SMAI-JCM
SMAI Journal of Computational Mathematics

# Optimizing Sensor Calibration in Open Environments: A Bayesian Approach for Non-Specific Multisensory Systems

Marine Dumon, Bérengère Lebental & Guillaume Perrin

Volume 10 (2024), p. 305-324.
https://doi.org/10.5802/smai-jcm.114



*The SMAI Journal of Computational Mathematics is a member of the Centre Mersenne for Open Scientific Publishing*
http://www.centre-mersenne.org/
Submissions at https://smai-jcm.centre-mersenne.org/ojs/submission
e-ISSN: 2426-8399

*SMAI Journal of Computational Mathematics*
Vol. 10, 305-324 (2024)

# Optimizing Sensor Calibration in Open Environments: A Bayesian Approach for Non-Specific Multisensory Systems

Marine Dumon [1]
Bérengère Lebental [2]
Guillaume Perrin [3]

[1] Université Gustave Eiffel, COSYS, F-77454 Marne-la-Vallée, France
*E-mail address*: marine.dumon@univ-eiffel.fr
[2] Université Gustave Eiffel, COSYS, F-77454 Marne-la-Vallée, France
*E-mail address*: berengere.lebental@univ-eiffel.fr
[3] Université Gustave Eiffel, COSYS, F-77454 Marne-la-Vallée, France
*E-mail address*: guillaume.perrin@univ-eiffel.fr.

**Abstract.** Air and water pollution present significant threats to public health, highlighting the need for precise environmental monitoring methods. Current solutions rely on multisensory systems with limited specificity. Their calibrations often struggle in real-world conditions, resulting in imprecise air and water quality measurements. This paper aim to address the challenge of calibrating non-specific multisensory systems deployed in open periodic environments. A data-driven calibration method is proposed within a Bayesian framework, which considers several sources of uncertainties that are often overlooked in sensor calibration.

The method combines a non-parametric approach, capturing correlations between pollutants and environmental variables, with a parametric method, that maximizes sensor-provided information. Unlike conventional sensor calibration, our method prioritizes the inclusion of input uncertainties and model errors during calibration, providing a comprehensive framework for robust sensor performance.

The theoretical foundations of the non-parametric approach are presented, and the coupling between non-parametric and parametric methods is detailed. The evaluation using synthetic data demonstrates the method's efficiency and limitations. Then the approach is validated in an experimental use case where a sensor array based on carbon nanotube is calibrated for monitoring ozone and carbon monoxide in an outdoor deployment.

**2020 Mathematics Subject Classification.** 65N35, 15A15.

**Keywords.** Bayesian Framework, Multisensory Systems, Air and Water Pollution, Data-driven Calibration, Uncertainty quantification.

## 1. Introduction

Air and water pollution are major public health challenges [3, 9, 13]. The development of effective environmental monitoring methods is essential to minimize people exposure and associated health risks. The use of innovative materials holds great promise for the development of affordable and highly sensitive sensors capable of detecting pollutants in both air [5] and water [2]. However, the use of low-cost or innovative instruments in real-world environments challenges current calibration methods, due to the variety of interfering factors, leading to inaccurate measurements of air and water quality. Therefore, it is essential to propose efficient calibration methods to improve the robustness of sensors against fluctuating environmental factors.

The authors acknowledge support from H2020 Project LOTUS number 820881 and from Agence National de la Recherche project CARDIF reference ANR-19-CE04-0010-05.
https://doi.org/10.5802/smai-jcm.114

Calibrating a sensor involves establishing a relationship between environmental variables and the outputs it provides. In general, there is no physical model for the sensor output, which means that the relationship between sensor inputs and outputs is completely unknown. While the literature tends to propose linear relationships to model this link between sensor inputs and outputs, these linear models are often not sufficiently relevant, which will be the case treated in this paper. In such cases, alternative models, such as machine learning methods [26, 27], in particular neural networks [21], are increasingly used. As well as dealing with non-linearity issues, they also aim to compensate for sensor cross-sensitivity, for example their sensitivity to undesirable factors. However, the data volume requirements of neural networks are generally an insurmountable hurdle for environmental sensors, as reliable calibration experiments are extremely time-consuming and there are many influencing variables. Other methods that require less data, such as multiple regression, decision trees [27], support vector regression [26], or Gaussian process regression [7], appear to be more viable options.

Several factors make the calibration of these non-linear models inaccurate. Firstly, both calibration and deployment datasets include noisy measurements, with a noise level that is unknown and often varies according to time and location. Secondly, the laboratory calibrations are poorly transferable to open environments due to the multiplicity of variable environmental factors. But when calibrating directly in an open environment (and not in laboratory), the high correlations between environmental data and pollutant concentrations limit the possibilities for individual characterisation of the interaction between pollutants and sensors. Moreover, the risk of overlearning is also significant when relying on data collected in an open environment, since the only conditions available are those measured during the sensor deployment phase, which cannot be forced to be exhaustive in terms of representing the full range of conditions under which the sensor will ultimately be deployed after calibration.

In this context, the objective of this paper is to propose a general data-driven method for calibrating non-specific (i.e. sensitive to several environmental variables in addition to their target measurand) multisensory systems (and sensor arrays) based on a limited set of noisy measurements collected in an open periodic environment. This calibration method is placed in a Bayesian framework in order to better take into account the different sources of uncertainty. Among the several studies reported in the literature on the topic of Bayesian calibration (see for instance [1, 12, 22]), [17] specifically focused on managing input uncertainties and model errors, which are generally neglected in calibration procedures. While this work has led in significant improvements in the predictive capability of sensor arrays in both synthetic and experimental datasets, the study was aimed at calibration in a controlled environment with a relatively limited number of measurands and possible interferents.

The present paper extends this work by considering an uncontrolled, open environment with periodic variations. This is representative of the conditions actually observed experimentally in environmental sensor deployments. The open environment results in a large increase in interfering parameters and source of uncertainties, while generally leading to highly correlated variables (due to the chemical process). This favours the development of a different methodology compared to the previously cited works, in particular through the coupling of non-parametric and parametric calibration.

In more detail, Section 2 presents the probabilistic formulation proposed for monitoring pollutant concentrations using innovative and non-specific multisensory systems. This method couples a non-parametric method to integrate the potential correlations between the pollutants to be predicted and environmental variables, and a parametric method to make maximum use of the information provided by the sensors. The theoretical foundations of the non-parametric approach are presented in Section 3, while the coupling between the non-parametric and parametric approaches is presented in detail in Section 4. The efficiency and the limitations of the proposed method are evaluated in Section 5 using synthetic data, before being applied to the calibration of a carbon nanotube-based sensor array from data collected during an actual measurement campaign.

## 2. Probabilistic formulation of the estimation problem

### 2.1. General framework

In this work, we are interested in the deployment of a non-specific multisensory system consisting of $d_y$ sensors. It is assumed that they all have similar properties, i.e. they are all sensitive to the same measurands and interferents, although their sensitivity may vary from sensor to sensor. This is often the case experimentally, where the same manufacturing process is used to produce the different sensors in the sensor array, resulting in sensor responses that are highly correlated.

