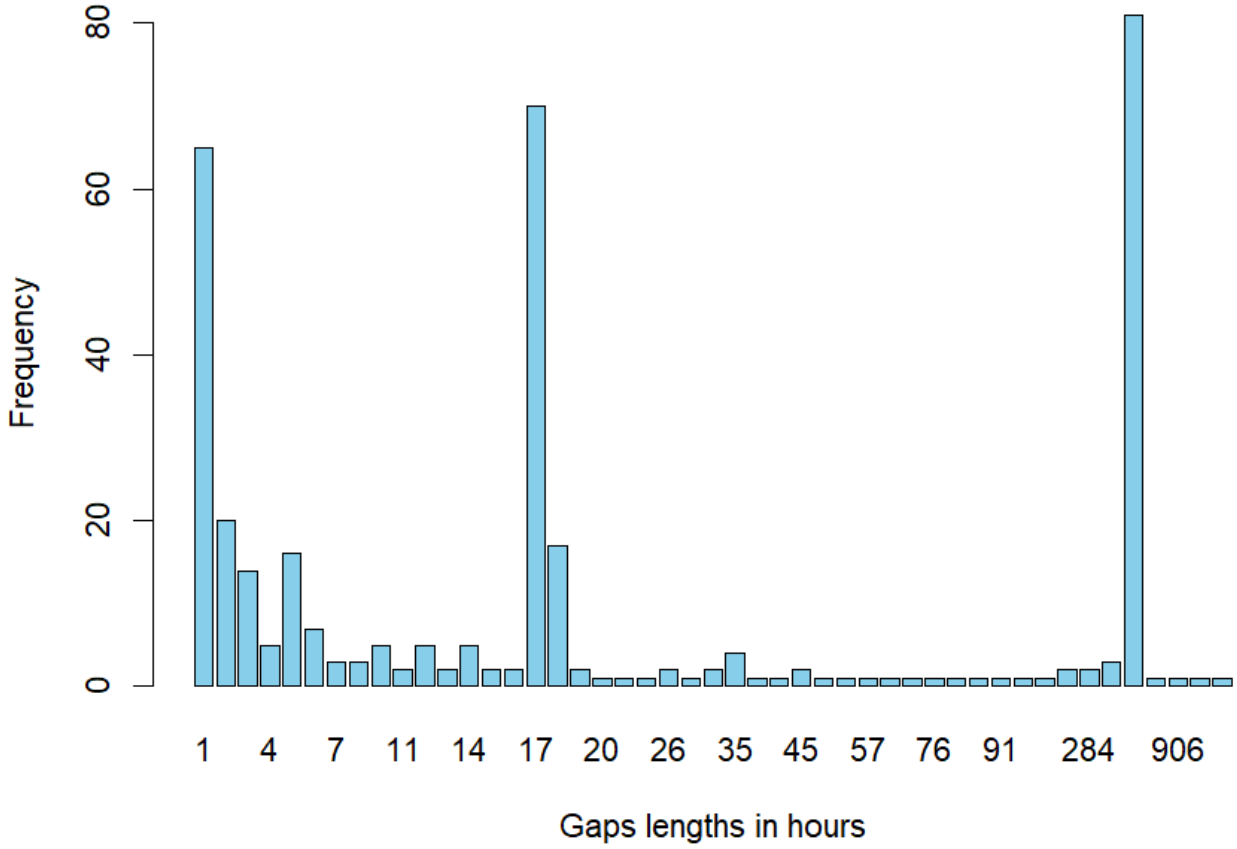


Figure 1.2: Number of gaps by gap length (hours) in the 2022 ARPA Lombardia dataset.

relatively rapidly.

The dataset exhibits several noteworthy characteristics, including variations in station density across the region, minimal natural obstacles due to the predominantly flat terrain, and a non-uniform spatial decay of correlation. These features make it particularly valuable for studying wind speed dynamics in a structured yet complex setting. The dataset combine data coming from two primary sources:

1. The ERA5 data which offers broad-scale view of wind speed trends.
2. A subset of 18 stations of the ARPA Lombardia monitoring network for high-fidelity wind speed data.

To create a unified dataset, a nearest neighbor spatial matching is used, where each monitoring station from the ARPA Lombardia network was matched with its closest corresponding grid cell in the ERA5 data. The resulting dataset contains 13392

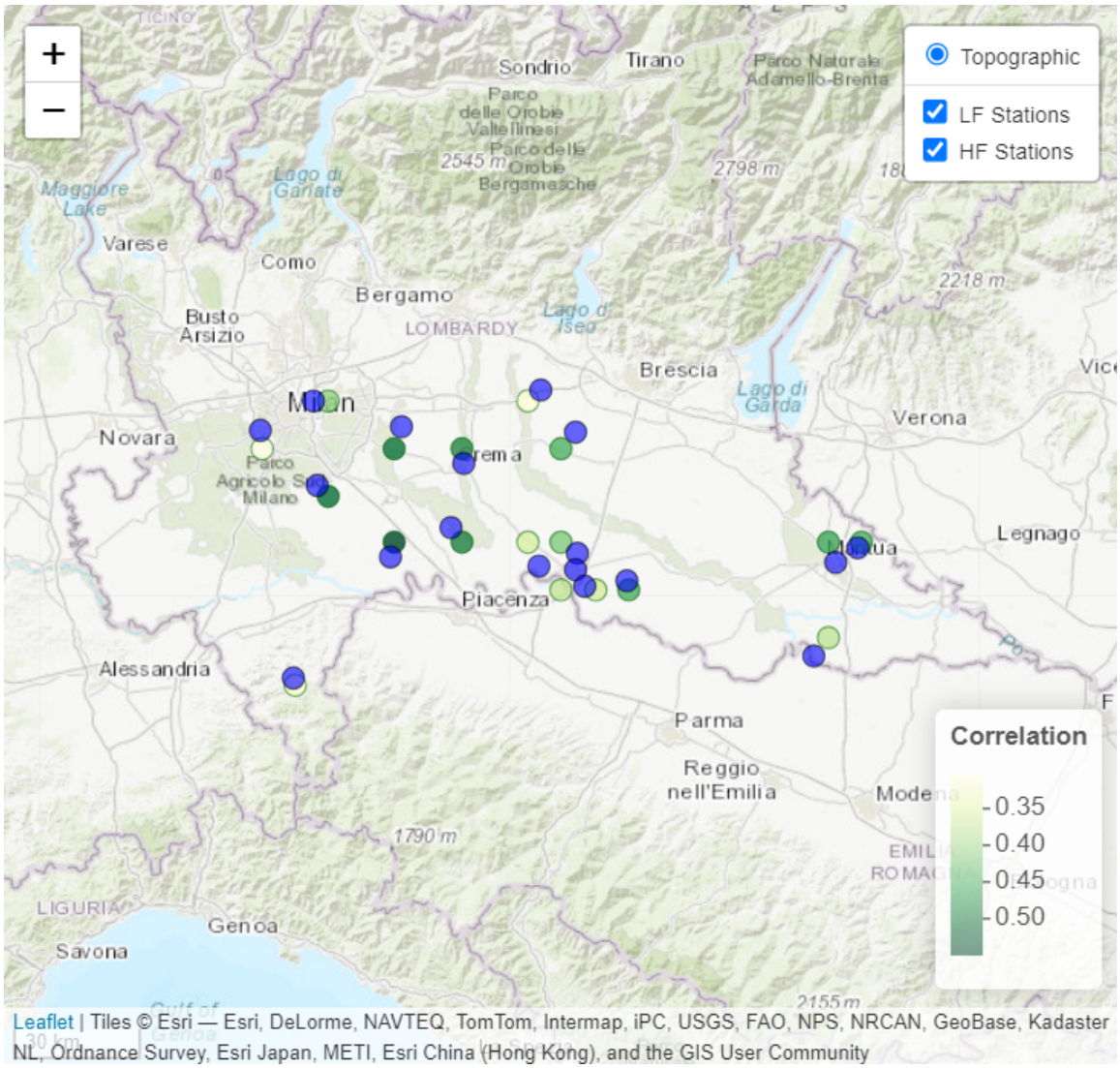


Figure 1.3: The figure shows a snapshot of the South Lombardy dataset. ARPA stations (HF) are marked in blue, while the centres of ERA5 grid cells are shown in green. The linear correlation coefficients between each station and its nearest grid cell centre are also provided.

records in time and spaces, each representing an hourly wind speed measurement. This merging process allows for a comprehensive analysis that combines the advantages of both high and low-fidelity data sources.

The merged dataset also enables the construction of a covariance matrix, which, in theory[1], spans 26784 rows, calculated as the sum of the low-fidelity data records (n_L) and the high-fidelity data records (n_H). The Figure 1.3 offers a snapshot of the selected stations with the corresponding reanalysis ERA5 grid cell. Notice that there is not a perfect alignment as the correlation coefficients spans from 0.3 to 0.55.

[1]During training some observations were removed.

1.6 Motivating data application – wind speed

Wind speed is used throughout this thesis as a motivating case study due to its relevance in environmental and energy applications, as well as its characteristic features such as skewness, strong temporal dynamics, and spatial heterogeneity. While the methods developed are motivated by this specific application, several of the challenges addressed—such as data sparsity, multi-source integration, and non-Gaussian behaviour—are common across a range of environmental variables. Nevertheless, the primary focus of this thesis remains on wind speed, and any broader applicability should be interpreted with this context in mind. First, wind speed data naturally come from sources with complementary characteristics, making them ideal for data fusion research. For example, wind speed can be measured in-situ using anemometers or Lidar (Light Detection and Ranging) technology, or through satellite-based retrievals. These measurement types differ in both resolution and reliability: satellite observations provide broad spatial coverage but lower accuracy, while in-situ measurements are highly accurate but geographically sparse.

Second, wind-speed data exhibit challenging statistical properties that complicate modelling. Skewness is common due to frequent gusts; time series often show weaker trends and less regularity than variables like temperature; and variability is strongly influenced by terrain morphology, with very different behaviour observed over sea, plains, or mountainous areas. These complexities make wind-speed modelling a demanding statistical problem.

Finally, wind speed data are of high relevance for both industry and public policy. Accurate wind-speed prediction is critical for wind farm resource assessment, guiding investment and planning. Moreover, wind speed is one of the most important predictors of air pollution dispersion (Otto et al., 2024). Understanding its spatio-temporal dynamics is thus crucial for highly polluted regions, such as Lombardy, one of the most polluted areas in Europe (Fassò et al., 2022; Otto et al., 2024).

1.7 Contribution

Multifidelity models are promising because they offer principled UQ and transparent assumptions. In Chapter 2, we systematically compare MF models against mono-fidelity baselines (HF-only, LF-only) and the classical AR1 MFGP across regimes of sample size, observational noise, data-generating processes, and multiple predictors—addressing **RQ1**. Results were presented at the International Workshop on Statistical Modelling (Dortmund, 2023) (Colombo et al., 2023).

Building on these foundations, Chapter 3 addresses skewness in wind-speed data—a fundamental obstacle for GP-based models (which assume independent Gaussian errors). Standard transformations often fail to normalise skewness while preserving cross-fidelity relations. We propose the *Warped Multifidelity Gaussian Process* (WMFGP), which predicts time series from heterogeneous sources while explicitly accounting for skewness and demonstrating partial independence from interpolation distance—addressing **RQ2**. Extensive simulations and an application to ARPA Lombardia show improved gap-filling and calibrated UQ for wind speed forecasting and network maintenance. A paper (Colombo et al., 2025) is published at *JRSS: Series C*.

Chapter 4 extends to the spatio-temporal domain, where high-resolution predictions are crucial. We introduce a framework for MFGPs that accommodates both stationary and non-stationary fidelity integration, combining broad but imprecise satellite reanalysis with precise but sparse in-situ measurements. To handle large-scale data, we integrate the Vecchia approximation for scalability and stability, and introduce a *spatially varying* integration parameter $\rho(x)$ for flexible fusion—addressing **RQ3**. Validation on synthetic and real data shows improved mean and variance fields versus standard GPs. A manuscript is under review at *Journal of Computational and Graphical Statistics*.

Finally, Chapter 5 concludes with a synthesis, methodological reflections, and future directions, including generalisation to other environmental variables.

1.8 Roadmap

This chapter has positioned multifidelity modelling within the broader data-fusion landscape and articulated the central challenges and research questions addressed in the thesis. The remainder of the thesis is organized as follows:

- **Chapter 2**: Background on competitive industry mono-fidelity methods and a comprehensive comparison with mono-fidelity alternatives across controlled regimes (addresses RQ1).
- **Chapter 3**: Literature review and introduction of the Warped MFGP for skewed data, including theoretical development, simulation studies, and application to the ARPA Lombardia dataset (addresses RQ2).
- **Chapter 4**: Background overview and development of scalable spatio-temporal MFGP models incorporating Vecchia approximations and spatially varying coupling, applied to ERA5 and ARPA data fusion (addresses RQ3).
- **Chapter 5**: Conclusions, limitations, and directions for future research.

For reference, Sections 1.3.1–1.3.6 provide the technical background on Gaussian processes and multifidelity modelling, while Section 1.5 describes the datasets (ERA5, ARPA Lombardia, and South Lombardy) that underpin the methodological developments presented in the subsequent chapters.

Chapter 2

A comparison between Univariate and Multivariate Methods

Chapter 1 introduced the aims and research questions of this thesis and positioned multi-fidelity modelling within the broader class of data fusion methods. This chapter presents a comparative analysis between a multi-fidelity approach, specifically the autoregressive MFGP, and two single-fidelity approaches: a standard Gaussian process and the Quantile Gradient Boosted Regression Tree (QGBRT). The primary objective is to identify the conditions under which multi-fidelity methods are most suitable for wind-speed applications.

The chapter opens with a case study where accurate wind-speed measurements are critical—the AGNES offshore project on Italy's Adriatic coast (Section 2.1)—and then describes the ERA5 dataset used in the analysis (Section 2.1.1). A concise methodological background follows, covering regression trees, boosting, and the QGBRT framework (Section 2.2). The experimental design is then introduced (Section 2.4.1), starting from a single-location baseline and subsequently adding a correlated covariate. Here, it is presented the evaluation metrics (Section 2.4.5) and the strategy for modelling the stochastic component via residual distribution selection (Section 2.4). The simulation results are presented next (Section 2.5), followed by a discussion of the study's limitations and their implications (Section 2.6). These limitations are then addressed through a revised experimental

design and additional analyses (Section 2.7). Finally, the chapter concludes with a synthesis of the main findings (Section 2.9). In doing so, this chapter directly addresses **Research Question 1 (RQ1)** introduced in Chapter 1, which investigates to which extent and under which conditions MFGP are are preferable to single fidelities models.

2.1 Motivational case study

The case study used to test and compare the proposed methods is based on the AGNES project, a multi-site initiative along the Adriatic coast (Ravenna, Emilia-Romagna, Italy) (see Figure 2.1), which involves the construction of offshore wind farms. Preliminary estimates suggest that offshore installations will account for approximately 83% of total AGNES energy production. Two farms are planned: *Romagna 1* and *Romagna 2*.

The planned work involves the construction of two wind farms of respectively 120 MW_e and 400 MW_e. The first wind farm takes the name of *Romagna*1, and it has the following characteristics:

- a farm extension of 52.92 km^2, approximately 18.5 km from the coast, between the cities of Punta Marina and Cervia.
- 15 wind turbines of 8 MW.
- The main layout is a single curved line formed by 15 wind turbines whose inter-distance is approximately 1.5 km.

The second wind farm takes the name of *Romagna*2, and it has the following characteristics:

- a farm extension of 197,5 km^2, with a minimum distance from the coast of 24 km, in front of the city of Marina Romea.
- 50 wind turbines of 8 MW.
- The layout is a cluster composed of 5 lines of 10 wind turbines, while the inter-distance between the different machines is 1.8 km.

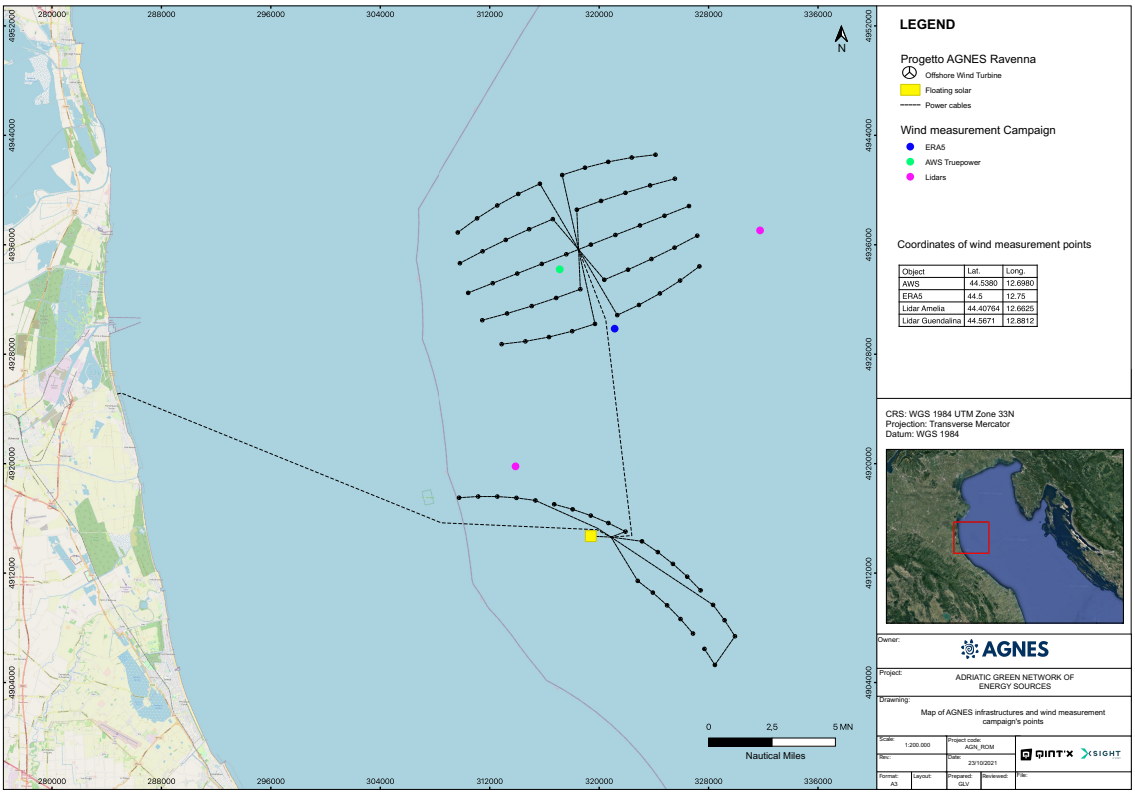


Figure 2.1: The figure represents the structure of the wind farms Romagna 1 and Romagna 2 in relation to the spatial location of the wind speed measurements: ERA5 (blue), AWS (green) and LIDAR (violet). This picture has been produced by Agnes srl, and it has been officially authorised for use in this report.

2.1.1 Data description: ERA5 at the AGNES site

Wind farm projects under approval are typically required to submit reproducibility analyses — long-term estimates of site profitability that combine reanalysis data with in-situ measurements. In this chapter, the ERA5 reanalysis dataset (see Section 1.5.1) at the AGNES wind farm site serves as the empirical foundation for subsequent experiments. 13 years of hourly wind speed measurements (2007–2020) at coordinates (44.5, 12.75) are analysed, focusing on their temporal structure and distributional properties, and discuss how these characteristics guide our later modelling choices. This time span provides a balance between statistical robustness and computational tractability when working with high-resolution (hourly) data. Figure 2.2 illustrates the interannual distribution of wind speeds. The median wind speed is approximately 4.5 m/s, with occasional observations exceeding 15 m/s. A gentle multi-year cycle is visible, particularly between 2007

and 2011, accompanied by a gradual increase in the overall mean wind speed. Figure 2.3 depicts the seasonal and trend decomposition (Cleveland et al., 1990) of the long time series. Such a decomposition does not reveal a strong or consistent long-term trend, it suggests mild multi-year fluctuations and a clear six-month seasonal pattern. Wind speeds tend to be lower during summer and winter, and higher in spring and autumn, reflecting well-known regional circulation patterns. These regular components contribute to the predictability of the system, whereas a substantial portion of the variance remains in the irregular component. This residual variation presents a key modelling challenge, as errors in representing it can significantly affect forecast accuracy.

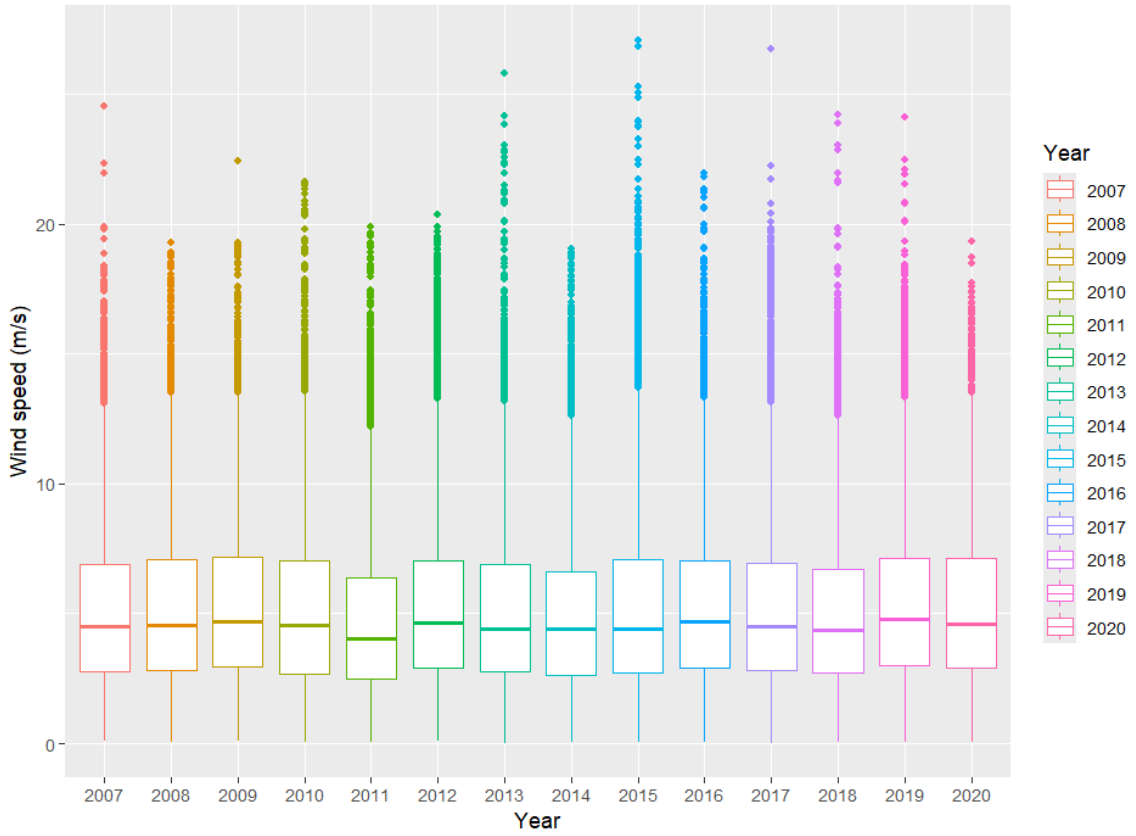


Figure 2.2: Boxplot of inter-annual wind speed distributions (2007–2020) for the ERA5 dataset at coordinates (44.5, 12.75).

To enable broader comparisons and avoid overly restrictive assumptions, the subsequent analysis does not impose strict stationarity. Instead, it allows for potential changes in the underlying distributional properties across years, as illustrated more clearly in the inter-annual distributions (Figure 2.2) and histograms (Figure 2.4).

The distributional structure of the data further complicates forecasting. Figure

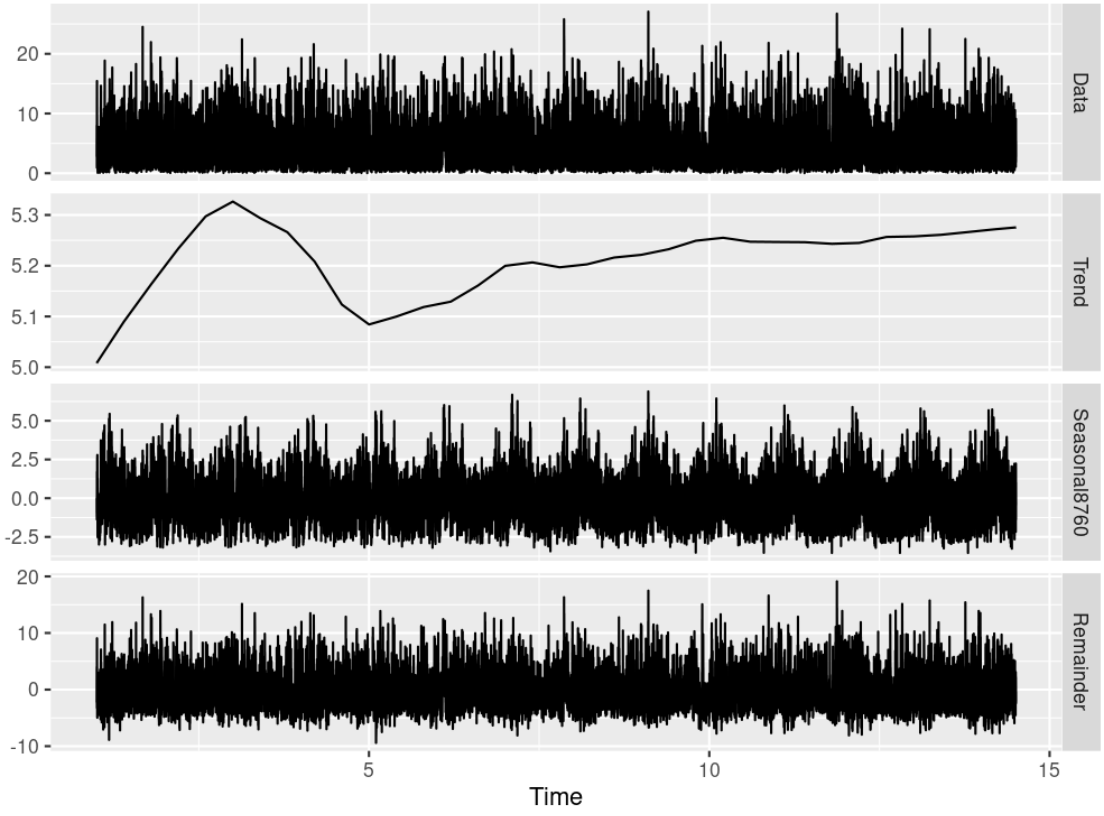


Figure 2.3: Seasonal and trend decomposition of 13 years of ERA5 wind speed data at coordinates (44.5, 12.75). The decomposition reveals long-term trend, seasonal variation, and residual irregularities.

2.4 presents annual histograms for the period 2007–2012, showing consistently positively skewed distributions with varying levels of kurtosis. The data also display substantial variability, with a mean standard deviation of approximately 3.3. Such features motivate the consideration of transformations or flexible probabilistic models that can accommodate skewness and heavy-tailed behaviour.

Monthly statistics (Figure 2.5) reinforce the presence of seasonal dynamics, with peaks in wind speed during March and November, and troughs during summer and winter. Variability itself appears to follow a seasonal pattern, being higher in periods of stronger winds.

In summary, the ERA5 dataset at the AGNES site exhibits an increasing mild long-term trend, a gentle 3–4-year fluctuation, clear six-month seasonality, and highly skewed distributions with considerable unexplained variability. These findings are not only descriptive but also methodological: they shape the simulation experi-

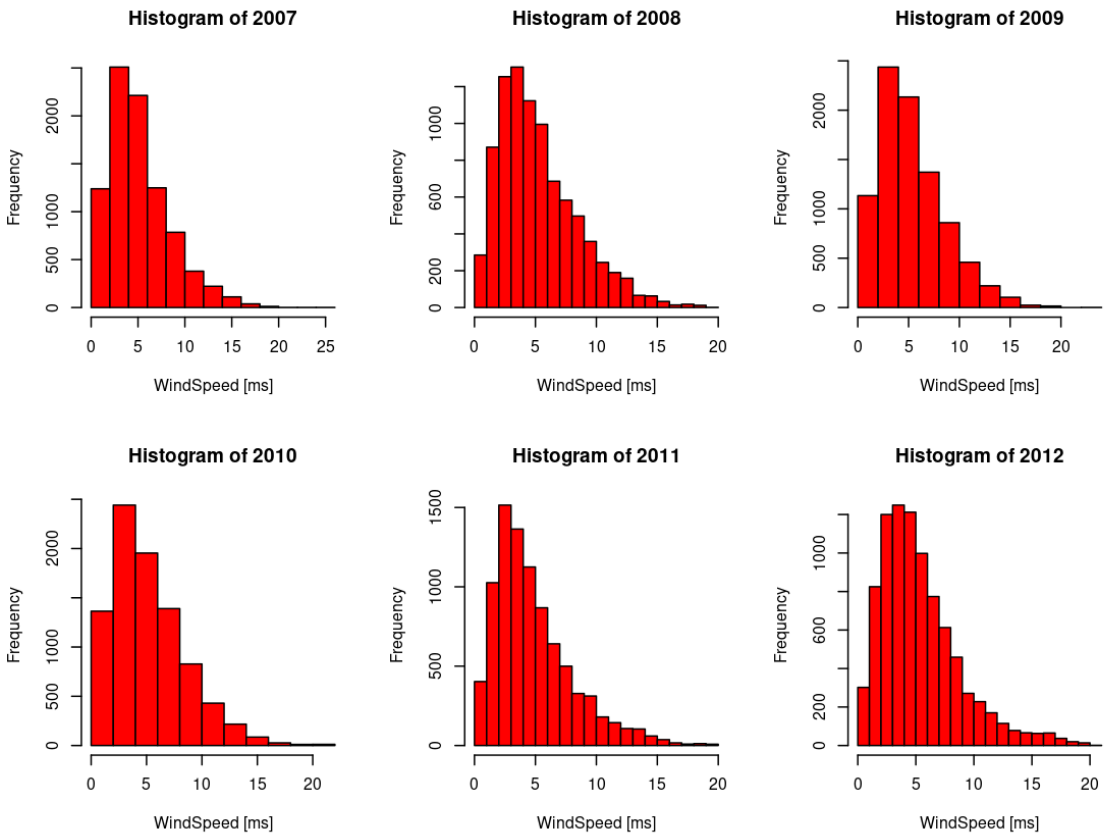


Figure 2.4: Histograms of hourly ERA5 wind speed measurements for the years 2007–2012 at coordinates (44.5, 12.75).

ments in Chapter 2, motivate the inclusion of seasonal and trend components in model design, and justify the use of data transformations to handle skewness and heavy tails, a topic at the core of Chapter 3. The substantial residual component underscores the complexity involved in modelling this dataset. The implications of these large residuals are discussed later in Section 2.6.

2.2 Methodological background

This section provides the background necessary to understand the QGBRT approach. It includes a brief overview of regression trees (Section 2.2.1), an explanation of the boosting mechanism (Section 2.2.2), and a description of the experiment conducted in this chapter.

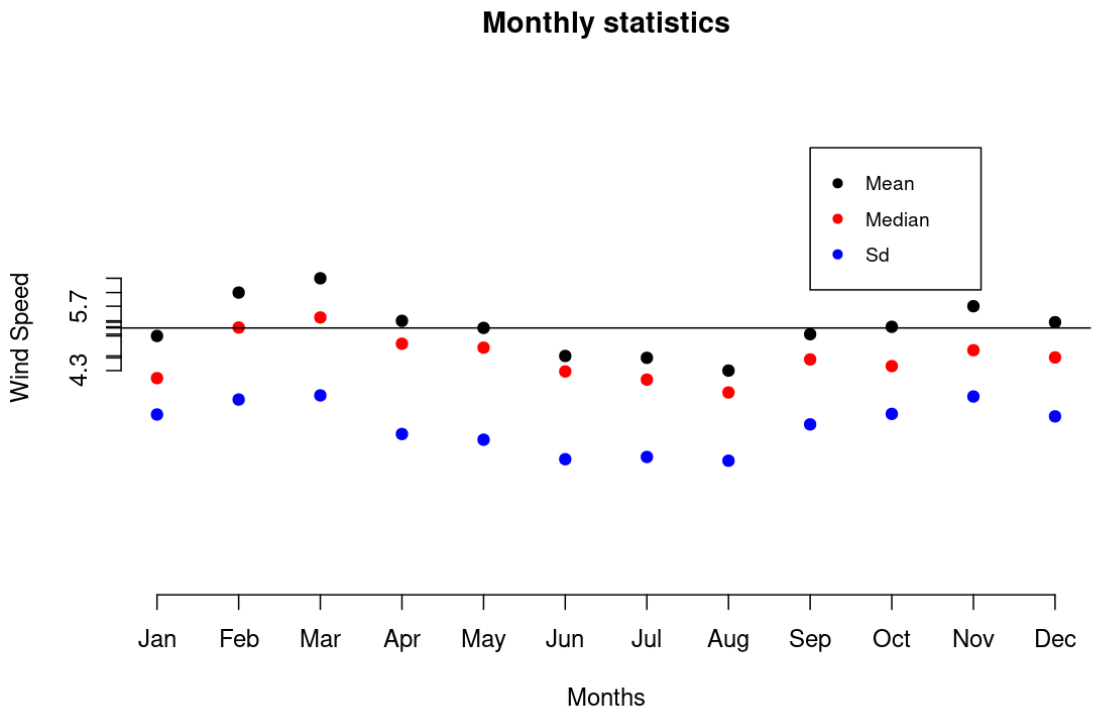


Figure 2.5: Monthly averages of ERA5 wind speed (2007–2020) at coordinates (44.5, 12.75). Stronger winds are observed in spring and autumn.

2.2.1 Regression trees

The QGBRT (Delcroix et al., 2021) is a boosting-based method constructed from sequential regression trees. Before introducing the full model, we briefly review regression trees themselves, as they form the core component from which QGBRT is assembled. Their core idea is to partition the input space into distinct regions, within which predictions are made as the average value of the observations in that region. These regions are known as *leaves* or *terminal nodes*. Consequently, all observations that fall within the same region receive the same predicted value. In theory, the regions can take on many possible shapes, leading to an extremely large number of potential partitions. The objective is to find the partition that minimises the residual sum of squares (RSS):

$$\min RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2, \tag{2.1}$$

where J is the number of regions, R_j denotes the j-th region, and $\hat{y}_{R_j}$ is the predicted value (mean response) for region R_j. Since there are many possible splits, the mini-

mization of the RSS is made by fixing the number of splits and iterating the process multiple times. Supposed we fix the number of splits to 2. At the first steps we find the two regions that give the minimal RSS. In the second stage, we iterate the process within the two found regions. This process is called *recursive-binary-spliting*, since it is based on a series of RSS minimizations. This recursive binary splitting approach can yield an excellent fit to the data, though it may occasionally overfit. To avoid overfitting regression trees are usually equipped with another stage called *pruning*, which increases the bias to reduce the variance. First, the best trees are fitted and then some "leafs" are "pruned". The tree is pruned maximising some performance metrics in the test set, using a cross-validation approach. In practice, the data are split, the best tree is fitted on the training dataset, and a subtree is selected using the validation set. A non-negative tuning parameter γ is introduced to control the complexity of the tree through a cost-complexity pruning criterion. Instead of minimising only the residual sum of squares, the objective function is modified as

$$RSS(T) + \gamma|T|, \tag{2.2}$$

where T denotes the tree and $|T|$ the number of terminal nodes. The parameter γ penalises larger trees, thereby controlling the trade-off between goodness of fit and model complexity.

A natural question is why regression trees typically employ piecewise constant predictions rather than more flexible local models such as linear regressions within each region. The main reason lies in the trade-off between flexibility and stability. While fitting linear models within each region could, in principle, capture local trends more accurately, it would also substantially increase the number of parameters to estimate, especially in small regions containing few observations. This can lead to high variance and unstable predictions. In contrast, piecewise constant models provide a more robust and parsimonious representation, as each region is summarised by a single parameter (the mean response), which can be reliably estimated even with limited data. Furthermore, the use of constant predictions simplifies the recursive splitting procedure, since the optimal split can be determined directly by minimis-

ing the residual sum of squares without requiring the estimation of additional local model parameters. More flexible variants, such as model trees, do exist, but the standard regression tree formulation favours piecewise constant approximations for their computational efficiency, interpretability, and strong empirical performance.

2.2.2 Boosting

Boosting generally refers to the concept of combining multiple weak learners,[1] where each model is trained to correct the errors made by its predecessors. In this thesis, boosting is used as a method to enhance the predictive performance of regression trees. Multiple trees are grown sequentially, with each tree exploiting the information provided by the one built in the previous stage. In particular, each tree is fitted to a modified version of the original dataset. The main idea is to learn slowly what is the best prediction tree. First, a tree is fitted, then some residuals are computed, and new regression tree is fitted on the residuals. This new tree is added to the regression function to update the residuals with some penalty term λ which controls the speed of the prediction improvement. In the next section, a full description of a boosting algorithm for regression trees is provided.

2.2.3 QGBRT: Quantile Gradient Boosted Regression Trees

QGBRT is an ensemble method combining regression trees with a quantile loss function. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, the objective is to estimate a function $\hat{f}(\mathbf{x})$ by minimizing the empirical quantile loss:

$$\Phi_\alpha(y, f(\mathbf{x})) = \sum_{i=1}^{N} w_i \, \rho_\alpha\big(y_i - f(\mathbf{x}_i)\big), \tag{2.3}$$

where the quantile loss function is defined as

$$\rho_\alpha(u) = u \left(\alpha - \mathbb{I}(u < 0)\right), \tag{2.4}$$

with $\alpha \in (0, 1)$ the target quantile and w_i fixed observation weights.

[1] A model that performs only slightly better than random guessing on a given classification or regression task.

Initialisation The model is initialised as a constant:

$$\hat{f}_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{N} \Phi_\alpha(y_i, \rho), \tag{2.5}$$

which corresponds to the weighted empirical α-quantile.

Boosting iterations For $t = 1, \ldots, T$:

1. Compute the negative gradient:

$$z_{ti} = \alpha - \mathbb{I}(y_i < \hat{f}_{t-1}(\mathbf{x}_i)). \tag{2.6}$$

   The quantity z_{ti} represents the negative gradient of the loss function with respect to the current model predictions, evaluated at iteration $t-1$. In the context of gradient boosting, these values are interpreted as pseudo-residuals. Rather than updating the model parameters directly using the gradient, a regression tree is fitted to the dataset $\{(\mathbf{x}_i, z_{ti})\}$ in order to approximate the direction of steepest descent in function space. The fitted tree $g_t(\mathbf{x})$ therefore provides a functional approximation to the negative gradient, which is subsequently used to update the model in an additive manner. In practise, the gradient represents the direction in which predictions should be adjusted, and the tree approximates this correction across regions of the input space, allowing the model to iteratively refine its predictions.

2. Fit a regression tree $g_t(\mathbf{x})$ to $\{(\mathbf{x}_i, z_{ti})\}$, partitioning the input space into regions $S_{kt},\ k = 1, \ldots, K_t$.

3. For each terminal node, compute:

$$\rho_{kt} = \arg\min_{\rho} \sum_{\mathbf{x}_i \in S_{kt}} \Phi_\alpha\big(y_i, \hat{f}_{t-1}(\mathbf{x}_i) + \rho\big). \tag{2.7}$$

4. Update the model:

$$\hat{f}_t(\mathbf{x}) = \hat{f}_{t-1}(\mathbf{x}) + \lambda \sum_{k=1}^{K_t} \rho_{kt}\, \mathbb{I}(\mathbf{x} \in S_{kt}), \tag{2.8}$$

where $\lambda \in (0, 1]$ is the learning rate.

2.3 Experimental design

2.3.1 Overview

The objective of this chapter is to perform a realistic comparison among the MFGP, QGBRT, and GP models. The selection of MFGP's competitors is deliberate: the GP represents its natural *mono-fidelity* counterpart, while the QGBRT is a widely adopted industrial method for wind speed prediction (Nagy et al., 2016).

To achieve this goal, a series of experiments were designed to evaluate the influence of fidelity level, sample size, and covariate dimensionality on model performance (see Section 2.4.5). Each experimental setup isolates one of these factors, enabling a clear assessment of its impact on predictive accuracy.

The experiments are based on reanalysis data from the AGNES wind farm project site (see Section 1.5.1). Although a high-fidelity measurement campaign using LIDAR systems was conducted at the site during 2022–2023, access to these data is restricted under a non-disclosure agreement (NDA) and therefore they cannot be presented in this thesis. Nevertheless, access to the data provided valuable insights—such as the average discrepancy between sources, its skewness, and the correlation between local reanalysis and lidar data. These findings are instrumental in designing realistic simulation scenarios that rely exclusively on publicly available reanalysis datasets. In particular, the LIDAR records revealed extended periods of missing observations, which motivated the inclusion of long missing sequences in the synthetic datasets to better reflect the real conditions at the AGNES site.

The following sections describe the simulation procedure (Section 2.4) and detail the various dimensions of the experiment (Sections 2.4.1, 2.4.3, and 2.4.2). The results for each setup (2.4.4) are then presented: the time-only configuration in Section 2.5.1, and the multidimensional configuration in Section 2.5.2.

2.4 Data decomposition and residual distribution modelling

To simulate high- and low-fidelity data resembling the wind speed at the site of interest (AGNES), a procedure based on time-series decomposition and stochastic reconstruction was employed. The underlying assumption is that wind speed observations can be represented as follows:

$$\mathbf{y} = f(\mathbf{X}) + \epsilon, \tag{2.9}$$

where $\mathbf{y}$ denotes the observed wind speed, $\mathbf{X} = (x_{it})_{i=1,\dots,p}$ represents the p-dimensional vector of exogenous regressors at time t, and ϵ corresponds to the unpredictable component. In this framework, a high-fidelity measurement is characterised by a smaller error term ϵ_H, whereas a low-fidelity measurement includes a larger error component, denoted by $\epsilon_L > \epsilon_H$. The simulation process therefore consists of two main stages: first, extracting the predictable component $f(\mathbf{X})$, and second, generating synthetic realisations of the unpredictable part ϵ. This approach allows for the reconstruction of realistic wind speed time series whose fidelity can be controlled by adjusting the level of stochastic variability.

The extraction of the predictable component was carried out using the classical Seasonal and Trend decomposition based on Loess (STL) method Cleveland et al. (1990).

The components $s(t)$, $T(t)$, and $r(t)$ are obtained using the STL (Seasonal-Trend decomposition using Loess) algorithm of Cleveland et al. (1990), as implemented in the `stl` function described in Hyndman and Athanasopoulos (2018). The method relies on an iterative sequence of locally weighted polynomial regressions (Loess smoothers) to separately estimate the seasonal and trend components.

Given the hourly resolution of the data, a seasonal structure is specified, and the seasonal component $s(t)$ is estimated by applying a Loess smoother to subseries corresponding to each seasonal cycle. In this study, a seasonal smoothing window of six months was adopted to capture medium-term periodic behaviour while allowing for gradual seasonal variation over time.

The trend component $T(t)$ is then obtained by applying a second Loess smoother to the seasonally adjusted series $y(t) - s(t)$, using a larger smoothing window to capture long-term dynamics. The remainder is subsequently computed as $r(t) = y(t) - s(t) - T(t)$.

The STL algorithm is iterated until convergence, and a robustness step is included, whereby observations are reweighted based on the magnitude of the residuals in order to reduce the influence of outliers.

This algorithm is particularly well suited for wind data because it applies iterative local smoothing operations that robustly separate the seasonal and long-term trend components while handling outliers effectively. Let $\{y(t) \in \mathbb{R}^+ : t = 1, \ldots, n\}$ represent the observed hourly wind speed series. The STL[1] procedure decomposes this series into three additive parts:

$$y(t) = s(t) + T(t) + r(t), \tag{2.10}$$

where $s(t)$ is the seasonal component, $T(t)$ represents the trend, and $r(t)$ is the remainder. The combination of $s(t)$ and $T(t)$ defines the structural, or predictable, component of wind speed, while $r(t)$ corresponds to the unpredictable term ϵ. The unpredictable component plays a central role in this analysis, as it governs the variability used to simulate different fidelity levels.

More precisely, the method relies on an iterative sequence of locally weighted polynomial regressions (Loess smoothers) to separately estimate the seasonal and trend components.

Given the hourly resolution of the data, a seasonal structure is specified, and the seasonal component $s(t)$ is estimated by applying a Loess smoother to subseries corresponding to each seasonal cycle. In this study, a seasonal smoothing window of six months was adopted to capture medium-term periodic behaviour while allowing for gradual seasonal variation over time.

The trend component $T(t)$ is then obtained by applying a second Loess smoother to the seasonally adjusted series $y(t) - s(t)$, using a larger smoothing window to

[1]The procedure is referenced in Chapter 1

capture long-term dynamics. The remainder is subsequently computed as $r(t) = y(t) - s(t) - T(t)$.

The STL algorithm is iterated until convergence, and a robustness step is included, whereby observations are reweighted based on the magnitude of the residuals in order to reduce the influence of outliers.

To reproduce realistic fluctuations in wind speed, it is crucial to identify an appropriate probability distribution for the remainder $r(t)$. For this purpose, the residuals[1] coming from the ERA5 hourly wind speed record from the grid point nearest to the wind farm, covering the period from 1979 to 2020 and consisting of approximately 365,000 observations, were selected. The empirical behaviour of the residuals was then analyzed to determine a suitable stochastic model. The initial exploration employed the Cullen and Frey graph (see Figure 2.6), which displays the empirical relationship between skewness and kurtosis together with the theoretical values corresponding to standard probability distributions. This graphical analysis provides a preliminary indication of which families of distributions may best capture the observed behaviour of $r(t)$. In this case, the residuals exhibited moderate kurtosis and relatively high skewness, consistent with either a Lognormal or a Gamma distribution. Given the established use of the Weibull distribution in wind modelling Pobočíková et al. (2017), it was also retained for comparison.

Following this exploratory analysis, the parameters of the candidate distributions were estimated through maximum likelihood estimation (MLE). For distributions involving multiple parameters, the Nelder–Mead algorithm was adopted because of its robustness to non-smooth likelihood surfaces, while the BFGS method was used for single-parameter cases where gradient information could be efficiently exploited. Once parameter estimates were obtained, the theoretical probability density functions, cumulative distributions, and quantiles were plotted against their empirical counterparts to visually assess the goodness of fit. An example of this comparison is presented in Figure 2.7, where the residual histogram is overlaid with the fitted Weibull, Gamma, and Lognormal distributions. Because the residual series contained a few negative values, a constant shift equal to the minimum of $r(t)$ was applied to

[1]Residuals obtained with the STL decomposition.

ensure all fitted values were positive; this adjustment can easily be reversed after simulation to preserve the physical interpretation of the data.