Let $\boldsymbol{y} \in \mathbb{R}^{d_y}$ be the vector gathering the sensor outputs. Since these sensors are assumed non-specific, they are likely to be sensitive not only to a single target measurand, but to several features of their environment, which are grouped into three distinct categories. First, we denote by $\boldsymbol{x} \in \mathbb{R}^{d_x}$ the vector gathering the $d_x$ target pollutant concentrations, e.g. those that the multisensory system aims to measure. We therefore implicitly assume that changes in $\boldsymbol{x}$ lead to changes in $\boldsymbol{y}$. Since the multisensory system is placed in an open environment, it is expected that other quantities will affect it. Among these quantities, we denote by $\boldsymbol{z} \in \mathbb{R}^{d_z}$ the vector gathering the environmental variables that are measured using additional, well-calibrated instruments and by $\boldsymbol{u} \in \mathbb{R}^{d_u}$ the vector gathering the environmental variables that remain unmeasured (due to lack of instrument or to lack of knowledge about the effect of this variable). Even if $\boldsymbol{u}$ is *by definition* unknown and unmeasured, it's important to include it in the general formalism to avoid being faced with an ill-posed problem.

To infer the link between $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$, we assume that these three vectors were measured at $n$ distinct times. A distinction is made between the "true" and "measured" values of $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$. For each $1 \leq i \leq n$, the vectors $\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{y}_i$ refer to the true values (the actual values that are never available) of $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$, while $\boldsymbol{x}_i^{\text{mes}}, \boldsymbol{z}_i^{\text{mes}}, \boldsymbol{y}_i^{\text{mes}}$ are the corresponding measured values. These measured values are gathered in $\mathcal{D}_n := (\boldsymbol{x}_i^{\text{mes}}, \boldsymbol{z}_i^{\text{mes}}, \boldsymbol{y}_i^{\text{mes}})_{i=1}^n$. The measurement noises, which characterise the differences between these two versions of $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$, are then noted:

$$\boldsymbol{\varepsilon}_i^x := \boldsymbol{x}_i - \boldsymbol{x}_i^{\text{mes}}, \ \boldsymbol{\varepsilon}_i^z := \boldsymbol{z}_i - \boldsymbol{z}_i^{\text{mes}}, \ \boldsymbol{\varepsilon}_i^y := \boldsymbol{y}_i - \boldsymbol{y}_i^{\text{mes}}. \tag{2.1}$$

The aim of this paper is to propose a method that fully exploits the information contained in $\mathcal{D}_n$ in order to 1) estimate pollutant concentrations at times other than the training times, and 2) associate credibility intervals to these estimates. To facilitate the reading and make a clear distinction with the training points, the test points are written with the subscript $\star$, and the same distinction is made between the "true" and "measured" values. In other words, $\boldsymbol{x}_\star, \boldsymbol{z}_\star, \boldsymbol{y}_\star$ denote the true values of $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$ when only the values of $\boldsymbol{z}$ and $\boldsymbol{y}$ are measured (they are denoted by $\boldsymbol{z}_\star^{\text{mes}}, \boldsymbol{y}_\star^{\text{mes}}$ respectively), and when we are interested in estimating the non-observed value of $\boldsymbol{x}_\star$.

### 2.2. Specificity and difficulties of the considered problem

Estimating the value of $\boldsymbol{x}$ using information on $\boldsymbol{y}$ and $\boldsymbol{z}$ requires an understanding of the relationship between these three quantities. For this work, it is important to note that this relationship is not known *a priori*, so it is necessary to propose a model for it, and then to adjust the model parameters for optimal fit to the available data. As with any learning problem, a major challenge in this construction is to find the best compromise between modelling bias (the more degrees of freedom a model has, the more likely it is to capture the potential non-linearities relating $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{y}$) and estimation variance (the more complex the model, the more difficult it will be to estimate its parameters correctly). The expected influence of several unmeasured quantities $\boldsymbol{u}$ makes this search for a compromise even more difficult. Let us note here that, conversely, since the environmental factors $\boldsymbol{x}$, $\boldsymbol{z}$ or even $\boldsymbol{u}$ are in practice highly correlated (daily cycles), it is challenging to determine which of the components of $\boldsymbol{x}$ and $\boldsymbol{z}$ do actually have a causal influence on $\boldsymbol{y}$.

The fact that all the data in the problem are noisy, with non-negligible noise of poorly known properties, is another major difficulty of the problem. On the one hand, this noise makes it difficult to identify cause-effect relationships between the different components of $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{y}$. On the other hand, as we shall see in next sections, it enhances the difficulty in estimating the true (rather than the noisy) value of $\boldsymbol{x}$.

### 2.3. **Bayes formulation**

Since the problem to be solved has many uncertainties, a resolution in a Bayesian framework is considered. It amounts to assuming that the value to be estimated, $\boldsymbol{x}_\star$, and the measurement errors can be modelled by random quantities. Under this formalism, estimating $\boldsymbol{x}_\star$ is equivalent to drawing from the probability distribution of $\boldsymbol{x}_\star$ conditioned by $\boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n$,

$$\pi[\boldsymbol{x}_\star | \boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n], \tag{2.2}$$

where to simplify the notations, we abusively write $\pi(\boldsymbol{a})$ the probability density function (PDF) of any random vector $\boldsymbol{a}$ taking the value $\boldsymbol{a}$.

Two methods for estimating this PDF are presented below. The first method, described in Section 3, is based on non-parametric inference using kernel density techniques [24]. The main advantages of this method are its ease and speed of use, and the fact that it does not require *a priori* information about the statistical properties of the measurement errors. However, since the non-noisy values of $\boldsymbol{x}$ are never observed, only the noisy value of $\boldsymbol{x}_\star$ can be estimated with this method.

To refine this estimation, a parametric method is then proposed in Section 4. By introducing several simplifying assumptions, it is used to estimate the non-noisy value of $\boldsymbol{x}_\star$, but at a much higher computational cost.

## 3. **Non-parametric approximation of the conditional PDF**

Since the true values of the pollutant concentrations $\boldsymbol{x}_\star$ are never observed, we first propose to neglect the differences between their true and noisy values:

$$\pi[\boldsymbol{x}_\star | \boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] \approx \pi[\boldsymbol{x}_\star^{\text{mes}} | \boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n], \tag{3.1}$$

where $\boldsymbol{x}_\star^{\text{mes}}$ is the noisy value of $\boldsymbol{x}_\star$ that would have been obtained if it had been possible to measure it at the same time as $\boldsymbol{y}_\star^{\text{mes}}$ and $\boldsymbol{z}_\star^{\text{mes}}$. The idea is then to proceed in three steps. The elements of $\mathcal{D}_n = (\boldsymbol{x}_i^{\text{mes}}, \boldsymbol{z}_i^{\text{mes}}, \boldsymbol{y}_i^{\text{mes}})_{i=1}^n$ are assumed to be $n$ realizations of a triplet $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$, where $\boldsymbol{X}, \boldsymbol{Z}$ and $\boldsymbol{Y}$ are three random vectors taking values in $\mathbb{R}^{d_x}, \mathbb{R}^{d_z}$ and $\mathbb{R}^{d_y}$ respectively. We then estimate the joint PDF of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ by its kernel density approximation, denoted $\widehat{f}_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}}$ (see [24] for further justification of this representation) and expressed by:

$$\widehat{f}_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) := \frac{\sum_{i=1}^n k_{d_x}(\boldsymbol{H}_x^{-1}(\boldsymbol{x} - \boldsymbol{x}_i^{\text{mes}})) k_{d_z}(\boldsymbol{H}_z^{-1}(\boldsymbol{z} - \boldsymbol{z}_i^{\text{mes}})) k_{d_y}(\boldsymbol{H}_y^{-1}(\boldsymbol{y} - \boldsymbol{y}_i^{\text{mes}}))}{n \times \det(\boldsymbol{H}_x) \det(\boldsymbol{H}_z) \det(\boldsymbol{H}_y)}, \tag{3.2}$$

with $\boldsymbol{H}_x, \boldsymbol{H}_z, \boldsymbol{H}_y$ three invertible matrices, and $k_x, k_z, k_y$ three positive functions whose integral over their domain of definition is 1. The choice of $\boldsymbol{H}_x, \boldsymbol{H}_z, \boldsymbol{H}_y$ is particularly important for this type of approximation. Indeed, too large values tend to smooth out and eliminate any singularity in the PDF of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$, while too small values can artificially increase the number of modes in it. Much work has focused on choosing the best estimates for $\boldsymbol{H}_x, \boldsymbol{H}_z, \boldsymbol{H}_y$ based on the available realizations of $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y})$ (see for instance [6, 8, 28]). Here, we restrict ourselves to the case where:

$$\begin{bmatrix} \boldsymbol{H}_x & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_z & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{H}_y \end{bmatrix} = h \begin{bmatrix} \boldsymbol{R}_x^{1/2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_z^{1/2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}_y^{1/2} \end{bmatrix}, \tag{3.3}$$

where $\boldsymbol{R}_x^{1/2}(\boldsymbol{R}_x^{1/2})^T, \boldsymbol{R}_z^{1/2}(\boldsymbol{R}_z^{1/2})^T, \boldsymbol{R}_y^{1/2}(\boldsymbol{R}_y^{1/2})^T$ are the empirical estimators of the covariance matrices of $\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{Y}$ taking the elements of $\mathcal{D}_n$ as statistically independent, and where $h$ is estimated using an adapted likelihood maximization procedure (see [16] for a precise description of this estimation method).

The other parameters to adjust for this approximation are the kernel functions. Again, many choices are possible [24], and in this paper we restrict ourselves to Gaussian kernels, so that:

$$k_d(\boldsymbol{u}) := \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\boldsymbol{u}\|^2}{2}\right), \quad d \in \{d_x, d_z, d_y\}, \quad \boldsymbol{u} \in \mathbb{R}^d. \tag{3.4}$$

Once the parameter $h$ has been estimated, and assuming that the dependence structure learnt from $\mathcal{D}_n$ is still valid for $(\boldsymbol{x}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \boldsymbol{y}_\star^{\text{mes}})$, the PDF $\pi[\boldsymbol{x}_\star^{\text{mes}}|\boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$ can finally be approximated by:

$$\boldsymbol{x} \mapsto \frac{\widehat{f}_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{x}, \boldsymbol{z}_\star^{\text{mes}}, \boldsymbol{y}_\star^{\text{mes}})}{\int_{\mathbb{R}^{d_x}} \widehat{f}_{\boldsymbol{X},\boldsymbol{Z},\boldsymbol{Y}}(\boldsymbol{u}, \boldsymbol{z}_\star^{\text{mes}}, \boldsymbol{y}_\star^{\text{mes}}) d\boldsymbol{u}} = \frac{1}{\det(\boldsymbol{H}_x)} \sum_{i=1}^{n} w_i k_{d_x}(\boldsymbol{H}_x^{-1}(\boldsymbol{x} - \boldsymbol{x}_i^{\text{mes}})), \tag{3.5}$$

where the weights $w_1, \ldots, w_n$ can be derived explicitly:

$$w_i := \frac{k_{d_z}(\boldsymbol{H}_z^{-1}(\boldsymbol{z}_\star^{\text{mes}} - \boldsymbol{z}_i^{\text{mes}})) k_{d_y}(\boldsymbol{H}_y^{-1}(\boldsymbol{y}_\star^{\text{mes}} - \boldsymbol{y}_i^{\text{mes}}))}{\sum_{j=1}^{n} k_{d_z}(\boldsymbol{H}_z^{-1}(\boldsymbol{z}_\star^{\text{mes}} - \boldsymbol{z}_j^{\text{mes}})) k_{d_y}(\boldsymbol{H}_y^{-1}(\boldsymbol{y}_\star^{\text{mes}} - \boldsymbol{y}_j^{\text{mes}}))}, \quad 1 \leq i \leq n. \tag{3.6}$$

It is interesting to note that the non-parametric approach presented above does not require any *a priori* information about the links between $\boldsymbol{x}, \boldsymbol{z}$, and $\boldsymbol{y}$, and can be seen as a smoothed version of a nearest neighbour method [23]. The dependence structure between $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$ is modelled in a purely empirical way, from the data only, and is based only on the estimation of a single scalar parameter $h$, which makes the method particularly easy to use.

**Remark.** In this non-parametric framework, the measurement noises on the vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ introduced in Eq. (2.1) are incorporated into the estimate of the pollutant concentrations transparently, as we are only manipulating noisy quantities. However, it should be emphasised that in this case, the estimate focuses on the noisy value of the pollutant concentrations, $\boldsymbol{x}_\star^{\text{mes}}$, and not on its noiseless value, $\boldsymbol{x}_\star$, which will be the focus of the next section.

## 4. Coupling of parametric and non-parametric approaches for the reduction of estimation uncertainties

In general, the approach presented in Section 3 is likely to be effective for configurations where the number of observation points $n$ is very large compared to the total dimension of the problem, which is here equal to $d_x + d_z + d_y$. As a consequence, it is not guaranteed to improve the estimation results by increasing the number of sensors as the total dimension of the inference problem also increases. Moreover, when the noise on the measurements of $\boldsymbol{x}$ is relatively high (which is often the case in environmental monitoring applications), it can prove to be limiting to only estimate the noisy value of $\boldsymbol{x}$, and not its denoised value. To overcome these two difficulties, a complementary estimation method is introduced in this section. The target PDF in Bayes' theorem is now rewritten as:

$$\pi[\boldsymbol{x}_\star|\boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] = c \times \pi[\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] \times \pi[\boldsymbol{x}_\star|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n], \tag{4.1}$$

where $c$ is a normalization constant. The PDF $\pi[\boldsymbol{x}_\star|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$ is approximated by its noisy version $\pi[\boldsymbol{x}_\star^{\text{mes}}|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$, which is then estimated, again, by a non-parametric approach as in Section 3, but

now this problem has dimension $d_x + d_z$ and not $d_x + d_z + d_y$:

$$\pi[\boldsymbol{x}_\star^{\text{mes}}|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] \approx \frac{1}{\det(\boldsymbol{H}_x)} \sum_{i=1}^{n} \widehat{w}_i k_{d_x}(\boldsymbol{H}_x^{-1}(\boldsymbol{x}_\star^{\text{mes}} - \boldsymbol{x}_i^{\text{mes}})), \tag{4.2}$$

$$\widehat{w}_i := \frac{k_{d_z}(\boldsymbol{H}_z^{-1}(\boldsymbol{z}_\star^{\text{mes}} - \boldsymbol{z}_i^{\text{mes}}))}{\sum_{j=1}^{n} k_{d_z}(\boldsymbol{H}_z^{-1}(\boldsymbol{z}_\star^{\text{mes}} - \boldsymbol{z}_j^{\text{mes}}))}, \quad 1 \leq i \leq n. \tag{4.3}$$