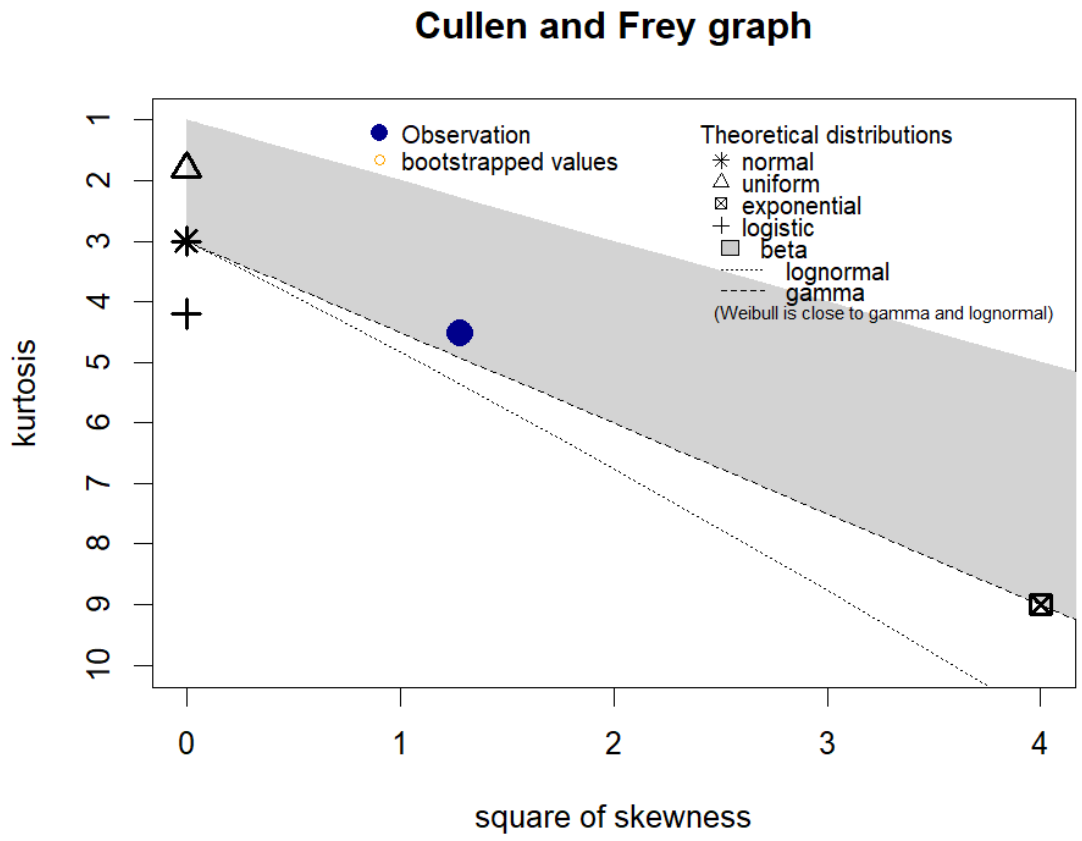


Figure 2.6: Cullen and Frey graph of the $r(t)$ remainder from the STL decomposition shown in Figure 2.3.

After the visual inspection, the fitted models were further evaluated using formal statistical tests and numerical measures of fit, including the Kolmogorov–Smirnov (KS) test, the root mean square error of the cumulative distribution ($RMSE_D$), and the coefficient of determination (R^2). The results are summarised in Table 2.1. Overall, no single distribution demonstrated a clearly superior fit across all metrics. The Gamma distribution exhibited slightly better alignment with the empirical density, whereas the Lognormal and Weibull distributions tended to over- or under-represent certain intervals of the data. To explore potential improvements, transformations based on the Box–Cox approach were also tested; however, the additional complexity introduced by such transformations offered limited gains in goodness of fit.

Despite the comparable performance of the candidate distributions, the Skew-Normal distribution was selected as a parsimonious model for the residual component. Unlike

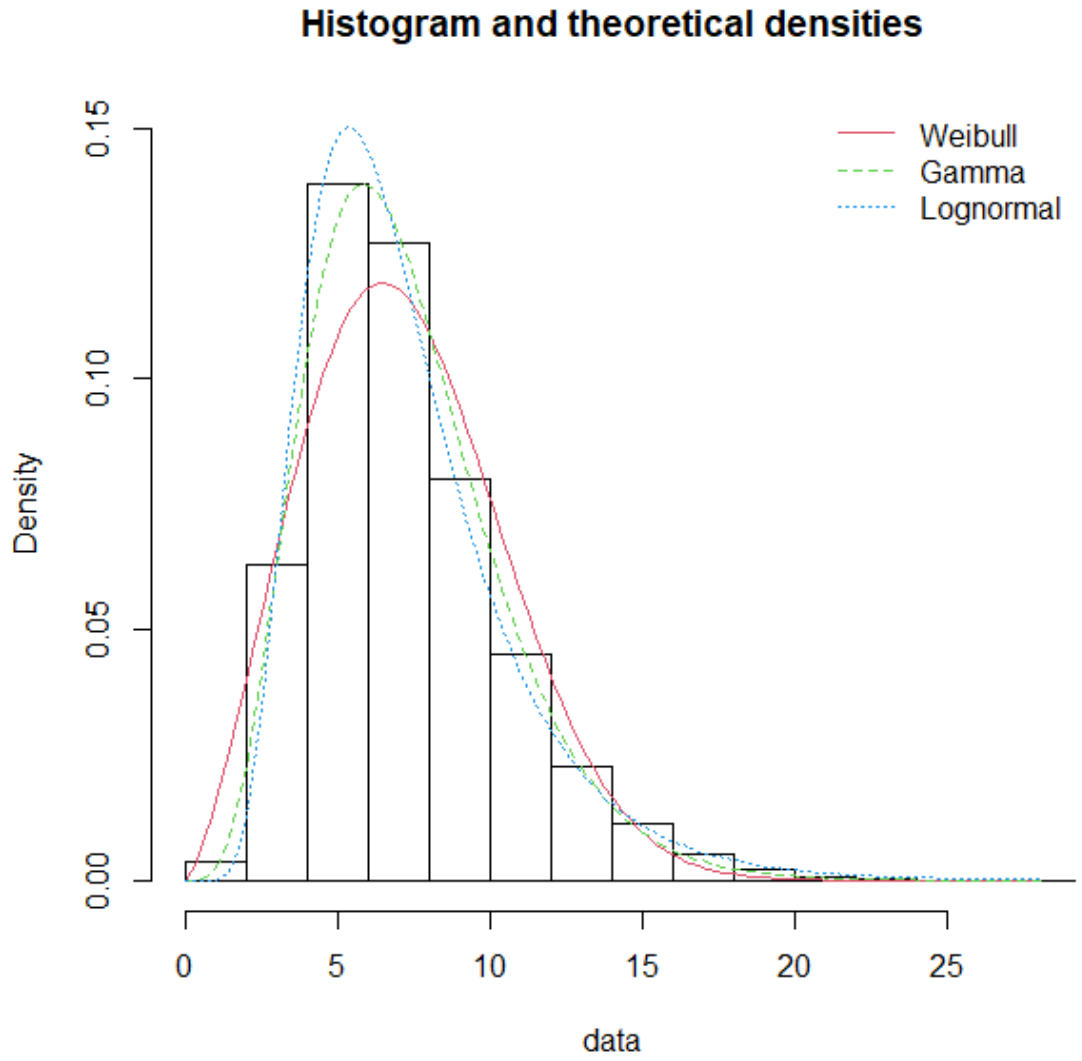


Figure 2.7: Histogram of $r(t)$ with fitted Weibull, Gamma, and Lognormal densities. The figure reveals the best fitting density for the $r(t)$ remainder.

Table 2.1: Goodness-of-fit results for candidate distributions of $r(t)$.

Distribution	R^2	**KS**	$RMSE_D$
Skew-Normal	0.9974	0.531	$7.72 \times e^{-8}$
Box–Cox	0.9965	0.540	$4.14 \times e^{-8}$
Weibull	0.9799	0.427	$3.47 \times e^{-8}$
Gamma	0.9970	0.520	$1.97 \times e^{-7}$
Lognormal	0.9892	0.536	$1.12 \times e^{-7}$

alternatives such as the Box–Cox transformation, the Skew-Normal formulation does not require any transformation of the data, thereby preserving the natural structure of the observed series. In terms of goodness-of-fit, the Skew-Normal distribution performs among the best candidates according to both the R^2 and $RMSE_D$ metrics, further supporting its suitability. However, it should be noted that the differences in

performance across the competing distributions are relatively small. As a result, the final choice is guided not only by numerical criteria, but also by considerations of interpretability and modelling simplicity. Consequently, the Skew-Normal distribution was selected to represent the stochastic behaviour of the unpredictable component $r(t)$.

Once the distributional form was determined, the synthetic wind speed series were reconstructed by combining the deterministic and stochastic components as follows:

$$y_{\text{sim}}(t) = s(t) + T(t) + r_{\text{SN}}(t), \tag{2.11}$$

where $r_{\text{SN}}(t)$ represents a random realisation drawn from a Skew-Normal distribution with adjustable variance. By modifying this variance, it becomes possible to generate high- and low-fidelity series that replicate the statistical characteristics of wind speeds at the site, while maintaining realistic temporal dynamics inherited from the ERA5 data.

2.4.1 Simulation experiment attributes

Exploring multiple dimensions of an experiment allows for a more comprehensive and nuanced understanding of the phenomenon under study. It also prevents the analysis from being constrained to a single, potentially unrepresentative scenario. In this simulation framework, several design factors were systematically varied to capture the main sources of variability that may affect model performance. Specifically, the experimental design differs along four dimensions.

First, the sample-size ratio between high- and low-fidelity data, defined as n_H/n_L, takes values in $\{0.03, 0.11, 0.18, 0.30, 0.48, 0.79\}$, with a fixed total sample size of $n = 850$. These ratios are chosen to span a broad range of practically relevant scenarios—from extremely limited high-fidelity availability (approximately 3% of the total data) to near-parity between fidelities. The lower end of this range is particularly informative, as it reflects the common situation in wind-energy applications where high-quality measurements are scarce and expensive, while low-fidelity reanalysis data are abundant. Testing this lower boundary allows us to examine the robust-

ness of the model when extrapolating from limited accurate information. Conversely, higher ratios are included primarily to verify the expected asymptotic behaviour as high-fidelity information becomes dominant. Second, the design includes a factor controlling the presence or absence of an additional correlated predictor P, which has a target correlation of approximately 0.70 with the response variable. This setting represents a moderately strong but not deterministic relationship, allowing us to assess how an informative auxiliary feature influences the model's capacity to exploit multi-fidelity structure. Third, two distinct noise regimes are considered—high and low—each characterised by different noise intensities for the high- and low-fidelity data, thus allowing an explicit evaluation of the models' robustness to measurement uncertainty. Finally, a *replication index* j is introduced to account for stochastic variability, with 100 independent Monte Carlo replications performed for each configuration.

A summary of the experimental design factors is provided in Table 2.2.

Table 2.2: Design factors considered in the simulation study.

Factor	**Description**
Sample-size ratio	$n_H/n_L \in \{0.03, 0.11, 0.18, 0.30, 0.48, 0.79\}$, with total sample size $n = 850$.
Correlated predictor	Presence or absence of an additional predictor P correlated with the target variable (target correlation $\approx$ 0.70).
Noise regime	Two noise levels (high and low), with distinct high-frequency (HF) and low-frequency (LF) noise settings.
Replication index	Monte Carlo replication index j, with 100 independent simulation runs per configuration.

A general $MFGP$ model for every possible condition would be :

$$\hat{y}_{MFGP} = MFGP(y_H(t)|D_H, D_L, n_H/n_L, j, \epsilon, GM, \boldsymbol{X}), \tag{2.12}$$

where D_H is an incomplete dataset of size $n_H < n_L$ of highly fidelity measurements of wind speeds, D_L, is a complete dataset of low fidelity measurements of wind speeds of size n_L, j is the design replication, ϵ is the noise level of the D_L low fidelity measurements, GM refers to the random generation mechanism of the noise level ϵ, while $\boldsymbol{X}$ is a matrix containing the models' predictors. For the quantile QGBRT

instead we have a different set of conditions:

$$\hat{y}_{QGBRT} = QGBRT(y_H(t)|D_L \vee D_H, j, \epsilon, GM, \boldsymbol{X}). \tag{2.13}$$

More precisely D_L and D_H cannot be used jointly.

2.4.2 Predictor construction with controlled correlation

The predictor matrix $\boldsymbol{X}$ constitutes another potential source of design variability. Two formulations were constructed for this matrix. The first formulation involves only time as a predictor, resulting in a matrix of dimension $n \times 1$. The second formulation has dimension $n \times 2$ and includes a pseudo-predictor correlated with $y_H(t)$. In the context of the wind case study, this pseudo-predictor may represent variables such as air density, temperature, or wave height, among others. The pseudo-predictor was simulated using a linear algebra procedure represented by the following equation:

$$\begin{aligned} \mathbf{L} &= chol(\mathbf{R}) \\ \mathbf{A} &= \mathbf{M} \times \mathbf{L}, \end{aligned} \tag{2.14}$$

Here, $\mathbf{M}$ denotes an $n \times 2$ matrix with the first column representing $\mathbf{y}_H$, and the second column representing a normally distributed vector of size n. This second column can represent any vector. $\mathbf{R}$ is a 2×2 design correlation matrix, wherein the extra-diagonal terms denote the desired design correlation coefficient (0.70 in this case), and the diagonal elements are equal to 1. The Cholesky factor, denoted by $\mathbf{L}$, is derived from the $\mathbf{R}$ matrix. The resulting matrix $\mathbf{A}$ is of dimension $n \times 2$ and contains the $\mathbf{y}_H$ vector of high-fidelity measurements in the first column, and the desired generic vector, correlated with $\mathbf{y}_H$ with an intensity close to the design correlation coefficient of 0.70, in the second column.

2.4.3 Noise regimes and replication

The analysis considers two noise regimes. Under the *high-noise* configuration, the LF data contribute roughly 60% of the total variance, whereas the HF data are comparatively less noisy. Under the *low-noise* configuration, both data sources display lower variance, though LF remains the noisier of the two. Skew-normal parameters (ξ, ω, α) are tuned to reproduce empirical skewness and kurtosis of the STL remainder. Such a generation completely depends on the random part $r(t)$. It is essential to select a distribution by which it is possible to generalise the results. Generating data from the correct distribution ensures that the simulated or experimental results accurately reflect the behaviour of the system being studied, improving the reliability and validity of the findings. The skew-normal distribution is described below, while the rationale for its selection is provided in Section 2.4.

Suppose that X is a continuous random variable with probability density function

$$f(x) = 2\phi(x)\,\Phi(\alpha x),$$

where $\phi(x) = \frac{\exp(\frac{-x^2}{2})}{2\pi}$ is the standard Gaussian density function and $\Phi(\alpha x) = \int_{-\infty}^{\alpha x} \phi(x), dx$ is the Gaussian cumulative distribution function of $\phi(x)$ evaluated at α shape parameter. Such a random variable has interesting properties: the data generated by such a function are generally skew; if the α shape parameter is equal to zero, then we obtain the standard normal distribution; if α increases in absolute value then the skewness of the data increases; if $\alpha \to \infty$ the density converges to the so-called half-normal (or folded normal) density function; if the sign of α changes, the density is reflected on the opposite side of the vertical axis. The above random variable X is the basis to construct the so-called skew-normal using a linear transformation of the following type:

$$Y = \xi + \omega X, \tag{2.15}$$

where Y is the skew-normal random variable, ξ is the location parameter, and ω is the scale parameter. In general, the skew-normal distribution depends on a series

of parameters, i.e. $Y \sim SN(\xi, \omega, \alpha)$.

Its mean is equal to:

$$E(Y) = \xi + \omega\sqrt{\frac{2}{\pi}}\delta, \tag{2.16}$$

where $\delta = \alpha/\sqrt{1+\alpha^2}$, while the variance of Y is:

$$V(Y) = \omega^2 + (1 - \frac{2\delta^2}{\pi}). \tag{2.17}$$

More information about a skew normal distributions can be found in Genton (2004).

2.4.4 Parameters setup: high and low noise

The simulation experiment incorporates two noise levels, two definitions of the predictor matrix $\boldsymbol{X}$, and three distinct time series (High-fidelity, Low-Fidelity, true) [1] The objective is to evaluate model performance under varying data quality and dimensionality conditions.

The first dataset represents the *true* signal, defined as the sum of its seasonal and trend components:

$$y_{\mathbb{T}}(t) = s(t) + T(t).$$

Here, the subscript $\mathbb{T}$ denotes the true underlying process, distinguishing $y_{\mathbb{T}}(t)$ from its noisy observations $y(t)$.

The second dataset corresponds to the high-fidelity (HF) data source and is generated as

$$y_{\text{HF}}(t) = y_{\mathbb{T}}(t) + r(t),$$

where $r(t)$ follows a Skew-Normal distribution with parameters $\xi = -1$, $\omega = 0.3$, and $\alpha = 5.2$. Under this specification, the noise term has variance approximately 0.0347 and skewness approximately 0.86, so that the HF observations remain close to the true signal in signal-to-noise terms.

The third dataset represents the low-fidelity (LF) data source. In this case, the

[1] In practical application the difference between HF and true is usually neglible, and the HF are usually assumed to represent the real target.

noise term follows a Skew-Normal distribution with parameters $\xi = -3.9$, $\omega = 3.5$, and $\alpha = 5.2$. This yields variance approximately 4.73 and skewness approximately 0.86. Therefore, the LF noise variance is about 136 times larger than that of the HF noise, resulting in a substantially lower signal-to-noise ratio.

Each experiment consists of 100 independent replications using hourly data generation. Every replication differs in the random noise realisation to ensure robustness of the results. Two different configurations of the predictor matrix $\boldsymbol{X}$ were considered:

1. **Single-predictor setup:** In this configuration, the matrix $\boldsymbol{X}$ contains only one predictor, time. The total sample size is $n = 850$ observations for the high-fidelity (HF) data and $n_L = 32$ for the low-fidelity (LF) data. The reduced sample size n_L represents a sparse low-fidelity observation regime, simulating conditions where only a limited number of low-cost or coarse measurements are available.

2. **Multidimensional experiment:** A second pseudo-predictor, denoted as P, was added to $\boldsymbol{X}$. This additional variable is designed to have a moderately high correlation (approximately 70%) with the target output. Such a setup introduces mild multicollinearity and a more complex dependency structure, providing a richer context for assessing model performance in multidimensional scenarios.

Three modelling approaches were compared: GP, MFGP, and QGBRT. Each method was tested under three distinct *information scenarios*[1]: (i) a mono-fidelity dataset of low quality (LF only), (ii) a mono-fidelity dataset of high quality but limited sample size (HF only), and (iii) a mixed-fidelity dataset combining both sources. This setup resulted in a total of five models: GP and QGBRT were each tested in mono-fidelity mode (yielding four models), while the fifth model—the MFGP—integrated both datasets to exploit the complementary strengths of high- and low-fidelity information. More precisely:

[1]By information scenario it is meant what type of data are fed into the model.

Model formulations

Let $\mathcal{D}_L = \{(\mathbf{x}_L^{(i)}, y_L^{(i)})\}_{i=1}^{n_L}$ denote the low-fidelity dataset and $\mathcal{D}_H = \{(\mathbf{x}_H^{(i)}, y_H^{(i)})\}_{i=1}^{n_H}$ the high-fidelity dataset.

1. $\mathbf{GP}_{LF}$

$$y_L(\mathbf{x}_L) = f_L(\mathbf{x}_L) + \varepsilon_L, \quad \varepsilon_L \sim \mathcal{N}(0, \sigma_L^2), \tag{2.18}$$

$$f_L(\mathbf{x}_L) \sim \mathcal{GP}(0,\ k(\mathbf{x}_L, \mathbf{x}_L'))\,, \tag{2.19}$$

with exponential covariance

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(- \sum_{j=1}^{d} \frac{|x_j - x_j'|}{\ell_j} \right). \tag{2.20}$$

2. $\mathbf{GP}_{HF}$

$$y_H(\mathbf{x}_H) = f_H(\mathbf{x}_H) + \varepsilon_H, \quad \varepsilon_H \sim \mathcal{N}(0, \sigma_H^2), \tag{2.21}$$

$$f_H(\mathbf{x}_H) \sim \mathcal{GP}(0,\ k(\mathbf{x}_H, \mathbf{x}_H'))\,. \tag{2.22}$$

$\mathbf{QGBRT}_{LF}$

$$\hat{y}_L(\mathbf{x}_L) = \hat{f}^{(L)}(\mathbf{x}_L), \quad \text{trained on } \mathcal{D}_L. \tag{2.23}$$

$\mathbf{QGBRT}_{HF}$

$$\hat{y}_H(\mathbf{x}_H) = \hat{f}^{(H)}(\mathbf{x}_H), \quad \text{trained on } \mathcal{D}_H. \tag{2.24}$$

5. MFGP (LF + HF) The multi-fidelity model follows an autoregressive structure:

$$f_H(\mathbf{x}_H) = \rho\, f_L(\mathbf{x}_H) + \delta(\mathbf{x}_H), \tag{2.25}$$

where

$$f_L(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \tag{2.26}$$

$$\delta(\mathbf{x}) \sim \mathcal{GP}(0, k_\delta(\mathbf{x}, \mathbf{x}')), \tag{2.27}$$

with $\delta(\mathbf{x})$ independent of $f_L(\mathbf{x})$.

Observations are given by

$$y_L(\mathbf{x}_L) = f_L(\mathbf{x}_L) + \varepsilon_L, \tag{2.28}$$

$$y_H(\mathbf{x}_H) = \rho\, f_L(\mathbf{x}_H) + \delta(\mathbf{x}_H) + \varepsilon_H, \tag{2.29}$$

and $k_\delta(\cdot, \cdot)$ is an exponential covariance function of the same form as $k(\cdot, \cdot)$.

2.4.5 Evaluation metrics

In this section, the metrics used to assess the simulation experiments are presented. For brevity, the true signal is denoted as $y_{\mathbb{T}}$, and the predictions from model m are represented as $\hat{y}_{mod}$. Performance metrics are computed using the residuals between these two quantities. Let $\{i \in \mathbb{Z} : i = 1, \ldots, n\}$ denote the index set of the input samples, where n is the total number of observations in the test set.

- **Mean Absolute Error (MAE):**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_{\mathbb{T},i} - \hat{y}_{mod,i}| \tag{2.30}$$

- **Root Mean Square Error (RMSE):**

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{\mathbb{T},i} - \hat{y}_{mod,i})^2} \tag{2.31}$$

- **Variance of Residuals:**

$$\text{Var}(res_{mod,m}) = \frac{1}{n} \sum_{i=1}^{n} (res_{mod,m,i} - \overline{res_{mod,i}})^2 \tag{2.32}$$

- **Bias:**

$$\text{Bias} = \overline{y}_{\mathbb{T}} - \overline{\hat{y}_{mod}} \tag{2.33}$$

While the evaluation framework adopted in this section focuses on point-wise error metrics (e.g., MAE, RMSE, and bias), it is important to recognise that such measures only assess the accuracy of the conditional mean and do not fully exploit the probabilistic nature of several models considered in this study i.e QGBRT. In this context, distribution-based evaluation criteria, and in particular proper scoring rules, offer a more comprehensive way to assess predictive performance.

Two natural candidates are the Pinball loss and the Continuous Ranked Probability Score (CRPS). The Pinball loss, commonly used in quantile regression settings, evaluates the accuracy of predicted conditional quantiles and is therefore well suited to models such as QGBRT, which directly target different parts of the conditional distribution. By contrast, the CRPS provides a strictly proper scoring rule defined over the entire predictive cumulative distribution function, effectively generalising the absolute error to probabilistic forecasts. It simultaneously rewards both calibration (statistical consistency between predictions and observations) and sharpness (concentration of the predictive distribution).

The inclusion of such metrics would be especially relevant in the present setting, where residual distributions are known to deviate from Gaussianity (further explored in Chapter 3). In particular, point-wise metrics may fail to detect systematic deficiencies in modelling distributional features such as asymmetry or tail behaviour, which are critical in environmental applications (e.g., extremes in wind speed). Proper scoring rules like CRPS would allow one to directly evaluate whether the predictive distributions produced by different models appropriately capture these features, rather than merely providing accurate point predictions.

Nevertheless, these metrics are not explicitly considered in the current analysis. The primary reason is to maintain comparability across a heterogeneous set of models, some of which do not naturally provide full predictive distributions or require additional assumptions or post-processing to do so. Incorporating distribution-based scoring rules would therefore introduce an additional layer of methodological complex-

ity that falls outside the scope of the present comparison. However, their inclusion represents a natural and valuable extension of the evaluation framework, particularly for future work aimed at fully exploiting the probabilistic outputs of multifidelity Gaussian process models.

2.5 Results

This section summarises performance in two settings: time-only, and time plus a correlated pseudo-covariate P.

2.5.1 Time-only setting

This section presents an evaluation of the performance of the MFGP model against two benchmark approaches: QGBRT and a standard GP. The QGBRT and GP models are trained exclusively on either abundant low-fidelity measurements (D_L) or incomplete high-fidelity measurements (D_H), whereas the MFGP model integrates all available information across fidelity levels. The results of the unidimensional time experiment are summarised in Table 2.5.1.

The results indicate that models incorporating high-fidelity information, even when temporally sparse, achieve superior performance. The difference between the autoregressive multi-fidelity approach and the single-fidelity models is relatively small. This outcome arises because the multi-fidelity framework functions as a smoothing mechanism across multiple data sources rather than as a filter or trend-extraction tool. Given that noise accounts for approximately 60% of the total variance in the low-fidelity datasets, the true signal becomes obscured, leading to negligible correlation between low-fidelity data and the true underlying signal. This represents an extreme case that is unlikely to occur in practical applications involving reanalysis wind-speed data and LIDAR measurements. Overall, models incorporating high-fidelity information exhibit minimal bias, with most of their predictive uncertainty attributable to variance. In contrast, low-fidelity models display substantial bias in addition to higher overall error.

MODELS	RMSE	MAE	BIAS	VARIANCE
$GP_{LF}(t)$	0.87	0.80	0.8	0.11
$GP_{HF}(t)$	0.34	0.27	0.02	0.11
$MFGP(t)$	0.34	0.28	0.01	0.12
$QGBRT_{LF}(t)$	0.95	0.84	0.81	0.25
$QGBRT_{HF}(t)$	0.36	0.30	-0.009	0.08

Table 2.3: Unidimensional results (time only). Metrics averaged across replications; lower is better for RMSE/MAE/Variance magnitude and Bias.The the HF noise is generated from $SN(-1, 0.3, 5.2)$, with mean approximately -0.77 and variance 0.0347, while the LF noise is generated from $SN(-3.9, 3.5, 5.2)$, with mean approximately -1.16 and variance 4.73.

The main take away is that from such an experiment is that a high level of noise could endanger the MFGP efficacy.

2.5.2 Time plus a correlated predictor

This section introduces the results when an additional correlated dimension is added to the simulation setup. The multidimensional time experiments results (see Table 2.5.2) do not show great differences in terms of metrics performance. However, Gaussian process-based models show a different bias-variance trade-off. The GP_{LF} resulted in smaller bias and higher variance, while the GP_{HF} models resulted in higher bias and smaller variance. The latter can be checked even graphically in Figure 2.8, where the predictions of two GP_{HF} are compared with the true signal. It is clear that information provided by the additional predictor prevents the GP from over-smoothing the data. Not much difference occurred instead in the multi-fidelity model.

MODELS	RMSE	MAE	BIAS	VARIANCE
$GP_{LF}(t, P)$	0.85	0.78	0.66	0.29
$GP_{HF}(t, P)$	0.27	0.21	-0.09	0.06
$MFGP(t, P)$	0.34	0.27	-0.08	0.11
$QGBRT_{LF}(t, P)$	0.92	0.81	0.67	0.40
$QGBRT_{HF}(t, P)$	0.38	0.31	-0.13	0.13

Table 2.4: Table of the summary results of the experiment described in section 2.5.1 using 2-dimension (Time and pseudo-variable).

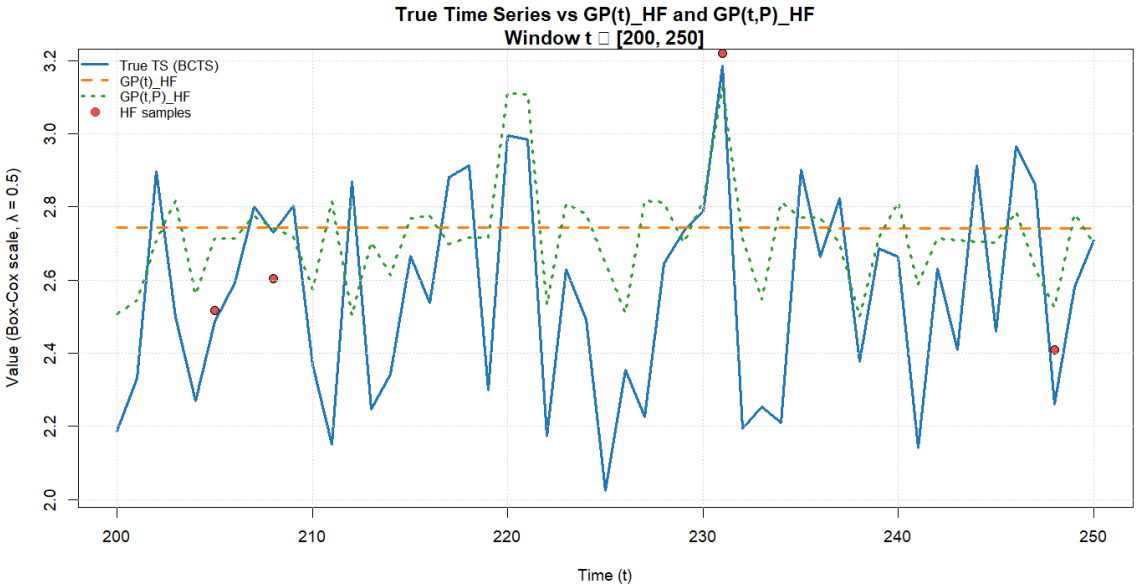


Figure 2.8: Comparison between the true HF time series and the predictions from two Gaussian processes models: *GP(t)_HF*, which uses time as the sole input, and *GP(t,P)_HF*, which incorporates both time and an auxiliary correlated predictor P. The plot shows the interval $t \in [200, 250]$, with red points representing the observed high-fidelity samples used for model training. The dashed and dotted lines correspond to the GP model predictions, while the solid blue line represents the true HF time series on the Box–Cox transformed scale ($\lambda = 0.5$).

The over-smoothing observed in the GP prediction using low-fidelity (LF) data primarily arises from the limited amount of informative data and the high noise level in the observations. These factors cause the GP to infer an overly smooth latent function. This issue can be mitigated by increasing the number or quality of LF samples, improving noise modelling, or by introducing additional explanatory variables, as illustrated in Figure 2.8.

Overall, the residuals of the fitted models (see Figure 2.10 and 2.11) appeared randomly scattered, indicating no evident systematic bias. However, the residuals of

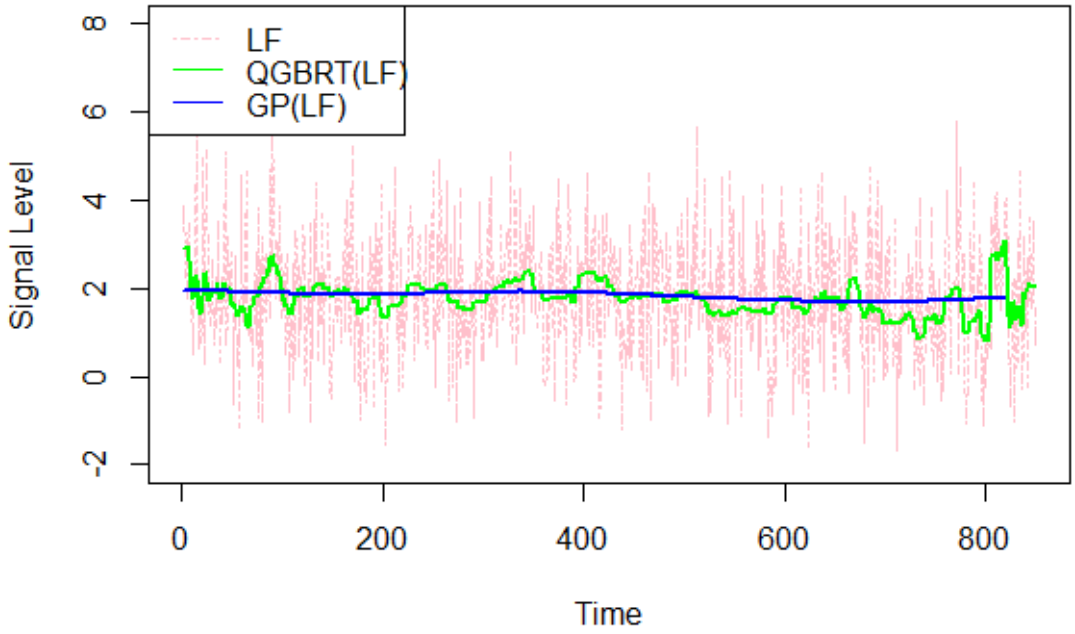


Figure 2.9: Comparison of the LF prediction with the LF signal. In blue the Gaussian process prediction resulted in a flat signal. The QGBRT, in green, shows a higher adaptability.

the $QGBRT_{LF}$ model displayed a slight upward trend. It is worth noting that the high-fidelity models produce relatively few residuals, whereas the low-fidelity models yield many more. This difference is a direct consequence of the respective sample sizes, $D_H = 32$ and $D_L = 850$.

2.6 Limitations and discussion

On the HF–LF relationship and noise ratio. The previous experimental design highlights a central limitation: multi-fidelity models require a nonzero (even nonlinear) relationship between HF and LF data to extract value. In the current setup (2.4.4), the HF–LF correlation is effectively zero; the LF corruption dominates the signal so strongly that LF becomes uninformative. Conceptually, the ratio $\epsilon/f(\mathbf{X})$ must be reduced. Two compact strategies follow: (i) decrease the LF corruption (reduce ϵ in 2.9); this, however, drifts away from the wind case where empirical skewness resides largely in the stochastic remainder, risking overly symmetric data;

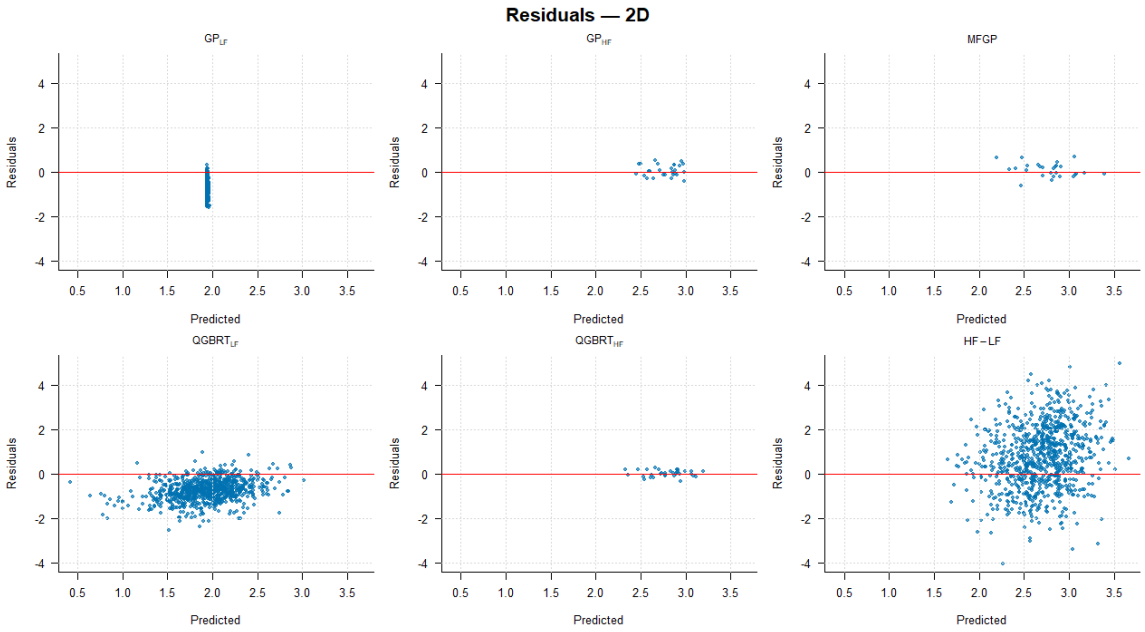


Figure 2.10: Residuals of the multidimensional models. Only the residuals for GP_{HF} and MFGP look randomly scattered. The last panel depicts the LF and HF signal discrepancy.

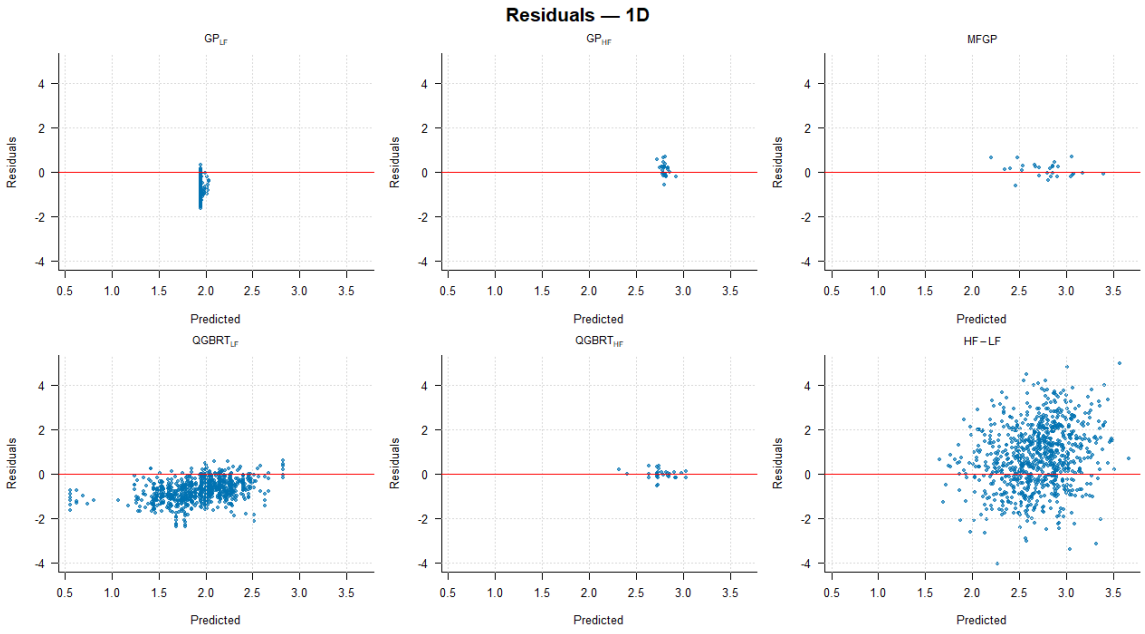


Figure 2.11: Residuals of the time models. The residuals for all models look randomly scattered with few outlier predictions, in the second, third and fifth panels. The last panel show the randomly scatter differences between the LF and HF.

(ii) reallocate part of the empirical skewness from the stochastic to the deterministic component by enriching $f(\mathbf{X})$ with shorter-period structure (e.g., daily seasonality) and inner cycles (e.g., weekly). Under this alternative approach, 2.3 becomes

$$y(t) \;=\; s(t) + T(t) + s(t)_{\text{inner}} + T(t)_{\text{inner}} + r(t), \tag{2.34}$$

where $s(t)_{\text{inner}}$ and $T(t)_{\text{inner}}$ capture intra-annual regularities that encode part of the skewness. A concrete estimation strategy for these inner components is needed; HF and LF would then remain differentiated via $r(t)$ yet share meaningful correlation through the deterministic structure.

On GP over-smoothing and kernel calibration. The over-smoothing observed in the $GP_{HF}(t)$ predictions may be attributed to a combination of kernel hyperparameters and experimental conditions, including limited sample size and high noise levels. At this stage, however, it is not yet clear which of these factors is the primary driver of the phenomenon.

In a Gaussian process model, the signal variance σ_s and the length-scale l play a central role in shaping the fitted function. The signal variance controls the amplitude of variation, while the length-scale governs the rate at which correlations decay with distance in the input space. In particular, larger values of l imply stronger long-range dependence and therefore smoother predictions, whereas smaller values allow for more local variability.

As shown in Table 2.5 and Figure 2.8, the model $GP_{HF}(t)$ is associated with a relatively large temporal length-scale, which may contribute to the observed smoothness. When an additional predictor P is introduced, the corresponding model $GP_{HF}(t, P)$ exhibits a smaller length-scale in the P-dimension, potentially allowing the model to capture more local variation and reduce over-smoothing.

Nevertheless, these observations remain indicative rather than conclusive. A more systematic investigation of the role of kernel choice and hyperparameter values is carried out in Section 2.7.1, where alternative covariance functions and parameter settings are explored. This analysis allows us to assess whether the observed over-smoothing arises primarily from model specification or from the characteristics of the data.

On skewness: scope and modelling routes. The Chapter 2 introduced skewness via the remainder, using the Box-Cox transformation as preprocessing step for dealing with asymmetry. However, such a transformation presents many limitations. In general, this creates a gap between data characteristics and likelihood assump-

Model	σ_s	l	g
$GP_{HF}(t, P)$	0.02	$[833, 0.69]$	1.90
$GP_{HF}(t)$	0.016	$[5.57, \text{N/A}]$	3.26

Table 2.5: Estimated parameters of the high-fidelity GP fit. σ_s denotes the signal variance, l the length scale, and g the nugget term.

tions. Chapter 3 will address this gap within the Gaussian-process framework along two established routes: (i) *direct modelling of skew distributions* (Sections 3.2.1 and 3.2.2), where skewness is handled in the residual model while retaining the GP prior on the latent function; and (ii) *Warping* (Sections 3.3.2 and 3.3.1) refers to applying a monotonic transformation to the response variable so that it becomes more Gaussian before GP modelling; the transformation is then inverted when making predictions. These approaches are assessed with a focus on their suitability for multi-source settings.

On practical emphasis and data fusion relevance. Within these two routes, the emphasis is on approaches that balance methodological rigour with practical applicability in data fusion. Chapter 3 therefore examines specific skew-aware GP formulations and warping strategies using multiple data sources, and introduces a data-driven warping method (Section 3.4) designed to preserve empirical skew while remaining compatible with multi-fidelity workflows.

2.7 Addressing the limitation of the first simulation experiment.

In Section 2.6, two main limitations of the experimental design presented in Section 2.4.1 emerged. First, the Gaussian process models produced over-smoothed predictions (Figures 2.8 and 2.9). Second, the level of corruption applied to the true signal was excessive. Such excessive corruption results in low- and high-fidelity data that are weakly correlated with the underlying true signal and therefore provide limited information for model training. To address these issues, two potential strategies are proposed: reducing the magnitude of corruption (ϵ) or increasing the proportion of the predictable component $f(X)$. In the following sections, a sensitivity analy-

sis of the Gaussian process model parameters is conducted to demonstrate that the observed over-smoothing arises from the experimental design conditions rather than from improper parameter estimation. Finally, a new experimental setup is introduced, based on a second decomposition aimed at removing residual structure in the remainder term and consequently reducing the noise-to-signal ratio $\epsilon/f(X)$.

2.7.1 Gaussian process sensitivity analysis

To connect with the limitations discussed in Section 2.6, this subsection examines whether the observed over-smoothing of the GP arises from *model mis-specification* (kernel choice or hyperparameters) or from *experimental conditions* (high noise, sparse HF samples). The analysis varies (i) the covariance function, (ii) key kernel hyperparameters (length-scale and signal variance), and (iii) the training sample size under the same noise regime, holding other factors fixed and estimating remaining parameters by MLE.

Sensitivity to the covariance function. The squared–exponential (Gaussian) kernel is compared with two Matérn kernels with smoothness parameters $\nu = 5/2$ and $\nu = 3/2$ (Rasmussen, 2004). Figure 2.12 presents GP_{HF}(Time) predictions with $n_H = 32$ fixed. All three kernels produce similarly flat fits; the Matérn 5/2 kernel matches the signal amplitude slightly better, although the differences are marginal. This suggests that the observed over-smoothing is not driven by the choice of kernel family, indicating that the use of the Matérn 5/2 kernel in the main analysis was appropriate.

Sensitivity to the length-scale. The length-scale l controls the decay of correlation over time. Fixing $l \in \{0.23, 0.74, 1.66, 2.36\}$ and re-estimating the remaining parameters by maximum likelihood yields the results shown in Figure 2.13: larger l values (light blue curve) induce smoother predictions, while smaller l (red) allow for faster local variation.

However, within a plausible range, the fitted trajectories remain relatively smooth and exhibit limited variation relative to the scale of the signal. This indicates that,

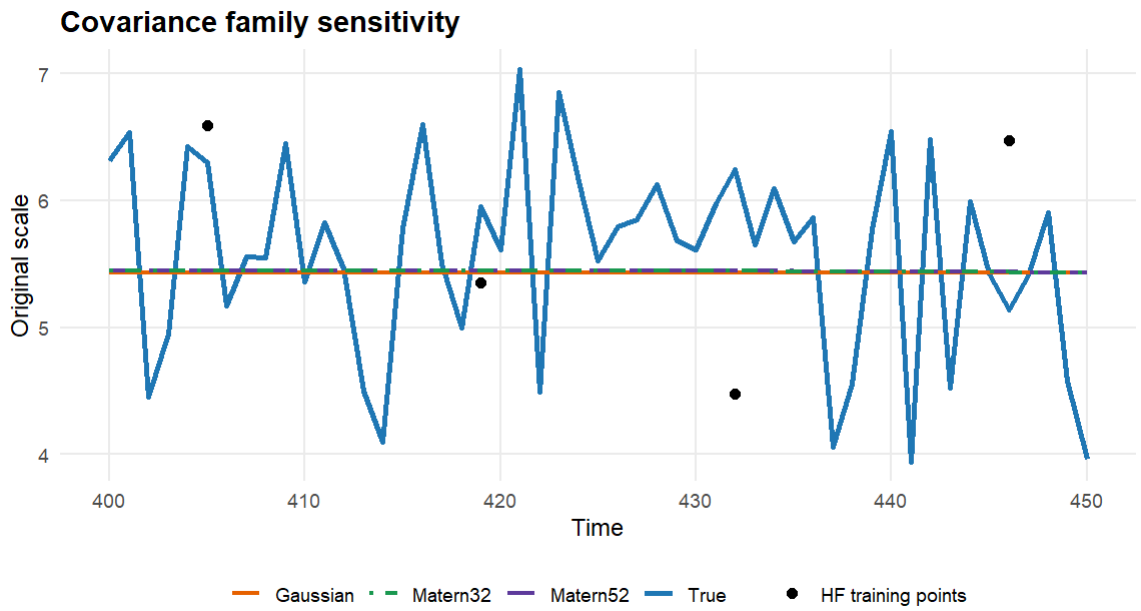


Figure 2.12: Sensitivity to covariance function for (Time) with $n_H = 32$. Prediction window: indices 400–450. Gaussian, Matérn 5/2, and Matérn 3/2 produce similarly flat fits indicating that the flatness of the GP prediction does not depend on covariance function choice.

under the current noise level and sample size, the model is only weakly sensitive to the choice of length-scale.