On the other hand, to approximate $\pi[\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$, a parametric representation is introduced to model the relationship between $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$ while integrating as precisely as possible all the sources of uncertainty. It is proposed to write the measured value of the $j^{\text{th}}$ sensor output, with $j \in \{1, \ldots, d_y\}$, under the following general relationship:

$$(\boldsymbol{y}^{\text{mes}})_j = \boldsymbol{h}_j(\boldsymbol{x}; \boldsymbol{z})^T \boldsymbol{\beta}_j + (\varepsilon^y)_j + \varepsilon_j^{\text{mod}}(\boldsymbol{x}; \boldsymbol{z}) + \delta_j^u, \tag{4.4}$$

where:

- $\boldsymbol{h}_j$ is a vector-valued function (typically a vector of polynomial or wavelet functions),
- $\{\boldsymbol{\beta}_j,\ 1 \leq j \leq d_y\}$ is a collection of vectors of parameters,
- $\boldsymbol{\varepsilon}^y$ characterizes the experimental uncertainties affecting the measurements of $\boldsymbol{y}$,
- $\varepsilon_j^{\text{mod}}$ is a first model error allowing to take into account the approximate character of the proposed function $\boldsymbol{h}_j$,
- $\delta_j^u$ corresponds to a second model error supposed to quantify the impact of not including the unobserved quantities into account in the models (hence the superscript letter $u$ to indicate this correspondence). This error is modelled as a constant that does not depend on time.

This relation is assumed to be valid not only for the training points, where the full triplet $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$ is measured, but also for test points where only $\boldsymbol{z}$ and $\boldsymbol{y}$ are measured and $\boldsymbol{x}$ is to be estimated. Replacing the true values of $\boldsymbol{x}$ and $\boldsymbol{z}$ by their noisy observations in Eq. (4.4), yields a $(n+1)$ *error-in-variables* (ERI) model [4, 11, 15, 17]:

$$(\boldsymbol{y}_i^{\text{mes}})_j = \boldsymbol{h}_j(\boldsymbol{x}_i^{\text{mes}} + \boldsymbol{\varepsilon}_i^x; \boldsymbol{z}_i^{\text{mes}} + \boldsymbol{\varepsilon}_i^z)^T \boldsymbol{\beta}_j + (\boldsymbol{\varepsilon}_i^y)_j + \varepsilon_j^{\text{mod}}(\boldsymbol{x}_i^{\text{mes}} + \boldsymbol{\varepsilon}_i^x; \boldsymbol{z}_i^{\text{mes}} + \boldsymbol{\varepsilon}_i^z) + \delta_j^u. \quad 1 \leq i \leq n, \tag{4.5}$$

However, the target is the noiseless value of $\boldsymbol{x}_\star$, hence for the prediction time, Eq.(4.5) is transformed into Eq. (4.6), where there is no error on $\boldsymbol{x}$:

$$(\boldsymbol{y}_\star^{\text{mes}})_j = \boldsymbol{h}_j(\boldsymbol{x}_\star; \boldsymbol{z}_i^{\text{mes}} + \boldsymbol{\varepsilon}_\star^z)^T \boldsymbol{\beta}_j + (\boldsymbol{\varepsilon}_\star^y)_j + \varepsilon_j^{\text{mod}}(\boldsymbol{x}_\star; \boldsymbol{z}_\star^{\text{mes}} + \boldsymbol{\varepsilon}_\star^z) + \delta_j^u. \tag{4.6}$$

To estimate the noiseless value of $\boldsymbol{x}_\star$ using Eq. (4.6), a Bayesian formalism is again adopted. It amounts to modelling the unknown quantities as random quantities, and then searching for their probability distributions conditioned on the available observations. The model parameters, the measurement errors and the model errors are thus modelled by random vectors. For the sake of tractability, only centred Gaussian probability distributions are considered with the exception of $\boldsymbol{\beta}_j$, which is centred on $\overline{\boldsymbol{\beta}}_j$, and the covariance functions are chosen to be independent of time. In practice, this assumption is not so strong since independence in time is almost always observed, while transformations of the inputs can be applied to make the errors Gaussian-like if it was not the case [18]. Let $\boldsymbol{C}_x, \boldsymbol{C}_z, \boldsymbol{C}_y$ be the covariance matrices of $\boldsymbol{\varepsilon}_i^x, \boldsymbol{\varepsilon}_i^z, \boldsymbol{\varepsilon}_i^y$ respectively, $\boldsymbol{\Gamma}_j$ be the covariance matrix of $\boldsymbol{\beta}_j$, $\boldsymbol{T}$ be the covariance matrix of $\boldsymbol{\delta}^u$, and $\boldsymbol{R}$ be the covariance function of $\boldsymbol{\varepsilon}^{\text{mod}}$.

The matrices $\boldsymbol{C}_x, \boldsymbol{C}_z, \boldsymbol{C}_y, \boldsymbol{\Gamma}_j$ and vector $\overline{\boldsymbol{\beta}}_j$ are assumed to be known ($\boldsymbol{C}_x, \boldsymbol{C}_z, \boldsymbol{C}_y$ are determined for each measurement device from dedicated measurements in controlled environments, $\boldsymbol{\Gamma}_j$ and $\overline{\boldsymbol{\beta}}_j$ are chosen by expert opinion. Making $\boldsymbol{\Gamma}_j$ tend to infinity, whatever the value of $\overline{\boldsymbol{\beta}}_j$, is a possibility to

model cases with almost no information about $\boldsymbol{\beta}_j$). In contrast, the quantities $\boldsymbol{T}$ and $\boldsymbol{R}$ are *a priori* unknown and have to be estimated from the data. In this work, we restrict ourselves to cases where these matrices are diagonal, i.e. we neglect any possible statistical dependence between the model errors, and write, for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_x}$ and all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{d_z}$:

$$\boldsymbol{T} = \begin{bmatrix} \theta_1^2 & 0 & \cdots & 0 \\ 0 & \theta_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \theta_{d_y}^2 \end{bmatrix}, \tag{4.7}$$

$$\begin{aligned} \boldsymbol{R}((\boldsymbol{x};\boldsymbol{z}),(\boldsymbol{x}';\boldsymbol{z}')) &= \mathrm{Cov}\left(\boldsymbol{\varepsilon}^{\mathrm{mod}}(\boldsymbol{x};\boldsymbol{z}), \boldsymbol{\varepsilon}^{\mathrm{mod}}(\boldsymbol{x}';\boldsymbol{z}')\right) \\ &= \begin{bmatrix} \sigma_1^2 k_x^{(1)}(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\ell}_1^x) k_z^{(1)}(\boldsymbol{z},\boldsymbol{z}';\boldsymbol{\ell}_1^z) & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_{d_y}^2 k_x^{(d_y)}(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\ell}_{d_y}^x) k_z^{(d_y)}(\boldsymbol{z},\boldsymbol{z}';\boldsymbol{\ell}_{d_y}^z) \end{bmatrix}, \end{aligned} \tag{4.8}$$

where for each $1 \leq j \leq d_y$, the functions $k_x^{(j)}$ and $k_z^{(j)}$ are Matern-5/2 covariance functions characterized by the correlation lengths $\boldsymbol{\ell}_j^x$ and $\boldsymbol{\ell}_j^z$ respectively (alternative choices for these covariance functions can be found in [20]), and the coefficients $\sigma_j^2$ correspond to variance coefficients. In this formalism, the statistical properties of the model errors are thus fully characterised by the two vectors $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_{d_y})$ and $\boldsymbol{\gamma} := (\sigma_1^2, \boldsymbol{\ell}_1^x, \boldsymbol{\ell}_1^z, \ldots, \sigma_{d_y}^2, \boldsymbol{\ell}_{d_y}^x, \boldsymbol{\ell}_{d_y}^z)$. Assuming that the measurement errors on the $\boldsymbol{x}$ and $\boldsymbol{z}$ inputs are small, and that the functions $h_j$ and $\boldsymbol{R}$ are differentiable, the probability distribution of the sensor outputs conditioned by $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is now approximated by Taylor expansion:

$$\boldsymbol{h}_j(\boldsymbol{x}_i^{\mathrm{mes}} + \boldsymbol{\varepsilon}_i^x; \boldsymbol{z}_i^{\mathrm{mes}} + \boldsymbol{\varepsilon}_i^z) \approx \boldsymbol{h}_j(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}}) + \frac{\partial \boldsymbol{h}_j}{\partial \boldsymbol{x}}(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}})\boldsymbol{\varepsilon}_i^x + \frac{\partial \boldsymbol{h}_j}{\partial \boldsymbol{z}}(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}})\boldsymbol{\varepsilon}_i^z, \tag{4.9}$$

$$\boldsymbol{h}_j(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}} + \boldsymbol{\varepsilon}_\star^z) \approx \boldsymbol{h}_j(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}}) + \frac{\partial \boldsymbol{h}_j}{\partial \boldsymbol{z}}(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}})\boldsymbol{\varepsilon}_\star^z, \tag{4.10}$$

$$\varepsilon_j^{\mathrm{mod}}(\boldsymbol{x}_i^{\mathrm{mes}} + \boldsymbol{\varepsilon}_i^x; \boldsymbol{z}_i^{\mathrm{mes}} + \boldsymbol{\varepsilon}_i^z) \approx \varepsilon_j^{\mathrm{mod}}(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}}) + \frac{\partial \varepsilon_j^{\mathrm{mod}}}{\partial \boldsymbol{x}}(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}})\boldsymbol{\varepsilon}_i^x + \frac{\partial \varepsilon_j^{\mathrm{mod}}}{\partial \boldsymbol{z}}(\boldsymbol{x}_i^{\mathrm{mes}}; \boldsymbol{z}_i^{\mathrm{mes}})\boldsymbol{\varepsilon}_i^z, \tag{4.11}$$

$$\varepsilon_j^{\mathrm{mod}}(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}} + \boldsymbol{\varepsilon}_\star^z) \approx \varepsilon_j^{\mathrm{mod}}(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}}) + \frac{\partial \varepsilon_j^{\mathrm{mod}}}{\partial \boldsymbol{z}}(\boldsymbol{x}_\star^{\mathrm{mes}}; \boldsymbol{z}_\star^{\mathrm{mes}})\boldsymbol{\varepsilon}_\star^z. \tag{4.12}$$

From a theoretical point, the smaller the measurement noises are, the more chance there is for the method to be efficient. However, as it will be discussed in further detail in the application section, this Taylor simplification appears to be very useful in situations where the measurement noises are significant, which is the situation of main interest in this work. Conditionally on the knowledge of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, and replacing the former expressions in (4.9) and (4.10), we then denote by $\boldsymbol{\mu}_{\mathrm{train}}$ and $\boldsymbol{\mu}_\star$ the expectations of $\boldsymbol{y}_{\mathrm{train}} := ((\boldsymbol{y}_1^{\mathrm{mes}})_1, \ldots, (\boldsymbol{y}_n^{\mathrm{mes}})_1, (\boldsymbol{y}_1^{\mathrm{mes}})_2, \ldots, (\boldsymbol{y}_n^{\mathrm{mes}})_{d_y})$ and $\boldsymbol{y}_\star^{\mathrm{mes}}$, by $\boldsymbol{C}_{\mathrm{train}}$ and $\boldsymbol{C}_{\star\star}$ their respective covariance matrices, and by $\boldsymbol{C}_\star$ the cross-covariance matrix between $\boldsymbol{y}^{\mathrm{train}}$ and $\boldsymbol{y}_\star^{\mathrm{mes}}$ (see Appendix A for the detailed expressions of these statistical moments). Finally, we approximate the probability distribution of $(\boldsymbol{y}_{\mathrm{train}}, \boldsymbol{y}_\star^{\mathrm{mes}})$ by a Gaussian PDF with the same parameters:

$$\begin{pmatrix} \boldsymbol{y}_{\mathrm{train}} \\ \boldsymbol{y}_\star^{\mathrm{mes}} \end{pmatrix} | \boldsymbol{\theta}, \boldsymbol{\gamma}, (\boldsymbol{x}_i^{\mathrm{mes}}, \boldsymbol{z}_i^{\mathrm{mes}})_{i=1}^n, \boldsymbol{x}_\star, \boldsymbol{z}_\star^{\mathrm{mes}} \approx \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathrm{train}} \\ \boldsymbol{\mu}_\star \end{pmatrix}, \begin{bmatrix} \boldsymbol{C}_{\mathrm{train}} & \boldsymbol{C}_\star \\ \boldsymbol{C}_\star^T & \boldsymbol{C}_{\star\star} \end{bmatrix}\right). \tag{4.13}$$

The Taylor expansions presented in Eqs. (4.9–4.12) enable to handle the products of Gaussian distributions rather than their compositions. This simplifies the formalism but does not lead to an

explicit expression for the joint probability distribution of $(\boldsymbol{y}_{\text{train}}, \boldsymbol{y}_\star^{\text{mes}})$ (the approximate distribution would be associated with a sum of Gaussian random variables and products of Gaussian variables, which could potentially be written as sums of random variables with a $\chi^2$ distribution). Manipulating such a distribution in a very high dimension (the dimension of $(\boldsymbol{y}_{\text{train}}, \boldsymbol{y}_\star^{\text{mes}})$ is equal to $(n+1) \times d_y$) would then be very complicated (if not numerically impossible), not to mention deriving conditional laws from it. Introducing the Gaussian approximation not only makes this joint distribution explicit, but also makes it possible to deduce explicitly the probability distribution of $\boldsymbol{y}_\star^{\text{mes}}$ conditional on $\boldsymbol{y}_{\text{train}}$.