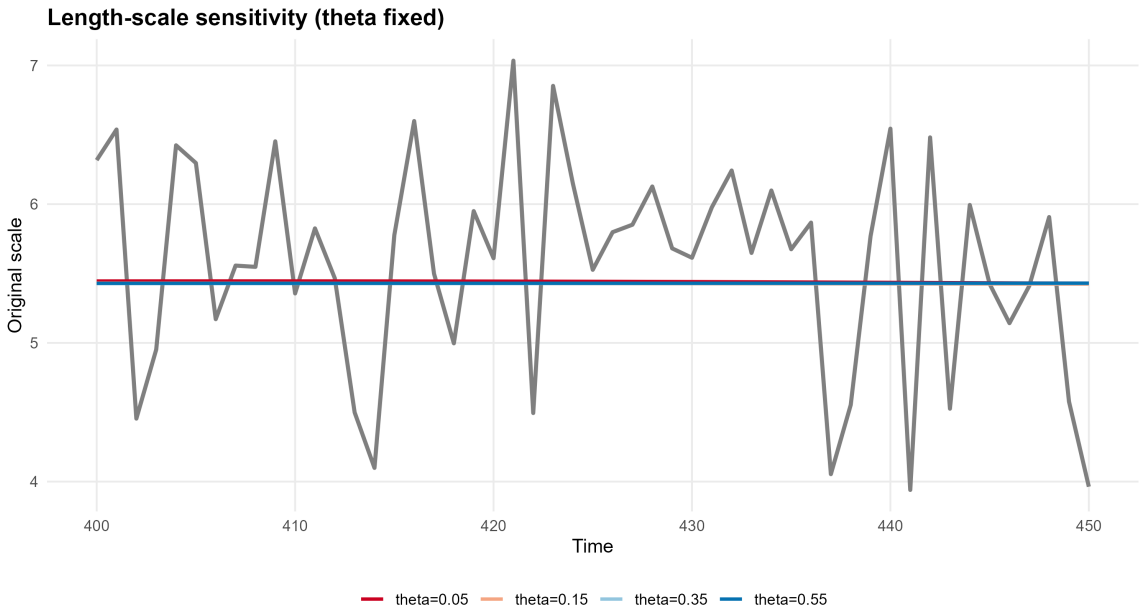


Figure 2.13: Sensitivity to the length-scale l (fixed), with other parameters refit by MLE. While smaller l allows for more local variation and larger l induces smoother predictions, the resulting trajectories remain similar across a wide range of values, indicating limited sensitivity to the length-scale.

To further investigate this behavior, Figure 2.14 reports the profile log-likelihood of l. The relatively flat shape over a broad range of values indicates that the data

provide limited information to precisely identify this parameter. This lack of identifiability explains the weak sensitivity observed in Figure 2.13.

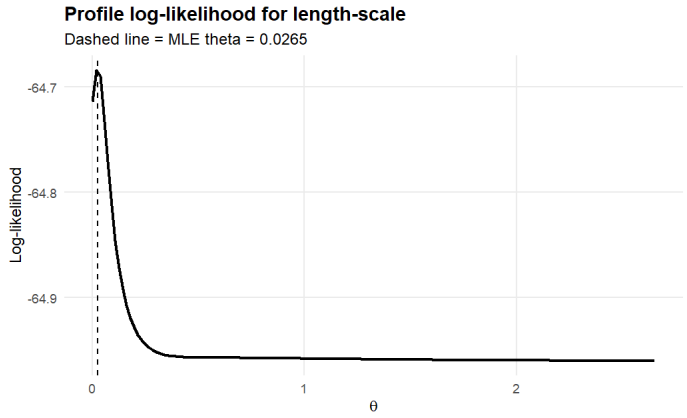


Figure 2.14: Profile log-likelihood as a function of the length-scale parameter l. The relatively flat shape over a broad range of values indicates weak identifiability of the length-scale, explaining the limited sensitivity observed in Figure 2.13.

Sensitivity to the signal variance. The signal variance (σ_s) controls the marginal variability of the Gaussian process prior, and therefore affects the amplitude of the fitted function. Figure 2.15 shows the posterior mean predictions obtained for different fixed values of (σ_s), with the remaining parameters re-estimated by maximum likelihood.

As (σ_s) increases, the model allows for larger deviations from the mean function, leading in principle to higher-amplitude fluctuations. However, in the present setting, the predicted trajectories remain relatively flat across all considered values. This indicates that, under the current noise level and sample size, the data do not provide sufficient information to exploit the increased flexibility induced by larger (σ_s).

Sensitivity to training sample size (and noise). The clearest driver of smoothness is the number of HF observations. Figure 2.16 illustrates that as the HF training size n_H increases, predictions become less smooth and track the target signal more closely, *holding the noise regime fixed*. This aligns with Section 2.6: when HF is sparse and noise is high, $GP_{HF}(\mathrm{t})$ is forced into conservative (smooth) estimates. In the noise experiments (see Section 2.4.4), lower noise levels also lead to noticeably more wiggly GP predictions under the same kernel.

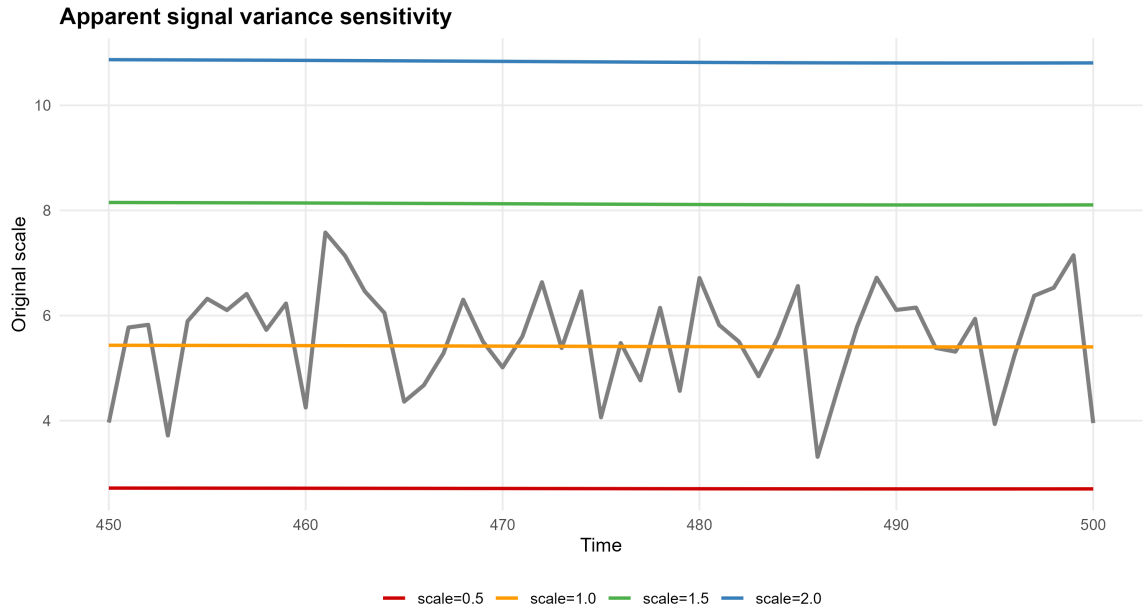


Figure 2.15: Sensitivity to the signal variance σ_s. Each curve represents the GP posterior mean obtained by fixing σ_s at different values and re-estimating the remaining parameters by maximum likelihood. Prediction window: indices 450–500. While larger σ_s increases the prior variance and allows for higher-amplitude fluctuations, the resulting predictions remain similar across values, indicating that the model output is weakly sensitive to σ_s under the current experimental conditions.

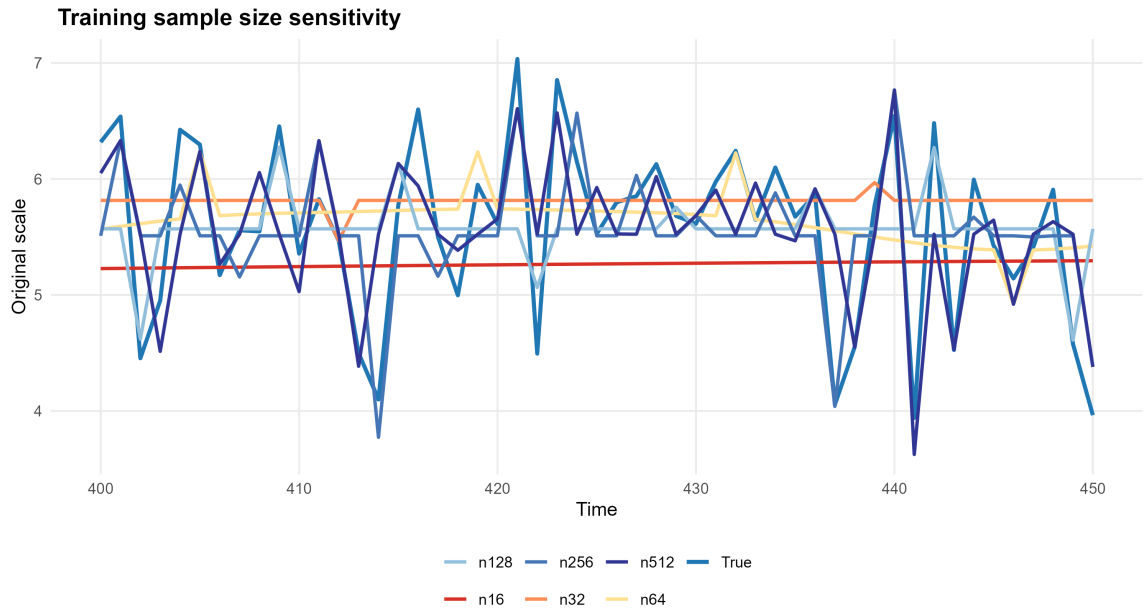


Figure 2.16: Length-scale $l = 0.1$.

Takeaway (link to Section 2.6). Across kernels (Gaussian, Matérn 5/2, Matérn 3/2) and reasonable hyperparameter ranges (l, σ_s), the $GP_{HF}(\mathrm{t})$ fits remain similarly smooth. The dominant factors behind underperformance are *experimental*: sparse HF samples and high noise. This supports the conclusion in Section 2.6 that re-

designing the experiment—by increasing HF coverage and/or reducing the effective noise-to-signal ratio (e.g., via inner seasonal components)—is more consequential than further kernel swapping or hyperparameter tinkering.

2.7.2 Exploring a new experimental design

For clarity, we refer to the experimental setup introduced in Section 2.4.1 as the *original design*, and to the modified procedure introduced here as the *new experimental design* (NES).

This subsection introduces a new experimental design that directly addresses the issues discussed in Section 2.6, namely the near-zero correlation between the high-fidelity and low-fidelity data and the consequent over-smoothing of the Gaussian process model based on the high-fidelity inputs, $GP_{HF}(t)$. The proposed approach aims to increase the deterministic component $f(\mathbf{X})$ shared by both fidelity levels, thereby reducing the remainder term $r(t)$. A lower noise-to-signal ratio is expected to yield $GP_{HF}(t)$ predictions that better capture the variability of the true signal. The workflow follows the same structure as outlined in Section 2.4.1: decomposition of the series, analysis of the remainder, data generation, and model comparison.

2.7.2.1 Greedy multi-seasonality decomposition (ILPE)

In this case a simple two-step approach was designed. Consecutive decompositions are applied with different seasonal and trend windows so that more structure is transferred from the remainder into the predictable component $f(\mathbf{X})$.

Step 1: STL with 6-month seasonality. As in Section 2.1.1 and Figure 2.3, first run STL to extract a six-month seasonal component and a long-term trend. This yields $s(t)$ and $T(t)$ in equation (2.10). The six-month window is chosen based on the descriptive evidence in Section 2.1.1.

Step 2: inner-level pattern extraction (ILPE). Then take the STL remainder and apply a faster second decomposition—either STL with shorter seasonal/trend windows or a moving average. We refer to this second pass as the *inner-level pattern*

extraction (ILPE). In STL, the "seasonal" and "trend" window parameters control how quickly these components can change; smaller windows capture shorter seasonalities and local trend shifts.

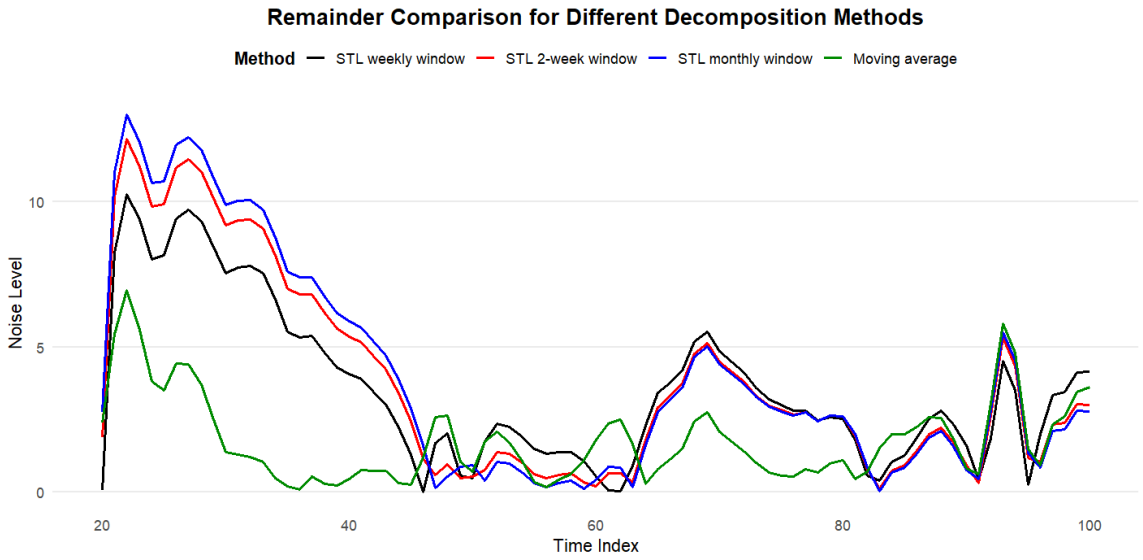


Figure 2.17: ILPE remainder comparison. The moving-average ILPE (green) leaves the smallest remainder; STL-based ILPEs with different seasonal/trend windows return similar patterns.

Based on Figure 2.17, the moving-average ILPE for the new experimental design (NES) was adopted, because it leaves the smallest remainder and thus strengthens the predictable component shared by HF and LF. The effect on the noise-to-signal ratio is shown in Figure 2.18: after the second decomposition the ratio stays below 1 (red line), indicating that noise is smaller than signal.

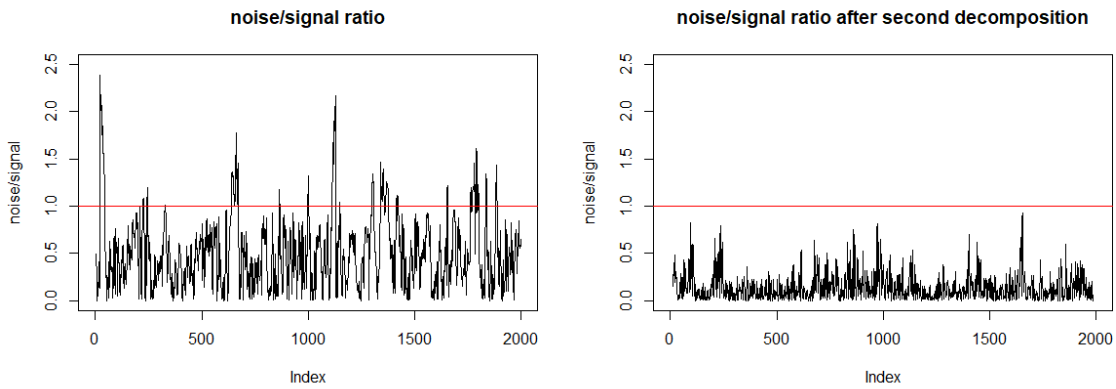


Figure 2.18: Noise-to-signal ratio before (left, STL only; cf. Figure 2.3) and after (right) the ILPE step. The red line marks ratio = 1. The second step keeps noise below signal.

A richer "true signal". We define a new "true signal" by adding inner components to the STL trend and seasonality:

$$\mathrm{TS}_2(t) = s(t) + T(t) + s_{\text{inner}}(t) + T_{\text{inner}}(t). \tag{2.35}$$

Figure 2.19 compares TS_2 with the original $\mathrm{y(t)} = s(t) + T(t)$ and with the raw data. The new TS_2 follows the original series more closely because it captures shorter seasonal/trend changes. In other words, we made $f(\mathbf{X})$ larger and left less to the remainder.

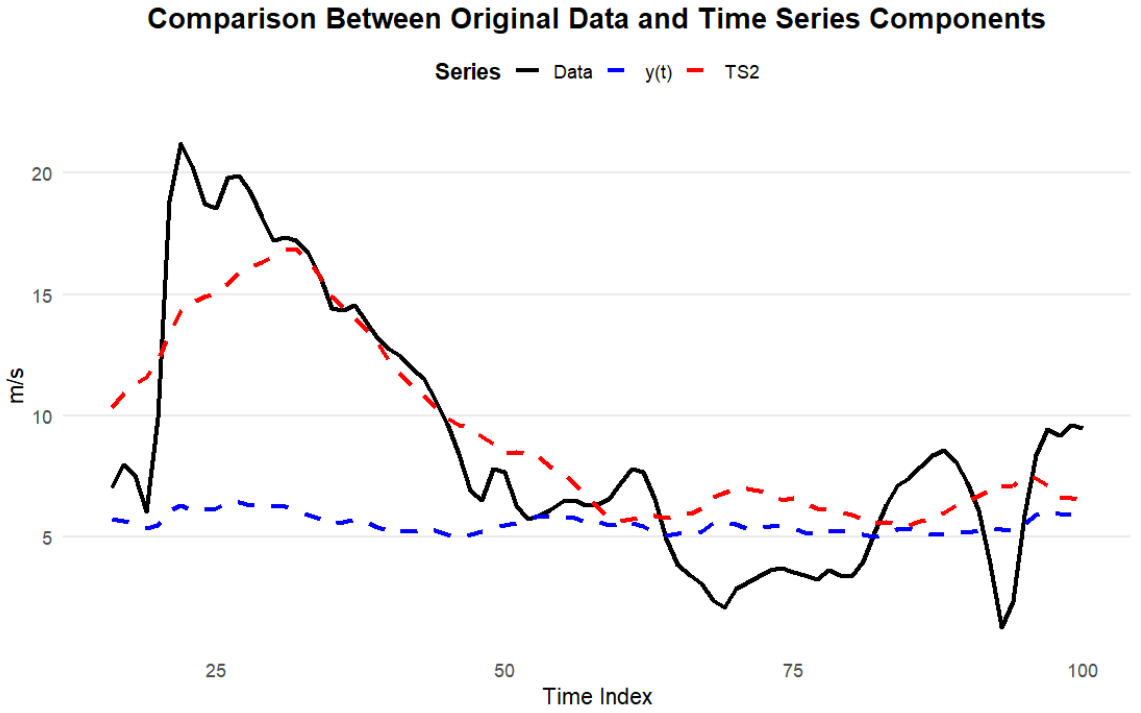


Figure 2.19: Window of 80 time points comparing y(t) (from equation 2.10), TS_2 (from equation 2.35), and the original data. The new TS_2 explains a larger share of the variability.

Remainder diagnostics and choice of distribution. As discussed in Section 2.4, the distribution of the remainder under the new experimental design (NES) is examined. Figure 2.20 presents Cullen–Frey plots for the inner components obtained through the two-stage (greedy) decomposition.

Compared to the original design (see Figure 2.6), the new remainder is only mildly skewed and much closer to normality. Most of the skewness previously present is absorbed by the trend component, while the seasonal component exhibits reduced kurtosis. Consequently, a Normal distribution provides a reasonable approximation

for simulating the remainder under the NES.

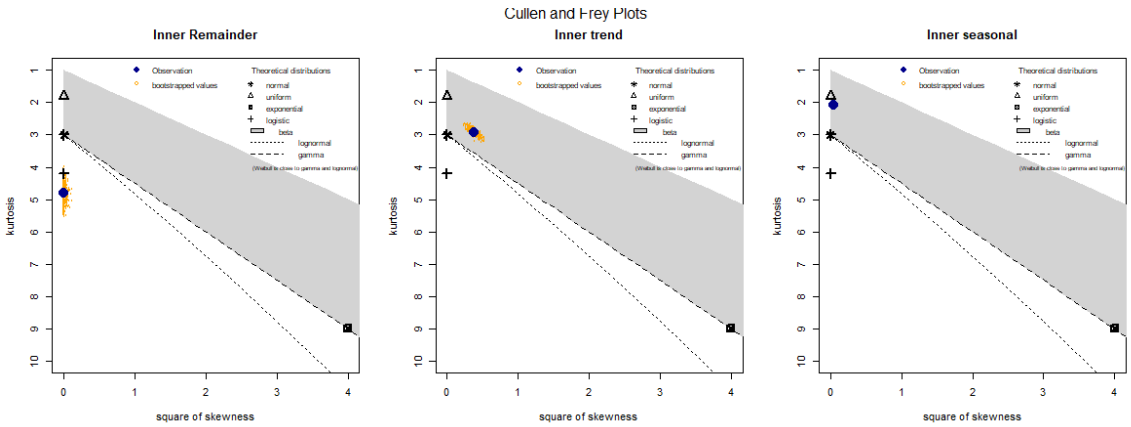


Figure 2.20: Cullen–Frey plots for the remainder under the new experimental design (NES), obtained via the two-stage greedy decomposition. Compared to the original design (Figure 2.6), the remainder is closer to Normal, with reduced skewness and kurtosis.

New simulation examples (HF and LF). For illustration, the simulated HF and LF series in Figure 2.21 include additive Gaussian noise with non-zero means, so the apparent vertical offset is intentional. The exact noise specifications are given below in Section 2.7.2.2.

Figure 2.21 presents two simple examples using Normal errors for both HF and LF. In the left panel, the HF and LF series differ mainly by the mean level of the random component, so their correlation with TS_2 is similar. In the right panel, the LF error has higher variance, producing a more wiggly series and a lower correlation with TS_2. The plot includes the correlation coefficients for clarity (yellow for TS_2–HF, blue for TS_2–LF).

Design space and expected impact on GP. By altering the mean and variance of the HF and LF noise, a set of controlled scenarios can be constructed, encompassing different noise-to-signal ratios and varying HF–LF correlations. The overall pattern remains simple and informative: as the variance of the LF noise increases, its correlation with TS_2 decreases. Importantly, since TS_2 is stronger (larger $f(\mathbf{X})$) and the remainder weaker, $GP_{HF}(t)$ is no longer constrained to overly smooth (flat) fits. This observation aligns with the findings in Section 2.7.1, which indicate that kernel swaps and small hyperparameter adjustments do not account for the over-smoothing; the primary factors are the noise level and the HF sample size. The NES effectively

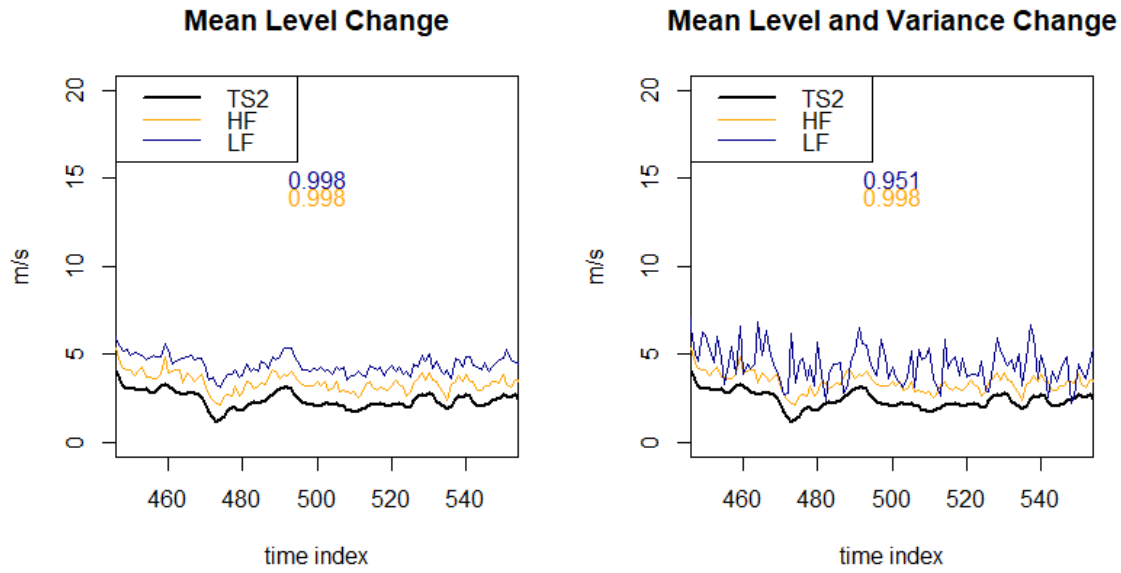


Figure 2.21: Illustrative HF and LF signals under the new experimental design, shown against TS_2 defined in equation (2.35). Black: TS_2; orange: HF; blue: LF. In panel (a), HF and LF differ mainly through a mean-level shift in the additive noise term, so their correlations with TS_2 remain similar despite the vertical offset. In panel (b), LF has both a mean-level shift and a larger variance, which reduces its correlation with TS_2.

reduces the noise-to-signal ratio and enhances the useful structure shared across fidelities, leading to clearer performance gains without modification of the GP model itself.

2.7.2.2 The new noise parameters

In the new simulation experiment, skewness is embedded directly within $f(\mathbf{X})$. Consequently, the noise term can be generated from a standard normal distribution. Specifically, the experiment employs an additive noise component defined as

$$r_{LF} \sim N(2, 0.2).$$

In addition, a second error structure is considered:

$$r_{LF} \sim N(2, 1),$$

which introduces an increased variance level in the observations. While the HF noise is:

$$r_{LF} \sim N(1, 0.2).$$

2.8 Results: under the new simulation design

The five models examined in the previous section are evaluated under the NES conditions. The models are tested for different n_H sample sizes. The average results from 100 replications and the mean level change error are reported in Table 2.6. Consistent with expectations, the models that use high-fidelity data (GP_{HF} and $QGBRT_{HF}$) improve their MAE performance as n_H increases. For $n_H > 160$, the performance delivered by the high-fidelity models is approximately equivalent to that of the multi-fidelity model. The table indicates that a multi-fidelity approach is advisable for small n_H sample sizes, while for larger n_H values, it becomes largely irrelevant.

The results for 100 replications of the experiment with both mean and variance level changes in the LF data are presented in Table 2.7. In this case, the advantage of the multi-fidelity model for $n_H = 32$ (the smallest sample size) is nullified by the noise in the LF data. For larger sample sizes, the results are comparable to those of the high-fidelity models, as observed in Table 2.6.

Figure 2.22 provides a graphical comparison of the predictions $MFGP(\text{t})$ and $GP_{HF}(\text{t})$ under NES conditions. In panel (a), the multi-fidelity model successfully reproduces the pattern of the true signal, yielding predictions that are, on average, 1 m/s[1] more accurate than those of the standard GP. In panel (b), however, the multi-fidelity advantage dissipates due to the high level of noise variance.

MODELS	$N_1 = 32$ (sd)	$N_1 = 96$ (sd)	$N_1 = 160$ (sd)	$N_1 = 256$ (sd)	$N_1 = 416$ (sd)	$N_1 = 672$ (sd)
$GP_{\text{LF}}(t)$	0.54(0.018)	0.54(0.018)	0.54(0.018)	0.54(0.018)	0.55(0.018)	0.55(0.018)
$GP_{\text{HF}}(t)$	0.43(0.048)	0.32(0.011)	0.29(0.008)	0.30(0.004)	0.30(0.007)	0.29(0.004)
MFGP(t)	0.29(0.047)	0.29(0.060)	0.29(0.015)	0.30(0.011)	0.30(0.019)	0.29(0.011)
$\text{QGBRT}_{\text{LF}}(t)$	0.55(0.02)	0.55(0.02)	0.55(0.02)	0.55(0.02)	0.55(0.02)	0.55(0.02)
$\text{QGBRT}_{\text{HF}}(t)$	0.52(0.005)	0.39(0.021)	0.34(0.016)	0.33(0.009)	0.32(0.015)	0.31(0.013)

Table 2.6: Models MAE summary from 100 replications and varying high-fidelity sample size n_H. Parentheses report the standard deviation. Under the NES, low-fidelity samples include an additive noise component with a mean-level shift (see Figure 2.21), which induces a persistent bias and explains the approximately constant MAE for LF-based models.

[1]The result is obtained through an inverse Box–Cox transformation of the predictions.

MODELS	$N_1 = 32$ (sd)	$N_1 = 96$ (sd)	$N_1 = 160$ (sd)	$N_1 = 256$ (sd)	$N_1 = 416$ (sd)	$N_1 = 672$ (sd)
$GP_{\text{LF}}(t)$	0.45(0.02)	0.46(0.02)	0.45(0.02)	0.45(0.02)	0.45(0.02)	0.46(0.02)
$GP_{\text{HF}}(t)$	0.41(0.048)	0.29(0.011)	0.30(0.008)	0.29(0.004)	0.30(0.007)	0.29(0.004)
MFGP(t)	0.44(0.010)	0.29(0.006)	0.31(0.007)	0.30(0.007)	0.31(0.006)	0.29(0.002)
$\text{QGBRT}_{\text{LF}}(t)$	0.53(0.021)	0.55(0.021)	0.53(0.021)	0.53(0.021)	0.53(0.021)	0.54(0.021)
$\text{QGBRT}_{\text{HF}}(t)$	0.52(0.005)	0.31(0.021)	0.36(0.016)	0.32(0.009)	0.31(0.015)	0.31(0.013)

Table 2.7: Models MAE summary from 100 replications and varying high-fidelity sample size n_H. Parentheses report the standard deviation. The data depend on the greedy decomposition scheme; low-fidelity samples include additive noise with both mean and variance distortions.

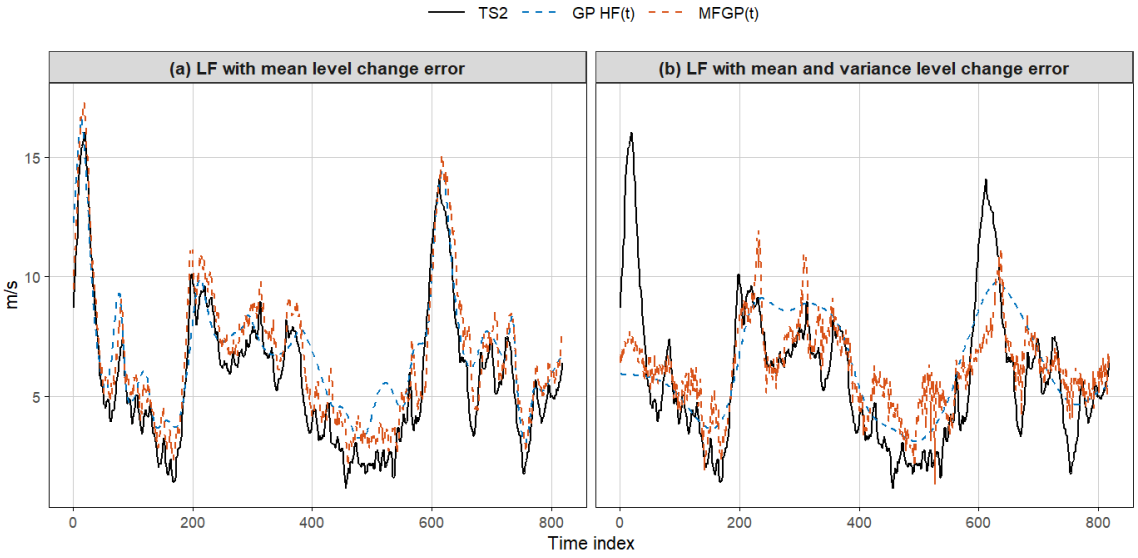


Figure 2.22: Comparison of the $MFGP(t)$ and $GP_{HF}(t)$ NES predictions and $y_{\mathbb{T}}$. In panel (a), the predictions in the case of LF with an error mean level change; in panel (b), the predictions in case of LF having an error with both mean and variance level change. The n_H HF sample size is fixed at 32.

2.9 Concluding discussion

This chapter compared a MFGP with single-fidelity baselines—a GP and QGBRT—within a wind-speed case study motivated by the AGNES project and ERA5 reanalysis data. The investigation began with an STL-based decomposition of the wind-speed series, which separated the long-term trend and seasonal components from an irregular remainder. A goodness-of-fit analysis of this remainder motivated the use of a skew-normal model, capturing the asymmetric nature of the stochastic fluctuations. Based on this decomposition, an initial simulation experiment was conducted to assess how the shape of the remainder distribution influenced predictive performance across model types. This first experiment served as a diagnostic baseline: it confirmed that execissive noise can make the MFGP useless.

A greedy, two-stage decomposition (ILPE) subsequently enhanced the predictable component by incorporating inner trend and seasonal structures, resulting in a refined signal TS_2 and a milder, near-Normal remainder. This "new experimental design" (NES) effectively reduced the noise-to-signal ratio without altering the model class, thereby enabling a clearer evaluation of multi-fidelity performance.

The results demonstrate that multi-fidelity modelling is most beneficial when high-fidelity data are limited but the cross-fidelity correlation remains substantial. Under NES conditions, the MFGP outperformed single-fidelity models for smaller n_H values, whereas for larger n_H (approximately $n_H > 160$ in the experimental design), high-fidelity single-source models achieved comparable mean absolute error (MAE). When low-fidelity noise was dominant, the correlation between the high- and low-fidelity components weakened, and the advantages of multi-fidelity learning largely disappeared. Sensitivity analyses further revealed that kernel substitutions (e.g., squared exponential versus Matérn) and minor hyperparameter adjustments were insufficient to mitigate over-smoothing in GP_{HF} under sparse and noisy conditions. The primary factors influencing performance were related to the experimental setup—specifically, the noise level and the size of the high-fidelity dataset—rather than to model tuning. In other words, most performance differences arose from the data conditions themselves rather than from deficiencies in the learning algorithm. The inclusion of a correlated predictor was found to improve the bias–variance trade-off of GP models and reduce over-smoothing, emphasizing the importance of informative covariates in conjunction with fidelity design.

From a practical standpoint, multi-fidelity models are advantageous when the high-fidelity sample size is small and a meaningful cross-fidelity correlation exists. For moderate or large n_H, simpler high-fidelity GP or QGBRT models achieve comparable accuracy. Experimental designs that reduce low-fidelity noise and enhance the deterministic component, such as trend and seasonal features, generally produce greater improvements than kernel or hyperparameter modifications. Nevertheless, these findings are subject to certain limitations: the decomposition and distributional assumptions, though data-informed, remain simplifications, and the

analysis was restricted to a single site without consideration of spatial dependence or data gaps typical of field deployments.

Future research should extend these findings by incorporating skew-aware likelihoods and warping strategies with inversion at prediction time, as discussed in Chapter 3. Additional directions include the development of richer fidelity link functions that account for heteroskedasticity, nonstationarity, or systematic bias; active experimental design for optimal placement of scarce high-fidelity measurements; and generalization of the framework to spatio-temporal contexts (see Chapter 4), such as wind farm arrays. Further work should also integrate probabilistic verification tools, such as the Continuous Ranked Probability Score (CRPS), and link predictive accuracy to economic performance indicators relevant to wind energy applications. The last chapter (see Chapter 5) is devoted to extension and further development.

In summary, the analysis confirms that multi-fidelity methods provide the greatest benefit when experimental conditions support meaningful cross-source correlation and when predictable structures are explicitly modeled. In contrast, when high-fidelity data are sufficiently abundant or low-fidelity information is highly noisy and uninformative, single-fidelity approaches remain adequate.

Chapter 3

Addressing skewness in environmental data

Chapter 1 situated multifidelity modelling within the broader context of data fusion, outlining the key challenges and research questions. Chapter 2 then compared multifidelity approaches with *mono-fidelity* methods across various dimensions, such as noise levels and sampling frequency, highlighting both the limitations and advantages of multifidelity modelling for wind speed data, thereby addressing **RQ1**.

This chapter investigates approaches for modelling skewed data (addressing the **RQ2**) in the context of Gaussian processes, with a focus on scenarios involving multiple data sources. The discussion begins by providing an overview of the challenges associated with skewness (Section 3.1). It then reviews the relevant literature on Gaussian processes for modelling skewed data (Section 3.1.2), highlighting the interactions between these approaches. Then specific models for dealing with skewness are examined, with their respective strengths and limitations discussed. Particular attention is given to their applicability in data fusion.

Two principal frameworks are investigated for dealing with skewed data. The first focuses on the direct modelling of skew distributions (see Sections 3.2.1 and 3.2.2), while the second—commonly referred to as warping—addresses skewness by transforming the data prior to modelling (Sections 3.3.2 and 3.3.1). Within these frameworks, particular emphasis is placed on approaches that balance methodological

rigour with practical applicability. Building on this foundation, a novel data-driven warping method is introduced (Section 3.4), whose performance is evaluated through extensive simulation studies and further demonstrated in a real-world application. The chapter concludes with general remarks and a synthesis of the key findings.

3.1 Approaches for handling skewness

As illustrated in Chapters 1 and 2, the distribution of raw wind-speed data is typically right-skewed. For this reason, in Section 2.8 a transformation from the Box–Cox family was applied to preprocess the data prior to modelling. A drawback of such transformations, however, is their reliance on a subjective choice of the λ parameter (the parameter that indicates the mathematical form of the transformation).

The literature has explored a range of less subjective approaches for handling skewness in a regression context. These include methods based on non-Gaussian copulas (Bárdossy and Li, 2008), scale-mixing skew Gaussian processes (Zareifard and Khaledi, 2013), skew Gaussian processes (Alodat and Shakhatreh, 2020b), transformed Gaussian random fields (Xu and Genton, 2017), and input-space warping (Peters et al., 2021; Snelson et al., 2003). Collectively, these approaches represent ideas that either remain within, or extend beyond, the GP framework, and they each come with certain limitations, which are briefly outlined below.

Copulas can be viewed as complementary to standard Gaussian Process models, since they allow flexible, non-Gaussian dependence structures by separating the modeling of marginals from their joint dependence. However, incorporating copulas into multi-fidelity settings is challenging. In particular, it is nontrivial to determine whether different data sources should share a common marginal transformation or be modeled with distinct marginals, as this choice affects both the strength and interpretability of the learned dependence structure.

More generally, copula-based interpolation is complex, since it requires fitting a multivariate copula to the observed data.

Other approaches can be considered extensions within the *GP machinery*, but they

face a number of distinct challenges. For example, approaches based on the skew Gaussian process are often difficult to derive in closed form; and even when closed forms exist, they tend to be computationally demanding. Scale mixing is potentially attractive, although it typically requires intensive procedures such as combining stochastic approximation of expectation–maximisation (SAEM) with MCMC (Zareifard and Khaledi, 2013). Input-space warping is often more practical, as it involves only the estimation of additional parameters for the warping function, though identifying a suitable function remains challenging. Another promising direction is the construction of parsimonious skew distributions—defined in terms of a limited number of parameters. This approach, in principle, avoids the over-parameterisation problem outlined above.

Despite the variety of available methods for addressing skewness, the present analysis focuses on approaches developed specifically for Gaussian processes, with particular attention to their suitability for multivariate extensions and, ultimately, for data fusion.

3.1.1 Handling skewness in data-fusion context

The fusion of data is essentially a means of modelling datasets of complementary nature, maximising their independent qualities while minimising computational cost. The multi-fidelity framework explored thus far in this thesis operates by learning the relationship between datasets, correcting LF data in locations where HF data are unavailable. The crucial element in this process is the inter-dataset relationship: if this relationship changes, an entirely new data-fusion model is required. This explains why simpler approaches, such as the Box–Cox transformation, are often suboptimal for data-fusion applications. When two datasets are present, the transformations required to normalise them may differ. For instance, applying a logarithmic transformation to high-fidelity data while using a square-root transformation for low-fidelity data may distort the relationship between datasets in the latent space. Preserving the ordering between datasets is therefore pivotal for effective data fusion. Conversely, enforcing the same transformation across all datasets may succeed in normalising one dataset but not the other. Practical examples of this issue are provided in Sec-

tion 3.5.1. In general, addressing skewness in a data-fusion context is non-trivial: it is necessary not only to normalise the datasets effectively, but also to preserve the inter-dataset relationships.

3.1.2 Gaussian processes and skewness

Modelling skewed data with Gaussian processes has historically posed a significant challenge. This difficulty arises largely from the identifiability issues inherent in skew Gaussian random fields, (Genton and Zhang, 2012; Kim and Mallick, 2004). To address this problem, two principal approaches have been developed. The first models skewness directly by employing flexible distributions, which we refer to as *direct modelling of skewness*, for example through the Closed Skew Normal (CSN) distribution, see Allard and Naveau (2007) for an early reference. The second approach mitigates skewness indirectly, by first transforming the data and then applying Gaussian processes to the normalised version. This is referred to as *indirect modelling of skewness*. In the following section, two approaches falling under the first category are discussed: Gaussian process with Skewed Error (GPRSE) introduced by Alodat and Shakhatreh (2020b), and Gaussian process with Specified Covariance Function (GPSCF) proposed by Khaledi et al. (2023). A detailed description of these methods is provided with a discussion of their respective advantages and disadvantages. Subsequently, two main methods for transforming skewed data (see Section 3.3) in the context of Gaussian processes are illustrated, the Parametric Warped GP (PWGP) (Snelson et al., 2003)[1] and the Data Driven Warped GP (DDWGP) (Agou et al., 2022).

These are only a few of the available approaches for modelling skewness within GPs, but they illustrate how specific challenges have been addressed within the GP framework. A practical way to look at these choices is the following: GPSCF resolves many of the limitations of GPRSE, just as DDWGP improves upon PWGP.

[1]Note that this PWGP acronym has never been used before, historically the method is referred simply Warped Gaussian process WGP.

3.2 Direct modelling of skewness

3.2.1 Model 1: GP with Skew Error (GPRSE)

The CSN distribution is often used as a building block for more complex models requiring skewed distributions, owing to its simplicity and ease of implementation. It also has several convenient mathematical properties, when compared to a skew normal distribution described in Section 2.4.3, including closed-form expressions for the moments and the cumulative distribution function.

Indeed, the model described below is based on the CSN distribution. Such a distribution is a specific form of the skew-normal distribution obtained by fixing the location parameter to zero and setting the shape parameter to one. This yields a simplified form that depends only on the scale parameter.

Consider a vector of observations $\mathbf{y}$ with a CSN distribution of dimension p; it has the following probability density function (PDF):

$$P_Y(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{v}, \mathbf{D}, \boldsymbol{\Delta}) = c_p \, \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \Phi_q\big(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \mathbf{v}, \boldsymbol{\Delta}\big). \tag{3.1}$$

The components of Equation 3.1 are:

- $\phi_p(\cdot; \cdot, \cdot)$ denotes the p-dimensional multivariate normal PDF with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- $\Phi_q(\cdot\,; \cdot, \cdot)$ denotes the CDF of the q-dimensional multivariate normal distribution.
- $\mathbf{D}$ is an arbitrary $q \times p$ matrix that typically controls skewness and scaling.
- $\boldsymbol{\Sigma}$ is a positive-definite matrix of dimension $p \times p$ representing the scale (covariance).
- $\boldsymbol{\Delta}$ is a positive-definite matrix of dimension $q \times q$.
- $\mathbf{v}$ is a vector in $\mathbb{R}^q$.
- $\boldsymbol{\mu}$ is the location parameter.

- $c_p = \Phi_q\big(0; \mathbf{v}, \boldsymbol{\Delta} + \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}^\top\big)$.

The notation for a random variable Y with a CSN distribution is $Y \sim CSN_{p,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{D}, \mathbf{v}, \boldsymbol{\Delta})$. If $Y_1, \ldots, Y_n$ are independent random vectors following a CSN distribution, their joint distribution is

$$(Y_1^T, \ldots, Y_n^T)^T \sim CSN_{p^+,q^+}(\mu^+, \boldsymbol{\Sigma}^+, \mathbf{D}^+, \mathbf{v}^+, \boldsymbol{\Delta}^+), \tag{3.2}$$

where the superscript + serves to distinguish the stacked multivariate case. Note that the CSN distribution involves a comparatively large number of parameters.
The Gaussian process Regression with Skewed Errors (GPRSE) is defined as a standard Gaussian process model:

$$y(t) = f(t) + \epsilon(t),$$

with $t = 1, \ldots, n$ and $t \in \mathbb{R}$, but the noise terms $\epsilon(t_1), \ldots, \epsilon(t_n)$ are assumed to follow independent skew-normal distributions $\mathrm{SN}(\mu, \sigma, \lambda)$, where μ is the location parameter, σ the scale parameter, and λ the skewness parameter. $f(t)$ is a smooth function defined by a Multivariate Gaussian distribution indexed at t. When $\lambda = 0$, the error distribution reduces to the standard normal and the model corresponds to a standard Gaussian process.
In particular, the error PDF is

$$p_\epsilon(\epsilon;\, \mu, \lambda, \sigma) = \frac{2}{\sigma}\, \phi\left(\frac{\epsilon - \mu}{\sigma}\right)\, \Phi\left(\lambda\, \frac{\epsilon - \mu}{\sigma}\right), \tag{3.3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal PDF and CDF, respectively (equivalently, $\phi(\epsilon; \mu, \sigma^2)$ is the $\mathcal{N}(\mu, \sigma^2)$ PDF). Here $\mu \in \mathbb{R}$ is the location, $\sigma > 0$ the scale, and $\lambda \in \mathbb{R}$ the skewness (shape) parameter;
The joint distribution of the error terms ϵ_i[1] is multivariate:

$$\boldsymbol{\epsilon} \sim CSN_{n,n}(\boldsymbol{\mu}\, \mathbf{I}_n, \boldsymbol{\Sigma}^2 \mathbf{I}_n, \mathbf{D}\, \mathbf{I}_n, \mathbf{0}\, \mathbf{I}_n, \boldsymbol{\Sigma}^2 \mathbf{I}_n),$$

[1] We drop the index notation since the error terms are independent across indices.

with PDF given by $\prod_{i=1}^{n} P_\epsilon(\epsilon_i; \mu, \lambda, \mathbf{\Sigma})$.