Based on this Gaussian approximation, we proceed in three steps to estimate $\boldsymbol{x}_\star$. First, we estimate the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ by log-likelihood maximization using only the data gathered in $\mathcal{D}_n$ (plug-in approach [25]):

$$(\boldsymbol{\theta}, \boldsymbol{\gamma}) \approx (\boldsymbol{\theta}^{\text{MLE}}, \boldsymbol{\gamma}^{\text{MLE}}) \in \arg\max_{(\boldsymbol{t},\boldsymbol{g})} \ell_n(\boldsymbol{t}, \boldsymbol{g}), \tag{4.14}$$

where, in our case, the log-likelihood function is simply defined by:

$$\ell_n(\boldsymbol{t}, \boldsymbol{g}) := -\frac{1}{2}\left(\log(\det(\boldsymbol{C}_{\text{train}})) + (\boldsymbol{y}_{\text{train}} - \boldsymbol{\mu}_{\text{train}})^T \boldsymbol{C}_{\text{train}}^{-1}(\boldsymbol{y}_{\text{train}} - \boldsymbol{\mu}_{\text{train}})\right). \tag{4.15}$$

We then estimate $\pi[\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$ by fixing the error parameters to their maximum likelihood estimators (MLE). Since the Gaussian distribution is stable by conditioning, the expression for this PDF is explicit,

$$\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{\theta}^{\text{MLE}}, \boldsymbol{\gamma}^{\text{MLE}}, \boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n \;\sim\; \mathcal{N}\left(\boldsymbol{\mu}_\star + \boldsymbol{C}_\star^T \boldsymbol{C}_{\text{train}}^{-1}(\boldsymbol{y}_{\text{train}} - \boldsymbol{\mu}_{\text{train}}),\; \boldsymbol{C}_{\star\star} - \boldsymbol{C}_\star^T \boldsymbol{C}_{\text{train}}^{-1} \boldsymbol{C}_\star\right). \tag{4.16}$$

Finally, since we cannot easily compute the value of $c$ in Eq. (4.1), sampling techniques, such as Markov Chain Monte Carlo (MCMC), can be used to generate a set of points that are approximately distributed according to $c \times \pi[\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{\theta}^{\text{MLE}}, \boldsymbol{\gamma}^{\text{MLE}}, \boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] \times \pi[\boldsymbol{x}_\star|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$, which we assume to be close to the PDF of interest $\pi[\boldsymbol{x}_\star|\boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$. Conventional Metropolis-Hastings algorithms could be employed here to generate these Markov chains, and the mode and credibility intervals could eventually be estimated from the generated points. Nevertheless, the fact that the prior distribution, $\pi[\boldsymbol{x}_\star|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$, is modeled using a non-parametric approach increases the chances for $\pi[\boldsymbol{y}_\star^{\text{mes}}|\boldsymbol{\theta}^{\text{MLE}}, \boldsymbol{\gamma}^{\text{MLE}}, \boldsymbol{x}_\star, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n] \times \pi[\boldsymbol{x}_\star|\boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$ to be multi-modal, which may require the use of more sophisticated MCMC procedures, and their initialization in different points of $\mathbb{X}$ (see [19] for more details).

In Section 5, in order to avoid these difficulties associated with the use of MCMC algorithms, which are not at the heart of this paper, we decided to project the estimation of the *a posteriori* distribution of $\boldsymbol{x}_\star$ onto a fine discretisation of the input space $\mathbb{X}$ (using approximatively $150^{d_x}$ points), then to calculate all the quantities of interest which will be presented from this discrete approximation of $\pi[\boldsymbol{x}_\star|\boldsymbol{y}_\star^{\text{mes}}, \boldsymbol{z}_\star^{\text{mes}}, \mathcal{D}_n]$.

## 5. **Application**

The aim of this section is to demonstrate the relevance of the proposed approaches when applied to the calibration of non-specific multisensory systems. While the final goal is to deal with data from actual sensors, numerical experiments based on simulated data allow a better quantification of the performance and analysis of the different phenomena. Therefore, the illustrations presented in this section include both a synthetic and an experimental dataset. The synthetic dataset includes the main features generally observed in experimental datasets from multisensory systems deployed in an open periodic environment. The experimental dataset demonstrates the relevance of this method for an innovative sensor array solution incorporating nanomaterials and applied to air quality monitoring.

### 5.1. **Presentation of the data sets**

As explained above, this section focuses on two types of data, the first one derived from real data and the second one from simulated data. They are presented in parallel in this section to highlight their similar characteristics. Using the same notations as in the previous sections, we denote by $\boldsymbol{x}$ the quantities of interest, and by $\boldsymbol{y}$ and $\boldsymbol{z}$ the sensor outputs and the measured environmental variables that can be used for the monitoring. In the real case, the vector $\boldsymbol{x}$ corresponds to the concentrations of carbon monoxide (CO) and ozone ($O_3$), while the components of $\boldsymbol{z}$ are temperature and humidity.

The generation of the simulated data is explained below, while the real data were obtained after deploying of a carbon nanotube-based multisensory system (CNT) in an open environment over a period of approximately one month. As detailed in [17], the system contains a twenty chemistors forming a 10×2 sensor array. The chemistors are made out of carbon nanotubes, either non-functionalized (10 devices) or functionalized with an active polymer (10 devices). Out of the twenty sensor outputs, it is only possible to extract two truly complementary outputs due to the very high degree of similarity between the outputs of sensors from the same type (eg. functionalized or not).

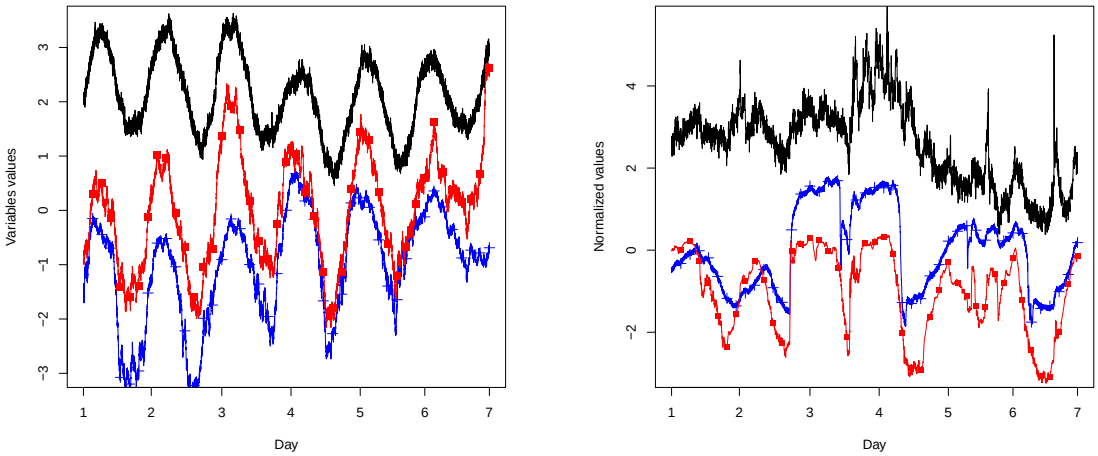
For the simulated data, we chose to consider Markov Chains to generate data that would share many of the characteristics of real data. The temporal evolution of $\boldsymbol{x}$, $\boldsymbol{z}$, and $u$ are thus associated with Gaussian random walk samples with a very strong correlation between two consecutive instants. We then added a trend to these random walks to model the day-night cycles for a 360-day period, and we multiplied the vector of these quantities by a matrix parameterized by $\rho \in (-1, 1)$ to correlate them. The larger $\rho$ is, the more the quantities $\boldsymbol{x}$, $\boldsymbol{z}$ and $u$ are correlated (positively or negatively). Each sensor output $y_i$ is then simulated using combinations of arc tangent, logarithmic, or linear functions:

$$y_j = (4 + \beta_j^1)\log(\beta_j^2 x_1 + 1 - \min(x_1)) + \beta_j^3 \arctan(\beta_j^4 z_1) + \beta_j^4 \arctan(\beta_j^5 x_2 + \beta_j^6 z_2) + \alpha_u u. \qquad (5.1)$$