As with any Gaussian process, the aim is to predict unobserved data points y^* at new locations t^*. This requires the predictive distribution, i.e. the distribution of y^* conditional on the observed values:

$$f^* \,\big|\, \mathbf{y}, t^* \sim CSN(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{\Sigma}}, \tilde{\mathbf{D}}, \tilde{\mathbf{v}}, \tilde{\mathbf{\Delta}}). \tag{3.4}$$

Although the predictive distribution involves a large number of parameters, these can be obtained through a sequence of tractable computations. The full derivation of such quantities can be found in Katzfuss and Guinness (2021). In particular, several covariance matrices are required: $\mathbf{K}_{11}$, the covariance matrix at the training locations; $\mathbf{K}_{12}$ and $\mathbf{K}_{21}$, the cross-covariances between training and test locations; and $\mathbf{K}^*_{22}$, the covariance matrix at the test locations. While such covariance structures are standard in Gaussian process models, the present framework additionally involves a number of predictive quantities, namely:

$$\tilde{\mathbf{D}} = \begin{pmatrix} \mathbf{0} \\ \lambda \mathbf{\Sigma}^2 \, \mathbf{T}_{n\times 1} \end{pmatrix},$$

$$\tilde{\mathbf{v}} = \begin{pmatrix} \mathbf{0}_{1\times n} \\ -\lambda \mathbf{\Sigma}^2 \, (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} (\mathbf{y} - \mu \, \mathbf{1}_n) \end{pmatrix},$$

$$\tilde{\mathbf{\Delta}} = \begin{pmatrix} 1 & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{\Sigma}^2(1+\lambda^2)\, \mathbf{I}_n - \lambda^2 \mathbf{\Sigma}^4 \, \mathbf{W} \end{pmatrix},$$

$$\mathbf{W} = (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} + \frac{(\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \mathbf{K}_{12} \mathbf{K}_{21} (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1}}{\mathbf{K}^*_{22} - \mathbf{K}_{12}^\top (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \mathbf{K}_{21}},$$

$$\mathbf{T} = \frac{(\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \mathbf{K}_{21}}{\mathbf{K}^*_{22} - \mathbf{K}_{12}^\top (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \mathbf{K}_{21}},$$

$$\tilde{\boldsymbol{\mu}} = \mathbf{K}_{12}^\top \, (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \, \mathbf{y},$$

$$\tilde{\mathbf{\Sigma}}^2 = \mathbf{K}^*_{22} - \mathbf{K}_{12}^\top (\mathbf{K}_{11} + \mathbf{\Sigma}^2 \mathbf{I}_n)^{-1} \mathbf{K}_{21}.$$

The final predictive equations for a new point t^* are therefore

$$E(f^* \,|\, \mathbf{y}, t^*) = \tilde{\boldsymbol{\mu}} + \boldsymbol{\Sigma}\, \tilde{\mathbf{D}}^\top \xi, \tag{3.5}$$

for the mean, and

$$V(f^* \,|\, \mathbf{y}, t^*) = \tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{\Sigma}}^2\, \tilde{\mathbf{D}}^\top\, \phi\, \tilde{\mathbf{D}} - \tilde{\boldsymbol{\Sigma}}^2\, \tilde{\mathbf{D}}^\top \xi \xi^\top \tilde{\mathbf{D}}, \tag{3.6}$$

for the variance. In Equations 3.5 and 3.6, two additional terms appear, ξ and ϕ, which are difficult to compute because they depend on derivatives of the normal CDF Φ. For example,

$$\xi = \frac{\dfrac{\partial}{\partial s} \Phi_{\mathbf{n+1}}\big(s; v; \tilde{\boldsymbol{\Delta}} + \tilde{\boldsymbol{\Sigma}} \tilde{\mathbf{D}} \tilde{\mathbf{D}}^\top\big)\Big|_{s=0}}{\Phi_{\mathbf{n+1}}\big(s; v; \tilde{\boldsymbol{\Delta}} + \tilde{\boldsymbol{\Sigma}} \tilde{\mathbf{D}} \tilde{\mathbf{D}}^\top\big)}. \tag{3.7}$$

(For the derivation of ϕ, see Alodat and Shakhatreh (2020b).) These expressions highlight the computational limitations of the method: evaluating the predictive terms is computationally demanding on a local machine, and a multi-fidelity extension—in which parameters are typically decoupled—would be even more prohibitive.

3.2.2 Model 2: GP with Specified Covariance Function (GPSCF)

The GPSCF exploits also the CSN distibution for addressing skewness in the context of GPs. However, it uses a completely different approach from Alodat and Shakhatreh (2020b). The idea is having both a *symmetric part* described by a Gaussian random field plus a latent truncated Gaussian vector that induces skewness. The backbone of such model is the covariance-adjustment.

To construct the covariance-adjusted skew-Gaussian (CASG) process, it is necessary to build a Gaussian structure that preserves a target covariance function $c(s, s')$ and is computationally tractable. Let $R = \{r_1, \ldots, r_m\}$ denote a set of reference

locations (knots). Define

$$\mathbf{C}_{RR} = \big[c(r_i, r_j)\big]_{m\times m}, \qquad \mathbf{c}_{Rs} = \big[c(r_1, s), \ldots, c(r_m, s)\big]^\top, \qquad \mathbf{B}(s) = \mathbf{C}_{RR}^{-1/2}\,\mathbf{c}_{Rs}.$$

Introduce a latent coefficient vector

$$\mathbf{\Gamma}_m \sim \mathcal{N}_m(\mathbf{0}, I_m),$$

and decompose the process as

$$p_0(s) = w_0(s) + \epsilon(s), \qquad w_0(s) = \mathbf{B}(s)^\top \mathbf{\Gamma}_m, \tag{3.8}$$

where $\epsilon(s)$ is an independent mean-zero Gaussian process chosen so that $\mathrm{Cov}\{p_0(s), p_0(s')\} = c(s, s')$. By construction,

$$\mathrm{Cov}\{w_0(s), w_0(s')\} = \mathbf{B}(s)^\top \mathbf{B}(s') = \mathbf{c}_{Rs}^\top \mathbf{C}_{RR}^{-1} \mathbf{c}_{Rs'},$$

and hence the residual covariance is

$$\mathrm{Cov}\{\epsilon(s), \epsilon(s')\} = c(s, s') - \mathbf{c}_{Rs}^\top \mathbf{C}_{RR}^{-1} \mathbf{c}_{Rs'}.$$

In particular, the residual variance is

$$\mathrm{Var}\{\epsilon(s)\} = c(s, s) - \mathbf{c}_{Rs}^\top \mathbf{C}_{RR}^{-1} \mathbf{c}_{Rs}. \tag{3.9}$$

A key property is exactness at the knots. Let $\mathbf{e}_j \in \mathbb{R}^m$ denote the j-th canonical (standard basis) vector, i.e. the vector with a one in position j and zeros elsewhere. If $s = r_j \in R$, then $\mathbf{c}_{Rr_j} = \mathbf{C}_{RR}\mathbf{e}_j$ and

$$\mathrm{Var}\{\epsilon(r_j)\} = c(r_j, r_j) - \mathbf{e}_j^\top \mathbf{C}_{RR} \mathbf{C}_{RR}^{-1} \mathbf{C}_{RR} \mathbf{e}_j = c(r_j, r_j) - c(r_j, r_j) = 0,$$

so $\epsilon(r_j) = 0$ and $p_0(r_j) = w_0(r_j)$. Stacking values at the knots yields

$$\mathbf{p}_0 = \big(p_0(r_1), \ldots, p_0(r_m)\big)^\top = \mathbf{C}_{RR}^{1/2}\,\mathbf{\Gamma}_m \sim \mathcal{N}_m(\mathbf{0}, \mathbf{C}_{RR}).$$

This decomposition shows which part is "covariance-adjusted": the latent Gaussian vector at the knots drives the structured part w_0, while the independent residual ϵ adjusts the covariance away from the knots so that the full backbone p_0 reproduces the target kernel $c(\cdot, \cdot)$ exactly. Once this Gaussian backbone is in place, skewness is introduced by replacing the standard normal coefficients $\mathbf{\Gamma}_m$ with a parsimonious closed skew-normal (CSN) vector having mean zero and identity covariance. This substitution preserves $\mathrm{Cov}\{p(s), p(s')\} = c(s, s')$ while endowing the finite-dimensional distributions with CSN/SUN skewness, thereby yielding the full CASG process.

3.2.2.1 The importance of a Parsimonious CSN Distribution for GPSCF

The parsimonious closed skew-normal (PCSN) distribution is introduced as a more tractable alternative to the standard closed skew-normal. While the CSN is a natural choice for introducing skewness, it suffers from a number of drawbacks: its mean and covariance expressions are difficult to evaluate, it is prone to overparametrisation with too many parameters (i.e. $\boldsymbol{\mu}$, $\mathbf{\Sigma}$,$\mathbf{D}$ $\mathbf{v}$, $\mathbf{\Delta}$) to estimate reliably, and it can suffer from identifiability issues in which different parameter combinations lead to indistinguishable distributions. In addition, the CSN is not always closed under linear transformations, which complicates its use in process modelling. These limitations make inference computationally demanding and conceptually unstable.

The PCSN addresses these concerns by reducing the dimensionality of the parameter space and simplifying the structure of the distribution. Its construction ensures that skewness can be introduced without distorting the second-order dependence structure of the Gaussian process, thereby preserving the covariance function. At the same time, redundant parameters are fixed, avoiding identifiability problems and leading to parameter estimates that are both interpretable and statistically valid. Overall, the PCSN provides a stable and computationally efficient framework for incorporating skewness into Gaussian processes.

Formal definition. The PCSN is built upon a latent Gaussian vector with identity covariance, leaving only a minimal set of skewness parameters. The objective is to

define a skew-normal random vector $\mathbf{\Gamma}_m$ that satisfies

$$E(\mathbf{\Gamma}_m) = 0, \quad \text{Cov}(\mathbf{\Gamma}_m) = \mathbf{I}_m, \tag{3.10}$$

conditions which are essential to maintaining the covariance structure in the covariance-adjusted skew-Gaussian (CASG) process.

A natural starting point is the univariate closed skew-normal (CSN) distribution, parameterised as

$$\eta_j \sim CSN_{1,1}(0, 1, \alpha, 0, 1), \tag{3.11}$$

where α is the skewness parameter. This yields a zero-mean, unit-variance skew-normal variable. Since the joint distribution of i.i.d. CSN variables is also CSN, one can extend this to the multivariate case:

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^T \sim CSN_{m,m}(0, \mathbf{I}_m, \alpha\mathbf{I}_m, 0, \mathbf{I}_m). \tag{3.12}$$

A crucial step is to standardise $\boldsymbol{\eta}$ so that its mean is zero and its covariance is the identity matrix. The transformed random vector is defined as

$$\mathbf{\Gamma}_m = \frac{1}{\sqrt{\nu}}\boldsymbol{\eta} - \frac{c}{\sqrt{\nu}}\mathbf{1}, \tag{3.13}$$

with

$$c = E(\eta) = \sqrt{\tfrac{2}{\pi}}\delta, \quad \nu = 1 - \tfrac{2}{\pi}\delta^2, \quad \delta = \frac{\alpha}{\sqrt{1+\alpha^2}}. \tag{3.14}$$

This construction ensures

$$\mathbf{\Gamma}_m \sim PCSN_m(0, \mathbf{I}_m, \alpha), \tag{3.15}$$

so that $\mathbf{\Gamma}_m$ is a zero-mean vector with identity covariance, where the single parameter α governs skewness.

In summary, the PCSN distribution provides a parsimonious and interpretable way to incorporate skewness into Gaussian processes. It avoids overparametrisation, preserves the spatial covariance structure, and eliminates identifiability problems,

thereby making inference both computationally efficient and statistically robust.

3.2.2.2 Replacing the CSN with SUN Distribution Process

The authors of the GPSCF model, rather than working directly with the CSN distribution, adopt the Unified Skew-Normal (SUN) distribution. The SUN is essentially a reparameterisation of the CSN, within which the PCSN appears as a special case. This reparameterisation is advantageous because the SUN distribution possesses several properties that the CSN lacks.

Most importantly, the SUN is closed under linear combinations: any linear combination of observations from a SUN vector also follows a SUN distribution. This closure property is crucial, since marginalisation and conditioning are the backbone of Gaussian process inference. Moreover, the SUN provides a clearer separation between covariance and skewness, thereby facilitating inference. Specifically, the distribution can be written as:

$$Y^* \sim SUN_{m,q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}^*, \boldsymbol{\nu}, \boldsymbol{\Delta}^*). \tag{3.16}$$

Here, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ retain their usual interpretation from the multivariate Gaussian case, while skewness is introduced explicitly through the parameter $\boldsymbol{\Gamma}^*$.

The asterisks in $\boldsymbol{\nu}$ and $\boldsymbol{\Delta}^*$ denote a crucial distinction from the standard CSN parameterization. In CSN, the analogous parameters $\mathbf{v}$ and $\boldsymbol{\Delta}$ are embedded nonlinearly through the matrix $\mathbf{D}$ in the density formula, leading to over-parameterization and identifiability problems. The SUN reparameterization explicitly decouples these components through the notation (*) to signal: (i) closure under linear transformations, essential for GP marginalization and conditioning; (ii) explicit separation of scale and skewness; and (iii) elimination of redundant parameters. This structural clarity and computational tractability are why the SUN is preferred over CSN for Gaussian process inference.

This separation ensures that the covariance structure remains unaffected by skewness, making the model both more interpretable and more stable in numerical esti-

mation. In practice, this means that skewness may be adjusted independently of covariance, thereby avoiding distortion of second-order dependence. Because the parameters are more clearly identifiable, optimisation is also more efficient.

Mathematical Distinction between CSN and SUN.

The Closed Skew-Normal (CSN) is defined by its probability density function:

$$P_Y(\mathbf{y}) = c_p \, \phi_p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \, \Phi_q\big(\mathbf{D}(\mathbf{y} - \boldsymbol{\mu}); \mathbf{v}, \boldsymbol{\Delta}\big), \tag{3.17}$$

where the density is a product of a p-dimensional multivariate normal PDF and a q-dimensional multivariate normal CDF. This structure embeds skewness through the matrix $\mathbf{D}$ and vectors $\mathbf{v}$, which control the interaction between the Gaussian part and the truncation part.

In contrast, the Unified Skew-Normal (SUN) follows a different construction:

$$Y^* = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \left(\mathbf{z} + \boldsymbol{\Gamma}^* \, w \right), \tag{3.18}$$

where $\mathbf{z} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$, $w \sim N^+$ (standard normal truncated to positive reals), and $\boldsymbol{\Gamma}^*$ is a $m \times 1$ vector controlling skewness.

The fundamental difference is that:

- **CSN**: Skewness enters through truncation of a multivariate Gaussian in the density formula (nonlinear mixing of parameters).
- **SUN**: Skewness enters as an additive perturbation in the stochastic representation (explicit separation of scale $\boldsymbol{\Sigma}$ and skewness $\boldsymbol{\Gamma}^*$).

Mathematically, this implies: if $\mathbf{A}Y$ is computed for CSN, the parameters of the resulting distribution change in complex, nonlinear ways. For SUN, $\mathbf{A}Y^* + \mathbf{b}$ remains SUN with transformed parameters, providing closure under linear operations. This closure property is a direct consequence of the additive structure in Equation 3.18. The CASG process inherits the closure properties of the SUN distribution, ensuring that the finite-dimensional joint distributions remain tractable for Gaussian process inference. For technical details on the construction and properties, see Khaledi et al. (2023).

3.2.3 Summary of the GPRSE and GPSCF frameworks

The two frameworks described above (GPRSE and GPSCF) provide a principled means of extending Gaussian processes to accommodate skewness. Both rely on the CSN distribution (or its generalisation, the SUN) owing to its flexibility, yet they differ in their respective limitations.

The GPRSE framework, while conceptually straightforward, suffers from identifiability problems arising from its highly parameterised form, and faces computational challenges due to the derivatives required in the mean function expression (see Equation 3.5). In contrast, the GPSCF framework addresses these difficulties by relying on parsimonious distributions and introducing skewness through a truncated Gaussian vector. By exploiting the closure of the SUN family under linear transformations, the GPSCF model ensures that both marginalisation and conditioning remain tractable, which is crucial for Gaussian process inference.

The principal contribution of the GPSCF approach lies in its explicit separation of covariance and skewness: the covariance kernel retains its conventional interpretation, while skewness is introduced through a small number of parameters. This leads to more stable inference, reduces identifiability concerns, and allows skewness to vary across space in a computationally efficient manner. Consequently, the GPSCF may be regarded as a natural generalisation of Gaussian processes, with the capacity to capture richer distributional features in spatial and spatio-temporal applications.

Nonetheless, thinking a multi-fidelity extension to such a framework is potentially complex. Questions may arise such as whether the truncated skewness vector should be introduced only in the LF process or also in the discrepancy process.

3.3 Indirect modelling of skewness (warping)

Another popular approach for dealing with skewness is transforming the observation space, an operation referred also as "warping". Warping can be done in a parametric fashion as explored by the paper of Snelson et al. (2003), or in a non-parametric way, as seen in Agou et al. (2022).

3.3.1 Parametric warped Gaussian process

The warped GP of Snelson et al. (2003) is a generalization of the standard Gaussian process, a transformation of the observation space (warping) is made automatically within the GP framework. The main idea is that the observations are mapped into a latent space such that the GP is best modeled in the latent space. A warped Gaussian process is based on the following classical statistical theorem of a transformation of a random variable (see details in Chapter 2 of Casella and Berger (2021)). Let Y being a random variable with PDF $f_k(y)$ and let the scalar a being $a = M(Y)$, where M is a monotone transformation. Let $f_k(y)$ be continuous over its domain and $M^{-1}(a)$ have continuous derivative on the a-domain $D_{\mathbf{a}}$. Then the PDF of $\mathbf{a}$ is:

$$\begin{cases} f_{\mathbf{a}}(\mathbf{a}) = f_k(M^{-1}(\mathbf{a}))\left|\frac{d}{da}M^{-1}(\mathbf{a})\right| & \mathbf{a} \in D_{\mathbf{a}} \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

The vector $\mathbf{a}$ can be consider a vector of latent target observations modeled by a GP. Consider also that M could be any monotonic increasing transformation parametrised by Ω such that:

$$a_i = M(y_i, \Omega), \tag{3.20}$$

where each observation y_i is mapped to a_i for each $i = 1, \cdots, n$. In this case, the negative log-likelihood of a Gaussian process modelling y now becomes:

$$-\log(P(\mathbf{a}_N|\mathbf{X}_n, \Omega)) = \frac{1}{2}\log(det(\mathbf{K}_n)) + \frac{1}{2}M(\mathbf{y})^\top \mathbf{K}_n^{-1} M(\mathbf{y}) - \sum_{n=1}^{N} \log(\frac{d}{dy}M(\mathbf{y}))\|_{y_i} + \frac{n}{2}\log(2\pi).$$

The negative log likelihood is identical to the standard Gaussian process likelihood, except for an additional transformation parameterized by Ω and its derivative. The $\mathbf{X}_n$ is an input matrix of n-rows. While, $\mathbf{K}_n$ is the variance covariance matrix, with subscript emphasizing the number of rows. In particular, the predictive density in the observation space is obtained by change of variables from the Gaussian

predictive density in the warped space:

$$p(y_{n+1} \mid \mathbf{X}_{n+1}, D, \Omega) = \frac{|M'(y_{n+1})|}{\sqrt{2\pi\, v_{n+1}}} \exp\left(-\frac{(M(y_{n+1}) - a_{n+1})^2}{2\, v_{n+1}}\right), \tag{3.21}$$

where a_{n+1} and v_{n+1} are the mean and variance of the standard GP predictive distribution in the warped space at $\mathbf{X}_{n+1}$, and M is the monotone warping function. The shape of the resulting distribution depends on the chosen warping function, which typically induces skewness and multimodality. An important point of discussion is therefore the selection of this warping function. A classical choice is the hyperbolic tangent:

$$M(t_i, \mathbf{\Omega}) = y + \tanh\left(\sum_{i=1}^{I} a_i, \tanh(b_i(y + c_i))\right), \tag{3.22}$$

where $a_i, b_i \geq 0$ for all i. In this case, $\mathbf{\Omega} = [a, b, c]$, where a, b, and c respectively control the size of the steps, their steepness, and their position. The parameter I denotes the number of steps: increasing I introduces additional tanh components, which makes the warping function more flexible and capable of representing more intricate shapes. The derivatives of this function with respect to the parameters $\mathbf{\Omega}$ or y_i are straightforward to compute. It should be noted that the tanh function is bounded, and hence its inverse does not exist outside the interval defined by these bounds. In the observation space this would not yield a proper density. The linear term y at the beginning of Equation 3.22 compensates for this issue.

In conclusion, parametric warping provides a viable approach for handling skewed distributions. By transforming the input space, it alters the evaluation of the covariance function, since observations that were initially close may appear farther apart after warping. Another key advantage is that the data preprocessing is learned rather than arbitrarily imposed, thereby minimising subjective bias in the analysis.

3.3.2 Non-parametric warping of Gaussian processes

Agou et al. (2022) propose a novel method for modelling environmental phenomena that may exhibit skewness. While their original work applied this technique

to rainfall data, its flexibility makes it equally suitable for wind speed analysis. This methodology is referred to as the Data-Driven Warped Gaussian process (DDWGP) model. This method involves kernel-based estimation of the cumulative distribution function (CDF) and spatial interpolation of normalised observations. Therefore, by construction, is non-parametric. The DDWGP model is specifically designed for spatio-temporal data and non-Gaussian distributions, and it is based on a non-parametric data-driven approach, providing the model with considerable flexibility for various application types. The model's initial steps involve utilizing a kernel density estimator, which can be expressed as follows:

$$\hat{f}_K = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_{[i]}}{h}\right), \tag{3.23}$$

where $\hat{f}_K$ is the estimated probability density function of a vector of observations y, $y_{[i]}$ is the i-th order statistic, h is the bandwidth and K is a kernel. Such estimate $\hat{f}_k$, is integrated as follows:

$$\hat{F}_K = \int_{-\infty}^{y} \hat{f}_k(y')dy', \tag{3.24}$$

to obtain the CDF estimate $\hat{F}_K$. Different choices for the kernel function are possible to obtain an explicit estimate of the CDF depending on the dataset under study. In general, when you apply the CDF to a variable Y the result is a new variable that follows a uniform distribution. While this uniform distribution does not keep the original data's values, shape, or spread, it does preserve the ranks — the order of the data points. These ranks contain important information and can help us understand and even reconstruct the original distribution. For example, if the original data was right-skewed, its CDF would increase slowly at first, then quickly in the middle, and slowly again near the end. On the other hand, if the original data was symmetric, the CDF would grow more evenly, almost in a straight line. This shows how the shape of the CDF reflects the structure of the original data, even though the values have been transformed.

The estimated CDF can be used to create a warping transformation from the original

data y to a new space of values a[1]. This transformation is written as $p(s) = M(y(s))$, where a is the transformed value and s is a point in the domain $\mathbb{R}^d$. More specifically, the transformation is given by:

$$a_i = \Phi^{-1}(p_i) = \Phi^{-1}(F_K(y_i)) = M(y_i(s)),$$

where Φ^{-1} denotes the inverse cumulative distribution function (quantile function) of the standard normal distribution, $F_K(y_i)$ is the marginal cumulative distribution function evaluated at the observation $y_i(s)$, and $p_i = F_K(y_{[i]}) = P(Y < y_{[i]})$ represents the probability level, or percentile, associated with y_i. This transformation converts the original observations into normal scores, meaning the transformed values a_i appear as if they come from a standard normal distribution. The use of the subscript i is for emphasising the idea of order statistics. This allows standard Gaussian regression to be applied in the a-space, producing predictions $\hat{\mathbf{a}}$. To recover predictions on the original data scale, the inverse transformation is applied:

$$\hat{\mathbf{y}} = F^{-1}(\Phi(\hat{\mathbf{a}})).$$

This inversion is valid due to the quantile invariance property, which states that the quantiles of a distribution remain unchanged under monotonic transformations. Since the data in the warped space is symmetric (as in a normal distribution), the mean and median are equal.

3.3.2.1 Practical aspects of the DDWGP

The values returned by the CDF correspond to probabilities p_i at n discretisation points. Although these points establish a mapping between the original space and the transformed space, the relationship is "discrete" or "piecewise-defined". Since the method itself does not provide an explicit transformation, a simple device can be employed to obtain a system that more closely resembles a continuous mapping. The idea is to increase the resolution by considering $N \gg n$ discretisation points. To

[1]Note that a has been used as a general symbol for a "transformed variable". Specifically, it represents the response of the GPSCF, where a is a linear transformation $\mathbf{\Gamma}$; it also denotes the response of the parametric warping, and here it refers to the response after the non-parametric transformations.

this end, a dense grid of s-locations is selected, and the corresponding p_i values are obtained by linear interpolation.
In practice, linear interpolation takes y_i as input and produces p_i as output, interpolating at the grid target locations within the range

$$[\mathbf{y}_{\min} - h, \mathbf{y}_{\max} + h].$$

These values are then stored in a lookup table, which is used to compute the transformed values as

$$a_i = \Phi^{-1}(p_i), \quad i = 1, \ldots, N^d.$$

The lookup table facilitates both direct and inverse transformations, meaning the normalisation process is implicitly derived rather than explicitly defined.

3.4 Developing a new data-fusion method for skewed data

In the previous sections, different methodologies for addressing skewness in the context of GP regression have been explored. This section will delve into the approach that is most suitable to be integrated with multi-fidelity framework. In making such a choice it is necessary to balance the complexity of the approach itself, i.e. if the approach involves a complex parametrisation, and the effectiveness, which is the relative advantage compared to the others.
In this regards, the GPRSE appears to be overly parametrised, a concern that could be exacerbated when extending the model to a multi-fidelity version. Similarly, the method GPSCF relies on the selection of inducing points, which could pose an additional challenge when integrating it with a multi-fidelity framework, where high-fidelity points inherently carry greater importance than low-fidelity points.
The warping technique PWGP introduced by Snelson et al. (2003) has shown inconsistent effectiveness in some preliminary experiments. Intuitively, this can be seen

as an identifiability problem since the joint derivation of warping parameters and Gaussian process parameters may not always result in a perfect normalisation.

In parametric warping (Section 3.3.1), the warping function parameters Ω and the GP hyperparameters (e.g., length scale, variance) are estimated jointly by optimizing a single likelihood function. This joint optimization creates an identifiability problem. In contrast, the data-driven non-parametric approach (Section 3.3.2) decouples these steps: the warping transformation is inferred *independently* from the Gaussian process parameters, eliminating this identifiability issue

As discussed in Agou et al. (2022), this method can normalise any type of data provided that a sufficient number of observations are available. The algorithm integrates naturally into a multi-fidelity framework as a straightforward pre-processing step, which also makes its implementation simple. Furthermore, because the method is based on quantile ordering, it mitigates the data-fusion issues concerning inter-dataset relationships highlighted in Section 3.1.1.

In summary, while existing approaches either suffer from excessive parametrisation (GPRSE), selection of inducing points (GPSCF) or issues of identifiability (PWGP), the data-driven non-parametric warping emerges as the only method that is both parsimonious and naturally compatible with multi-fidelity frameworks. Its simplicity, robustness, and ability to preserve inter-dataset relationships make it particularly well suited for our problem setting. Motivated by these advantages, we propose the Warped Multi-Fidelity Gaussian process (WMFGP), which combines the flexibility of non-parametric warping with the structure of the MFGP described in Chapter 1.

3.4.1 A new non-Parametric Warped Multi-fidelity GP

By integrating the multi-fidelity Gaussian process with the data-driven warping approach, we introduce a new procedure termed the Warped Multi-fidelity Gaussian process (WMFGP). The method is designed to perform data-fusion tasks in a multi-fidelity context in the presence of skewed data.

The complete procedure is summarised in Algorithm 1. Note that the procedure involves a bandwidth estimation. In Step 1, the bandwidth h is estimated using

Algorithm 1 Warped Multi-fidelity Gaussian process

Require: Input data: $D = [\mathbf{y_H}, \mathbf{y_L}]$, n_D=number of datasets.

1: **for** $i = 1$ to n_D **do**
2: Perform kernel density estimation and estimate the bandwidth h.
3: Compute the kernel-based estimate of the CDF to obtain the probability levels p_i, using the sample values $\mathbf{y}$ of the time series, the selected kernel, and the estimated h.
4: Compute the normal scores by inverting $\Phi^{-1}(p_i)_{i=1}^{N} = \mathbf{a}$.
5: Interpolate the estimated CDF values (p_i) at the data points (either $\mathbf{y_H}$ or $\mathbf{y_L}$) to produce a dense grid of p_i across a range ($[\mathbf{y}_{\min} - h, \mathbf{y}_{\max} + h]$).
6: Generate a lookup table linking the actual data to their probability levels. The table contains a vector of a-scores corresponding to the interpolated p_i.
7: **end for**
8: Run the MFGP (see Section 1.3.4) on the normal scores $\mathbf{a}_L$ and $\mathbf{a}_H$.
9: Use the lookup table to back-transform the MFGP estimates.
10: **return** $\mu_H(\mathbf{s}^*)$, conditional multi-fidelity mean at the prediction location.

the **adaptive plug-in method** Botev et al. (2010). This method automatically selects an optimal bandwidth by minimizing the integrated squared error of the kernel CDF estimate:

$$h^* = \arg\min_h \mathbb{E}\left[\int \left(\hat{F}_K(x;h) - F_X(x)\right)^2 dx\right]. \tag{3.25}$$

The plug-in principle estimates bandwidth-dependent functionals of the distribution from the data itself, yielding a bandwidth that adapts to the sample's scale, skewness, and higher-order moments. This is particularly important for strongly skewed distributions like precipitation, where naive rule-of-thumb methods (e.g., Silverman's rule) provide inadequate smoothing. The computational implementation requires no manual tuning or cross-validation, making the procedure fully automatic and reproducible **?**.

It involves applying a non-parametric warping function to both the high-fidelity vector $\mathbf{y_H}$ and the low-fidelity observations $\mathbf{y_L}$. This warping process consists of several steps: kernel density estimation, reconstruction of the cumulative distribution function (CDF) based on the estimated kernel bandwidth, and inversion of the CDF probabilities p_i to obtain normalised observations $\mathbf{a}$. As outlined in Section 3.3.2, the method establishes a one-to-one correspondence between each y_i and its normalised counterpart a_i. However, it does not provide an explicit

functional form for the normalisation, and therefore no analytical inverse transformation exists. Instead, the inverse transformation is constructed implicitly by generating a dense grid over the interval $([\mathbf{y}_{\min} - h, \mathbf{y}_{\max} + h])$. This grid maps each value in the original space to its corresponding value in the normalised space via linear interpolation, with the query points given by the grid and the input points by the probability levels. In this way, the warping function is obtained implicitly. Both the forward mapping and the inverse mapping are enabled through this grid, providing a consistent framework for transformation between the original and normalised spaces. The procedure is repeated independently for each data sources. Since it is based on a quantile ordering the inter-dataset relationship is mostly preserved. Meanwhile, the disjoint normalisation procedure ensures an effective result for both data sources.

3.5 Data and motivating examples

In this section, the test-beds employed to assess the proposed approach are introduced. These comprise two phases: (i) controlled experiments based on simulated datasets, and (ii) an application to a real-world case study using the ARPA Lombardia dataset (see Section 1.5). The ARPA data are particularly challenging, as they contain long sequences of missing values, making them a suitable benchmark for evaluating a gap-filling procedure tailored to skewed data such as WMFGP.
Before the simulation experiments are presented, a motivating example based on the ARPA data is provided. This example illustrates how standard methods may fail to achieve adequate normalisation and, in doing so, distort inter-dataset relationships. Subsequently, two simulation experiments are described, followed by the ARPA case study. The simulations are designed to evaluate the efficacy of the procedure in the temporal domain, while the spatial extension will be explored in Chapter 4.

3.5.0.1 ARPA Lombardia data

To evaluate the proposed method, we use the ARPA Lombardia dataset described in Section 1.5.2. Not all monitoring stations provide complete records, and several

lack specific variables such as wind speed. Since this study focuses on wind-speed data as a representative case, we restrict the analysis to a subset of 94 stations. Their spatial distribution across the Lombardia region is shown in Figure 3.10. The stations collect high-quality measurements at 10-minute intervals. For the present analysis, however, we aggregate the data to hourly averages to emphasise broader temporal patterns, although higher-resolution studies would also be feasible.

3.5.1 Difficulties in the joint normalisation of multiple data sources

A main methodological contribution of this work is the development of a joint normalisation approach for skewed data across several datasets in a multi-fidelity framework. In standard preprocessing, transformations such as logarithmic or Box–Cox functions are often applied to reduce variance and approximate normality. However, in multi-fidelity settings, where datasets can differ in their distributional shape, applying the same transformation to all datasets is rarely effective. For example, a transformation that normalises one dataset well may fail to normalise another. On the other hand, applying different transformations to each dataset—for instance, a logarithmic function to one and a square–root function to another—can disturb the statistical relationships that are essential for multi-fidelity modelling.

This issue is especially clear in the Lombardia wind–speed datasets analysed here. In 38% of the cases, the Box–Cox λ parameters that worked best for normalisation were very different (indicating different transformations) between low–fidelity and high–fidelity datasets. While applying separate transformations may improve normality within each dataset, it also tends to distort the relationships between them, which are crucial for hierarchical and multi-fidelity regression models.

To overcome this problem, we propose a normalisation method based on order statistics. Instead of forcing the same transformation on all datasets, or letting each one have a completely different transformation, the method aligns the marginal distributions while keeping their rank–based structure consistent. In practice, this means that the normalised values retain similar quantile positions across datasets,

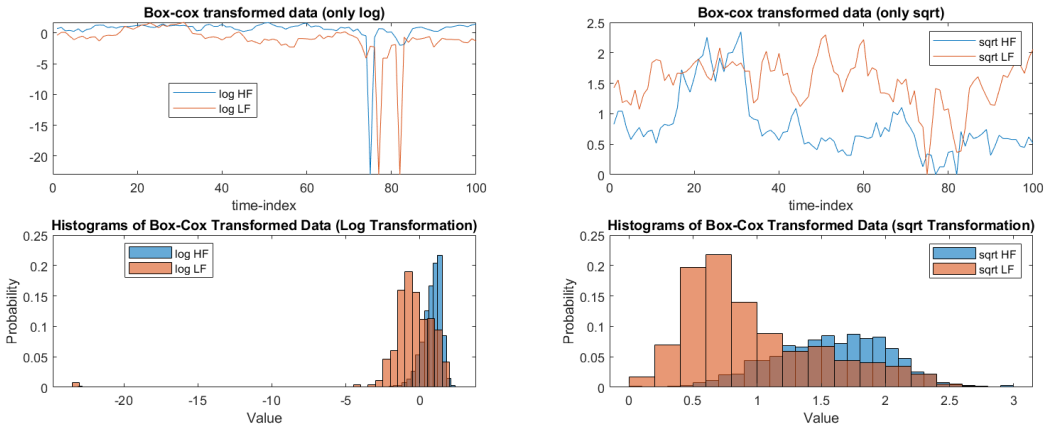


Figure 3.1: Effects of applying the same transformation (logarithmic or square root) to both HF (SAN) and LF (CO) dataset (see Section 3.5.1). The top panels show preserved relationships between the signals, while the histograms in the lower panels reveal that only one dataset is effectively normalised.

preserving the statistical connections needed for effective multi-fidelity integration.

The effect of standard transformations on correlation can be seen in a simple example. Two stations—San Siro Alpe Recascia (SAN) and Colico–Via La Madoneta (CO)—show a pre–transformation linear correlation coefficient of 0.53. Based on the Box–Cox method, SAN is best normalised using a logarithmic transformation ($\lambda = 0.2$), while CO is best with a square–root transformation ($\lambda = 0.5$). If these transformations are applied independently, the correlation falls to 0.34, showing a clear loss of inter–dataset structure. Using the same transformation for both datasets maintains the correlation better, but then one of the datasets remains poorly normalised (see Figure 3.1).

Our proposed method achieves both goals at once. It preserves the relationship between datasets, with correlation changing only slightly (from 0.53 to 0.52), and it normalises both datasets effectively. Figure 3.2 shows this effect. The left panel displays the original signals, with overlapping patterns around time indices 30–40 and 80. After independent Box–Cox transformations, these overlaps are lost. With our order–statistics method, the overlapping patterns remain, and both datasets show improved marginal symmetry. These results demonstrate that the proposed approach provides a reliable alternative to classical transformations for normalisation in multi-fidelity analysis.

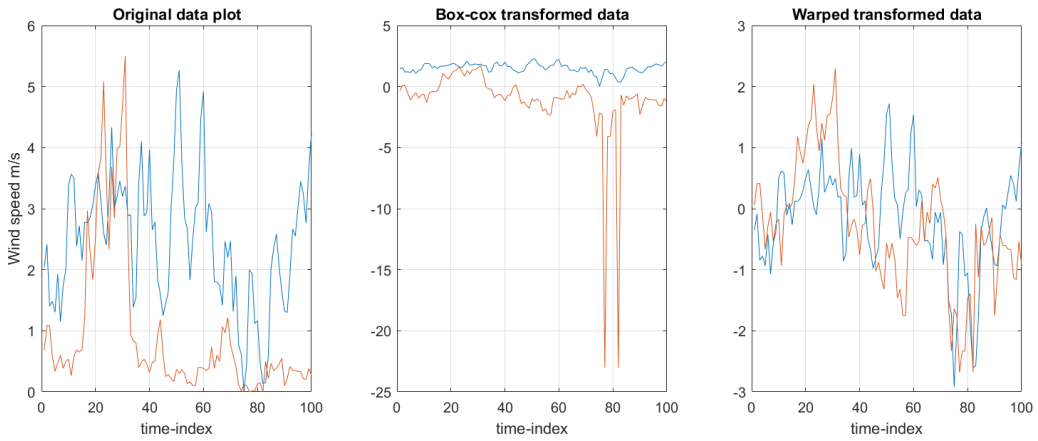


Figure 3.2: Comparison of original signals (left), signals normalised independently using Box-Cox transformations (center), and signals normalised using the proposed method (right). The proposed method maintains overlapping structure while achieving effective normalisation.

3.6 Validation: simulations and real world case study

In this section, we present two simulation studies demonstrating the effectiveness of the WMFGP in handling skewed data. The first experiment evaluates the model's performance in the presence of randomised patterns of missing information in a HF time-series, referred to as "randomised missingness". The randomised missingness can be thought as the presence of random occuring missing values in a time-series, which is a very common problem in many application, see Colombo and Fassò (2022) and Fassò et al. (2020) for some examples. The second experiment addresses structural missingness, involving long sequences of missing data in a time-series, particularly focusing on data from ARPA Lombardia monitoring stations.

3.6.1 First simulation design: randomised missingness

Five models were tested across different levels of skewness using a data generation procedure designed to replicate wind–speed data. The models include the standard Gaussian process (GP), which is a classic univariate method for gap filling (see Colombo and Fassò (2022), Fassò et al. (2020)); the Warped Gaussian process (WGP), which allows us to assess whether the univariate version equipped with warping is advantageous; the classic multi–fidelity GP (MFGP), since here an extension of this model

is developed; a multi–fidelity model integrated with the Box–Cox transformation (BCMF), to provide a comparison with other transformation–based approaches; and the proposed WMFGP.

Tests were conducted on two distributions—the Weibull and the CSN—and two levels of skewness (high and low), resulting in four experimental scenarios. An example of the noise generated from the Weibull distribution is shown in Figure 3.3, while an example of noise generated from the CSN distribution is shown in Figure 3.4. For each scenario, the experiment was replicated 200 times, with a unique missing-data pattern simulated in every run. The ratio $\frac{D_{HF}}{D_{LF}}$ was fixed at 10%, which represents the approximate upper bound for which the advantage of the MFGP approach was highlighted in Section 2.4.1. The time series were generated as follows:

$$\mathbf{y_L} = \mathbf{T} + w_L,$$
$$\mathbf{y_H} = \mathbf{T} + w_H,$$

where $\mathbf{T}$ denotes the deterministic component, obtained by averaging the ERA5 data (see Section 2.4). The term w_L is the error for the low–fidelity data, generated either from a Weibull or a CSN distribution, and w_H is the corresponding error for the high–fidelity data, defined to be smaller by construction. Skewness is therefore introduced explicitly in the error component.

This design is not arbitrary; it is deliberately chosen to highlight properties of Gaussian processes. Although the preceding discussion has focused on ways of overcoming the so-called "skewness limitation" of GPs, it is important to emphasise that Gaussian processes can, in fact, be used to model skewed raw data. The normality assumption in GPs applies to the joint distribution of the latent functions (in this case, those generating each t_i) together with the noise term. Consequently, the marginal distribution of the observed data may still appear skewed. In GP modelling, the location of the skewness is of particular theoretical importance. Modifying the error terms in the data-generating process is therefore the most

Table 3.1: Parameters of the CSN distribution for the different skewness scenario.

Parameters	**Low skewness** w_H , w_L	**High skewness** w_H , w_L
μ	-0.25,-0.5	-0.25,-0.5
Σ_1	0.04,0.8	0.8,2.4
Γ	4,4	50,50
ν	2	2
δ	3	3

Table 3.2: Summary of statistical measures and generating parameters of the errors generated from the Weibull distribution, for different skewness scenario.

High Skewness	**Error**	**Scale**	**Shape**	**Mean**	**SD**	**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

direct means of introducing skewness and, in turn, of testing the Gaussian assumptions of the model. Tables 3.1 and 3.2 report the exact parameters used for error generation.

The results of such a simulation experiment are depicted in Figures 3.5 and 3.6. The two figures compare the five selected models taking the box plots of each Mean Absolute Error (MAE) of the experiments. The experiment found that the WMFGP was the best model in scenarios with high skewness. But when the error skewness was low, it performed similarly to the standard MFGP. This indicates the advantage of the method is particularly prominent when the skewness is high. The outliers in the boxplot represent runs with numerical instability.

The model based on Box-Cox transformation was less robust to data generated with Weibull error compared to data generated with a CSN distribution. This is because the Weibull error was noisier and more skewed than the CSN counterpart, so, while Box-Cox transformation can work well in some situations, it struggles with more complicated ones. Naturally, the simpler univariate GP showed very poor performances in any of the analysed scenarios.

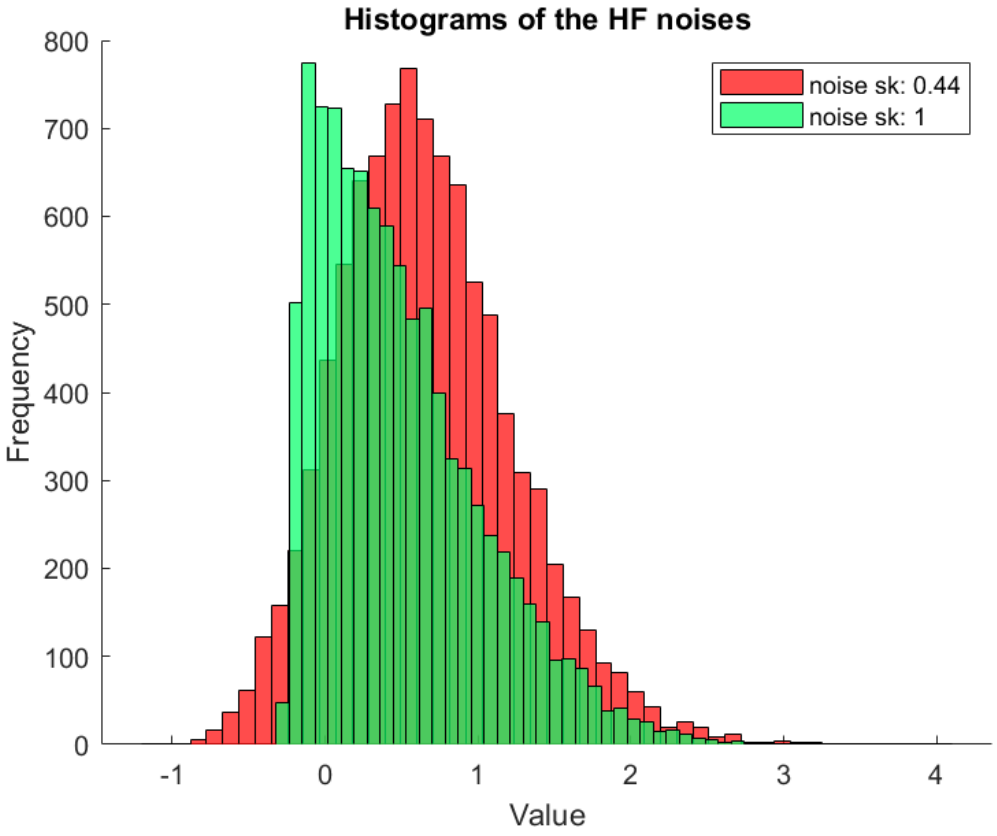


Figure 3.3: Examples of the noises generated from a CSN distribution for high and low skewness scenarios. In green the high skewness and in red the high skewness. Sk stands for skewness.

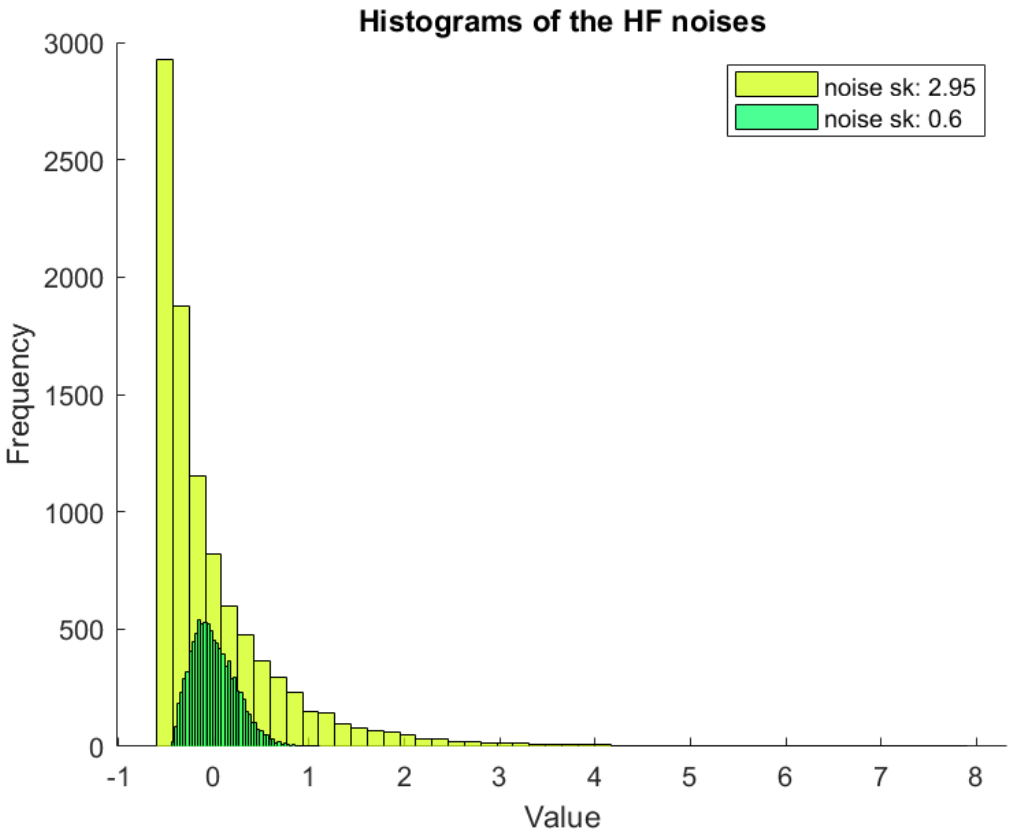


Figure 3.4: Examples of the noises generated from a Weibull distribution for high and low skewness scenarios. The figure reports also the skewness of noise distributions. In green the low skewness. In yellow the high skewness. Sk stands for skewness

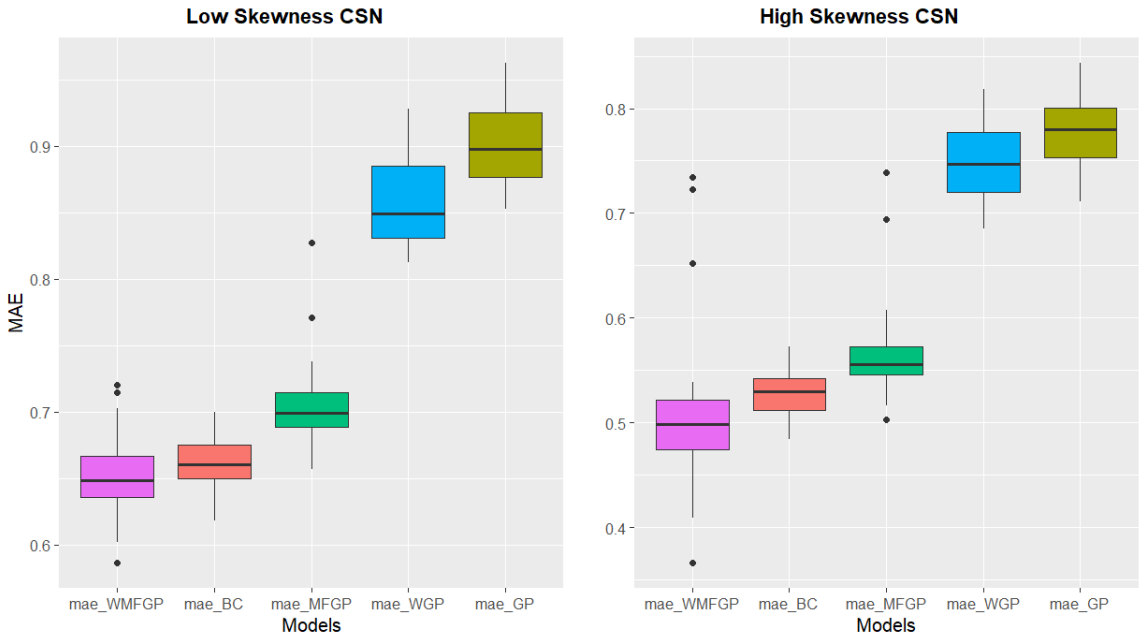


Figure 3.5: The figure depicts the results of the simulation experiment of Section 3.6.1 conducted using error generated from a CSN distibution. The models are ranked based on their MAE. On the x-axis the labels of each model: Warped Multifidelity (WMFGP), BOX COX Multifidelity (BC), Multifidelity (MFGP), Warped Gaussian Process (WGP) and Gaussian process (GP).

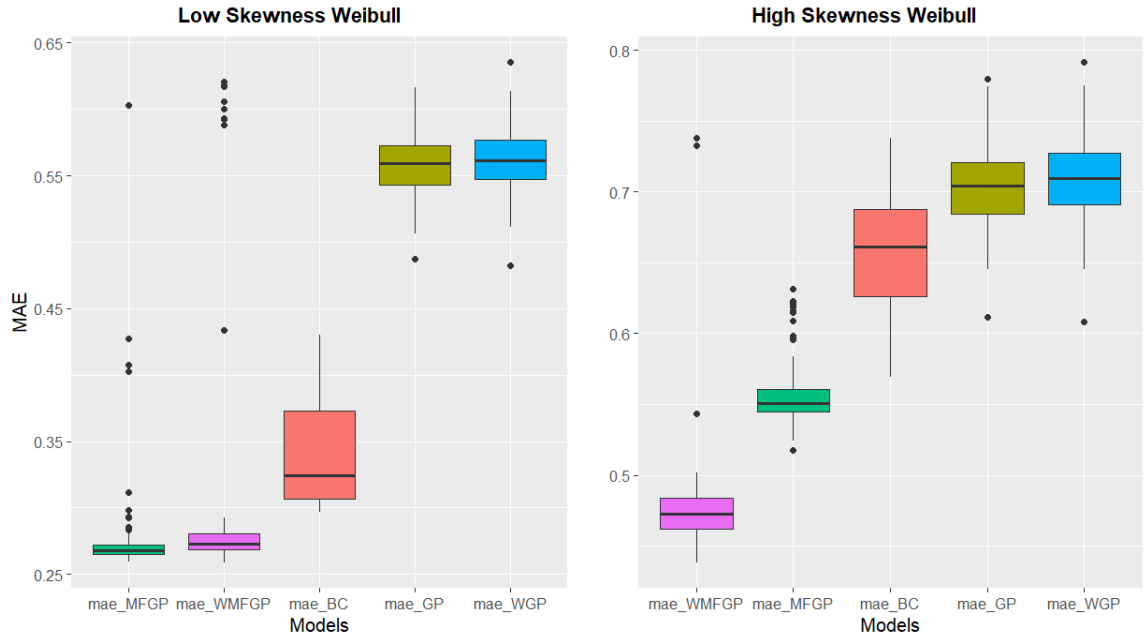


Figure 3.6: The Figure depicts the results of the simulation experiment of Section 3.6.1 using error generated from Weibull distribution. The models are ranked based on their MAE on the y-axis. On the x-axis the labels of each model: Warped Multifidelity (WMFGP), BOX COX Multifidelity (BC), Multifidelity (MFGP), Warped Gaussian Process (WGP) and Gaussian process (GP).

3.6.2 Second simulation design: structural missingness

The second simulation experiment is meant to deal with the problem of structural missingness or in other words the presence of long missing sequences in a given time-series. This problem is particularly challenging as standard interpolation, i.e. Gaussian process regression fail to give accurate estimates for data falling outside the Gaussian process range[1]. A practical example of such limitation has been shown in Figure 2.16, where it is shown that for the different ranges the prediction returned from GP regression after a certain gap distance (distance from the last observation) fall to just an average. This is coherent with the interpolation study of Colombo and Fassò (2022), where it is shown that the interpolation uncertainty increases with the gap size. In this second simulation study the following models are compared: GP, Surrogate LF, MFGP, WMFGP, and Simple Imputation (SI). The Box-Cox model was removed as it was proven to be not effective in the previous simulation experiment and replaced with SI, which is a fast imputation method based moving average window of the nearest neighbor observations. We include also the surrogate LF, which

[1]The range defined through the length scale parameters.

Table 3.3: The table contains medians of MAE performances and their standard deviations for each simulated missing sequence. ML stands for missing sequence length.

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

is simply the difference recorded between the HF and LF signal in the test region. This latter score represents a benchmark performance, as ideally we would like our methods to be at least better then the original discrepancy between HF and LF data. For this experiment, we worked with ARPA Lombardia dataset, focusing on hourly resolution as we interested in recovering general patterns. We simulated gaps of five fixed dimensions: 24, 48, 72 , 96 and 192 hours, and replicated the experiment 500 times. To find a time-frame free of real gaps, we got a sub-sample of 72 stations out 94. The results of this experiment are contained in table 3.3, while an example of gap generated using such framework is given in Figure 3.7. In such a figure, in blue the LF data, in red HF data, in yellow the interpolation area.

The table presents the median values of MAE performances along with their standard deviations for each simulated missing sequence length. For a missing sequence length of 24, WMFGP performed the best with a MAE median of 0.46, closely followed by MFGP with an MAE median of 0.55. For a missing sequence length of 48, again WMFGP exhibited the lowest MAE median of 0.49, with MFGP closely trailing at 0.57. For lengths of 72 and 96, WMFGP consistently showed the lowest MAE medians among all methods. As the missing sequence length increased to 196, the performance of all methods degraded (beside the multi-fidelity), with Simple Imputation exhibiting the highest MAE median of 1.71, followed by GP with an MAE median of 0.84. Overall, WMFGP consistently outperformed other methods in handling missing data across different lengths of missing sequences, showing its robustness in such scenarios. These experiments demonstrate that MF models are particularly well suited for gap filling. Their effectiveness relies on the observation that relationships between datasets over time tend to remain relatively stable. An empirical analysis of the datasets, examining how pairs of

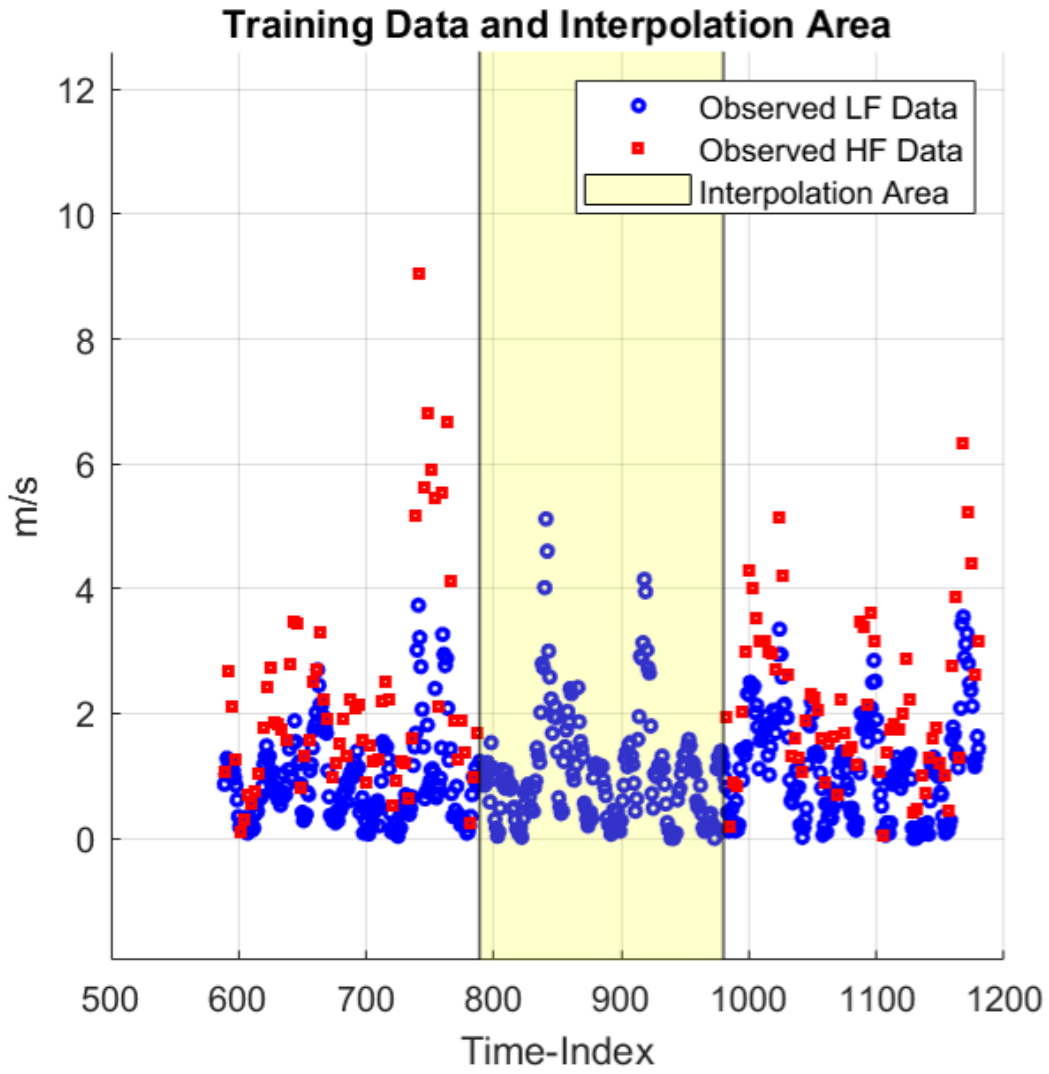


Figure 3.7: Example of gap generated in the second simulation experiment. The yellow area highlight the interpolations target.

time series evolve jointly, supports this assumption. Moreover, the simulation experiments (see Table 3.6.2) indicate that the uncertainty in the performance of both MF and WMFGP models is largely independent of the gap size. For instance, a performance value of 0.46 for WMFGP with a missing length of 24 is essentially equivalent to 0.45 when the missing length increases to 196.

The differences between MFGP and WMFGP are not striking in magnitude, yet they are consistent across cases. Figure 3.8 presents a comparison of the two approaches, applied to the interpolation gap shown in Figure 3.7. It can be observed that WMFGP remains systematically closer to the test data (red squares) than MFGP, particularly in regions where the wind–speed series exhibits sharp spikes (e.g., between time indices 845–850, 850–855, 870–875, 890–900, and 915–920).

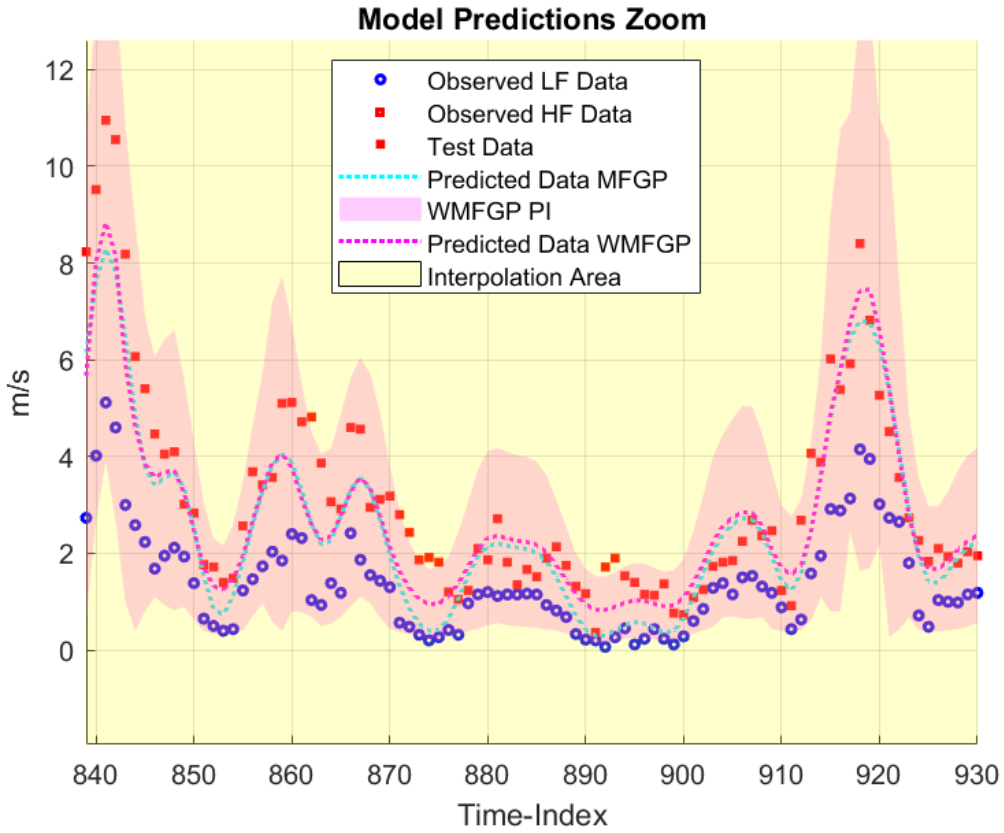


Figure 3.8: Comparison of MFGP and WMFGP interpolation performance over the gap illustrated in Figure 3.7. Red filled squares (■): high-fidelity (HF) observed data. Blue circles (○): low-fidelity (LF) observed data. WMFGP predictions are shown as magenta dashed lines with asymmetric credible intervals (shaded magenta), while MFGP predictions are shown as light blue dashed lines.

3.6.3 Real world case study

The WMFGP and MFGP were applied to recover the real missing sequence in the ARPA Lombardia dataset. The method capitalises on the inherent temporal correlation among time-series data observed in close proximity, leading to a highly efficient computational strategy that is easy to implement. However, by excluding the spatial dimension from the multi-fidelity aspect, we risk encountering endogeneity issues, as external factors such as orography, geography, and land cover can significantly impact the model's performance. To mitigate this potential concern, we conducted a clustering experiment based on the latitude, longitude, and altitude of the monitoring stations.

The strategy can be summarized in the following steps:

1. Performed a constrained K-means (Bradley et al., 2000) to identify clusters based on spatial information (latitude, longitude, and altitude of the stations). A constrained K-means allows users to define groups with predefined maximum and minimum numbers of elements. We constrained clusters to contain between 2 and 6 stations: a cluster with fewer than two stations would be uninformative, while more than six stations might introduce excessive noise from distant or dissimilar locations.

2. Once clusters were identified, select a station within a cluster where missing data recovery is desired. The time-series measured at this location represents the high-fidelity (HF) dataset.

3. From the same cluster, randomly select another station; this time-series represents the low-fidelity (LF) dataset.

4. Repeat the procedure for any desired time-series.

This procedure ensures that geographical characteristics guide the pairing of stations, thereby reducing the risk of endogeneity. Simultaneously, the random selection of low-fidelity series within each cluster prevents systematic pairing biases, ensuring results are not driven by arbitrary or deterministic station matches.

Optimal Cluster Determination: To determine the optimal number of clusters, we employed three complementary metrics: the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The **Silhouette Score** (ranging from -1 to 1) measures cluster cohesion by comparing intra-cluster to inter-cluster similarity; higher values indicate better-separated clusters. The **Calinski-Harabasz Index** quantifies the ratio of between-cluster to within-cluster variance, with higher values indicating superior cluster separation. The **Davies-Bouldin Index** (lower is better) measures the average similarity between each cluster and its nearest neighbor, emphasizing cluster distinctness.

Our analysis evaluated cluster counts from $k = 3$ to $k = 26$. Results reveal a robust performance plateau across the range $k \in [22, 26]$, with metric differences less than 3%:

- Silhouette Score: 0.4520 ($k = 22$) to 0.4578 ($k = 24$), declining modestly to 0.4098 at $k = 26$
- Calinski-Harabasz Index: peaks at $k = 24$ (92.08), remaining competitive at $k = 26$ (91.90)
- Davies-Bouldin Index: reaches its minimum (best value) at $k = 26$ (0.6082)

Based on this comprehensive analysis, supported by the original elbow plot indicating an inflection point at $k = 26$, we selected **26 clusters** as the optimal partition. This choice is justified by: (1) Davies-Bouldin optimality for cluster distinctness, (2) negligible degradation in other metrics ($< 3\%$), and (3) consistency with the elbow method. Figure 3.9 presents the diagnostic plot across all evaluated metrics.

The 26-cluster partition yields an average cluster size of 3.6 stations, with individual cluster sizes ranging from 3 to 6 elements. These cluster sizes are advantageous for two reasons. First, they provide redundancy: if the same missing sequence occurs simultaneously at two nearby stations, an alternative station within the same cluster can still be selected as the auxiliary low-fidelity source. Second, restricting the fusion step to one randomly selected companion station per run avoids introducing an informed or deterministic pairing strategy that could bias the assessment of the method. In other words, the purpose of clustering is not to exploit all stations

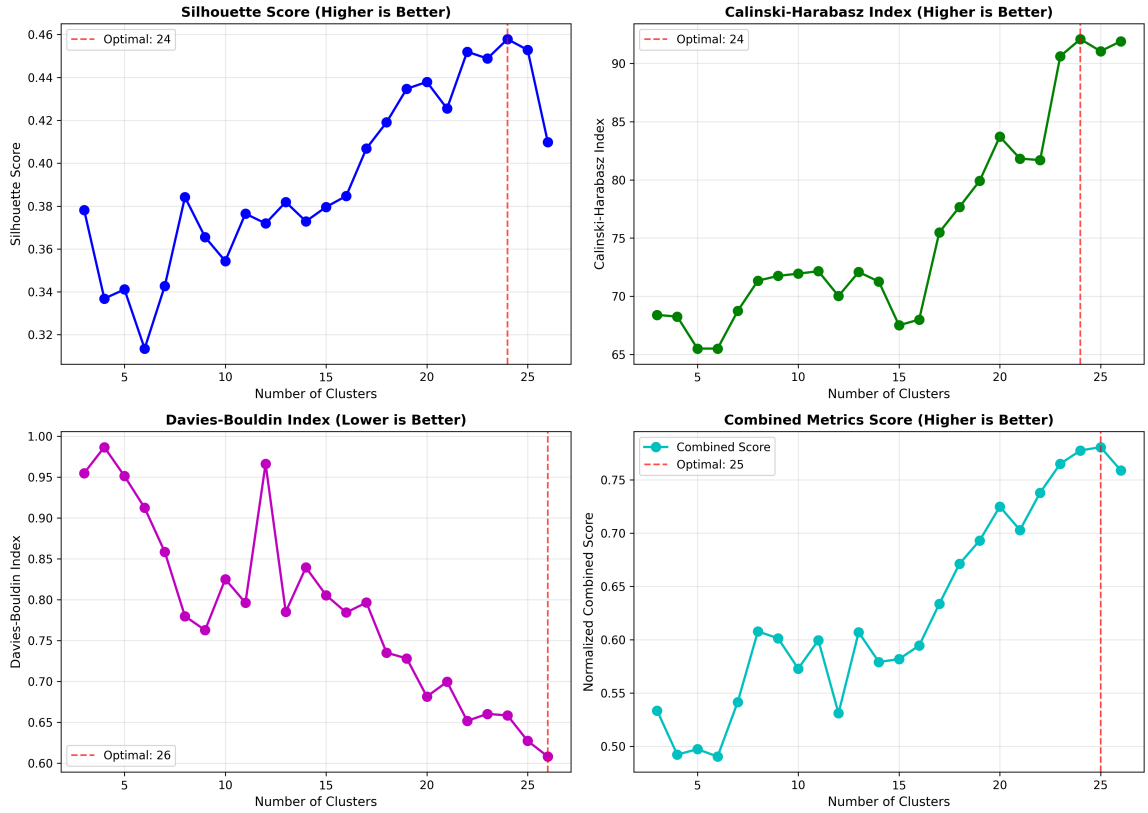


Figure 3.9: Clustering quality metrics for $k \in [3, 26]$ clusters. *Top-left:* Silhouette Score (higher is better). *Top-right:* Calinski-Harabasz Index for cluster separation (higher is better). *Bottom-left:* Davies-Bouldin Index for cluster distinctness (minimum at $k = 26$). *Bottom-right:* Combined normalized score. The robust plateau region ($k \in [22, 26]$) demonstrates consistent high performance with metric differences $< 3\%$.

simultaneously, but to define a pool of geographically comparable candidates from which a low-fidelity series can be sampled. This design makes the evaluation more conservative, as it prevents performance gains from being driven by ad hoc station selection rather than by the robustness of the fusion framework itself. Additionally, smaller clusters reduce the likelihood of grouping stations influenced by different environmental phenomena.

An example of real imputation performed using this strategy is shown in Figure 3.11, which illustrates the reconstruction of observational gaps for the Veddasca Monte Cadrigna station. In the figure, the blue points correspond to the LF data, the red points to the HF data, while the predictions are represented by the MFGP (light blue line), the WMFGP (magenta line), and SI (light green line). As this case depicts an actual interpolation of missing data, the true values remain unknown; nevertheless, some observations can be made.

Based on a simulation experiment (see Section 3.6.2), where we collected informa-

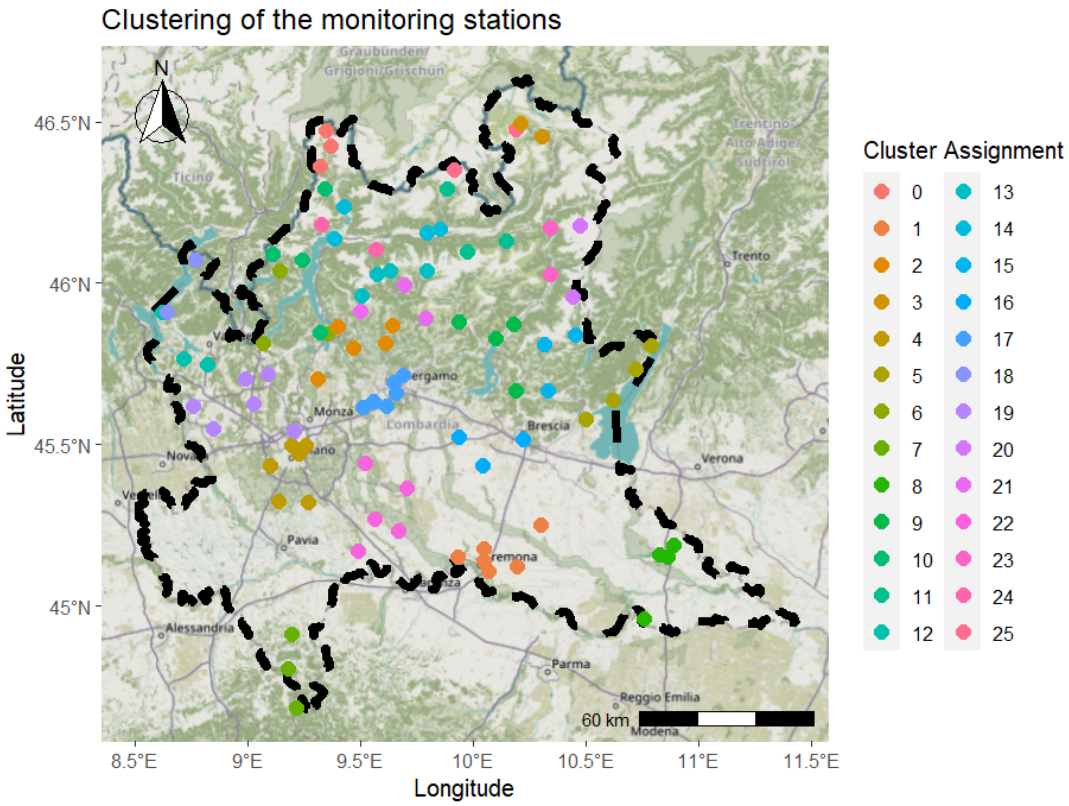


Figure 3.10: Lombardia boundary and results of the clustering experiment. Each dot depicts the position of a monitoring station, colored by cluster assignment.

tion about the missing interpolation, it was found that, for stations with similar altitude, longitude, latitude, and LF–HF correlation, the Mean Absolute Error (MAE) of WMFGP and MFGP was approximately 30% lower than that of the Surrogate. Moreover, WMFGP achieved a 13% lower MAE compared with MFGP. These results are consistent with Figure 3.11, where both multi-fidelity models successfully captured the seasonal pattern, but WMFGP demonstrated greater adaptability than MFGP. For instance, in the yellow-highlighted interpolation region, WMFGP reproduced two harmonics—consistent with the LF data—whereas MFGP produced a stronger smoothing effect with a single harmonic. The SI produced erratic predictions due to its inherent method construction as an average of nearby observations.

3.6.4 Linking model Performance to Geographical Properties

The purpose of this subsection is to investigate the relation between models performances and geographical properties. In other terms, which geographical aspect

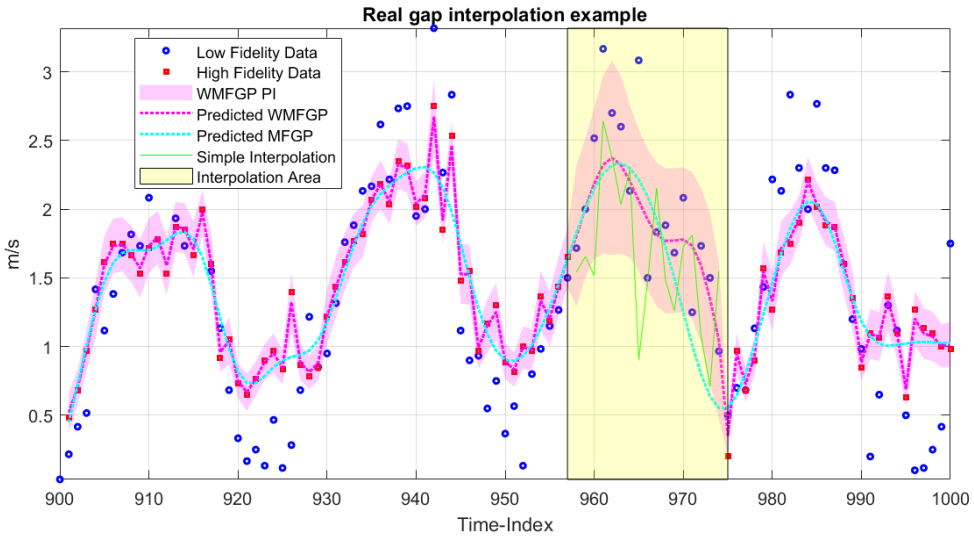


Figure 3.11: Example of a missing sequence imputation using WMFGP, MFGP, and SI methods.

explains most of the models performances. This is because when performing real imputation it is not possible to check for its quality. However, we can employ the simulation experiment described in Section 3.6.2 to gain an idea of the expected outcome under certain "experimental conditions", (i.e. geographical properties or correlation coefficients). In particular, Tables 3.5 and 3.4 depict the results of an auxiliary linear regression experiment. This experiment regresses the ratio of MAE of WMFGP (MFGP) to the MAE of the surrogate model, against Latitude, Longitude, Altitude, and the linear correlation coefficient between HF and LF data. In practise, the ratio $\frac{MAE_{WMFGP(MFGP)}}{MAE_{SURROGATE}}$ tells us how much we gain in terms of a benchmark (LF SURROGATE), while the regressors tells us if this gain is driven by geographical or data properties.

The obtained coefficient has a straightforward interpretation: the smaller it is, the greater error reduction we obtain by adopting specific models. Therefore, this experiment provides performance measures relative to the experimental conditions. Table 3.5 illustrates the impacts on the MAE of MFGP of Latitude, Longitude, Altitude, and the correlation coefficient with LF data. The intercept, Latitude, and Longitude returned statistically insignificant coefficients. However, the altitude coefficient was negative, indicating that as altitude increases, the ratio coefficient tends to decrease, signaling better performance of MFGP. As the values[1] in this

[1] All the variables are standardised, therefore more importance is not given to one variable with respect to another.

Table 3.4: Standardised coefficients table for the WMFGP performances, of the regression task performed in section 3.6.4 for analysing geographical properties. The asterisk indicate the different level of significance of the coefficients.

Variable	Estimate	Std. Error	t value	Pr(>—t—)
Intercept	-0.019	0.038	-0.508	0.611
Latitude	-0.059	0.048	-1.229	0.219
Longitude	-0.055	0.039	-1.405	0.160
Altitude	-0.110	0.053	-2.060	**0.039***
Corr	-0.095	0.043	-2.181	**0.029***

Table 3.5: Standardised coefficients table for the MFGP performances, of the regression exercise performed in section in section 3.6.4 for analysing geographical properties. The asterisk indicate the different level of significance of the coefficients.

Variable	Estimate	Std. Error	t value	Pr(>—t—)
Intercept	-0.017	0.038	-0.451	0.652
Latitude	0.046	0.048	0.965	0.334
Longitude	-0.095	0.039	-2.428	**0.015***
Altitude	-0.187	0.053	-3.504	**0.0004*****
Corr	-0.077	0.043	-1.773	**0.076.**

table are standardised, they cannot be interpreted directly. However, if they are converted to their original scales, a more precise interpretation arises. Specifically, for every 1000 meters increase in altitude, we expect an error reduction of 0.063 m/s. The linear correlation coefficient is also significant, with a more pronounced impact on error reduction. Similarly, when analysing the performance of WMFGP, the linear correlation coefficient and the altitude were found to be significant. However, the impact of the correlation coefficient has been found bigger, while the impact of altitude is reduced. These results overall suggest that warping the data makes the linear correlation between LF and HF data more important, and the altitude less relevant, a factor that might contribute to the overall better performance of the WMFGP.

3.7 Strengths and insights from the Warped multi-fidelity GP

The results of our two simulation experiments demonstrate the effectiveness of the Warped multi-fidelity Gaussian process (WMFGP) in handling data fusion tasks for skewed data distributions. We observed that the model performed well under

both random missingness scenarios, where only a few samples of HF data were available, and structural missingness scenarios, characterized by long sequences of missing data. The method successfully imputed missing sequences in the ARPA Lombardia network dataset, showcasing its applicability to data sources with similar missing data structures. The key advantages of the WMFGP, empowered by data-driven warping, are multiples.

Firstly, the non-parametric warping enables effective normalisation of diverse datasets given a sufficient amount of data. Second, when combined with the clustering strategy, the method offers a computationally efficient approach to gap-filling, capable of automatically adapting to both seasonal and subseasonal variability. This efficiency arises from the decomposition of the spatio-temporal problem into subtasks involving paired time series, rather than treating the information as a single, monolithic structure. As a result, computational cost depends primarily on temporal length, while scaling only linearly with spatial dimensionality.

Remarkably, our imputation method demonstrated robustness across gap sizes, contrasting with findings from other interpolation studies, see (Colombo and Fassò, 2022; Fassò et al., 2020). In our application strategy, we embraced a clustering approach grounded in the Third Law of Geography, positing that similarity in geographic configurations implies similarity in target variable values (Zhu et al., 2018). This rationale guided our choice of an isotropic model, focusing on clusters sharing similar geographical features.

Nevertheless, a more computationally intensive model incorporating spatial features could potentially enhance overall predictions. For instance, a Gaussian process model that treats longitude and latitude as an additional spatial dimension could exploit the spatial correlation to smooth over spatial domain were stations are not present. Finally, a regression analysis was conducted to examine how the performance of the WMFGP and MFGP models (measured by MAE) relates to key variables: latitude, longitude, altitude, and correlation (Corr). In other words, this analysis identifies which variables are most important in explaining model performance. As shown in Tables 3.4 and 3.5, the correlation coefficient (LF and HF) emerged as the most influential factor, with a much stronger effect in the

WMFGP than in the MFGP. This result is consistent with the better performance of the WMFGP, since both models depend heavily on linear correlation assumptions. The finding also helps explain why the warping approach outperformed the standard multi-fidelity Gaussian process.

Our findings suggest promising avenues for addressing multi-fidelity models in the presence of skewness, with the WMFGP emerging as a viable solution. Exploring spatial extensions holds potential for developing models capable not only of filling gaps in monitoring networks but also of predicting values in uncovered areas. One such application could involve integrating reanalysis data with in-situ samples, which would be pertinent in wind farm development for accurate resource assessment, integrating Lidar measurements with reanalysis data.

3.7.1 Limitations

A limitation of the WMFGP framework concerns the propagation of uncertainty arising from subjective modelling choices within the nonparametric normalization step. In particular, the choice of kernel used to perform the normalization introduces assumptions on the local structure of the data, such as smoothness and regularity. These assumptions directly affect the quality of the transformation, and consequently the effectiveness of the overall procedure. In practice, the adequacy of the kernel can be assessed through the resulting normalization quality: when the transformed data exhibit a more regular and homogeneous structure, the normalization can be considered successful. This provides a pragmatic criterion for kernel selection, although it remains heuristic and does not explicitly quantify uncertainty.

In addition, numerical aspects of the implementation—such as the discretization grid used during normalization—introduce approximation errors. While grid density does not alter the underlying modelling assumptions, it determines the accuracy of the numerical procedure and may therefore impact the stability of the results. This effect is closely linked to the number of available observations and to the chosen kernel. In particular, when the sample size is limited (e.g., fewer than 200 observations), linear interpolation over the grid may fail to capture nonlinear features of the normalized function, and more flexible interpolation schemes may be preferable.

The uncertainty associated with the interpolation step can be assessed via cross-validation strategies, for instance by removing a subset of observed points and evaluating how accurately the interpolation reconstructs them. More generally, these sources of uncertainty are not explicitly propagated through the WMFGP framework and may therefore lead to overconfident predictions. Empirical evidence from preliminary experiments suggests that, when a sufficiently large number of observations is available (e.g., more than 500 data points), both normalization and interpolation errors become negligible, and their impact on the overall uncertainty is substantially reduced.

3.8 Conclusion

This chapter illustrates the challenges that arise when applying Gaussian process models to skewed data (and by construction the MFGP), a common feature in environmental and geospatial applications. Standard GP models assume that residuals and the data generating process are normally distributed, an assumption that can limit their effectiveness when data distributions are asymmetric. In the presence of skewness, GP models may yield biased predictions, poor uncertainty quantification, and unreliable interpolation, especially in areas with limited observations. To overcome this limitation, several modelling strategies were evaluated aimed at either capturing or transforming the skewness in the data. First the chapter examined Gaussian process Regression with Skew Error (GPRSE). This model explicitly introduces skewness in the error term, allowing the predictive distribution to become asymmetric. While this approach improves flexibility, it comes with notable trade-offs. The model requires a larger number of parameters to characterize the skew distribution, which increases its complexity in the identifiability. Furthermore, the inference procedure is computationally intensive even for small dataset, and as consequence totally infeasible when applied to large datasets. These characteristics make GPRSE less practical for large-scale environmental applications where efficiency is critical (see Equation 3.7). Next, the Gaussian process with Specified Covariance Function (GPSCF) was considered.

This approach maintains Gaussian errors but modifies the covariance structure to reflect features such as non-stationarity or anisotropy, which may be linked to skewed spatial behaviour . GPSCF offers a more stable and computationally efficient alternative to GPRSE and can better adapt to complex spatial patterns. However, its structure can become nested and difficult to manage, especially when the specified covariance incorporates multiple interacting components. This added complexity may reduce interpretability and make model tuning and parameter estimation more difficult, even more in a multi-fidelity context.

Methods based on data transformation have also been explored, such as the parametric warping GP, where a known parametric transformation is incorporated in the likelihood to reduce skewness before fitting a standard GP. Although appealing in theory, this method relies on selecting a suitable parametric form, which is often insufficient for environmental data due to their irregular, nonlinear nature. These transformations can be sensitive to outliers and may require manual tuning or assumptions that do not hold in practice.

To address these issues, a non-parametric warping GP approach was adopted. This method learns the transformation directly from the data using a flexible, data-driven framework. Unlike parametric transformations, the non-parametric approach does not assume any specific functional form, making it highly adaptable to a wide range of skewness patterns. It is also computationally efficient, as the warping function can be estimated with minimal cost. Most importantly, this method generalizes well across different datasets and sources. Because it operates through quantile alignment, it enables consistent normalisation of multiple datasets without modifying their inter-dataset relationship. This makes it a particularly attractive solution for joint modelling of heterogeneous environmental data, as shown in the normalisation framework in 3.5.1.

In conclusion, while models like GPRSE and GPSCF offer valuable tools for incorporating skewness into GP models, they each present trade-offs in terms of complexity, scalability, and interpretability. The non-parametric warping method overcomes many of these limitations by offering a flexible, robust, and scalable approach to normalisation. Its ability to accommodate multiple data sources and

adapt to complex data structures makes it well suited for modern environmental application.

Chapter 4

Constructing a space time model

In Chapter 2, multi-fidelity models were introduced as a suitable approach for handling data containing long sequences of missing values. In Chapter 3, this class of models was further extended to address skewed data, while also demonstrating stable and reliable predictive uncertainty when interpolating across extended gaps. These findings, however, are based exclusively on time-dependent models, and the role of spatial information remains largely unexplored. This limitation motivates the need to consider whether incorporating spatial relationships could further enhance model performance and applicability.

In many real-world applications, spatial relationships are just as crucial as temporal dynamics. For example, in weather forecasting, the temperature at a given location depends not only on its own historical records but also on the conditions of neighbouring areas. This phenomenon is commonly referred to as the *spatial spillover effect*. A recent discussion of this concept can be found in Yin et al. (2024), while a more comprehensive treatment is provided in the seminal work of Anselin (1988). Importantly, this effect is not confined to a small number of domains; rather, it is widespread. Examples include air quality monitoring (Chen and Ye, 2019; Feng et al., 2020), traffic modelling (Guo et al., 2022), and the studies about disease outbreaks (Ulimwengu and Kibonge, 2021; Wang et al., 2023). In such contexts, neglecting spatial dependence risks overlooking valuable information about the interactions that occur across locations.

Incorporating spatial information generally enhances both the accuracy and interpretability of statistical models. This integration can be achieved in several ways. One approach is to include location-related variables directly as predictors, which improves interpretability by explicitly showing how specific geographical features, as seen, for example, in Section 3.7, influence the response. However, this strategy may overlook hidden but spatially structured influences—for instance, two areas of soil that share a common water supply—leading to residuals that remain spatially correlated. An alternative is to introduce spatially structured random effects, which capture spatial correlation by modelling unobserved variation across groups or locations. Spatial information can also be incorporated by specifying spatially correlated error structures, by including spatial weights or adjacency matrices as regressors, by employing multilevel (hierarchical) models, or by modelling the response surface directly—for example, using GPR, which is particularly well-suited when the aim is to predict values at unsampled locations.

Vice-vera, omitting spatial information could result in the problem of endogeneity. If spatial variation is omitted, part of the correlation structure may be incorrectly attributed to temporal effects, resulting in a well-known statistical pathology[1] known as *omitted variable bias*. In the context of wind speed, for example, spatial correlation may be mistakenly codified as temporal persistence, leading to an inconsistent estimation of temporal effects.

Beyond improving statistical consistency, the spatial dimension provides a richer understanding of the contextual relevance of phenomena. This flexibility is crucial for imputing values in areas where observations are sparse, unevenly distributed, or altogether missing due to logistical, economic, or political constraints. By leveraging spatial dependencies, models can uncover patterns that remain hidden in time-only frameworks. Such insights not only strengthen environmental assessments but also enable more targeted interventions in areas such as pollution control, land-use planning, or biodiversity conservation.

Recent work illustrates the potential of this perspective. For instance, Fassò et al.

[1]The word *pathology* is defined in medicine as *structural and functional abnormalities in cells, tissues, or organs that result from a disease process*. In the same way, the meaning of *structural malfunctioning* is borrowed here.

(2023) investigate the role of ammonia emissions in shaping air quality across Lombardy, revealing a link between intensive livestock farming and the degradation of air quality. Their findings highlight how spatially explicit modelling can provide critical evidence for policy and environmental regulation.

Multi-fidelity models are particularly useful when prediction at unsampled design points is required. In this context, response-surface-based surrogates, such as Gaussian processes, are well suited because they provide the best linear unbiased predictor (BLUP) at unobserved locations and minimise the mean squared prediction error within the class of linear predictors. Extending multi-fidelity models to the spatial dimension, however, is not trivial, as they are essentially hierarchical models based on Gaussian processes. Consequently, they rely upon inverting large spatio-temporal covariance matrices. Adding the spatial dimension inevitably increases the computational complexity of the matrix inversion, which may become infeasible without appropriate strategies. Many methods have been introduced in the spatial statistics literature for dealing with large spatio-temporal datasets and Gaussian processes; see Heaton et al. (2019) for a review. The various approaches offer alternatives for log-likelihood approximation. One of the most successful methods is the so-called *Vecchia approximation* (Katzfuss and Guinness, 2021), which approximates the precision matrix. Recently, Rambelli and Sigrist (2025) systematically compared the accuracy of different Gaussian process approximations and found that Vecchia emerged as the most accurate in most experimental settings.

In this chapter, the multi-fidelity models previously introduced used for predicting unsampled time location are extended to explicitly incorporate the spatio-temporal dimension. To achieve this, a new likelihood-based scalable estimation framework is designed, built upon the integration of the Vecchia approximation within the hierarchical structure of multi-output Gaussian processes. This development advances the earlier contributions of Chen et al. (2021) and Cheng et al. (2024), and provides the first fully likelihood-based implementation of the Vecchia approximation in the context of multi-fidelity modelling.

This development constitutes a novel framework, as the proposed implementation

makes it possible to estimate models with either a stationary ρ parameter, the parameter that controls the contribution of LF data toward the HF (see Equation 1.5), or a non-stationary counterpart (i.e., $\rho(s, s')$). In doing so, it extends earlier spatio-temporal multi-fidelity models, such as that of Babaee et al. (2020) for environmental applications. Consequently, a source of spatial non-linearity arises that must be explicitly accounted for. While the notion of non-linearity is not new in the multi-fidelity literature, with several models following this rationale (see, for example, Non-Linear Autoregressive Gaussian Processes (NARGP) (Perdikaris et al., 2017) and Deep Multi-Fidelity Gaussian Processes (Cutajar et al., 2019)), this represents the first attempt to incorporate a non-linear ρ within a spatio-temporal multi-fidelity framework, thereby extending the work of Babaee et al. (2020).

The chapter is organised as follows. Section 4.1 introduces the background necessary for the development of the proposed spatio-temporal multi-fidelity framework. Section 4.2 then presents the main characteristics of the new framework. The datasets employed as testbeds are described in Section 4.3. Section **??** reports the results of applying the framework to both synthetic and real data. Finally, Section **??** provides an extended discussion on the spatio-temporal scalability of the proposed approach.

4.1 Background

The new framework presented in this chapter introduces a spatio-temporal multi-fidelity model that is fitted using a scalable and robust likelihood-based estimation method founded on the Vecchia approximation. In what follows, the essential components required for this framework are introduced. We begin with a description of the Vecchia approximation and its key properties, followed by a brief discussion of spatio-temporal kernels.

4.1.1 The Vecchia approximation

The Vecchia approximation is a well-established method for approximating large covariance matrices in the GPs literature (Katzfuss and Guinness, 2021). In general, maximum-likelihood estimation in GPs involves factorizing or inverting an $n \times n$ covariance matrix, which costs $O(n^3)$.

Consider a realization of a latent GP $w(x)$, with $x \in \mathcal{D} \subset \mathbb{R}^d$ and $d \in \mathbb{N}$. For convenience, write $w_i := w(x_i)$ for the latent value at location x_i. We observe noisy measurements $y_i = w_i + e_i$, where $e_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

The main idea behind the Vecchia approximation is to factorize the joint density via the chain rule and then replace long conditioning sets with small, local ones. Let $o = (o_1, \ldots, o_n)$ be an *ordering* (a permutation) of the indices $\{1, \ldots, n\}$. Define the *history* under this ordering as $h_o(i) \subseteq \{o_1, \ldots, o_{i-1}\}$, with $h_o(o_1) = \emptyset$ (Katzfuss and Guinness, 2021). Let $\mathbf{y}_o := (y_{o_1}, \ldots, y_{o_n})^\top$ denote the data reordered by o. By the chain rule, the exact joint density is

$$p(\mathbf{y}_o) \;=\; \prod_{i=1}^{n} p\big(y_{o_i} \,|\, \mathbf{y}_{h_o(i)}\big). \tag{4.1}$$

The *Vecchia approximation* replaces each history $h_o(i)$ with a small conditioning set $g_o(i) \subseteq h_o(i)$ (often the m nearest predecessors in $\{o_1, \ldots, o_{i-1}\}$), yielding

$$\tilde{p}(\mathbf{y}_o) \;=\; \prod_{i=1}^{n} p\big(y_{o_i} \,|\, \mathbf{y}_{g_o(i)}\big), \qquad |g_o(i)| \le m \ll n. \tag{4.2}$$

Here, o denotes an *ordering* (not "observed"); if all entries are observed, o simply reorders the full data vector.

The set of variables conditioning at location i is called the *conditioning set*, and more precisely, it is a subset of observations called $g(i)$, which is found in a neighbourhood surrounding location i. The corresponding vector of conditioning values is denoted as $\mathbf{y}_{g(i)} = y_1, ..., y_{i-1}$. The approximated density of equation 4.2 is referred to as the Vecchia density. The *conditioning set* restricts the dependence of each y_i observation to a small set of nearest neighbors, which in turns simplifies computations while approximating the joint distribution.

4.1.2 Ordering of the Vecchia approximation

The ordering determines the sequence in which the corrupted realizations $y_1, \ldots, y_n$ are processed and, therefore, affects the composition of the conditioning sets for $i = 1, \ldots, m$, where m is the neighbour size. The ordering is not obvious especially if multiple dimensions are involved, but more importantly, as explicitly described by Guinness (2018) is relevant for the accuracy of the Vecchia approximation. The most common approach is the so called *coordinate ordering* which typically means arranging (or indexing) the data points according to their spatial (and possibly temporal) coordinates, the points which are physically (or temporally) close to each other in the real world also end up being "close" in the index ordering.

4.1.3 Conditioning strategy and neighbour size

The conditioning strategy specifies which previously ordered observations each point is conditioned on, while the neighbour size m governs the accuracy–cost trade-off. Increasing m typically improves the approximation but raises the cost. *Nearest–neighbour (NN) conditioning* selects, for each location, the m closest *predecessors* under a chosen metric (e.g., a space–time Mahalanobis distance using the model length-scales). NN is simple and fast, but its accuracy and numerical stability depend strongly on the ordering: poor orderings can yield redundant, highly correlated neighbours. A widely used remedy is *maximum–minimum–distance (maximin) ordering* (Guinness, 2018). Starting from a set of points i.e. A, B, D, E, F, G, a seed point is randomly chosen, let's say $S = \{A\}$, then compute "distance to the set" for every unselected point = distance to its nearest point in S. The farthest from A is G, so pick G. S become $S = \{A, G\}$. For each remaining point, find its distance to $S = \min(\text{distance to } A, \text{distance to } G)$. The one with largest such minimum is D (it sits far from both edges), so pick D. $S = \{A, G, D\}$. This produces a spread-out ordering, so that each point's nearest predecessors are informative and less collinear. In practice, one *combines* the two ideas: compute a maximin ordering, then apply NN conditioning to pick the m nearest *predecessors* in that order. This combination typically improves stability and accuracy at the same m

size. A schematic representation of such a strategy is depicted in figure 4.1.