The coefficients $\beta_j^m$ are fixed and chosen as $\pm(1+\alpha_j^m)$, where $\alpha_j^m$ is chosen uniformly between $-0.5$ and 0.5. Finally, a centred Gaussian noise is applied to the data to define the measured quantities, the noise-to-signal ratio is chosen equal to 5%. As an illustration, Figure 5.1 shows the temporal evolution of three measured quantities for the simulated case and the real case (the real outputs are normalised so that they can appear on the same graph), and Figure 5.2 shows the correlations between all the variables.

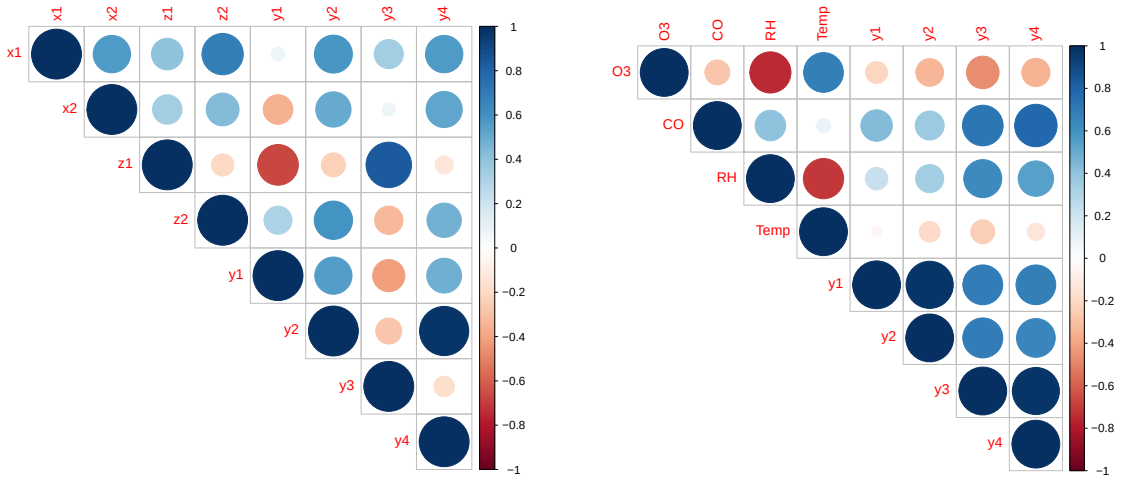
These figures illustrate the main features of the data under consideration and highlight the similarities between the simulated and real data. In both cases, the daily cycles of all the measured variables can be observed. Mechanically, there is a high statistical correlation between them. Focusing on the real data, the reference measurements of the target pollutants (especially CO) are very noisy, which is accounted for in the simulated data. While it is possible to arbitrarily define the number of interferent factors $\boldsymbol{z}$ in the simulated case, in real use cases (including the present one) some factors known to be influential cannot be easily measured (here, for instance the concentration of volatile organic compounds). This may explain certain fluctuations in sensor response that are not necessarily reflected in the evolution of $\boldsymbol{x}$ or $\boldsymbol{z}$. Finally, the very strong similarity between sensors of the same type can also be seen in the correlation matrix of the real data, where a correlation of 1 is observed between $y_1$ and $y_2$, as well as between $y_3$ and $y_4$.

In Figure 5.1, the real data are normalised so that they appear on the same graph. In reality, the non-standardised values vary from 7 to 8 ppm for CO, from 0 to 83 ppb for $O_3$, from 8 to 32 degrees for temperature, and from 30% to 100% for relative humidity. However, we will restrict ourselves to the points where $O_3$ is greater than 15 ppb. This choice is explained by the fact that below 15 ppb, the measured $O_3$ values are no longer really significant, as they are below the detection thresholds of the reference measuring instruments.




(A) Simulated data (B) Real data

FIGURE 5.1. Representation of the time evolution (left: simulated data; right: real data) over one week of one sensor output $y_1$ (in red with squares), one environmental variable $z_1$ (in blue with crosses) and one target pollutant (in black). For the real data (in the right figure), $x_1$ is the carbon monoxide concentration and $z_1$ is the relative humidity.




(A) Simulated data (B) Real data

FIGURE 5.2. Correlation matrices (left: simulated data; right: real data) between two pollutants of interest, $x_1$ and $x_2$ (ozone and carbon monoxide for the real case), two known environmental variables, $z_1$ and $z_2$ (relative humidity and temperature for the real case), and four sensor outputs $y_1$, $y_2$, $y_3$, $y_4$ (for the real case, $y_1$ and $y_2$ are two representatives of the non-functionalized sensors, $y_3$ and $y_4$ of the functionalized ones.

### 5.2. **Presentation of the performance indicators**

The two datasets presented in Section 5.1 serve to illustrate the interest of the methods presented in Sections 3 and 4. First, each dataset is divided into a training set of $N$ points, and a test set of 250 points. The dimension of the test set is chosen to be as large as possible to give a good idea of the predictive ability of the models, without excluding too much of the data for the training set. As the data are highly correlated in time, the test points are chosen far enough away from the training points to avoid any possible overlearning. Regarding the training points, it may seem interesting to take $N$ as high as possible. The calculation cost of the proposed methods nevertheless increases significantly with $N$ (in particular it is necessary to invert a matrix of size $N \times N$ many times during the construction of the models). Hence, the value of $N$ results from a compromise between prediction performance and numerical cost, and is chosen equal to 300 for the real case (this value will be specified for the simulated data, as it will vary depending on the context). To maximize the information provided by these $N$ points, the training points are chosen so as to best fill the input space $\mathbb{X} \times \mathbb{Z}$ using clustering methods [10, 14]. To obtain more consistent results, this random selection is repeated 10 times for the simulated case, and 5 times for the real case.

To evaluate the relevance of the methods for estimating the values of $x_i$, several of the standard indicators are used:

- $R_i^2$ is the coefficient of determination,
- $\text{MAE}_i$ is the mean average error,
- $\mathcal{L}_i^\alpha$ is the difference between the $1 - \alpha/2$ and the $\alpha/2$ quantiles, which defines a credibility interval of level $\alpha$ (without necessarily being the smallest credibility interval of level $\alpha$),
- $\mathcal{I}_i^\alpha$ is the percentage of true value belonging to these intervals.