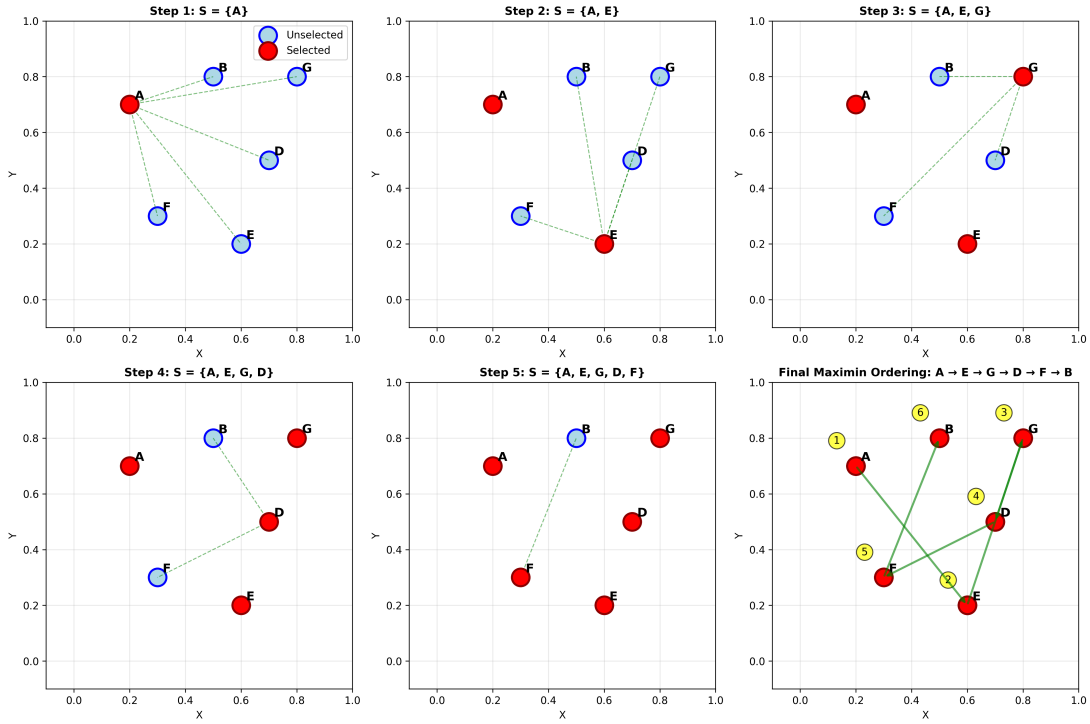


Figure 4.1: Schematic representation of the maximin ordering procedure. Steps 1–5 show the iterative selection of points maximizing minimum distance to previously selected points, producing a spread-out ordering. The final panel (bottom-right) displays the complete ordering sequence A → E → G → D → F → B with numbered steps and connecting arrows, ensuring that each point's nearest predecessors are well-distributed and non-collinear.

More recently, the *correlation-based ordering* (CVecchia) has been introduced, in which both site ordering and neighbor selection are guided by a correlation-based distance. For example, the ordering can be based on the following distance:

$$d_{\text{corr}}(\mathbf{s}_i, \mathbf{s}_j) = 1 - \phi(\mathbf{s}_i, \mathbf{s}_j), \tag{4.3}$$

where $\phi(\mathbf{s}_i, \mathbf{s}_j)$ represents the (modeled) correlation between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. Conditioning each site on those that are most correlated typically yields more accurate conditional densities. Although computing this distance increases the complexity of building the approximation, the overall computational burden remains in the order of $O(nm^3)$.

The main advantage of using a correlation-based structure is that CVecchia adapts better when the covariance structure varies over space or direction, making it a more flexible approach for spatial modelling.

4.1.4 The Vecchia approximation algorithm

The Vecchia algorithm computes an approximation of the precision matrix $\mathbf{K}^{-1}$, given some chosen kernel function, a neighbour size and the input of the kernel. The procedure is iterative, and it serves to derives the matrix $\mathbf{B}$ which is a sparse Cholesky factor of the precision matrix $\mathbf{K}^{-1}$, merely $\mathbf{K}^{-1} = \mathbf{B}\mathbf{B}^{\top}$. In other words, a random variable $W \sim N(0, (\mathbf{B}\mathbf{B}^{T})^{-1})$. The completion of the j and i elements of the matrix $\mathbf{B}$ is based on the following scheme:

$$B_{i,j} = \begin{cases} -A_{i,j}, & \text{if } j \in g(i) \\ 1, & \text{if } j = i \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

In practise, the Vecchia approximation is computed iteratively on the row of the matrix $\mathbf{B}$. Specifically, to compute each $\mathbf{B}_i$, we solve the linear system:

$$\mathbf{K}_{g(i),g(i)}\mathbf{A}_i = \mathbf{K}_{i,g(i)}, \tag{4.5}$$

where:

- $\mathbf{K}_{g(i),g(i)}$ is the covariance matrix of the neighbors,
- $\mathbf{K}_{i,g(i)}$ is the covariance vector between y_i scalar and its neighbors,
- $\mathbf{A}_i$ is the solution, representing the factor loadings.

Additionally, the diagonal precision matrix is then computed as:

$$\mathbf{D}_i = \frac{1}{\mathbf{K}_{i,i} - \mathbf{K}_{i,g(i)}\mathbf{A}_i^{T}}. \tag{4.6}$$

A compact version of the algorithm is expressed below:

4.1.5 Spatio-Temporal kernels function

Spatio-temporal GPs are characterized by the presence of spatio-temporal kernels. Here, a brief introductions to spatio-temporal kernels is illustrated. Choosing the

Algorithm 2 Vecchia Approximation Procedure

Require: Let $\ell_i = (t_i, \mathbf{s}) \in \mathbb{R}^3$, i = 1,. . . ,n.; covariance kernel k; number of neighbours m; small jitter ϵ for numerical stability during inversion of neighbour covariance matrices.

Ensure: Sparse matrix $\mathbf{B}$ and diagonal matrix $\mathbf{D}$ for approximation.

for $i = 1, \ldots, n$ **do**
  if $i = 1$ **then**
    $\mathbf{D}_i(i) \leftarrow \frac{1}{\mathbf{K}_{1,1}}$ ▷ First point has no neighbours
  else
    Identify the m nearest neighbours of point i among $\{1, \ldots, i-1\}$.
    Form the covariance matrix $\mathbf{K}_{g(i),g(i)}$ for these neighbours and add jitter ϵ.
    Compute cross-covariance $\mathbf{K}_{i,g(i)}$ between point i and its neighbours.
    $\mathbf{A}_i \leftarrow \mathbf{K}^{-1}_{g(i),g(i)}\mathbf{K}_{i,g(i)}$ ▷ Linear coefficients
    Update the i-th row of $\mathbf{B}$ with $-\mathbf{A}_i^\top$.
    $\mathbf{D}_i(i) \leftarrow \frac{1}{\mathbf{K}_{i,i} - \mathbf{K}_{i,g(i)}\mathbf{A}_i^\top}$.
  end if
end for
$\mathbf{B} \leftarrow \mathbf{B} + \mathbf{I}$
$\mathbf{D} \leftarrow \text{diag}(D_i(1), \ldots, D_i(n))$
return $\mathbf{B}, \mathbf{D}$

spatio-temporal kernels (covariance functions) of a spatio-temporal model signifies also to make assumptions regarding the process of interest behaviour in time and space. Given some spatial locations **s**, **s**' and time input t, t' the following provide a broad classification of the covariance function modelling space and time:

1. **Separable**
   (a) Multiplicative
   (b) Additive
2. **Non-separable**
   (a) Joint kernel
   (b) Mixed kernel

4.1.6 Separable kernels

The separable covariance functions have a series of assumptions that can simplify the modelling, but these may sometimes be unrealistic. Conversely, the non-separable covariance functions are more flexible, but they increase the number of parameters, so their estimation is not always straightforward.

The additive separable kernel for a spatio-temporal process assumes that the spatial and temporal components contribute to the overall covariance independently. Therefore, the covariance function equation takes the following form:

$$k(\mathbf{s}, \mathbf{s}', t, t') = k_s(\mathbf{s}, \mathbf{s}') + k_t(t, t'). \tag{4.7}$$

In other words, there is no explicitly modeled interaction between the spatial and temporal components. In addition, the contributions of each covariance are additive. The additive structure implies that the variance of the spatio-temporal process at any point $(\mathbf{s}, t)$ is the sum of the variances from the spatial and temporal components, with no cross-correlation terms. This highlights another aspect: the spatial covariance structure is invariant across time (spatial stationarity), and the temporal covariance structure is invariant across space (temporal stationarity). This also means that the covariance depends only on relative distances (not absolute positions) in space or time. The scalar observation $y(\mathbf{s}, t)$, at locations $\mathbf{s}$ and t, is implicitly modeled as:

$$y(\mathbf{s}, t) = w_s(\mathbf{s}) + w_t(t) + \epsilon(\mathbf{s}, t),$$

where:

- $W_s(\mathbf{s}) \sim \mathcal{GP}$ modelling the spatial effect.
- $W_t(t) \sim \mathcal{GP}$ modelling the temporal effect.
- $\epsilon(\mathbf{s}, t)$: Independent noise (often Gaussian).

Note again that the stochastic nature of W is emphasised by writing as a capital letter. In synthesis, the observed value at any $(\mathbf{s}, t)$ is a sum of independent spatial and temporal contributions. There is no direct spatio-temporal interaction term $W(\mathbf{s}, t)$.

The multiplicative separable kernel for a spatio-temporal process assumes that the spatial and temporal components contribute to the overall covariance as a product. This means that the covariance is the result of the interaction between the spatial

and temporal components, and the model equation takes the following form:

$$k(\mathbf{s}, \mathbf{s}', t, t') = k_s(\mathbf{s}, \mathbf{s}') \cdot k_t(t, t'). \tag{4.8}$$

In this structure, the covariance depends on both the spatial and temporal correlations, with the overall effect being multiplicative rather than additive. This implies that:

1. The spatial correlation, $k_s(\mathbf{s}, \mathbf{s}')$, scales the temporal covariance, $k_t(t, t')$, and vice-versa.

2. There is no explicit cross-term between space and time, but the covariance reflects joint dependencies via the product.

The variance of the spatio-temporal process at any point $(\mathbf{s}, t)$ is given by:

$$\text{Var}(y(\mathbf{s}, t)) = k_s(\mathbf{s}, \mathbf{s}') \cdot k_t(t, t'),$$

which indicates that the contributions from the spatial and temporal components are multiplicatively combined. The observations $y(\mathbf{s}, t)$ are implicitly modeled as:

$$y(\mathbf{s}, t) = w(\mathbf{s}, t) + \epsilon(\mathbf{s}, t),$$

where $w(\mathbf{s}, t)$ is the realization of a latent process jointly dependent on space and time, represented by $k_s \cdot k_t$. While, $\epsilon(\mathbf{s}, t)$ is an independent noise process (often Gaussian). This structure assumes: spatial and temporal components interact multiplicatively, making the covariance a function of both space and time. The kernel is separable in the sense that it can be factored into independent spatial and temporal components, but their joint effect is through a product.

4.1.7 Non-separable kernel

In a non-separable kernel, the spatial and temporal dimensions cannot be treated independently; instead, the kernel also captures their interaction. A joint non-separable kernel $k((\mathbf{s}, t), (\mathbf{s}', t'))$ directly models spatio-temporal interactions:

$$k((\mathbf{s},t),(\mathbf{s}',t')) = f(||\mathbf{s}-\mathbf{s}'||,|t-t'|),$$

where:

$$f(||\mathbf{s}-\mathbf{s}'||,|t-t'|)$$

is a covariance function that depends jointly on spatial distance $||\mathbf{s}-\mathbf{s}'||$ and temporal lag $|t-t'|$. For example, a squared exponential non separable kernel could be as follows:

$$k((\mathbf{s},t),(\mathbf{s}',t')) = \sigma^2 \exp\left(-\sqrt{\frac{||\mathbf{s}-\mathbf{s}'||^2}{\theta_s^2}+\frac{|t-t'|^2}{\theta_t^2}}\right),$$

.

A non separable kernel is the one composed of sum of independent kernels for space and time and an interaction terms. This type of kernel is know, as mixed kernel, and it looks as follows:

$$k((\mathbf{s},t),(\mathbf{s}',t')) = k_{\mathrm{s}}(\mathbf{s},\mathbf{s}')k_{\mathrm{t}}(t,t') + k_{\mathrm{interaction}}((\mathbf{s},t),(\mathbf{s}',t')),$$

where the $k_{\mathrm{interaction}}$, using a squared exponential kernel, is:

$$k_{\mathrm{interaction}}((\mathbf{s},t),(\mathbf{s}',t')) = \sigma_{\mathrm{int}}^2 \exp\left(-\frac{|t-t'|^2\cdot||\mathbf{s}-\mathbf{s}'||^2}{\theta_{\mathrm{int}}^2}\right).$$

Notice that in the interaction kernel the spatio and temporal distances $|t-t'|$ and $||\mathbf{s}-\mathbf{s}'||$, are multiplied.

4.2 Proposed framework: a scalable multi-fidelity spatio-temporal GP based on Vecchia approximation

4.2.1 The framework

In this section, a new scalable implementation of a fully likelihood based spatio-temporal MFGP is presented . The following representation recasts the multi-fidelity model equations presented in Chapter 1 as a composition of two independent Gaussian processes, W_1 and W_2, in its matrix-vector form[1]. Recalling that $\mathbf{y}_H$ is the vectoring containing the HF observations, while $\mathbf{y}_L$ is the vector containing LF observations, the new MF model equations are:

$$\mathbf{y}_L = \mathbf{Z}_1 \mathbf{w}_1 + \boldsymbol{\epsilon},$$
$$\mathbf{y}_H = \boldsymbol{\rho} \mathbf{Z}_{21} \mathbf{w}_1 + \mathbf{w}_2 + \boldsymbol{\epsilon},$$

with $\mathbf{w_1} \sim \mathcal{GP}(m(\cdot), k_{LL}(\cdot,\cdot))$ and $\mathbf{w_2} \sim \mathcal{GP}(0, k_D(\cdot,\cdot))$. Even if the above equations are very similar to those presented in Chapter 1, here below the major differences (and their implication) are listed:

1. First, ρ is a vector and not a parameter, reflecting the non-stationary nature of this model.

2. Second, $\mathbf{Z}_1 \in [0,1]^{n_L \times N}$ is an indicator matrix. The entries of $\mathbf{Z}_1$ are 1 when the LF data are present and 0 when the spatial location is considered but lacks LF data (e.g., because it includes a HF datum). However, in practice, we often assumes a nested design, where each HF datum has a corresponding LF but not vice-versa, and this latter case $\mathbf{Z}_1$ is often a diagonal matrix.

3. $\mathbf{Z}_{21} \in [0,1]^{n_H \times N}$ is an indicator matrix for the positions of the HF data.

The rest of the model is similar to the standard multi-fidelity model described in Chapter 1 where $\mathbf{y}_L \in \mathbb{R}^{n_L}$ and $\mathbf{y}_H \in \mathbb{R}^{n_H}$. With new representation mediated

[1]The scalar/process notation is dropped in favour a clear representation of the algebraic tricks.

by the use of the matrix $\mathbf{Z}_1$ and $\mathbf{Z}_2$, the Vecchia approximation is implemented independently in each process.

4.2.1.1 Independent treatment of the covariance

The Vecchia approximation cannot be applied directly to $\mathbf{K}$, since it is unclear how to treat the cross-covariance structure between the two processes W_1 and W_2. However, the model representation introduced in Section 4.2.1 enables the Vecchia approximation to be implemented independently for each process, after which the full covariance matrix $\mathbf{K}$ can be reconstructed. This is done introducing few additional matrices, in particular the matrix $\mathbf{A}$, defined as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{Z}_1 & 0 \\ \boldsymbol{\rho}\mathbf{Z}_{21} & I \end{bmatrix}.$$

$\mathbf{A}$ is a matrix that connects the low and high-fidelity components. Then the matrix

$$\boldsymbol{\Sigma}_w = \begin{bmatrix} \mathbf{K}_{LL} & 0 \\ 0 & \mathbf{K}_D \end{bmatrix},$$

where $\mathbf{K}_D$ is the covariance matrix of $\mathbf{w_2}$ discrepancy process. Given these new matrices, the matrix $\mathbf{K}$ can be alternatively decomposed as:

$$\mathbf{K} = \mathbf{A}\boldsymbol{\Sigma}_w\mathbf{A}^\top + \mathbf{D}_\epsilon,$$

$\mathbf{D}_\epsilon$ in this decomposition represents a diagonal matrix containing the noise variances σ_L^2 and σ_D^2 . We can also say that $\mathbf{K}^{-1}$ can be derived using the woodbury identity:

$$\mathbf{K}^{-1} = \mathbf{D}_\epsilon^{-1} - \mathbf{D}_\epsilon^{-1}\mathbf{A}\left(\boldsymbol{\Sigma}_w^{-1} + \mathbf{A}^\top\mathbf{D}_\epsilon^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{D}_\epsilon^{-1}.$$

If $\mathbf{K}_{LL}^{-1\text{Vecchia}}$ and $\mathbf{K}_D^{-1\text{Vecchia}}$ are the precision matrices of $\mathbf{w}_1$ and $\mathbf{w}_2$ obtained using the Vecchia approximation, then $\boldsymbol{\Sigma}_w^{-1\text{Vecchia}}$ is a block-diagonal precision matrix corresponding to $\mathbf{w}_1$ and $\mathbf{w}_2$ under the same approximation. Replacing $\boldsymbol{\Sigma}_w^{-1}$ with $\boldsymbol{\Sigma}_w^{-1\text{Vecchia}}$ in the Woodbury identity allows us to derive an approximate precision

matrix for $\mathbf{K}$ using the Vecchia approximation.

In practise, each Gaussian process is treated independently, the Vecchia approximation is used and applied independently to each fine realization of the process, and the full model is reconstructed using the woodbury identity.

4.2.2 Computational complexity and stability

The matrix

$$\mathbf{H} = \left(\boldsymbol{\Sigma}_w^{-1} + \mathbf{A}^\top \mathbf{D}_\epsilon^{-1} \mathbf{A}\right)^{-1},$$

is referred to as an intermediate matrix. It is a sparse banded matrix, with the degree of sparsity directly influenced by $\boldsymbol{\Sigma}_w^{-1\text{Vecchia}}$ for both $\mathbf{w_1}$ and $\mathbf{w_2}$. The number of diagonal elements of the bands reflects the correlation structure of the data, generally around 10 for fast decaying correlation up to 30 for slower decaying correlation. This formulation also provides the flexibility to selectively apply approximations to the sub-Gaussian processes. For instance, one could choose to approximate only the LF precision matrix while leaving the discrepancy process precision matrix unapproximated, as the latter is computed only at the HF data locations, which are inherently sparse. The total computation cost (TC) would be approximately:

$$\text{TC} = \mathcal{O}(n_L \cdot m^3) + \mathcal{O}(n_H \cdot m^3) + \mathcal{O}(n \cdot nb), \tag{4.9}$$

where $n = n_L + n_H$ and nb represents the bandwidth of the band in the matrix in $\mathbf{H}$. An example of the computational adavantages of the new multi-fidelity model is depicted in Figure 4.2, that illustrates a comparison of the computation times between a standard multi-fidelity and the new model multi-fidelity likelihood. The overall speed of computation depends on numbers of cores used and the neighbour size. An example of the code use for obtaining these computation time is available on the github repository of the project (https://github.com/Pietrostat193/Public-Vecchia-Approximation-for-multifidelity-models). Notice that, since the standard model is based on a dense variance–covariance matrix, datasets with more than 20,000 observations cause the MATLAB implementation to run out of memory. This issue does not occur with the new approximated version, for

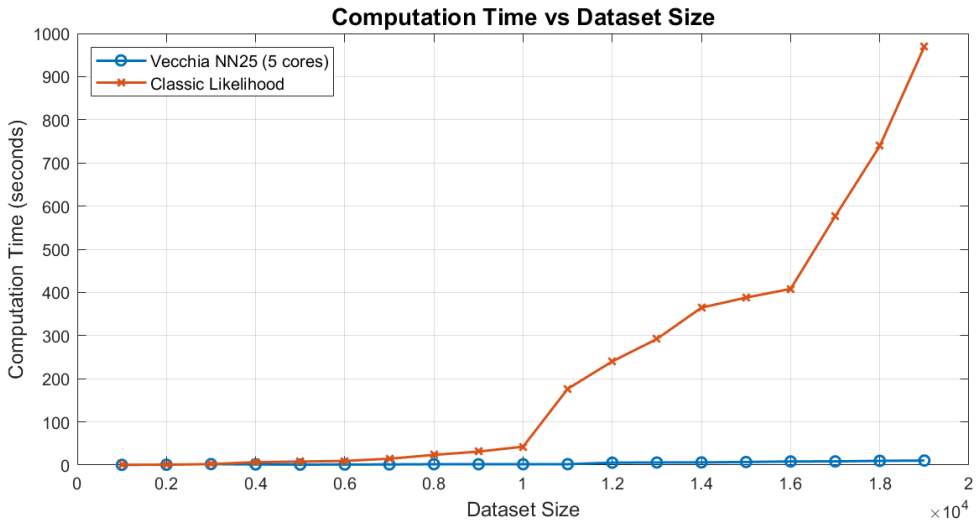


Figure 4.2: Comparison of the computation time in seconds between the classic likelihood and our Vecchia approximated likelihood for increasing dataset size. The dataset used for this comparison is the South Lombardia dataset see Section 1.5.3.

which the largest dataset considered (20,000 observations) requires approximately 7 seconds, compared to the 15 minutes required by the classical likelihood implementation.

4.3 Properties of the experimental dataset

Here, we illustrate the properties of the experimental dataset that motivate the modelling choices adopted in this chapter. For this study, we use the South Lombardia wind speed dataset introduced in Section 1.5.3 of Chapter 1. The dataset combines low-fidelity ERA5 reanalysis data with high-fidelity measurements from the ARPA Lombardia monitoring network, matched using a nearest-neighbour spatial procedure. It consists of hourly observations across multiple monitoring stations spanning a predominantly flat region with heterogeneous station density and moderate spatial alignment between datasets (correlation coefficients approximately ranging from 0.3 to 0.55). These characteristics create a setting where temporal dynamics are strong and coherent across sites, while spatial correlations are comparatively weak and irregular — making the dataset particularly well suited for exploring spatio-temporal multi-fidelity modelling strategies.

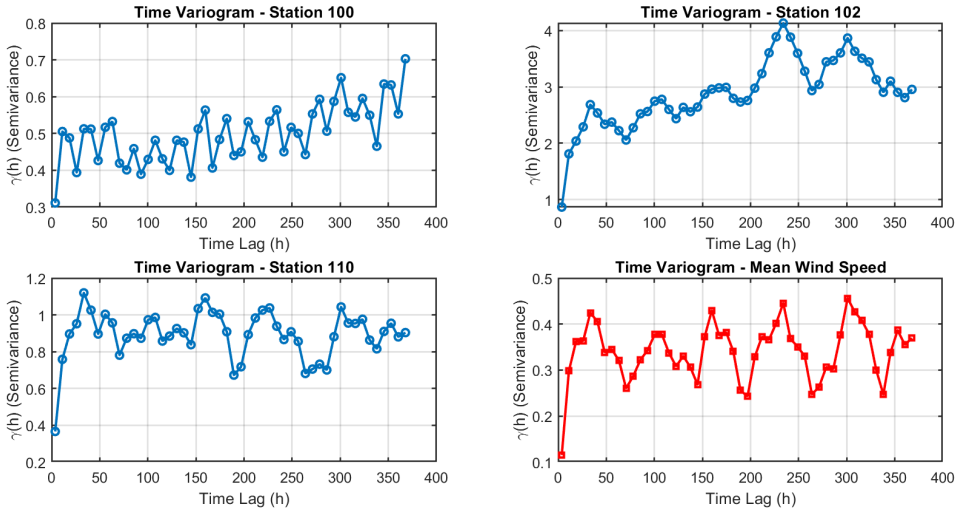


Figure 4.3: The figure shows the empirical variograms computed at different station from the South Lombardia dataset. It highlights how the time correlation patterns differs at different point in space, even in a relatively small and homogeneous region.

4.3.1 Spatio-Temporal aspects of south Lombardia dataset

The variogram can be a valuable tool to analyze space-time data as it quantifies the degree of similarity for increasing distance. Let $Y(\mathbf{x})$ be a random field at location $\mathbf{x}$ in a space or time domain. The semivariogram $\gamma(\mathbf{h})$ for a lag vector $\mathbf{h}$ is defined as

$$\gamma(\mathbf{h}) = \frac{1}{2} E\Big[\big(Y(\mathbf{x}+\mathbf{h}) - Y(\mathbf{x})\big)^2\Big].$$

Looking at the variogram on the raw data can provide some intuition of the decay of correlation across different station, even though these should be considered as "preliminary" observations since, as illustrated in section 4.4.1, the model may contains a mean function[1].

The multipanel plot 4.3 contains the time variogram of multiple stations from the South Lombardia dataset, except for the last panel, which depicts the average of all stations. The figures show that all stations exhibit a cyclic pattern in wind speed. Given the hourly resolution, this corresponds to a pattern recurring approximately every two days (48 hours). However, some stations (e.g., station 100 in Figure 4.3) show a faster decay in correlation.

The spatial heat map (see Figure 4.4) instead shows, that on average, the wind speed

[1]Note that the mean function is always computed as space-dependent; therefore, the primary source of correlation remains persistent, particularly over time.

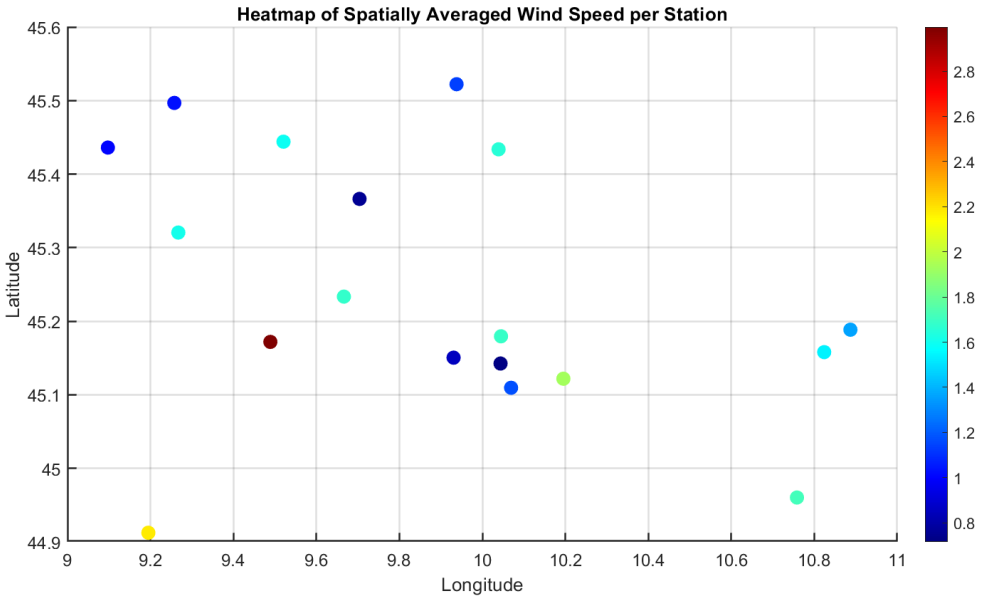


Figure 4.4: The spatial heatmap highlights how mean level of the south Lombardia dataset changes across space.

mean level is not driven much by proximity. For example, the cluster around the longitude levels of 9.8-10.2 shows different shades of colors indicating a consistent difference in the wind speed mean level across the stations.

This suggests that while wind speeds observed at different stations have a relatively coherent temporal pattern (i.e. a seasonal component), with variability in the rate of decay of correlation, the spatial coherence is quite small. In other words, increasing spatial distance does not necessarily correspond to increasingly different wind speeds. This latter point suggests a potentially independent model of spatial and temporal correlation.

4.4 Experiments

Two experiments are designed to test the limitations and effectiveness of the multi-fidelity model for spatio-temporal studies. The section begins by illustrating in details the newly adopted spatio-temporal model, subsequently the experimental design is introduced.

4.4.1 Model specification

Consider a spatio-temporal domain $\mathbf{s} \in \mathbb{R}^2$ and $t \in \mathbb{R}$, where t is an index having hourly resolution. The observations from the reanalysis ERA5 data are generally of low quality compared to the monitoring stations of ARPA, and they are modelled as follows:

$$y_L(\mathbf{s}, t) = w_1(\mathbf{s}, t) + \epsilon_L(\mathbf{s}, t), \quad t \in \{1, 2, \ldots, n_L\},$$

where the process W_1 which underlies all scalar realizations w_1:

$$W_1(\mathbf{s}, t) \sim \mathcal{GP}(m_1(\mathbf{s}, t), k_{LL}((\mathbf{s}, t), (\mathbf{s}', t'))),$$

and:

$$\epsilon_L(\mathbf{s}, t) \sim \mathcal{N}(0, \sigma_L^2).$$

The wind speed data of southern Lombardy are modelled using a squared exponential kernel $k(\cdot)$ with a multiplicative structure, see Section 4.1.6. The choice of squared exponential is motivated by both theoretical and practical considerations. First, it ensures coherence with the experiments in Chapter 2, allowing for direct comparability of the spatio-temporal results with those of the temporal study. Second, while kernel selection was not the primary focus of this work, squared exponential kernels demonstrate important advantages in the Vecchia approximation framework. In particular, although the exponential kernel was initially evaluated (Matérn with $\nu = 0.5$) for its fast correlation decay controlled through the lengthscale parameter, the squared exponential kernel was ultimately selected for superior convergence properties. Compared to exponential kernels, the squared exponential yields covariance matrices with better condition numbers $\kappa(\mathbf{K}) = \lambda_{\max}/\lambda_{\min}$, reducing numerical instability during matrix factorizations and inverse computations. Moreover, the squared exponential's infinitely smooth correlation decay enables more efficient information propagation in the Vecchia conditioning scheme, where information from neighbouring observations propagates through the sequence. This smoothness property requires smaller neighbour sizes m to achieve comparable accuracy, rendering the approximation more effi-

cient. Additionally, the smoother likelihood surface of the squared exponential kernel facilitates faster and more reliable convergence during likelihood optimization over hyperparameters, with stable gradients enabling robust parameter estimation. These numerical and computational advantages were confirmed through cross-validation experiments, which demonstrated superior overall fit and convergence compared to exponential and higher-order Matérn formulations.

The assumption of independent spatial and temporal components represents a starting point of the analysis. However, preliminary experiments systematically comparing independent (additive) and dependent (multiplicative) structures revealed that the multiplicative formulation, wherein temporal dynamics scale spatially, provides substantially better model fit and predictive performance. This transition from independence to dependence is supported by the observed structure of the spatio-temporal process: as illustrated in Figure 4.3, temporal variability exhibits strong spatial heterogeneity rather than uniform temporal behavior across locations. The multiplicative structure naturally accommodates this localized dependence, allowing the temporal component to modulate differently at each location, thereby capturing the complex interaction between spatial and temporal dynamics inherent in wind speed fields. Consequently, the analysis proceeds with the multiplicative factorization, which more faithfully represents the data characteristics while maintaining interpretability through the structured factorization. The South Lombardia dataset depicts stations that are separated by several kilometres and located in flat terrain, where the influence of hidden variables-such as terrain elevation and climatic factors-is minimized. The HF data from the ARPA monitoring network have been modelled as follows:

$$y_H(\mathbf{s}, t) = \rho(\mathbf{s})\, w_1(\mathbf{s}, t) + w_2(\mathbf{s}, t) + \epsilon_H(\mathbf{s}, t),$$

where:

$$W_2(\mathbf{s}, t) \sim \mathcal{GP}(0, k_D((\mathbf{s}, t), (\mathbf{s}', t'))),$$

and:

$$\epsilon_H(\mathbf{s}, t) \sim \mathcal{N}(0, \sigma_H^2).$$

The main difference with the model introduced in Section 4.2.1 is that ρ is not anymore a parameter but a function. Such a treatment of ρ introduces an inherent non-stationary spatially dependent rescaling of the low fidelity process. If $\rho(\mathbf{s}, \mathbf{s}')$ is a linear function of $\mathbf{s}$ and $\mathbf{s}'$, it can be written as:

$$\rho(\mathbf{s}, \mathbf{s}') = a + b\mathbf{s} + c\mathbf{s}',$$

where a, b, and c are constants that define the linear relationship. Another possible functional form is:

$$\rho(\mathbf{s}, \mathbf{s}') = \sum_{i=0}^{n} \sum_{j=0}^{m} a_{ij} \mathbf{s}^i (\mathbf{s}')^j,$$

where a_{ij} are the coefficients of the polynomial, and n, m are the degrees of the polynomial in s and s', respectively. A totally different approach for modelling ρ is starting from the empirical values:

$$\rho(\mathbf{s}, \mathbf{s}') = \frac{\text{cov}(y_H(\mathbf{s}), y_L(\mathbf{s}))}{\text{var}(y_L(\mathbf{s}))}. \tag{4.10}$$

These empirical values can ultimately be modelled using a smooth function, such as an additional Gaussian process, which provides a high degree of flexibility in capturing complex patterns.

Apart from its implementation—which, as demonstrated in Section 4.2.1, is fully likelihood-based—the model described in this section can be considered a non-linear multi-fidelity Gaussian process, similar to Deep and Verma (2024) and Perdikaris et al. (2017). However, its nonlinearity is explicitly targeted, enhancing the model's interpretability. In practice, this approach allows users to specify both the functional form of ρ and its input variables.

Notice that in this regard ρ could also be a function of time, $\rho(\mathbf{s}, \mathbf{s}', t, t')$. However, this additional flexibility is not usually required in our context, as the relationship between datasets, as seen in Chapter 3[1], and therefore the ρ values, is usually constant across time. Allowing ρ to vary over both space and time introduces a highly flexible, non-stationary scaling of the low-fidelity process. While this

[1]This is evidenced by the stable interpolation uncertainty across different gap sizes.

increases model expressiveness, it also weakens the smoothness induced by the underlying Gaussian process, as the effective covariance structure becomes locally modulated. In practice, this can lead to overfitting and excessively irregular (i.e., 'wiggly') posterior predictions unless additional regularization is imposed.

4.4.2 Synthetic data experiment

The first experiment is based on synthetic wind speed data. These data are made of two main components. A deterministic component, which is derived from yearly daily averages of the South Lombardia dataset (see 1.5.3 for another example of this approach) and a stochastic component which is a spatio-temporal Gaussian process. In other words, the deterministic component is an average from a real dataset, while the stochastic component is generated using a spatio-temporal Gaussian process (see Section 4.4.3). In this way, we ensure that the data resemble realistic wind-speed patterns while maintaining known spatio-temporal variability. In this experiment, we compare the performance of different Gaussian processes and the classic multi-fidelity model with zero mean and constant ρ parameters. The Gaussian processes are designed with different input combinations.

- The first model, GP(x), is a Gaussian process trained on spatio-temporal coordinates, i.e., (latitude, longitude, time).
- The second model, GP(y_L), is a Gaussian process that takes the low-fidelity data as input and attempts to recover the unknown function $f(x)$ that maps $y_L \xrightarrow{f(x)} y_H$.
- The final model, GP(x, y_L), is a Gaussian process that uses both the spatio-temporal coordinates and the low-fidelity data as input.

The models are compared across three different sample sizes of y_H and three different noise levels (high, medium, low). Specifically, y_H is available at one, three, or five simulated monitoring stations. The model validation and comparison is made with a classic cross-validation approach:

1. The data are partitioned into two sets such that $D_t = [D_{train}, D_{test}]$ where D_{train} and D_{test} are the training and test sets, respectively;

2. Only the training set is used to calculate the parameters for the GPs and or the multi-fidelity;

3. The estimated models is used to predict values at the locations of measurements in the test set.

4. Finally, the metrics of performance such as mean absolute error (MAE), root mean squared error (RMSE), normalised Estimation Error (NEE) [1] are used to compared the models performance.

4.4.3 Simulation procedure for synthetic spatio-temporal data

The generation of the synthetic dataset is designed to mimic spatio-temporal wind speed with a multi-fidelity structure. To achieve this goal, we average the yearly observations from South Lombardia[2] across all stations. This average constitutes the mean function of our data. We then applied a spatio-temporal corruption using a Gaussian process.

In practice, assume $w(t, \mathbf{s})$ to be the wind speed observed at a particular combination of spatio-temporal coordinates t and $\mathbf{s}$. The high-fidelity data generation proceeds as follows. First, a spatio-temporal Gaussian process is modelled as:

$$w(t, \mathbf{s}) \sim \mathcal{GP}(m(t, \mathbf{s}), k_{\text{st}}(t, \mathbf{s})),$$

where $m(t, \mathbf{s})$ is the mean function extracted from real wind speed data, and k_{st} is the spatio-temporal covariance function with a multiplicative structure:

$$k_{\text{st}}(\mathbf{s}, \mathbf{s}', t, t') = k_{\text{s}}(\mathbf{s}, \mathbf{s}') \cdot k_{\text{t}}(t, t').$$

[1] A sort of mean absolute percentage error without the percentage, defined as $\frac{1}{n}\sum|\frac{y-\hat{y}}{y}|$).
[2] A subset of the ARPA Lombardia dataset. See Section 4.3 for more information.

Using the Cholesky decomposition of this covariance matrix:

$$L = \text{chol}(K_{\text{st}} + \epsilon I),$$

a random GP sample $g \sim \mathcal{N}(0, K_{\text{st}})$ is drawn:

$$\mathbf{g} = L\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, I).$$

The high-fidelity wind speed realisation is then given by:

$$w_H(t, \mathbf{s}) = g(t, \mathbf{s}) + m(t, \mathbf{s}) + 2 + \eta, \quad \eta \sim \mathcal{N}(0, \sigma^2_{\text{noise}}), \tag{4.11}$$

where the constant 2 acts as a safety offset term that shifts the generated series away from 0 to prevent negative values. Equation (4.11) defines the baseline high-fidelity process without spatial rescaling.

The low-fidelity data are then generated through a two-step process: rescaling of the high-fidelity data and addition of an interference field. The high-fidelity realisations are rescaled using a space-dependent factor $\rho(\mathbf{s})$:

$$w_L(t, \mathbf{s}) = \frac{1}{\rho(\mathbf{s})} w_H(t, \mathbf{s}) + w_I(t, \mathbf{s}), \tag{4.12}$$

where $w_I(t, \mathbf{s})$ is an interference field simulated as a multivariate normal distribution. The rescaling factor $\rho(\mathbf{s})$ serves two purposes: (i) to introduce spatial dependence at the model level, and (ii) to generate an additional source of spatial variability independent of the underlying correlation structure. An example of the simulated time series obtained with this procedure is shown in Figure 4.5. The black line represents the mean function, while the coloured lines correspond to nine realisations at different spatial locations. Notice that the realisations exhibit different mean levels due to the spatial rescaling inherent in Equation (4.12).

The rescaling factor is space-dependent and is derived empirically through a para-

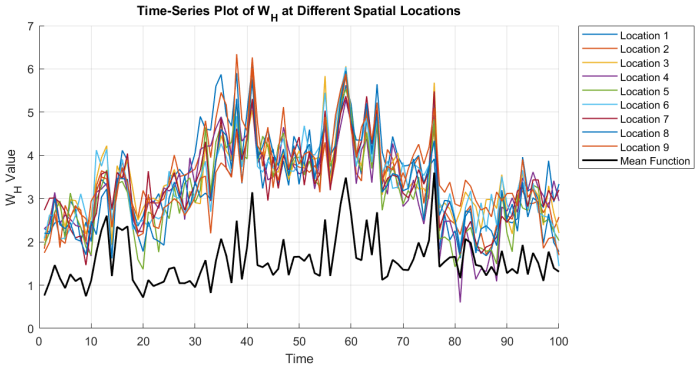


Figure 4.5: Simulated time series for the synthetic data experiment. The black solid line represents the mean function $m(t, \mathbf{s})$ (constant across all spatial locations). The coloured lines show the rescaled high-fidelity realisations $\frac{1}{\rho(\mathbf{s})} w_H(t, \mathbf{s})$ at nine different spatial locations. The varying mean levels across locations are a direct consequence of the spatial rescaling factor $\rho(\mathbf{s})$, which modulates the amplitude at each location independently. The constant offset term $+2$ contributes identically to all realisations.

metric two-dimensional sinusoidal model:

$$\rho(s_1, s_2) = b_1 \sin(b_2 s_1 + b_3) + b_4 \cos(b_5 s_2 + b_6) + b_7, \tag{4.13}$$

where s_1 and s_2 are spatial coordinates, and $\mathbf{b} = [b_1, \ldots, b_7]$ is the parameter vector. Here, b_1 and b_4 are the amplitudes of the sine and cosine components, b_2 and b_5 their frequency scaling factors, b_3 and b_6 their phase shifts, and b_7 a constant offset.

The function can be interpreted as a sum of sinusoidal components:

- $b_1 \sin(b_2 s_1 + b_3)$ introduces oscillatory behaviour along the s_1-axis.
- $b_4 \cos(b_5 s_2 + b_6)$ introduces oscillatory behaviour along the s_2-axis.
- The constant term b_7 acts as a baseline shift.

This model is useful for representing wave-like patterns in two dimensions and is commonly applied in surface fitting, wave modelling, and signal processing. The parameters of the sinusoidal model are estimated on the empirical ρ, see Equation 4.10. Figure 4.6 shows the spatial variability of the ρ coefficients.

When compared with the model in Section 1.3.4, a key difference emerges. That model assumes HF data to be a rescaled version of the LF data plus a discrepancy process. In the synthetic generation procedure, however, we generate HF data first (Equation (4.11)), then construct LF data by rescaling HF by $1/\rho(\mathbf{s})$ and adding interference (Equation (4.12)). Despite this directionality reversal, the two

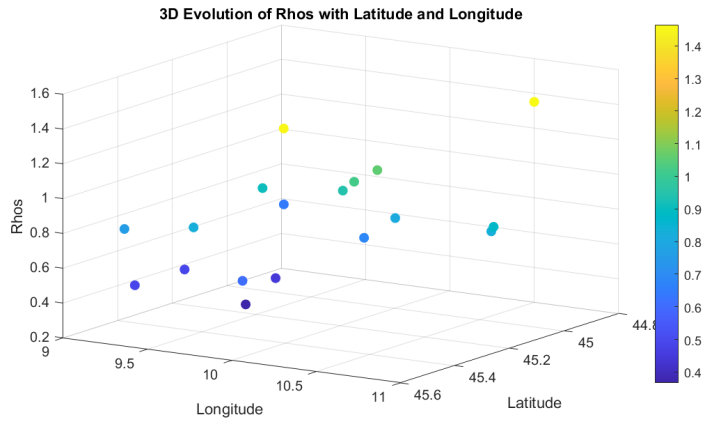


Figure 4.6: Spatial distribution of the empirical ρ parameters in the South Lombardia dataset. Values of ρ computed using equation 4.10.

formulations are mathematically equivalent: rearranging Equation (4.12) yields $w_H(\mathbf{s}, t) = \rho(\mathbf{s})[w_L(t, \mathbf{s}) + w_I(t, \mathbf{s})]$, which matches the model structure of Section 1.3.4 with the discrepancy process w_D replaced by the interference field w_I. The motivation for this reversed order was primarily practical: to achieve realism, it sufficed to incorporate into the HF generating process a mean function derived from empirical data. More generally, generating from HF data facilitates the regulation of correlation levels between HF and LF.

4.4.4 Results of the synthetic data experiment

The synthetic data experiment results, presented in Table 4.1, demonstrate that the multi-fidelity model consistently outperforms all Gaussian process configurations in recovering wind speed across all metrics. Notably, the multi-fidelity model also shows a stronger improvement when the number of HF data points increases. For instance, increasing the HF sample size from 3 to 5 locations yields a 9% improvement in MAE for the MF model, accross all levels of noises. Additionally, the GP models exhibit greater sensitivity to noise. Taking the baseline GP(y_L) model as an example, when the HF sample size is 3, increasing the noise level from low to high degrades the MAE from 0.563 to 0.627. With 5 HF samples, the same transition results in a rise from 0.571 to 0.638. In contrast, the multi-fidelity model demonstrates greater robustness: its performance remains relatively stable between average and high noise levels. This robustness can be attributed to the

Table 4.1: Performance metrics (MAE, RMSE, NEE) for the synthetic data experiment for different sample sizes (3, 5) and noise levels (Low, Avg, High) across five models: Gaussian process using low fidelity data as input GP(y_L, Gaussian process using both low-fidelity, space and time as input GP4D), Gaussian processes using space and time GP3D as input and Multifidelity gaussian process with zero mean function a constant ρ function.

Sample	**Noise**	**Metric**	**GP(y_L)**	**GP4D(y_L,x)**	**GP3D(x)**	**MF0c**
3	Low	MAE	0.563	0.685	0.679	**0.489**
3	Low	RMSE	0.171	0.219	0.208	**0.142**
3	Low	NEE	0.695	0.871	0.865	**0.618**
3	Avg	MAE	0.610	0.695	0.681	**0.514**
3	Avg	RMSE	0.181	0.208	0.205	**0.147**
3	Avg	NEE	0.754	0.866	0.875	**0.652**
3	High	MAE	0.627	0.685	0.694	**0.512**
3	High	RMSE	0.195	0.209	0.205	**0.155**
3	High	NEE	0.790	0.853	0.858	**0.651**
5	Low	MAE	0.571	0.689	0.687	**0.448**
5	Low	RMSE	0.171	0.202	0.198	**0.133**
5	Low	NEE	0.727	0.867	0.869	**0.571**
5	Avg	MAE	0.605	0.687	0.681	**0.463**
5	Avg	RMSE	0.180	0.209	0.203	**0.131**
5	Avg	NEE	0.743	0.855	0.864	**0.599**
5	High	MAE	0.638	0.686	0.678	**0.461**
5	High	RMSE	0.186	0.208	0.199	**0.139**
5	High	NEE	0.801	0.861	0.857	**0.587**

MF model's ability to model cross-correlation between fidelities. When the low-fidelity data are noisy, the model tends to absorb that noise into the discrepancy process δ, rather than allowing it to propagate directly into the final prediction. Figure 4.8 shows the prediction results from a randomly selected high-noise synthetic data experiment, using 5 high-fidelity training locations. In particular, the figure shows the comparison of the MF0c against the standard Gaussian process trained on space time data.

Figure 4.7, illustrates the spatial layout of the selected test locations. Only 9 spatial locations are used in total, with 5 designated for training. In each simulation, both the training and test locations, as well as the data realizations, are randomly varied.

4.4.5 Real data experiment

The second experiment is developed using directly the data from the South Lombardia dataset, see section 1.5.3. In this occasion, the flexibility of the multi-fidelity is

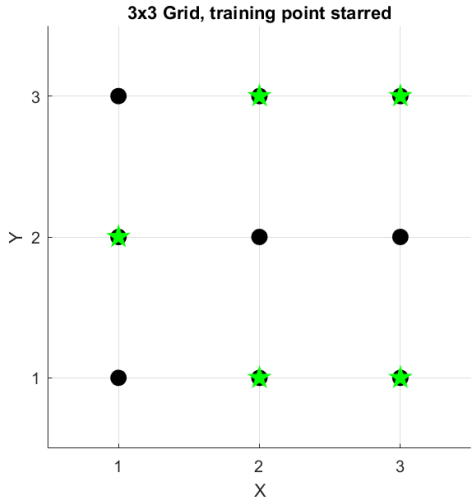


Figure 4.7: Spatial distribution of the training and test sets used in the experimental run shown in Figure 4.8. The starred locations refer to training locations.

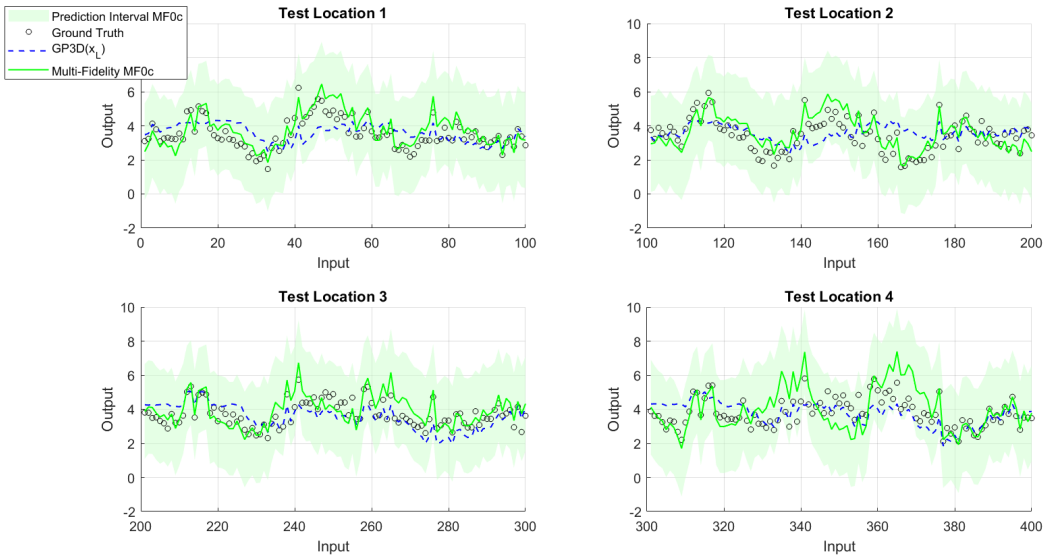


Figure 4.8: High noise prediction at different test locations for the synthetic data experiment described in section 4.4.4.

Table 4.2: Description of the models. The first column lists the acronym for each tested model, where MF stands for multi-fidelity. The second column specifies the mean function used for the low-fidelity process, while the third column indicates the function used for ρ.

Model	$\mu_L(\mathbf{s}, \mathbf{s}')$	$\rho(\mathbf{s}, \mathbf{s}')$
WMF gpgp	*GP*	*eGP*[1]
MF cc	*constant*	*constant*
MF cl	*constant*	*linear*
MF lc	*linear*	*constant*
MF cp	*constant*	*polynomial*
MF ll	*linear*	*linear*
MF 0c	*zero*	*constant*
GP	NA	NA

increased by including a $\rho(\cdot)$ function as described in Section 4.2.1. The latter choice is motivated by two major factors, first, that we empirically observed that the standard multifidelty, the one having a unique ρ parameter, did not model the data in a satisfactory way. Second, the dataset had a size that allows the estimation for a function $\rho(\cdot)$. Intuitively, if little spatial locations are available it is harder to get an estimate for the function for ρ. The estimated models are depicted in Table 4.2. Notice that there are various combinations of the mean function and the ρ function. We also tested a warped MFGP model, using a Gaussian process for the mean function μ and another Gaussian process for ρ, based on its empirical values.

For the validation of the results, leave one station out approach is used. This is the same approach adopted in Otto et al. (2024). The main idea is to use all the station of the dataset minus one to train the models parameters, to predict at the missing station. Using such a strategy it is possible to assess if there are areas where the models under-perform compared to others.

4.4.6 Results of the real data experiment

Table 4.3 presents the performance of various prediction methods based on Mean Absolute Error, Normalised Estimation Error, and Correlation Coefficient, computed as linear correlation coefficient between prediction and true data.

The best overall performance is achieved by MF ll, which records the lowest MAE (0.621) and NEE (0.750) among all methods, followed closely by WMF gpgp. Interestingly, these two models demonstrate their superiority in fundamentally

different ways. The performance of WMF gpgp is highly dependent on the signal amplitude: it tends to produce low errors when the signal is weak and higher errors when the signal is strong[1]. This pattern is primarily due to a misspecification of the mean function, which at times overestimates or underestimates the local mean level. This issue appears to stem more from the mean structure than from the multi-fidelity component of the model, which is well estimated as indicated by the correlation coefficient. Although, its correlation coefficient (0.55) is slightly lower than others MF models, the reduction in error is significant. MF cp and MF cc have the highest correlation coefficient (0.61), suggesting a stronger linear relationship with true values. However, they exhibit slightly higher NEE and MAE compared to MF ll. The MF cl also shows a good balance between MAE (0.624), NEE (0.737), and correlation (0.60), making it a competitive alternative. The weakest methods were the MF 0c and the GPR, which performed the worst across all metrics. GPR has the highest NEE (0.959), high MAE (0.762), and the lowest correlation (0.40), indicating poor predictive quality. Most MF-based methods (cp, ll, cl, lc, cc), however, perform consistently well with relatively low MAEs (0.62–0.66), moderate-to-high correlation (0.55–0.61), and acceptable NEE values. In conclusion, MF-based models clearly outperform GPR, with MF ll offering the best trade-off between low error and correlation strength. Minimizing both MAE and NEE while maintaining a decent correlation makes it the most reliable option. Figure 4.9 depicts a snapshot of the predictions of 629 of the leave one station out experiment. Notice that the Gaussian process prediction is too smooth making, lacking the capacity of capturing local variability. Comparing these performances with those of the simulated data, it is clear both models had and overall worsening as expected, due the irregularity of the predictions grid and the unevenness of the correlation between different locations.

[1] As exemplified, by the NEE error very low and MAE error generally higher.

Table 4.3: Prediction performance metrics (MAE, NEE, Correlation Coefficient) for different methods for real data experiment described in section 4.4.5. For the labels of models tested check table 4.2.

Prediction Type	MAE	NEE	Correlation Coefficient
WMF gpgp	0.872	0.618	0.55
MF cp	0.660	0.700	0.61
MF ll	0.621	0.750	0.55
MF cl	0.624	0.737	0.60
MF lc	0.645	0.771	0.56
MF cc	0.662	0.700	0.61
MF 0c	0.795	0.872	0.47
GP	0.762	0.959	0.40

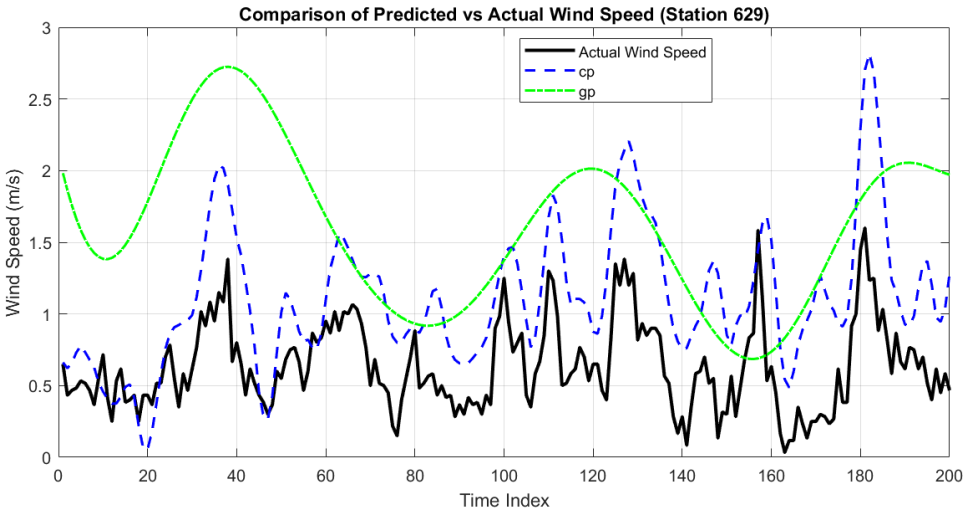


Figure 4.9: The plot depicts a snapshot of the predictions at station 629 of the Gaussian process and and MFcp.

4.5 Limitations of the experimental framework

While the proposed experimental framework provides encouraging results, several limitations must be acknowledged.

First, the analysis does not fully clarify whether the approximation error introduced by the proposed covariance decomposition propagates in a controlled (i.e., non-destructive) manner to the individual model components. In particular, it remains unclear how approximation inaccuracies in the latent processes affect the overall predictive performance when recombined through the multi-fidelity structure.

Second, the joint estimation of the mean and covariance functions may lead to identifiability issues. In practice, this can result in part of the large-scale structure of the data being absorbed by the covariance function rather than the mean, potentially degrading interpretability and predictive stability.

Third, uncertainty quantification has not been thoroughly addressed. In particular, no predictive intervals are reported, making it difficult to assess the reliability and calibration of the model outputs.

To address these limitations, a modified multi-fidelity Gaussian process framework is introduced. The key idea is to decouple mean estimation from covariance modelling through a generalized least squares (GLS) approach, thereby improving identifiability and enhancing numerical stability.

4.5.1 Modified MFGP with GLS mean removal and separable spatio-temporal covariance

To address the limitations discussed in the previous section, we extend the multi-fidelity Gaussian process (MFGP) framework by explicitly incorporating a generalized least squares (GLS) mean-removal step within the likelihood-based inference procedure. The resulting method preserves the decomposed Vecchia structure while ensuring that systematic offsets between fidelity levels do not distort the covariance estimation.

Let $\boldsymbol{y} = [\boldsymbol{y}_L^\top, \boldsymbol{y}_H^\top]^\top$ denote the stacked observation vector. The model relies on the covariance decomposition

$$\boldsymbol{K} = \boldsymbol{A}\boldsymbol{\Sigma}_w\boldsymbol{A}^\top + \boldsymbol{D}_\epsilon,$$

where $\boldsymbol{\Sigma}_w$ is block-diagonal and contains the latent covariance structures of the low-fidelity process f_L and the discrepancy process δ. The key idea is to apply the Vecchia approximation independently to these latent components, yielding sparse precision representations $\boldsymbol{\Sigma}_L^{-1}$ and $\boldsymbol{\Sigma}_\delta^{-1}$.

The full multi-fidelity precision matrix is never formed explicitly. Instead, inference is carried out through the Woodbury identity, leading to the auxiliary sparse system

$$\boldsymbol{H} = \boldsymbol{\Sigma}_w^{-1} + \boldsymbol{A}^\top\boldsymbol{D}_\epsilon^{-1}\boldsymbol{A}.$$

All required linear solves and log-determinant computations are performed using sparse Cholesky factorization of $\boldsymbol{H}$, ensuring computational scalability.

Before evaluating the marginal likelihood, the mean structure is estimated via GLS

under the implied covariance $\boldsymbol{K}$. In its simplest form, the model assumes fidelity-specific intercepts, such that $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{G}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\beta_L, \beta_H)^\top$ and

$$\boldsymbol{G} = \begin{bmatrix} \mathbf{1}_{n_L} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_H} \end{bmatrix}.$$

This formulation accounts for global differences in baseline levels between low- and high-fidelity observations.

To capture more complex discrepancies, we further consider an adaptive GLS specification in which the mean depends on spatial location. Denoting the spatial coordinates by $\boldsymbol{s} = (s_{lat}, s_{lon})$, the mean structure is defined as

$$\mathbb{E}[y_f(\boldsymbol{s})] = \beta_{f,0} + \beta_{f,1} s_{lat} + \beta_{f,2} s_{lon}, \quad f \in \{L, H\}.$$

This induces a block design matrix $\boldsymbol{G}_{\text{gls}}$ where each fidelity is associated with its own spatially varying linear trend, allowing the model to account for geographically heterogeneous biases between data sources.

In both cases, the regression coefficients are estimated via the GLS estimator

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{G}^\top \boldsymbol{K}^{-1} \boldsymbol{G})^{-1} \boldsymbol{G}^\top \boldsymbol{K}^{-1} \boldsymbol{y},$$

and the centred residual vector is defined as

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{G}\hat{\boldsymbol{\beta}}.$$

All likelihood computations are subsequently performed using $\tilde{\boldsymbol{y}}$, ensuring that the covariance structure is estimated on residual variation rather than systematic offsets.

The negative log-marginal likelihood is therefore given by

$$\mathcal{NLML} = \frac{1}{2} \tilde{\boldsymbol{y}}^\top \boldsymbol{K}^{-1} \tilde{\boldsymbol{y}} + \frac{1}{2} \left(\log |\boldsymbol{\Sigma}_w| + \log |\boldsymbol{H}| + \log |\boldsymbol{D}_\epsilon| \right) + \frac{N}{2} \log(2\pi),$$

with an additional correction term $\log |\boldsymbol{G}^\top \boldsymbol{K}^{-1} \boldsymbol{G}|$ in the restricted maximum likelihood formulation.

Both latent processes are modelled using separable spatio-temporal covariance functions. Let $x = (\boldsymbol{s}, t)$ denote a spatio-temporal input. The covariance is specified as

$$k((\boldsymbol{s}, t), (\boldsymbol{s}', t')) = k_s(\boldsymbol{s}, \boldsymbol{s}') \cdot k_t(t, t'),$$

where k_s and k_t represent spatial and temporal covariance components. In the implementation, these are chosen as either squared-exponential (RBF) kernels or Matérn kernels, depending on the application. For instance, the RBF specification takes the form

$$k_s(\boldsymbol{s}, \boldsymbol{s}') = \sigma_s^2 \exp\left(-\frac{\|\boldsymbol{s} - \boldsymbol{s}'\|^2}{2\ell_s^2}\right), \quad k_t(t, t') = \sigma_t^2 \exp\left(-\frac{(t - t')^2}{2\ell_t^2}\right).$$

The full covariance matrix is then constructed via the element-wise product $\boldsymbol{K} = \boldsymbol{K}_s \odot \boldsymbol{K}_t$, which corresponds to a separable covariance structure.

This formulation implicitly assumes that spatial and temporal dependencies factorize and that both components are stationary. Furthermore, the low-fidelity and discrepancy processes are assumed independent, while the Vecchia approximation introduces a local dependence structure by conditioning each observation on a limited neighbourhood.

From a computational perspective, the dominant cost arises from constructing the Vecchia factors for the latent processes and factorizing the sparse matrix $\boldsymbol{H}$. This yields an overall complexity of order

$$\mathcal{O}(n_L m_L^2) + \mathcal{O}(n_\delta m_\delta^2) + \mathcal{O}(\mathrm{nnz}(R)),$$

which is substantially more efficient than the cubic complexity required for exact multi-fidelity Gaussian process inference.

Overall, the procedure can be interpreted as a sequence in which the latent processes are approximated via Vecchia, the joint precision structure is reconstructed through the Woodbury identity, and the mean structure is removed via GLS prior to likelihood evaluation. This separation ensures that the covariance model captures only the spatio-temporal dependence and cross-fidelity interaction, while systematic

offsets are handled independently at the mean level.

4.6 Revised experimental framework

Following the methodological corrections introduced in the previous sections, the experimental design originally presented in this chapter has been substantially revised. The initial version of Chapter 4 was built around a formulation that, although empirically promising, was not fully consistent with the corrected multi-fidelity Gaussian process framework developed later in the accompanying paper. In particular, the earlier chapter did not clearly disentangle three separate issues: first, whether the approximation error introduced by the decomposed Vecchia strategy propagates in a controlled manner through the Woodbury reconstruction; second, whether the synthetic-data experiment is truly aligned with the corrected hierarchical multi-fidelity formulation; and third, whether the real-data comparison reflects the role of the new generalized least squares (GLS) mean-removal step and the revised interpretation of the cross-fidelity structure.

For this reason, the experimental section has been reformulated. The objective is not merely to replace the previous results with a more polished version, but rather to correct the logic of the empirical validation so that it becomes fully coherent with the updated methodology. The revised framework is organised into three complementary experiments. The first experiment is methodological and is designed to test whether the decomposed Vecchia approximation remains numerically stable after reconstructing the full multi-fidelity likelihood. The second experiment revisits the synthetic-data study under a corrected data-generating mechanism, allowing the approximated model to be compared directly with the exact multi-fidelity Gaussian process and with single-fidelity Gaussian process alternatives. The third experiment redefines the real-data validation under the corrected framework, introducing model variants that differ in mean specification, cross-fidelity scaling, and warping.

This revised organisation is important because the original version of the chapter implicitly treated all experiments as predictive comparisons, whereas the corrected

framework reveals that the first experiment must instead be interpreted as a validation of the approximation itself. Similarly, the synthetic-data experiment in the original chapter acted only as a general benchmark, while in the revised version it becomes a consistency check between exact and approximated multi-fidelity inference. Finally, the real-data experiment is no longer a generic model comparison, but a structured investigation of which modelling components remain stable and useful when the theoretical corrections of the framework are properly implemented.

4.6.1 Revised Experiment 1: uncertainty propagation in the decomposed Vecchia framework

The first revised experiment is specifically designed to address a limitation that was not resolved in the original chapter, namely whether the approximation error introduced by the decomposition

$$\boldsymbol{K} = \boldsymbol{A}\boldsymbol{\Sigma}_w\boldsymbol{A}^\top + \boldsymbol{D}_\epsilon$$

propagates non-destructively through the final Woodbury-based reconstruction of the full multi-fidelity precision structure. In the earlier version of the chapter, this point was effectively assumed rather than tested. This was problematic, because the entire computational framework depends on the fact that approximation is applied separately to the latent low-fidelity and discrepancy processes, rather than directly to the full covariance matrix. If the approximation errors generated at the latent level accumulated in an uncontrolled manner after reconstruction, then the framework would no longer be trustworthy, even if it appeared to perform well empirically on a few predictive tasks.

For this reason, the revised experiment is not framed as a standard prediction study. Its primary purpose is methodological. The experiment isolates approximation effects from parameter estimation noise by fixing the hyperparameters at values obtained via exact multi-fidelity inference. In this way, any discrepancy between the exact and approximated versions can be attributed to the decomposed Vecchia reconstruction itself, rather than to differences in optimization or unstable

likelihood estimation. The comparison is then carried out at the level of quantities that directly determine the likelihood and the predictive equations, namely the reconstructed inverse action $\boldsymbol{K}^{-1}\boldsymbol{y}$, the quadratic form $\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}$, and the log-determinant $\log|\boldsymbol{K}|$. In addition, predictive root mean squared error is measured on held-out high-fidelity observations in order to verify whether the reconstruction error also translates into practically relevant predictive distortions.

Relative to the earlier version of the chapter, this revised validation study is therefore much more informative. It does not merely say that the approximated model works reasonably well; rather, it explains *why* the framework is computationally reliable. It also makes it possible to compare different neighbour-selection strategies and to quantify the extent to which approximation quality improves as the conditioning size increases. This experiment therefore provides the missing internal justification for the corrected framework and plays a foundational role for the two experiments that follow.

The results of this revised validation study are reported in Table 4.4. These results show a clear pattern. First, for both nearest-neighbour conditioning and correlation-based conditioning, all relative approximation errors decrease as the neighbour size m increases. This is important because it demonstrates that the decomposition does not create unstable behaviour: approximation error can be systematically controlled by increasing the amount of local information used in the Vecchia approximation. Second, correlation-based conditioning performs substantially better than a simpler geometric nearest-neighbour strategy. This indicates that, in the present multi-fidelity spatio-temporal setting, selecting neighbours according to empirical dependence is more informative than selecting them purely by distance. Third, the decrease in approximation error is accompanied by an improvement in predictive RMSE, which confirms that the likelihood reconstruction is not only numerically accurate but also relevant for prediction.

The main conclusion of this first revised experiment is therefore that the corrected decomposition is statistically and numerically viable. In the context of the original chapter, this is a substantial improvement: rather than treating the approximation as an implementation trick, the revised chapter now validates it directly as an

Table 4.4: Replicated validation of the decomposed Vecchia approximation. Results are averaged over 20 replications. Average exact RMSE is 0.688. This comparison does not involve GLS adjustment.

Neighbour Selection	m	n_{rep}	Mean rel. $\|\boldsymbol{K}^{-1}\boldsymbol{y}\|$ (SD)	Mean rel. $\log\|\boldsymbol{K}\|$ (SD)	Mean rel. $\boldsymbol{y}^\top\boldsymbol{K}^{-1}\boldsymbol{y}$ (SD)	Mean RMSE
Nearest-Neighbor	10	20	0.464 (0.045)	0.789 (0.144)	0.0378 (0.0202)	1.130
Nearest-Neighbor	20	20	0.345 (0.026)	0.223 (0.045)	0.0326 (0.0226)	1.033
Nearest-Neighbor	30	20	0.315 (0.025)	0.189 (0.040)	0.0338 (0.0222)	0.985
Nearest-Neighbor	40	20	0.298 (0.022)	0.175 (0.037)	0.0275 (0.0202)	0.990
Nearest-Neighbor	60	20	0.208 (0.021)	0.066 (0.011)	0.0193 (0.0166)	0.888
Corr	10	20	0.345 (0.026)	0.247 (0.050)	0.0221 (0.015)	0.957
Corr	20	20	0.220 (0.017)	0.032 (0.007)	0.0269 (0.021)	0.883
Corr	30	20	0.172 (0.019)	0.0084 (0.006)	0.0152 (0.010)	0.790
Corr	40	20	0.124 (0.015)	0.0055 (0.004)	0.0131 (0.010)	0.767
Corr	60	20	0.0748 (0.010)	0.0074 (0.003)	0.0088 (0.006)	0.720

object of study.

4.6.2 Revised Experiment 2: synthetic data experiment

The second revised experiment updates the synthetic-data analysis. In the original chapter, the synthetic experiment provided an initial comparison between spatio-temporal Gaussian process models and multi-fidelity models, but it was not fully aligned with the corrected framework. In particular, the previous version did not explicitly separate the role of the exact multi-fidelity model from that of the approximated Vecchia-based version, and it did not place sufficient emphasis on uncertainty calibration. As a consequence, the original synthetic experiment was useful as a first benchmark but only partially informative about whether the new scalable framework remains faithful to the exact model.

The revised synthetic experiment addresses these issues by redefining the data-generating process so that it directly matches the corrected autoregressive multi-fidelity formulation. The data are generated on a regular spatio-temporal grid, with the high-fidelity process defined as a scaled version of the low-fidelity process plus an independent discrepancy component. Both the low-fidelity latent process and the discrepancy process are modelled using separable squared-exponential covariance kernels. The experimental design therefore provides a controlled setting in which the true dependence structure is fully known and the exact multi-fidelity Gaussian process remains computationally feasible.

This revised experiment differs from the original chapter in three important ways. First, it now compares the corrected scalable approximation directly against the

exact multi-fidelity Gaussian process, rather than treating all methods as equally generic competitors. This is essential, because the main methodological question is whether the approximated framework preserves the characteristic advantage of the multi-fidelity model. Second, the experiment explicitly includes uncertainty calibration through empirical 95% coverage. This is an important addition, since one of the limitations of the earlier chapter was that predictive uncertainty was not systematically assessed. Third, the benchmark set is now formulated more clearly in terms of the information available to each competing Gaussian process model: a model based only on the low-fidelity signal, a model based only on spatio-temporal coordinates, and a model combining both.

The interpretation of the revised synthetic results is given in Table 4.5. Across both noise settings, the exact multi-fidelity model (Classic) achieves the best MAE and RMSE. This is expected, because it uses the correct multi-fidelity structure without any approximation. The key point, however, is that the Vecchia-based approximation remains close to the exact model and clearly outperforms all the single-fidelity Gaussian process alternatives. This result is central for the thesis: it shows that the scalable framework does not simply reduce computation, but does so while retaining most of the predictive benefit of exact multi-fidelity inference.

The revised synthetic experiment also clarifies the role of noise. As the discrepancy variance increases from $\sigma_d^2 = 2$ to $\sigma_d^2 = 4$, all models deteriorate, but the deterioration is much more pronounced for the single-fidelity GP baselines. By contrast, the exact and approximated multi-fidelity models remain more robust. This confirms that, when the relationship between the two fidelities is correctly represented, the multi-fidelity structure provides a form of protection against increased noise that standard Gaussian process baselines do not possess.

Perhaps the most important addition relative to the original version of the chapter concerns uncertainty quantification. The GP baselines exhibit severe undercoverage, especially at higher noise levels. This means that their predictive intervals are too narrow and therefore overconfident. In contrast, both the exact multi-fidelity model and the Vecchia approximation provide empirical coverage close to the nominal 95% level, with the approximated model in fact showing

the best calibration in both scenarios. This result is methodologically important because it confirms that the corrected scalable framework is not only accurate in point prediction but also reliable in its uncertainty statements. The original chapter did not make this point clearly enough.

Noise Level	Metric (sd)	GP-3D	GP-L	GP-4D	Classic	Vecchia_v4
$\sigma_d^2 = 2$	MAE	2.143 (0.265)	2.346 (0.334)	2.257 (0.309)	**1.343 (0.131)**	1.527 (0.183)
	RMSE	2.681 (0.320)	2.935 (0.399)	2.831 (0.370)	**1.699 (0.170)**	1.923 (0.224)
	COV95	0.71 (0.058)	0.604 (0.069)	0.645 (0.057)	0.939 (0.033)	**0.953 (0.039)**
$\sigma_d^2 = 4$	MAE	3.463 (0.461)	4.009 (0.547)	3.926 (0.523)	**2.265 (0.314)**	2.661 (0.463)
	RMSE	4.324 (0.554)	5.023 (0.655)	4.919 (0.626)	**2.885 (0.398)**	3.345 (0.575)
	COV95	0.61 (0.06)	0.46 (0.06)	0.49 (0.05)	0.933 (0.048)	**0.959 (0.055)**

Table 4.5: Performance comparison for the revised simulation study. Values are reported as mean (standard deviation). The neighborhood size is set to 40, observations are ordered temporally, and neighbour selection is based on correlation conditioning.

Overall, the revised synthetic-data experiment plays a much stronger role in the chapter than the original version. It now serves as a direct consistency check between the corrected theory and the corrected implementation, while also demonstrating that the scalable approximation retains both predictive accuracy and calibrated uncertainty under controlled conditions.

4.6.3 Revised Experiment 3: real data experiment

The third revised experiment concerns the application to real wind-speed data in southern Lombardy. This part of the chapter required the most substantial reformulation. In the original version, the real-data experiment compared different models, but it did not fully reflect the corrected methodological insights developed later in the paper. In particular, the original experiment did not properly separate the role of the mean structure from that of the covariance structure, and therefore did not make clear whether the estimated cross-fidelity relationship was genuinely capturing shared spatio-temporal dependence or merely compensating for systematic level differences between the two data sources.

The corrected framework resolves this issue by incorporating the GLS mean-removal procedure before likelihood evaluation and prediction. This is a substantial conceptual improvement over the original version of the chapter. Once fidelity-specific offsets are removed, the covariance structure is no longer forced to absorb large-

scale mean discrepancies between low- and high-fidelity signals. As a consequence, the cross-fidelity parameter ρ can be interpreted more cleanly, and the comparison between model variants becomes much more meaningful.

Another important difference relative to the original chapter is that the revised real-data experiment is no longer presented as a single comparison between one Gaussian process baseline and one multi-fidelity model. Instead, it is organised around a set of model configurations that reflect the main methodological choices introduced by the corrected framework. These choices concern three aspects: the treatment of the mean structure, the form of the cross-fidelity scaling, and the possible use of warping. This is essential, because the paper shows that these components lead to qualitatively different behaviours in terms of accuracy, robustness, and uncertainty calibration.

The revised experimental configurations are summarised in Table 4.6. The benchmark is an approximated GP-3D model trained only on high-fidelity data. The multi-fidelity models then differ according to whether the GLS step is global or adaptive, whether the cross-fidelity parameter is constant or spatially varying, and whether warping is applied. This table should replace the simpler setup used in the earlier version of the chapter, because it is now directly tied to the corrected methodological framework.

Model ID	**Mean Structure**	$\rho(\boldsymbol{s})$	Input Warping
GP-3D (Approx)	NA	NA	NO
$MFGP_{gc}$	Global GLS	Constant	NO
$MFGP_{ac}$	Adaptive GLS	Constant	NO
$MFGP_{gWc}$	Global GLS	Constant	YES
$MFGP_{aWc}$	Adaptive GLS	Constant	YES
$MFGP_{gGP}$	Global GLS	GP-based (empirical)	NO
$MFGP_{aGP}$	Adaptive GLS	GP-based (empirical)	NO

Table 4.6: Revised experimental configurations for the real-data application.

The aggregated predictive results are reported in Table 4.7. These results provide the strongest practical support for the corrected framework. The first point to note is that all the best-performing models are multi-fidelity models, and they clearly outperform the single-fidelity GP benchmark in both MAE and correlation. This confirms that, even in a large and irregular real-world dataset where exact inference is infeasible, the multi-fidelity structure remains highly beneficial.

Among the revised models, the simplest configuration, namely $MFGP_{gc}$ with global GLS and constant ρ, emerges as the most reliable overall in terms of point prediction accuracy and correlation. This is an important finding, and it changes the interpretation of the chapter relative to the earlier version. The original framework tended to suggest that additional flexibility would necessarily improve the model. The corrected results show instead that, in a real monitoring network with sparse and uneven spatial support, disciplined structure can be more effective than added flexibility. In other words, the best model is not the most complex one, but the one that most cleanly separates offsets, shared dependence, and discrepancy.

The adaptive GLS variants illustrate this point even more clearly. Their point prediction performance remains competitive, but the prediction interval coverage deteriorates sharply, especially for $MFGP_{ac}$. This indicates overconfidence. The likely interpretation is that the spatially adaptive mean structure can overfit regional trends when spatial support is weak, thereby shrinking the residual uncertainty too aggressively. This is precisely the kind of issue that the original version of the chapter could not explain well, because the role of the mean structure had not yet been fully formalised.

The warped variants provide a different kind of insight. Their point prediction accuracy is slightly weaker than that of the best unwarped model, but they provide better uncertainty calibration, with $MFGP_{gWc}$ yielding the best coverage among the multi-fidelity models. This suggests that non-Gaussianity in wind-speed data has a measurable impact on predictive uncertainty, even when the point forecasts are already quite accurate. This is also an important conceptual bridge with Chapter 3, where warping was introduced in the temporal setting.

Finally, the models with spatially varying empirical $\rho(\boldsymbol{s})$ underperform in this application. This is a useful negative result. It indicates that, although non-stationary cross-fidelity coupling is theoretically appealing, empirical estimation of $\rho(\boldsymbol{s})$ may be too noisy when the number of stations is small and spatial support is limited. This finding is especially valuable for the thesis because it shows that the corrected framework is not simply adding complexity for its own sake; rather, it allows one to evaluate which types of flexibility are actually justified by the data.

Model ID	Count	MAE	RMSE	Corr	$PICP_{95}$	NLML
$GP-3D$	18	0.3974	0.5245	0.77	93.84%	
$MFGP_{gc}$	18	**0.1906**	**0.2429**	**0.9351**	89.06	2790.6
$MFGP_{ac}$	18	0.1935	0.2477	0.9312	58.28	2738.5
$MFGP_{gWc}$	18	0.1949	0.2487	0.9307	**91.39**	3323.3
$MFGP_{aWc}$	18	0.2005	0.2552	0.9304	76.94	3274.5
$MFGP_{gGP}$	18	0.3203	0.4007	0.8729	83.78	2729.3
$MFGP_{aGP}$	18	0.4446	0.5598	0.8854	83.17	2786.3

Table 4.7: Aggregated performance metrics across 18 validation stations in the revised real-data experiment. Bold values indicate the best performance in each category.

Taken together, the revised real-data experiment provides a much more coherent empirical conclusion than the original version of the chapter. It shows that the corrected framework delivers substantial gains over a single-fidelity benchmark, that the GLS adjustment is crucial for meaningful multi-fidelity estimation, that excessive flexibility in the mean structure may reduce uncertainty calibration, and that warping can improve reliability of predictive intervals when skewness is relevant.

4.6.4 Overall interpretation of the revised experiments

The three revised experiments should therefore be read together. The first experiment validates the internal numerical logic of the corrected approximation. The second shows that, in a controlled setting where the truth is known, the approximated framework closely reproduces exact multi-fidelity behaviour while preserving uncertainty calibration. The third demonstrates that these advantages translate into practical gains in a realistic environmental reconstruction task.

This revised structure is substantially stronger than the original version of the chapter. In the earlier framework, the experiments were informative but not fully aligned with the theoretical model. In the revised framework, each experiment has a clearly defined role. The first justifies the approximation, the second validates it under known conditions, and the third tests its practical usefulness under realistic complexity. This progression makes the empirical part of the chapter fully consistent with the corrected methodology and explains why some of the ambiguities of the original version are resolved only after introducing the new framework.

4.7 Discussion and alternative approaches

This chapter presented a revised framework for computationally efficient spatio-temporal multi-fidelity Gaussian processes (MFGPs). Compared to the earlier formulation, the key modification consists in a clearer separation between the mean structure and the covariance model, achieved through the integration of a generalized least squares (GLS) mean-removal step within the likelihood-based inference procedure. This adjustment ensures that systematic discrepancies between fidelity levels are not absorbed into the covariance structure, thereby improving both interpretability and numerical stability.

From a computational perspective, scalability is achieved through the decomposition of the multi-fidelity covariance into latent components and the application of the Vecchia approximation at the level of the underlying processes. This allows the likelihood to be evaluated without explicitly forming the full covariance matrix, while the use of sparse linear algebra substantially enhances numerical stability. In contrast to the original framework, where approximation and modelling choices were more intertwined, the revised approach provides a more principled separation between approximation (Vecchia), covariance modelling, and mean specification.

The experimental results presented in Section **??** highlight the practical implications of these modifications. In particular, the revised framework demonstrates that approximation errors introduced at the latent level do not propagate destructively through the model, confirming the validity of the decomposition strategy. Moreover, the introduction of GLS significantly improves the identifiability of the cross-fidelity relationship, preventing the scaling parameter ρ from compensating for simple baseline differences between low- and high-fidelity data. This represents a key conceptual improvement over the initial formulation.

Despite these advancements, several limitations remain. First, the effectiveness of the framework depends on the adequacy of the separable spatio-temporal covariance assumption. While this choice provides a favourable balance between flexibility and computational tractability, it may not fully capture complex interactions between space and time in highly heterogeneous environments. This issue becomes

particularly evident in large-scale applications such as the Lombardy case study, where the coexistence of mountainous areas, urban regions, and flat terrain induces markedly different local dynamics.

Second, although the GLS step successfully removes systematic offsets, it introduces an additional layer of modelling decisions, particularly in its adaptive formulation. As observed in the empirical analysis, overly flexible mean structures may lead to overfitting and reduced uncertainty calibration, especially in regions with sparse observations. This highlights a fundamental trade-off between bias correction and model stability that was not explicitly addressed in the original framework.

Another limitation concerns the sensitivity of the model to anomalies in the data. Since the framework relies on Gaussian assumptions at multiple stages, extreme observations in the low-fidelity signal can still propagate through the model and affect the inferred dependence structure. This motivates the extension of the current approach towards robust multi-fidelity Gaussian processes, where heavy-tailed distributions or alternative likelihood formulations could be employed to mitigate the influence of outliers.

Furthermore, while the proposed approach improves scalability, it remains inherently global in nature. As a consequence, it may struggle to fully capture strongly localized non-stationary behaviour without introducing additional complexity in the covariance structure. Two possible directions emerge from this limitation. One consists in developing surrogate modelling strategies based on local approximations, where multiple models are trained over subregions and subsequently combined through a smoothing procedure. Another promising direction involves spatially adaptive modelling approaches, such as those based on clustering or spatially varying coefficients, which allow the dependence structure to change across the domain in a data-driven manner.

Finally, the issue of physical constraints, such as the positivity of environmental variables, remains relevant. While the warping approach introduced in Chapter 3 provides an effective solution in practice, it operates as an external transformation rather than being fully integrated into the probabilistic structure of the model. A more principled treatment of such constraints within the multi-fidelity framework

represents an additional avenue for future research.

Overall, the revised framework clarifies the respective roles of mean specification, covariance modelling, and approximation, leading to a more robust and interpretable methodology. At the same time, the experimental results indicate that careful control of model flexibility is essential in practical applications, particularly when dealing with heterogeneous spatial domains and multi-source data.

Chapter 5

Overview, Reflections and Future Developments.

This final chapter brings together the main findings, reflections, and methodological contributions developed throughout this thesis. Building upon the preceding chapters, it revisits the original research objectives and questions, synthesizes the key outcomes, and situates them within the broader context of environmental data fusion and statistical modelling.

Together, these discussions aim to consolidate the thesis's main contributions while identifying pathways for future advancements in multifidelity modelling for environmental applications. The chapter concludes with a summary of the answers to the three research questions (RQ1–RQ3), emphasizing how the proposed methods collectively enhance the reliability, and scalability of multifidelity Gaussian process models in real-world environmental monitoring contexts.

5.1 Thesis overview

This thesis illustrated how environmental monitoring networks present various challenges related to missing data and undercoverage with a specific example of wind speed applications. Classical interpolation methods, such as GP regression, can offer a viable solution for filling these gaps; however, they remain constrained by several difficult conditions, including long sequences of missing data, strong normality as-

sumptions, and the limited availability of highly reliable observations. These factors often undermine the effectiveness of traditional, single-fidelity modelling approaches. Data fusion has emerged as a promising strategy to address these issues (**RQ1**). By leveraging multiple correlated data sources, it becomes possible to exploit their complementary strengths. When high-quality data are missing, one can learn the error structure linking them to more abundant but lower-quality sources, and then use this relationship to reconstruct high-fidelity information from low-fidelity observations. The first research question of this thesis therefore investigates whether the use of multifidelity Gaussian processes can reduce the reliance on abundant HF samples and to which extent under realistic noise and sparsity conditions.

Since environmental variables such as wind speed often deviate from Gaussianity, the assumption of normal errors in MFGP models becomes unrealistic. To ensure that model predictions remain well-calibrated, it is necessary to develop a strategy capable of capturing such skewed distributions without disrupting the cross-fidelity relationships that underlie MF learning (**RQ2**). The second research question therefore focused on identifying approaches that can robustly handle non-Gaussian behaviour in a MF setting, while not distorting the relation between across fidelity levels.

Finally, environmental datasets are often inherently large and complex, spanning wide spatial domains and long temporal horizons. Under these conditions, scaling MFGP models becomes computationally demanding, and ensuring numerical stability can be challenging. The third research question (**RQ3**) thus examined the capacity of MFGPs to operate effectively in large spatio-temporal contexts, and motivated the development of a new spatio-temporal MFGP framework based on a Vecchia-approximated likelihood and spatially varying coupling. This framework is assessed in Chapter 4 in terms of numerical stability, computational scalability, and flexibility in representing spatial heterogeneity across multiple sites and time periods.

5.1.1 Chapter 2: a comparison with relevant univariate methods

Multifidelity models were multi-response methods that organized responses into distinct levels of fidelity.[1] Whether multifidelity modelling was the most appropriate approach depended strongly on the relative availability of data across fidelity levels. In particular, the ratio between the number of high- and low-fidelity observations ($\frac{n_H}{n_L}$) largely determined the potential benefit of multifidelity models over simpler, single-fidelity alternatives. When high-fidelity data were abundant (i.e., the ratio was large), univariate methods could perform equally well or even better. Therefore, the practical effectiveness of multifidelity approaches proved to be highly application dependent.

This chapter adopted an unconventional structure. It presented a sequence of experiments whose outcomes did not necessarily demonstrate the superiority of multifidelity methods but instead defined their limitations and potential for future development. The work progressed from an initial unsuccessful experiment, through a preliminary discussion, to a redesigned simulation study incorporating new modelling considerations.

The application domain investigated in this thesis involved wind speed as a representative environmental variable. Wind speed was particularly suitable for this analysis because of its intermittent gusts, pronounced skewness, and strictly positive values—all of which made spatial interpolation challenging. Moreover, wind speed was a key variable in applications such as wind resource assessment and air-pollution prediction (Otto et al., 2024).

The chapter compared the performance of multifidelity models with two benchmark methods: Gaussian process regression, the univariate counterpart of the multifidelity Gaussian process, and QGBRT, a machine-learning approach commonly used in industry for wind-speed forecasting. The comparison considered various wind-speed generation mechanisms, different sets of predictors, and—most impor-

[1] Although this thesis focused on high-fidelity predictions, the same framework could have been extended to low-fidelity predictions. In practice, high-fidelity process parameters were estimated jointly with low-fidelity parameters.

tantly—different levels of noise and high-to-low fidelity data ratios (n_H/n_L).

The initial set of simulations illustrated that excessive noise in the low-fidelity data could obscure the true relationship between fidelity levels. When noise dominated, two datasets that were in fact correlated appeared uncorrelated, making a multifidelity approach ineffective. In the first experiment, skewness was introduced solely through the error term—that is, through the stochastic component. As a result, increasing skewness corresponded to increasing noise levels, which weakened the observable relationship between high- and low-fidelity data.

These initial findings motivated a second simulation study, in which skewness was instead incorporated into the deterministic component of the data-generation process. This revised design revealed that the performance of univariate methods, such as Gaussian process regression and quantile gradient boosted regression trees, depended strongly on the size of data gaps. In contrast, multifidelity methods exhibited a degree of robustness: both their predictive uncertainty and their overall performance remained relatively stable across varying high-fidelity sample sizes.

By contrast, Gaussian process regression performed poorly under highly sparse conditions, often yielding overly flat predictions. This behaviour was consistent with findings in the literature (Colombo and Fassò, 2022; Fassò et al., 2020), which indicated that Gaussian process regression learned correlation structures from the relative proximity of data points—when points were too far apart, there was insufficient information to capture meaningful relationships.

Although the multifidelity Gaussian process models generally performed well, highly skewed data could violate their underlying assumptions. These violations were not always explicit, as normality was assumed in the latent error and conditional distributions—neither of which were directly observed. In practice, Gaussian process regression could accommodate moderate skewness through the non-linearity of its latent function, $f(x)$. However, when the observed data were strongly skewed, the normality of the error term was likely violated, which limited model performance.

Overall, this chapter explored the influence of skewness and noise on multifidelity modelling and established a foundation for future work aimed at extending these methods to non-Gaussian settings and more complex environmental applications.

5.1.2 Limitations of Chapter 2

The analysis presented in Chapter 2.1 revealed several important limitations, most of which were driven by data characteristics rather than by the modelling framework itself. When the noise level in the low-fidelity data was high, or when the correlation between the high- and low-fidelity sources was weak, the multifidelity Gaussian process model struggled to capture meaningful cross-fidelity relationships. High noise and weak inter-fidelity correlations emerged as the primary factors influencing multifidelity performance. Because these factors are inherent properties of the data rather than flaws of the method, any resulting performance degradation should be attributed to the experimental conditions rather than to limitations of the modelling approach.

A second limitation concerns the behaviour of Gaussian process models under sparse sampling. The results highlight that GPs tend to oversmooth when HF data are limited, particularly if the length-scale is not tuned to capture adequate temporal structure. This behaviour, while well known in the literature, remains a practical challenge in real-world environmental applications where dense, high-quality data are rarely available.

Another simplifying assumption in the simulations is that residuals follow a skew-normal (and later, a normal) distribution. Although empirically motivated, this assumption oversimplifies the complex characteristics of real wind speed data, which often exhibit time-varying skewness and heavy tails. This may limit the generality of the conclusions regarding uncertainty quantification and predictive calibration.

Finally, although the QGBRT method provides quantile-based forecasts and the GP models produce variance estimates, the chapter does not explicitly assess the calibration or sharpness of predictive uncertainty—for instance, through Continuous Ranked Probability Scores (CRPS) or reliability diagrams. Overall, these limitations highlight the importance of realistic data conditions and careful model validation. They do not undermine the contributions of the chapter but rather delineate the empirical boundaries within which the presented findings hold, motivating the methodological extensions developed in the following chapters.

5.1.3 Chapter 3: Warped Multifidelity methodology

In the multifidelity framework, discrepancies between low- and high-fidelity data are typically modelled as a Gaussian process. This means the conditional distribution of this discrepancy is also assumed to be Gaussian. When the underlying data distributions are strongly non-normal, particularly skewed or heavy-tailed, this assumption becomes problematic and can lead to poor performance or miscalibrated uncertainty estimates. There are generally two approaches to addressing non-normality in GP models: modelling the skewness directly, often by introducing skewed likelihoods or more complex priors, or transforming the data to approximate normality before applying standard GP methods. Chapter 3 examined how to effectively model skewness within the context of MFGP modelling. This discussion was far from trivial, as it addressed the challenges of handling skewed data in GPs, as well as the additional complexities introduced by multiple data sources.

First, it is well studied that directly modelling skewness tends to introduce challenges, including increased model complexity, identifiability issues, and greater computational burden due to intractable posteriors. On the other hand, transformation-based methods—especially when datasets from multiple fidelities are involved- are also problematic. For example, different datasets might requires separate transformations to be normalized. Using different transformations for each data source can distort the inter-dataset relationships that are central to effective multifidelity modelling. Using instead a single transformation (i.e. log) usually is sub-optimal for one of the data sources. Using the same transformation could, in principle, exacerbates the problem: the discrepancy between the two datasets, which is central for MFGP modelling might become more skewed if both data sources are not appropriately normalized. To address this, the thesis proposes a novel strategy, a joint transformation that preserves inter-dataset relationships while normalizing the data. One promising solution is the use of nonparametric warping transformations, which map data to an approximately normal space using empirical cumulative distribution function (CDF) estimates. This approach leverages quantile invariance, ensuring that monotonic transformations preserve rank-based relation-

ships between datasets—a crucial feature for coherent data fusion.

This idea is formalized in the paper Warped Multifidelity Gaussian process (WMFGP) Colombo et al. (2025), which introduces a flexible nonparametric warping layer into the multifidelity GP framework. The method allows for independent warping functions, effectively normalizing multiple datasets with differing marginal distributions. By learning a joint warping that preserves cross-fidelity structure, the WMFGP allows for accurate prediction and uncertainty quantification even in the presence of severe skewness and non-Gaussian noise. Nonparametric warping provides a powerful and practical way to extend Gaussian process models to non-normal, skewed data, while retaining the interpretability and flexibility of the multifidelity GP framework. This makes it especially suitable for environmental applications, where skewness, noise heterogeneity, and multiple data sources are the norm rather than the exception. This approach was applied to fill wind speed data gaps in the monitoring network of ARPA Lombardia, the primary environmental agency in Italy. The results revealed that interpolation uncertainty did not increase with the size of the data gaps. An explanation for the outcome lies in the temporal stability of relationships between datasets, ensuring that the error structure learned across fidelities remains highly predictable.

This work indicates that, multifidelity methods—which rely on learning inter-dataset relationships as a core mechanism—are capable of accurately predicting long data gaps without a corresponding increase in uncertainty. This behaviour stands in contrast to traditional interpolation methods, where uncertainty typically grows with gap length. Until this stage of the thesis, the results were obtained only for temporal studies, hence it was necessary to extend the modelling framework to the spatio-temporal case.

5.1.4 Limitations of Chapter 3

The proposed Warped multifidelity Gaussian process relies on simplifying assumptions of separability between fidelity levels and spatial isotropy. While these assumptions enhance interpretability and computational tractability, they may constrain the model's capacity to represent more complex, anisotropic, or non-stationary depen-

dencies across datasets. A further open question concerns the warping design: it remains to be investigated whether employing a single, shared warping function for both datasets could offer performance comparable to, or even exceeding, that of independent warpings. Moreover, the framework follows a multi-stage structure, which, although practical, does not formally propagate uncertainty across steps — a feature that limits its statistical elegance and theoretical coherence.

Finally, spatial dependencies were approximated through a clustering strategy rather than by full spatial modelling. While this choice improves scalability, it restricts the model's ability to extrapolate beyond observed regions. This limitation, along with possible spatial extensions, is further discussed in Chapter 4.

5.1.5 Chapter 4:a computationally efficient spatio-temporal model

To enable a comprehensive treatment of environmental data, it was necessary to extend the application of multifidelity Gaussian process models to spatio-temporal settings. This extension offers two primary advantages: prediction in unmonitored areas — allowing estimates where no monitoring stations are present and enhanced robustness through spatial correlations — leveraging the spatial dimension to stabilize predictions and improve generalization. However, introducing spatio-temporal modelling with Gaussian processes brings major computational challenges. Chief among them, is the cubic scaling of standard GP inference, which requires $O(n^3)$ operation for covariance matrix inversion. To address this, the Vecchia approximation was integrated into the multifidelity GP framework.

This method introduces sparsity in the precision matrix by approximating the joint distribution using a set of conditional independence assumptions. The result is a scalable, efficient, and accurate method, especially suitable for large datasets. Implementing the Vecchia approximation within a multifidelity setting is non-trivial as the multifidelity covariance matrix includes cross-level interactions that complicate standard conditioning schemes. To resolve this, our framework decomposes the full covariance matrix into independent components—one for each

fidelity level—allowing the Vecchia approximation to be applied separately to the low- and high-fidelity processes. This design not only permits tailored approximations for each fidelity level (e.g., denser conditioning for sparse HF data), but also improves numerical stability. The independent precision matrices are later combined through the Woodbury matrix identity, enabling efficient inversion of the full system while preserving the computational gains of sparsity. Notably, this decomposition maintains the physical relationships between fidelity levels, and the structure allows for nonstationary, spatially-varying integration parameters ρ, which significantly enhances flexibility in heterogeneous environments. Simulation studies and real-world experiments—particularly on wind speed data in Lombardy—demonstrated that the proposed approach is not only significantly faster and more scalable than standard multifidelity GPs, but also more robust. The Vecchia-based multifidelity model consistently delivered accurate predictions with reduced numerical instability, outperforming baseline GP implementations, especially as data size increased. This framework, by jointly addressing computational limitations and spatial heterogeneity, marks a significant advancement for scalable, high-resolution environmental modelling using multifidelity data fusion.

5.1.6 Limitations of Chapter 4

The model developed in Chapter 4 to capture local heterogeneity introduces an increasing number of parameters, which in turn makes parameter estimation considerably more complex. In this sense, increasing the modelling flexibility might be not the best strategy for addressing local heterogeneity. More efficient and perhaps even more precise schemes could be available, which will be presented in Sections 5.3.3 and 5.3.4.

The presence of anomalous observations—whether in the low-fidelity or high-fidelity data sources—poses a substantive concern for multifidelity modelling. Since these models fundamentally rely on the stability of the relationship between fidelity levels, any disruption to this relationship can lead to degraded model performance and unreliable inference. Outliers or systematic errors in either data stream can distort the estimated covariance structure or the cross-fidelity scaling parameter ρ,

leading to biased predictions and inflated uncertainty estimates. In other words, when HF and LF data are strongly correlated but the LF data contain outliers that co-move with valid HF observations, the model interprets these co-movements as meaningful signals.

5.2 Reflections on the difference between Temporal and Spatio-Temporal results

A direct comparison between the models presented in Chapter 3 (temporal) and Chapter 4 (spatio-temporal) cannot be outlined. Such a comparison would only be meaningful under standardised experimental conditions—for instance, if both models were evaluated on identical datasets, forecast horizons, and contextual information. Since this was not the case, any apparent difference in performance between the two settings must be interpreted with caution. However, several contextual observations that explain the differences can be performed.

In particular, the forecast horizon differs substantially between the two experiments. In the temporal setting of Chapter 3, the longest prediction gap was approximately one week of wind-speed data. In contrast, the spatio-temporal experiment in Chapter 4 involved predicting up to a month of data. Because prediction error generally scales with the length and variability of the underlying signal—especially in domains like wind speed where longer periods increase the likelihood of rare, high-magnitude events (such as gusts)—the spatio-temporal setting naturally incurs higher error. Moreover, while the temporal task focused on gap filling, the spatio-temporal task addressed under-coverage—predicting time series for under-monitored or entirely unmonitored sites. Temporal gap filling benefits from partial observations at the target location, whereas spatial inference depends solely on more distant or indirectly related data.

Therefore, the apparent discrepancy in performance between the two settings (time and space-time) should not be attributed to model quality alone but rather to the inherent differences in problem formulation and difficulty. The temporal task leverages continuity over time, whereas the spatio-temporal task requires gen-

eralization across space, representing a more ambitious and inherently complex modelling challenge.

5.3 Future research Directions

5.3.1 Local and Unified Modelling Approaches

The methodology presented in this thesis employ sequential two-stage approaches dealing with skewness where input space transformations (warping) are applied as a preprocessing step prior to Gaussian process modelling. While this separation of concerns provides interpretability and computational tractability, alternative frameworks merit consideration for future work.

5.3.1.1 Local Gaussian Process Models

A principled alternative to global GP modelling is the local Gaussian process framework, extensively developed by Gramacy and collaborators (Gramacy, 2020). Rather than assuming a global non-stationary structure, local GPs partition the input domain into regions and fit separate local models to each partition. This approach offers natural advantages for applications exhibiting piecewise heterogeneous behavior: computational efficiency through smaller local design matrices, flexibility in accommodating abrupt structural changes without explicit parametric assumptions, and reduced complexity in hyperparameter estimation within each local region.

The trade-offs involve managing boundary effects between partitions and increased flexibility that may require careful regularization. For the wind speed application studied here, where regional climate patterns and localized temporal dynamics are evident, a local GP approach could potentially capture these heterogeneities more directly than the global parametric non-stationarity structure employed in this work.

5.3.1.2 Deep Gaussian Processes for Unified Warping and Modelling

The warping method presented in this thesis (non-parametric approach) follow a decoupled paradigm: the transformation is optimized separately in a preprocessing stage, followed by standard GP modelling on the transformed data. An alternative is

to employ deep Gaussian processes (deep-GPs) (Macaulay et al., 2025; Sauer et al., 2023), which learn input space transformations jointly with the model through stacked layers of Gaussian processes, similarly to the parametric approach. In this unified framework, successive GP layers induce non-linear warping of the input space, with all parameters (including warping) estimated simultaneously through a single likelihood optimization.

Deep-GPs offer conceptual advantages: uncertainty in the learned transformations propagates naturally through predictive uncertainty estimates, the model has flexibility to discover complex or simple transformations as dictated by the data rather than user specification, and there is no artificial separation between transformation and modelling objectives. Conversely, computational costs are substantially higher, interpretability of learned transformations can be limited, and the inference procedures are more complex than standard GPs.

For future investigation, comparing the two-stage warping-then-GP approach against a deep-GP implementation would illuminate whether joint optimization provides practical benefits in predictive accuracy and whether the interpretability gain from decoupled methods justifies the potential performance cost.

5.3.2 Robust multifidelity Gaussian process modelling

As discussed in Section 5.1.6, MFGP models in their current form are sensitive to anomalous observations, as they lack the ability to distinguish between outliers and valid data points. The following section presents a simple and practical approach to mitigate this limitation. The proposed method is first introduced conceptually and then illustrated through a series of toy examples.

A promising direction for future work is the incorporation of a robust loss function—such as the Huber loss—to further reduce sensitivity to outliers. The Huber loss behaves like a squared loss for small errors, preserving accuracy, and like an absolute loss for large errors, limiting the influence of outliers. This property makes it particularly suitable for applications where data contamination is expected. Even though, a comprehensive investigation of this approach lies beyond the scope of this thesis, here below the general intuitions around this methods are provided.

To ensure robustness against outliers in high-fidelity observations, it is possible to minimize a Huber loss function applied to the weighted residuals of the high-fidelity predictions. Let $\hat{\mathbf{y}}_H$ denote the predicted high-fidelity outputs and $\mathbf{r} = \mathbf{y}_H - \hat{\mathbf{y}}_H$ the residuals. These are scaled by the inverse square root of the predictive variance to obtain standardized residuals:

$$\mathbf{z} = \mathbf{r} \odot \sqrt{\text{diag}(\mathbf{W}_{\text{inv}})}, \tag{5.1}$$

where $\mathbf{W}_{\text{inv}}$ is a block of the inverse covariance matrix corresponding to held-out points. A robust scale estimate (Huber, 1964) is computed using the Median Absolute Deviation (MAD):

$$\delta = 1.345 \cdot \frac{\text{MAD}(|\mathbf{z}|)}{0.6745}. \tag{5.2}$$

The final objective is the sum of Huber losses:

$$\mathcal{L}_{\text{Huber}} = \sum_i \begin{cases} \frac{1}{2} z_i^2 & \text{if } |z_i| \leq \delta, \\ \delta(|z_i| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases} \tag{5.3}$$

This robust objective provides resilience to anomalies in high-fidelity data and guides hyperparameter optimization of the MFGP model. A simple illustrative example is presented to demonstrate the behaviour of a robust implementation of the Multifidelity Gaussian process model. The experiment is conducted on a toy dataset, shown in Figure 5.1, which consists of simulated wind speed data. The dataset includes six low-fidelity and six high-fidelity locations, each with 100 time points. A spatial layout of the dataset is provided in Figure 5.1.

The robust implementation intentionally underfits the training data, as reflected in the log-likelihood values: the standard model achieved a log-likelihood of -98, while the robust variant yielded a significantly higher (less negative) value of 589.632. This underfitting is also visually evident in Figure 5.2, where the standard model (blue dotted line) closely follows the training data, whereas the robust model (orange line with crosses) maintains a greater distance from the training observations. Interestingly, this underfitting does not persist at the test location. On the contrary,

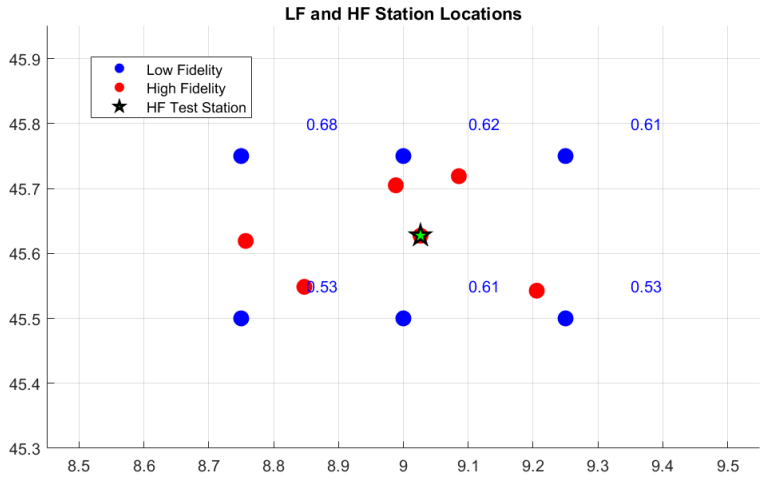


Figure 5.1: Toy dataset used for preliminary testing of the robust multifidelity model. Blue points indicate LF locations, red points denote HF locations, and the test location is marked with a star. Pairwise correlation coefficients between datasets are also shown.

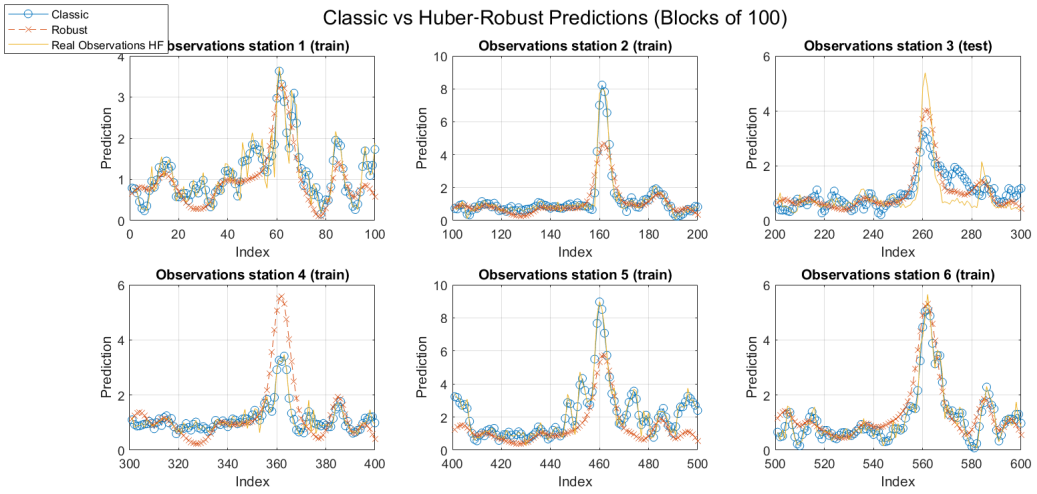


Figure 5.2: Comparison of Robust model with Classic model predictions. In this experiment the test station is the number 3.

the robust model provides superior generalization performance. This suggests that the robust model is better suited to mitigate overfitting, particularly when faced with noisy or anomalous training data. Preliminary tests confirm that this is not an isolated occurrence. As shown in Table 5.1, the robust model consistently outperforms the standard model across all tested stations, indicating a generalizable improvement in predictive accuracy.

The advantages of the robust multifidelity model become especially apparent when the training data include anomalous observations. To illustrate this, a simple yet effective test was conducted: the low-fidelity (LF) data at one station were artificially scaled by a factor of 10, rendering them anomalous relative to the sur-

Station	RMSE_Classic	MAE_Classic	RMSE_Robust	MAE_Robust
1	0.620	0.474	0.551	0.425
2	1.090	0.568	0.939	0.466
3	0.619	0.450	0.439	0.315
4	0.900	0.605	0.914	0.654
5	1.593	1.098	1.592	1.094
6	0.682	0.583	0.708	0.548

Table 5.1: Comparison of Classic and Robust Forecasting Metrics.

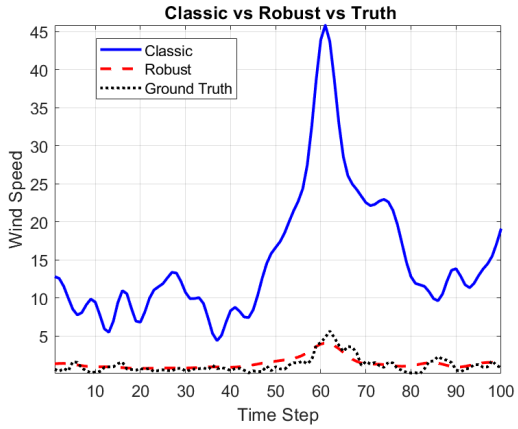
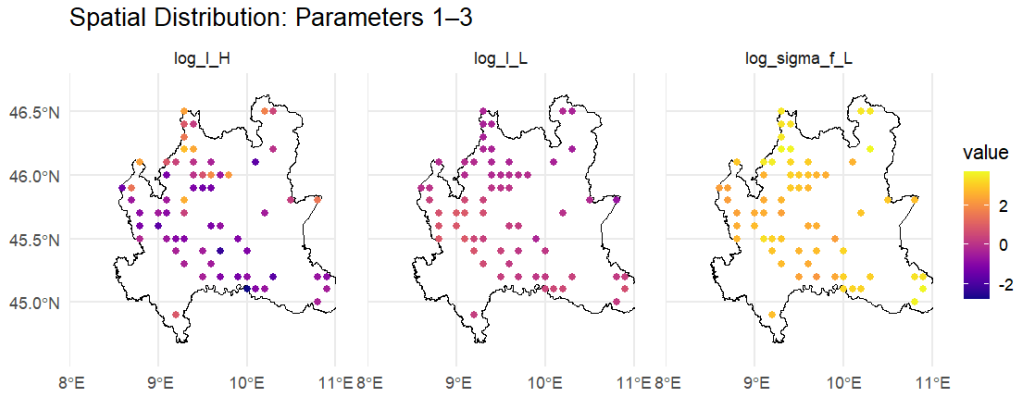


Figure 5.3: Comparison between robust and classic model predictions in the test set. For such an example anomalies are introduced by multiplying LF observations by a factor of 10. In blue we can observe the "Classic", MF predictions, in red the Robust MF prediction, while the black dotted line is the true signal.

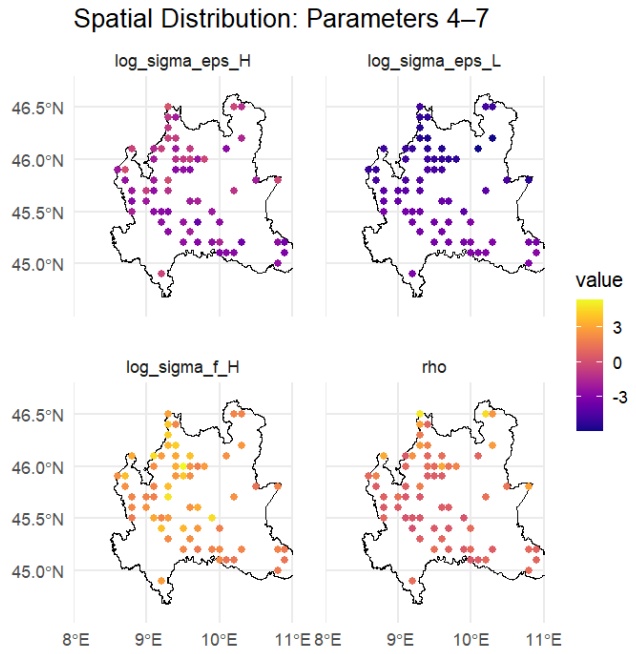
rounding locations. As expected, this distortion had a dramatic impact on the classic multifidelity (MF) model, whose predictions at the test station deviated significantly from realistic values. Figure 5.3 highlights this behaviour , showing how the classic MF model produces excessively high predictions as a direct consequence of the outlier contamination. In contrast to the classic model, the robust model remains stable, producing forecasts comparable to those observed under the baseline conditions. Quantitatively, in the baseline (non-anomalous) scenario at test station 6, the robust model achieved a Mean Absolute Error (MAE) of 0.548, slightly outperforming the classic MF model, which recorded an MAE of 0.583. However, under the extreme (anomalous) scenario, the robust model maintained a low MAE of 0.510, whereas the classic model's error surged dramatically to 13.0. This stark contrast highlights the robust model's resilience and its ability to preserve predictive accuracy in the presence of outliers.

5.3.3 Surrogate modelling and parameters space smoothing

The evaluation of a global model on a large spatial area can be considered the evaluation of costly function $f(x) : \mathcal{R}^p \rightarrow \mathcal{R}$. An idea for reducing the computational complexity is choosing multiple local models rather than a single global. The model parameters are then smoothed or interpolated over space, effectively creating spatial fields (or surfaces) that describe model behaviour. This spatial smoothing enables spatio-temporal prediction, allowing outputs to be estimated at new times and locations. The entire pipeline functions as a surrogate for a complex process or simulation that varies across both space and time. A simple example in which 64 local models have been trained for the whole Lombardy region is shown below in Figure 5.4.



(a) Plot multifidelity surrogated models parameters. From left to right, length scale of low fidelity data, length scale of high fidelity, noise signal of low fidelity.



(b) Plot Multifidelity surrogated models parameters. From left to right, nugget of high fidelity data, the nugget of low fidelity, noise signal of high fidelity and ρ.

Figure 5.4: Estimated spatial distribution of multididelity models parameters using a surrogate models approach.

Because Figure 5.4 displays the geographical distribution of seven parameters with different numerical meanings and scales, direct comparison between them is challenging. Each parameter operates on its own scale and lacks a shared baseline, making inter-comparisons difficult. A full comparative analysis also lies beyond the scope of this thesis. However, a more localized examination—for instance, of

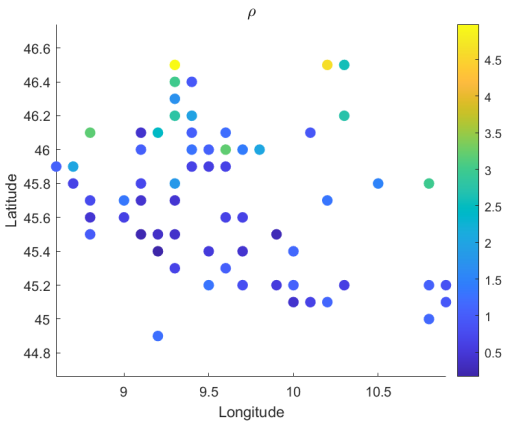


Figure 5.5: Surrogate model approach estimates of ρ across the whole Lombardy region.

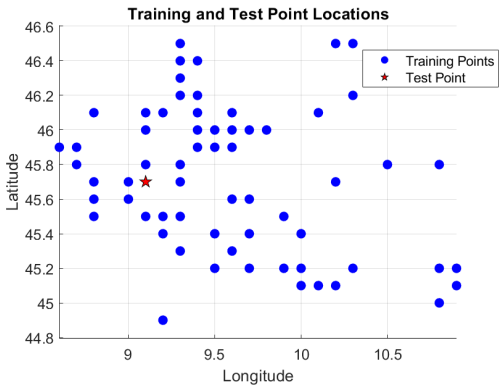
the parameter ρ (see Figure 5.5), which was treated as non-stationary in Chapter 4—indicates that the estimated values of ρ exhibit substantially greater variability in the northern part of the region than in the south. This evidence again how complex was attempted of Chapter 4 of defining a unique function for ρ.

Figure 5.6 illustrates an example where the parameters were inferred using the surrogate strategy and predictions were made at the starred location 5.6a. The second panel (Figure 5.6b) shows the predicted values at that location, which appear reasonable for this particular test case. However, in other instances, the predictions were not as accurate.

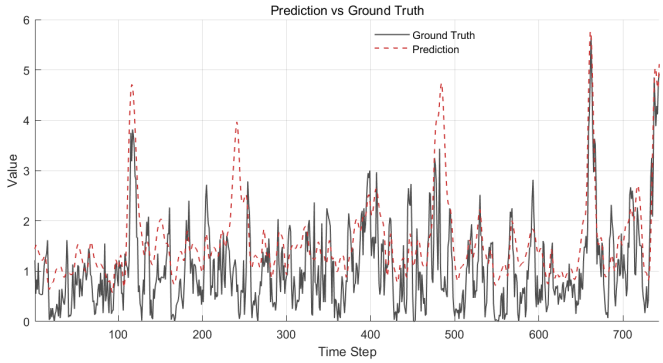
Although effective in delivering good predictive performance, this method is characterized by two distinct phases: the "local model training" and the "smoothing over the parameter space". These phases are not integrated into a single procedure; therefore, future work in this direction should aim to develop a framework that transfers uncertainty from "Phase 1" to "Phase 2". A Bayesian hierarchical model appears to be a promising approach.

5.3.4 Spatially clustered regression

Another extension to address local heterogeneity and scalabilty at the same time would involve looking to possible integration with Spatially Clustered Regression (SCR) a model introduced by Sugasawa and Murakami (2021) to capture spatial heterogeneity in regression analysis. Unlike traditional Geographically Weighted Regression (GWR), which allows coefficients to vary smoothly across space, SCR—similar



(a) Training and test location for the surrogate modelling example. Parameters are first estimated at the blue location independently the GP is used to smooth at the starred locations.



(b) Example of the prediction at the test location of the above panel.

Figure 5.6: Example of predictions obtained with the surrogate approach. The top panel shows the training and test stations, while the bottom panel displays predictions at the test station.

in spirit to the model described in Section 4.2—assumes that groups of nearby locations share the same regression parameters. This leads to more interpretable and discrete spatial patterns.

The model includes a spatial regularization term inspired by the Potts model (Potts, 1952), a concept from statistical physics. This penalty encourages neighbouring locations to be assigned to the same group or "cluster". The idea is similar to how, in physics, neighboring particles prefer to be in the same state—here, spatial locations that share similar behaviour are more likely to belong to the same group. The objective function of the Spatially Clustered Regression (SCR) model is given by:

$$Q(\theta, g) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \theta_{g_i}) + \phi \sum_{i<j} w_{ij} \cdot \mathbb{I}(g_i = g_j) \tag{5.4}$$

where:

- $\theta = (\theta_1, \ldots, \theta_G)$ are group-specific regression parameters,
- $g = (g_1, \ldots, g_n)$ is the vector of group assignments,
- $w_{ij} \in [0, 1]$ are spatial weights between location i and j,
- $\mathbb{I}(g_i = g_j)$ is the indicator function, equal to 1 if $g_i = g_j$, 0 otherwise,
- ϕ is a tuning parameter controlling spatial smoothness.

The model can be integrated with a Gaussian process framework since it is based on likelihood function estimation. In principle, SCR should identify spatial regions with similar behaviour . Within each cluster (group), fit a separate GPR model (MFGP), assuming stationarity or local smoothness within that region. This will further increase the scalability of the algorithm.

5.4 Conclusion

This thesis addressed critical challenges in environmental data fusion, scarcity, and computational scaling through a series of methodological advancements built upon the MFGP framework. The collective work establishes MFGP—now enhanced for realism and scalability—as a statistically rigorous and robust solution for operational environmental monitoring networks.

The research questions (RQs) posed at the outset were fully addressed: It was confirmed that MFGPs effectively reduce the reliance on high-fidelity samples under appropriate noise conditions. Through comparative simulation and real-world experiments, the MFGP consistently delivered superior predictive accuracy and, crucially, exhibited a robustness to long data gap size not found in traditional single-fidelity baselines like standard GPR or industry-standard QGBRT. This is attributed to the successful learning of the predictable cross-fidelity error structure.

The development of the Warped multifidelity Gaussian process successfully resolved the conflict between normalizing highly skewed environmental data (like wind speed) and preserving the crucial inter-fidelity correlation. By introducing a joint,

rank-preserving nonparametric warping layer, the WMFGP maintains coherent uncertainty quantification while meeting the Gaussian assumptions required for effective covariance modelling.

The final objective was met by integrating the Vecchia approximation into the MFGP framework. This approach provides a scalable and numerically stable Spatio-Temporal MFGP model, transforming the framework from a purely temporal tool into a solution capable of handling large-scale monitoring under-coverage. This method successfully captured complex, spatially varying coupling (ρ) in a computationally efficient manner.

The collective body of work in this thesis moves the MFGP framework beyond theoretical application to establish it as a leading-edge, statistically coherent solution for addressing the dual challenges of data sparsity and non-Gaussianity in high-stakes environmental monitoring and prediction. By jointly addressing realism (Warping) and complexity (Vecchia scaling), the methodologies presented here offer a pathway toward fully autonomous, high-resolution environmental data reconstruction.

Looking ahead, the research opens several promising avenues for future exploration. The immediate focus is on integrating robust estimation techniques, such as the Huber loss, to safeguard the cross-fidelity relationship against data anomalies—a development that is already realized in a manuscript submitted to the Journal of the American Statistical Association: Applications and Case Studies. Further work involves formally unifying the concepts of spatially clustered regression and surrogate modelling into a coherent framework.

Bibliography

ERA5 hourly data on single levels from 1940 to present. https://cds.climate.copernicus.eu/cdsapp!/dataset/reanalysis-era5-single-levels?tab=overview.

V. D. Agou, A. Pavlides, and D. T. Hristopulos. Spatial modeling of precipitation based on data-driven warping of gaussian processes. *Entropy*, 24(3):321, 2022. 69, 79, 81, 85

D. Allard and P. Naveau. A new spatial skew-normal random field model. *Communications in Statistics—Theory and Methods*, 36(9):1821–1834, 2007. 69

M. Alodat and M. K. Shakhatreh. Gaussian process regression with skewed errors. *Journal of Computational and Applied Mathematics*, 370:112665, 2020a. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2019.112665. URL https://www.sciencedirect.com/science/article/pii/S0377042719306703.

M. Alodat and M. K. Shakhatreh. Gaussian process regression with skewed errors. *Journal of Computational and Applied Mathematics*, 370:112665, 2020b. 67, 69, 73

M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

L. Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media, 1988. 110

R. B. Arellano-Valle and A. Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006.

ARPA Lombardia. Dati e indicatori ambientali. https://www.arpalombardia.it/dati-e-indicatori/, 2025. Accessed: 05 December 2025. 15

H. Babaee, C. Bastidas, M. DeFilippo, C. Chryssostomidis, and G. Karniadakis. A multifidelity framework and uncertainty quantification for sea surface temperature in the massachusetts and cape cod bays. *Earth and Space Science*, 7(2): e2019EA000954, 2020. 113

A. Bárdossy and J. Li. Geostatistical interpolation using copulas. *Water resources research*, 44(7), 2008. 67

A. Benavoli, D. Azzimonti, and D. Piga. Skew gaussian processes for classification. *Machine Learning*, 109(9):1877–1902, 2020.

Z. I. Botev, J. F. Grotowski, and D. P. Kroese. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5), 2010. doi: https://doi.org/10.1214/10-AOS799. 86

P. S. Bradley, K. P. Bennett, and A. Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000. 99

J. Browell and C. Gilbert. Probcast: Open-source production, evaluation and visualisation of probabilistic forecasts. In *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–6. IEEE, 2020. 5

J. Browell, I. Dinwoodie, and D. McMillan. Forecasting for day-ahead offshore maintenance scheduling under uncertainty. 2016. doi: DOI:10.1201/9781315374987-171).

L. Cai, J. Gu, J. Ma, and Z. Jin. Probabilistic wind power forecasting approach via instance-based transfer learning embedded gradient boosting decision trees. *Energies*, 12(1):159, 2019.

R. Calandra, J. Peters, C. E. Rasmussen, and M. P. Deisenroth. Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345. IEEE, 2016. doi: doi.org/10.48550/arXiv.1402.5876. 3

G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021. 80

H. Chen, Y. Birkelund, S. N. Anfinsen, R. Staupe-Delgado, and F. Yuan. Assessing probabilistic modelling for wind speed from numerical weather prediction model and observation in the arctic. *Scientific Reports*, 11(1):1–11, 2021. doi: doi.org/10.1038/s41598-021-87299-4. 112

X. Chen and J. Ye. When the wind blows: Spatial spillover effects of urban air pollution in china. *Journal of Environmental Planning and Management*, 62(8): 1359–1376, 2019. 110

S. Cheng, B. A. Konomi, J. L. Matthews, G. Karagiannis, and E. L. Kang. Hierarchical bayesian nearest neighbor co-kriging gaussian process models; an application to intersatellite calibration. *Spatial Statistics*, 44:100516, 2021.

S. Cheng, B. A. Konomi, G. Karagiannis, and E. L. Kang. Recursive nearest neighbor co-kriging models for big multi-fidelity spatial data sets. *Environmetrics*, 35(4): e2844, 2024. 112

A. T. Chief Operating Officer. Offshore Wind Farm in the Adriatic Sea. https://www.agnespower.com/en/eolico-offshore-adriatico/, 2022. Accessed:2022-04-10. 12

V. Christelis, G. Kopsiaftis, R. G. Regis, and A. Mantoglou. An adaptive multi-fidelity optimization framework based on co-kriging surrogate models and stochastic sampling with application to coastal aquifer management. *Advances in Water Resources*, 180:104537, 2023. 3

R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A Seasonal-Trend Decomposition. *J. Off. Stat*, 6(1):3–73, 1990. doi: 10.2514/1.j057750. 24, 32

P. Colombo and A. Fassò. Quantifying the interpolation uncertainty of radiosonde humidity profiles. *Measurement Science and Technology*, 33(7):074001, 2022. 90, 95, 105, 159

P. Colombo, C. Miller, R. O'Donnell, and X. Yang. A multifidelity framework for wind speed data. In *Proceedings of the 37th International Workshop on Statistical*

Modelling (IWSM 2023), Dortmund, Germany, July 2023. ISBN 978-3-947323-42-5. 16–21 July 2023. 19

P. Colombo, C. Miller, X. Yang, R. O'Donnell, and P. Maranzano. Warped multifidelity gaussian processes for data fusion of skewed environmental data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, page qlaf003, 2025. 19, 162

Copernicus Climate Change Service (C3S). ERA5 hourly data on single levels from 1940 to present. https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels, 2025. Accessed: 05 Dec 2025. 13

F. S. Costabal, P. Perdikaris, E. Kuhl, and D. E. Hurtado. Multi-fidelity classification using gaussian processes: accelerating the prediction of large-scale computational models. *Computer Methods in Applied Mechanics and Engineering*, 357:112602, 2019. 9

A. C. Cullen, H. C. Frey, and C. H. Frey. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media, 1999.

K. Cutajar, M. Pullin, A. Damianou, N. Lawrence, and J. González. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*, 2019. doi: https://doi.org/10.48550/arXiv.1903.0732. 113

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

G. Deep and J. Verma. Deep learning models for fine-scale climate change prediction: Enhancing spatial and temporal resolution using ai. In *Big Data, Artificial Intelligence, and Data Analytics in Climate Change Research: For Sustainable Development Goals*, pages 81–100. Springer, 2024. 130

B. Delcroix, S. Sansregret, G. L. Martin, and A. Daoud. Quantile regression using gradient boosted decision trees for daily residential energy load disaggregation. In

Journal of Physics: Conference Series, volume 2069, page 012107. IOP Publishing, 2021. 27

F. Ding, H. Peng, J. Zhang, and A. Kareem. Parallel multifidelity design of experiment strategy considering low-fidelity simulation feasibility. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 11(2): 04025012, 2025. 3

A. Fassò, M. Sommer, and C. von Rohden. Interpolation uncertainty of atmospheric temperature profiles. *Atmospheric Measurement Techniques*, 13(12):6445–6458, 2020. 90, 105, 159

A. Fassò, P. Maranzano, and P. Otto. Spatiotemporal variable selection and air quality impact assessment of covid-19 lockdown. *Spatial Statistics*, 49:100549, 2022. 18

A. Fassò, J. Rodeschini, A. Fusta Moro, Q. Shaboviq, P. Maranzano, M. Cameletti, F. Finazzi, N. Golini, R. Ignaccolo, and P. Otto. Agrimonia: a dataset on livestock, meteorology and air quality in the lombardy region, italy. *Scientific Data*, 10(1): 143, 2023. 111

T. Feng, H. Du, Z. Lin, and J. Zuo. Spatial spillover effects of environmental regulations on air pollution: Evidence from urban agglomerations in china. *Journal of Environmental Management*, 272:110998, 2020. 110

M. G. Fernández-Godino, C. Park, N. H. Kim, and R. T. Haftka. Issues in deciding whether to use multifidelity surrogates. *AIAA Journal*, 57:2039–2054, 2019. doi: 10.2514/1.j057750. 2, 3

A. I. Forrester and A. J. Keane. Recent advances in surrogate-based optimization. *Progress in aerospace sciences*, 45(1-3):50–79, 2009. 8

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis (3rd ed.)*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 2000 Corporate Blvd N.W. Boca Raton, FL 33431, USA, 2013. ISBN 978-0-4291-1307-9.

M. G. Genton. *Skew-elliptical distributions and their applications: a journey beyond normality*. CRC Press, 2004. 41

M. G. Genton and H. Zhang. Identifiability problems in some non-gaussian spatial random fields. *Chilean Journal of Statistics*, 3(2):171–179, 2012. 69

K. Giannoukou, S. Marelli, and B. Sudret. Uncertainty-aware multifidelity surrogate modeling with noisy data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 11(3):04025037, 2025. 3

C. Goncalves, L. Cavalcante, M. Brito, R. J. Bessa, and J. Gama. Forecasting conditional extreme quantiles for wind energy. *Electric Power Systems Research*, 190: 106636, 2021.

C. A. Gotway and L. J. Young. Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, 97:632–648, 2002. doi: doi.org/10.1198/016214502760047140. 2

J. C. Gower and G. J. S. Ross. Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics*, 18:54, 1969. ISSN 00359254. doi: 10.2307/2346439.

R. B. Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020. 166

J. Guinness. Permutation and grouping methods for sharpening gaussian process approximations. *Technometrics*, 60(4):415–429, 2018. 115

Y. Guo, Q. Lu, S. Wang, and Q. Wang. Analysis of air quality spatial spillover effect caused by transportation infrastructure. *Transportation Research Part D: Transport and Environment*, 108:103325, 2022. 110

M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of agricultural, biological and environmental Statistics*, 24(3):398–425, 2019. 112

M. Heinonen, H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Artificial Intelligence and Statistics*, pages 732–740. PMLR, 2016.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1):73–101, 1964. 168

R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018. 32

N. Il Fatto Quotidiano. Senato Approva il Decreto sul Fondone Modifiche per 700 milioni. https://www.ilfattoquotidiano.it/2021/06/17/il-senato-approva-il-decreto-sul-fondone-modifiche-per-700-milioni-fondi-per-6233203/, 2022. Accessed:2022-26-09. 13

J. D. Jakeman, M. Perego, D. T. Seidl, T. A. Hartland, T. R. Hillebrand, M. J. Hoffman, and S. F. Price. An evaluation of multi-fidelity methods for quantifying uncertainty in projections of ice-sheet mass-change. *EGUsphere*, 2024:1–38, 2024.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

M. Katzfuss and J. Guinness. A general framework for vecchia approximations of gaussian processes. 2021. 72, 112, 114

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. URL http://www.jstor.org/stable/2673557. 8, 9

M. J. Khaledi, H. Zareifard, and H. Boojari. A spatial skew-gaussian process with a specified covariance function. *Statistics & Probability Letters*, 192:109681, 2023. 69, 78

H.-M. Kim and B. K. Mallick. A bayesian prediction using the skew gaussian distribution. *Journal of Statistical Planning and Inference*, 120(1-2):85–101, 2004. 69

M. Klapacz. Multifidelity gaussian processes for uncertainty quantification. 2021.

M. Landry, T. P. Erlinger, D. Patschke, and C. Varrichio. Probabilistic gradient boosting machines for gefcom2014 wind forecasting. *International Journal of Forecasting*, 32(3):1061–1066, 2016.

L. Le Gratiet and J. Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014. doi: 10.1615/Int.J.UncertaintyQuantification.2014006914. 9

J. Lee, M. Lee, B. J. Lee, and I. Lee. A comprehensive multi-fidelity surrogate framework based on gaussian process for datasets with heterogeneous responses. *Knowledge-Based Systems*, 295:111827, 2024. 3

L. Lin, N. Mu, P. Cheung, and D. Dunson. Extrinsic gaussian processes for regression and classification on manifolds. *Bayesian Analysis*, 14(3):887–906, 2019. 2

P. Ma. Arcokrig: Autoregressive cokriging models for multifidelity codes. r package version 0.1. 2, 2019.

P. Ma, G. Karagiannis, B. A. Konomi, T. G. Asher, G. R. Toro, and A. T. Cox. Multifidelity computer model emulation with high-dimensional output: An application to storm surge. *arXiv preprint arXiv:1909.01836*, 2019. doi: doi.org/10.48550/arXiv.1909.01836.

C. Macaulay, D. Husmeier, and V. Davies. A comparative analysis of deep gaussian processes and multivariate bayesian spline-based methods for simulating multidimensional surfaces. 2025. 167

P. Maranzano and A. Algieri. 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, 2024.

D. Maraun. Bias Correcting Climate Change Simulations-a Critical Review. *Current Climate Change Reports*, 2(4):211–220, 2016. doi: doi.org/10.1007/s40641-016-0050-x. 2, 4

D. Maraun and M. Widmann. *Statistical downscaling and bias correction for climate research*. Cambridge University Press, 2018. 4

D. Maraun, F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. Rust, T. Sauter, M. Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, 48(3), 2010. doi: doi.org/10.1007/s40641-016-0050-x. 2, 4

E. Medina-Lopez, D. McMillan, J. Lazic, E. Hart, S. Zen, A. Angeloudis, E. Bannon, J. Browell, S. Dorling, R. Dorrell, et al. Satellite data for the offshore renewable energy sector: synergies and innovation opportunities. *Remote Sensing of Environment*, 264:112588, 2021. doi: doi.org/10.1016/j.rse.2021.112588. 13

T. Meng, X. Jing, Z. Yan, and W. Pedrycz. A Survey on Machine Learning for Data Fusion. *Information Fusion*, 57:54, 2020. doi: doi.org/10.1016/j.inffus.2019.12.001.

C. . L. G. Ministry of Housing. National statistics: English indices of deprivation 2019 (file 10). https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019, 2019. Accessed: 2022-02-10.

G. I. Nagy, G. Barta, S. Kazi, G. Borbély, and G. Simon. Gefcom2014: Probabilistic solar and wind power forecasting using a generalized additive tree ensemble approach. *International Journal of Forecasting*, 32(3):1087–1093, 2016. 31

P. Otto, A. Fusta Moro, J. Rodeschini, Q. Shaboviq, R. Ignaccolo, N. Golini, M. Cameletti, P. Maranzano, F. Finazzi, and A. Fassò. Spatiotemporal modelling of pm 2.5 concentrations in lombardy (italy): a comparative study. *Environmental and Ecological Statistics*, 31(2):245–272, 2024. 18, 138, 158

P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017. doi: https://doi.org/10.1098/rspa.2016.0751. 9, 113, 130

G. W. Peters, I. Nevat, S. G. Nagarajan, and T. Matsui. Spatial warped gaussian processes: Estimation and efficient field reconstruction. *Entropy*, 23(10):1323, 2021. 67

I. Pobočíková, Z. Sedliačková, and M. Michalková. Application of four probability distributions for wind speed modeling. *Procedia Engineering*, 192:713–718, 2017. ISSN 1877-7058. doi: https://doi.org/10.1016/j.proeng.2017.06.123. URL https://www.sciencedirect.com/science/article/pii/S1877705817326693. 12th international scientific conference of young scientists on sustainable, modern and safe transport. 34

D. Poole and A. E. Raftery. Inference for deterministic simulation models: the bayesian melding approach. *Journal of the American Statistical Association*, 95 (452):1244–1255, 2000. 10

R. B. Potts. Some generalized order-disorder transformations. In *Mathematical proceedings of the cambridge philosophical society*, volume 48, pages 106–109. Cambridge University Press, 1952. 174

P. Z. Qian and C. J. Wu. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2):192–204, 2008. doi: DOI: 10.1198/004017008000000082.

F. Rambelli and F. Sigrist. An accuracy-runtime trade-off comparison of scalable gaussian process approximations for spatial data. *arXiv preprint arXiv:2501.11448*, 2025. 112

J. Ramon, L. Lledó, P.-A. Bretonnière, M. Samsó, and F. J. Doblas-Reyes. A perfect prognosis downscaling methodology for seasonal prediction of local-scale wind speeds. *Environmental Research Letters*, 16(5):054010, 2021. doi: doi.org/10.1088/1748-9326/abe491.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2004. 5, 53

M. Rummukainen. *Methods for statistical downscaling of GCM simulations*. SMHI, 1997. 2

M. P. Rumpfkeil, K. Hanazaki, and P. S. Beran. Construction of multi-fidelity locally optimized surrogate models for uncertainty quantification. In *19th AIAA Non-Deterministic Approaches Conference*, page 1948, 2017.

A. Sauer, A. Cooper, and R. B. Gramacy. Vecchia-approximated deep gaussian processes for computer experiments. *Journal of Computational and Graphical Statistics*, 32(3):824–837, 2023. 167

V. Sella, T. O'Leary-Roseberry, X. Du, M. Guo, J. R. Martins, O. Ghattas, K. E. Willcox, and A. Chaudhuri. Improving neural network efficiency with multifidelity and dimensionality reduction techniques. In *AIAA SciTech 2025 Forum*, page 2807, 2025. 3

E. Snelson, Z. Ghahramani, and C. Rasmussen. Warped gaussian processes. *Advances in neural information processing systems*, 16, 2003. 67, 69, 79, 80, 84

K. Stengel, A. Glaws, D. Hettinger, and R. N. King. Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117(29):16805–16815, 2020. doi: doi.org/10.1073/pnas.1918964117.

S. Sugasawa and D. Murakami. Spatially clustered regression. *Spatial Statistics*, 44: 100525, 2021. 173

J. Ulimwengu and A. Kibonge. Spatial spillover and covid-19 spread in the us. *BMC Public Health*, 21(1):1765, 2021. 110

S. J. Villejo, J. B. Illian, and B. Swallow. Data fusion in a two-stage spatio-temporal model using the inla-spde approach. *Spatial Statistics*, 54:100744, 2023. ISSN 2211-6753. doi: https://doi.org/10.1016/j.spasta.2023.100744. URL https://www.sciencedirect.com/science/article/pii/S2211675323000192. 3

K. Wang, X. Han, L. Dong, X.-J. Chen, G. Xiu, M.-p. Kwan, and Y. Liu. Quantifying the spatial spillover effects of non-pharmaceutical interventions on pandemic risk. *International Journal of Health Geographics*, 22(1):13, 2023. 110

Y. Wang, S. C. Warder, E. F. Benmoufok, A. Wynn, O. R. Buxton, I. Staffell, and M. D. Piggott. Geographic variability in reanalysis wind speed biases: A high-resolution bias correction approach for uk wind energy. *Energy Conversion and Management*, 352:121066, 2026. 14

Q. Wen, J. Gao, X. Song, L. Sun, H. Xu, and S. Zhu. Robuststl: A robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5409–5416, 2019.

C. J. Wilkie, C. A. Miller, E. M. Scott, R. A. O'Donnell, P. D. Hunter, E. Spyrakos, and A. N. Tyler. Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, 30(3):e2549, 2019. 3

A. Winstral, T. Jonas, and N. Helbig. Statistical downscaling of gridded wind speed data using local topography. *Journal of Hydrometeorology*, 18(2):335–348, 2017. doi: 10.2514/1.j057750. 4

G. Xu and M. G. Genton. Tukey g-and-h random fields. *Journal of the American Statistical Association*, 112(519):1236–1249, 2017. 67

J. Xu, Y. Wang, and L. Zhang. Interpolation of extremely sparse geo-data by data fusion and collaborative bayesian compressive sampling. *Computers and Geotechnics*, 134:104098, 2021. doi: doi.org/10.1016/j.compgeo.2021.104098. 2

F. Yin, Y. Qian, J. Zeng, and X. Wei. The spatial spillover effects of transportation infrastructure on regional economic growth—an empirical study at the provincial level in china. *Sustainability*, 16(19):8689, 2024. 110

H. Zareifard and M. J. Khaledi. Non-gaussian modeling of spatial data using scale mixing of a unified skew gaussian process. *Journal of Multivariate Analysis*, 114: 16–28, 2013. 67, 68

S. Zen, E. Hart, and E. Medina-Lopez. The use of satellite products to assess spatial uncertainty and reduce life-time costs of offshore wind farms. *Cleaner Environmental Systems*, 2:100008, 2021. doi: doi.org/10.1016/j.cesys.2020.100008. 13

Y. Zhang, J. F. Schutte, W. P. Seneviratne, N. H. Kim, and R. T. Haftka. Sampling by exploration and replication for estimating experimental strength of composite structures. *AIAA Journal*, 55(10):3594–3602, 2017. doi: doi.org/10.2514/1.J055862.

A.-X. Zhu, G. Lu, J. Liu, C.-Z. Qin, and C. Zhou. Spatial prediction based on third law of geography. *Annals of GIS*, 24(4):225–240, 2018. 105