In general, a method is said to be relevant if it is able to estimate the target values with reasonable confidence, i.e. with the smallest possible MAEs and credibility interval lengths, and with coefficients of determination close to 1. In the case of simulated data, since the true (non-noisy) values are known, we can also quantify the relevance of the credibility intervals by using the $\mathcal{I}_i^\alpha$ indicators, which should be as close to $\alpha$ as possible (on average a proportion $\alpha$ of the predictions are expected to be in $\mathcal{L}_i^\alpha$). In the real case, where the true (non-noisy) value of $\boldsymbol{x}$ remains unknown, this indicator is not as meaningful, as we can expect the number of noisy values belonging to the credibility interval associated with the unnoisy values of $x_i$ to be lower than $\alpha$.

We also propose to compare the capabilities of the proposed methods with respect to more standard models from the literature. In the following, we therefore refer to NP as the non-parametric method presented in Section 3, and to GPR+IU (IU for input uncertainties) as the parametric method presented in Section 4. In addition to these two, we include several parametric methods with different levels of complexity. We refer to GLR as the Generalised Linear Regression without taking into account input uncertainties, GLR+IU as the Generalised Linear Regression when taking into account input uncertainties, and GPR as the Gaussian Process Regression without input uncertainty consideration. Model errors are also be added in each case, in the same way as for GPR+IU. The functions $\boldsymbol{h}_j$ are selected from classes of polynomials with degrees optimized by cross-validation. The optimization was conducted empirically by incrementally elevating the total degree of the considered polynomials (more advanced techniques using different kinds of penalization to avoid overlearning could also be considered). For GLR, degree 2 polynomials are selected, while for GPR, only degree 1 polynomials are considered.

### 5.3. **Analysis of the two datasets**

The performance indicators achieved by the different prediction methods are listed in Table 5.1. Similar trends are observed for simulated and real data:

- The best results for each indicator are mostly obtained with either NP or GPR+IU.
- An increase in model complexity from GLR to GPR results in a decrease in MAE and an increase in $R^2$ values.
- The addition of input uncertainties in both GLR and GPR cases results in a decrease in the length of the credibility intervals, but does not automatically translate into an improvement in the MAE.
- The NP method produces almost equivalent results to GPR in terms of MAE with reduced computational cost, but with wider credibility intervals. This result is to be expected, given that the NP method predicts the noisy value, whereas the GPR method seeks to predict the denoised value.
- By focusing on the standard deviations of the obtained results rather than the means of the different splits, we generally observe better consistency when adding the uncertainties on the methods. This is particularly clear in the credibility intervals for both simulated and real cases. This is even more true for the percentages presented for the simulated case.

Table 5.1. Compared performance of different methods for the monitoring of pollutants in the air with application to the simulated data and the real data. The values correspond to the empirical averages of the indicators obtained after 10 repetitions of the complete approach for the simulated data, and 5 repetitions for the real data. It also shows the standard deviation of the performance indicators over the repetitions. For the real data, the index 1 is for $O_3$ (in ppb) while the index 2 is for CO (in ppb). The pollutants vary from 7 to 8 ppm for CO (with a mean of 7.4 and a standard deviation of 0.077), from 15 to 83 ppb for $O_3$ (with a mean of 34 and a standard deviation of 17). The best value observed for each indicator is displayed in bold letters.

| Method | $R_1^2$ | $R_2^2$ | $\text{MAE}_1$ | $\text{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|
| GLR | 0.87 ± 0.019 | 0.83 ± 0.021 | 0.39 ± 0.028 | 0.49 ± 0.038 | 1.8 ± 0.086 | 2.7 ± 0.18 | 87 | **95** |
| GLR+IU | 0.88 ± 0.017 | 0.81 ± 0.022 | 0.36 ± 0.025 | 0.51 ± 0.029 | 1.5 ± 0.070 | 2.6 ± 0.14 | 94 | 94 |
| GPR | **0.98 ± 0.0052** | **0.88 ± 0.019** | **0.14 ± 0.013** | **0.38 ± 0.023** | 0.95 ± 0.043 | 1.9 ± 0.16 | 94 | 91 |
| GPR+IU | **0.98 ± 0.0048** | **0.88 ± 0.030** | 0.15 ± 0.013 | **0.38 ± 0.021** | **0.86 ± 0.033** | **1.8 ± 0.13** | **97** | 93 |
| NP | 0.92 ± 0.011 | 0.87 ± 0.020 | 0.29 ± 0.016 | 0.40 ± 0.027 | 1.6 ± 0.062 | 1.9 ± 0.13 | **96** | 93 |

(a) Simulated data

| Method | $R_1^2$ | $R_2^2$ | $\text{MAE}_1$ | $\text{MAE}_2$ | $\mathcal{L}_1^{90\%}$ | $\mathcal{L}_2^{90\%}$ |
|---|---|---|---|---|---|---|
| GLR | 0.55 ± 0.072 | 0.74 ± 0.020 | 5.1 ± 0.19 | 0.030 ± 0.0022 | 18 ± 1.34 | 0.083 ± 0.011 |
| GLR+IU | 0.55 ± 0.057 | 0.74 ± 0.021 | 5.0 ± 0.13 | 0.028 ± 0.0021 | 18 ± 0.35 | 0.081 ± 0.0063 |
| GPR | 0.65 ± 0.041 | 0.79 ± 0.0082 | 4.4 ± 0.14 | 0.023 ± 0.0017 | 18 ± 0.52 | 0.079 ± 0.038 |
| GPR+IU | **0.73 ± 0.017** | 0.79 ± 0.017 | **4.2 ± 0.088** | **0.022 ± 0.014** | **17 ± 0.33** | **0.073 ± 0.0038** |
| NP | 0.68 ± 0.068 | **0.80 ± 0.056** | 4.5 ± 0.34 | **0.022 ± 0.0019** | 20 ± 0.45 | 0.097 ± 0.0073 |

(b) Real data

It is worth focusing briefly on parameter 1 of the real dataset, which strongly exemplifies the added value of the NP, GPR and GPR+IU compared to GLR. While parameter 1 of the real data set is not

predicted by the standard GLR method ($R^2 = 0.55$), the application of either GPR and NP has a significant effect on the $R^2$ values ($> +20\%$), then further consideration of IU in GPR increases the $R^2$ values by more than 30% compared to simple GLR.

Overall, while the advantage of each of the methods presented here is clear from Table 5.1, the specific added value of each method depends on the specific use case; it differs between each parameter of each use case. In order to understand their advantages and disadvantages in more details, and to propose more general conclusions, a wide range of configurations is now studied using the simulated data generator.

### 5.4. **Parametric study using simulated data**

Compared to the simulated data results presented in Table 5.1, we use the data generator to modulate the following parameters of the use case:

- the number of sensors $d_y$, from 2 to 10,
- the amplitude of the input noise, the standard deviation ranging from 1% to 20% of the mean of each input $x_i$,
- the presence or not of a non-observed interferent $u$,
- the size of the training set $N$, from 50 to 1000,
- the value of the correlation factor $\rho$ that links statistically the different environmental variables, from 0 to 0.9.

For the sake of brevity, we have included the detailed results of these different analyzes in Appendix B, and here we only list the main conclusions of these analyses.

- The performance of NP and GPR+IU improves as the number of sensors increases and tends towards similar predictions. However, the results suggest that the NP method may require more sensors and training data to achieve convergence compared to GPR+IU.
- Increasing the size $N$ of the training set also shows improvements in prediction errors and credibility interval lengths for both methods. GPR+IU shows superior performance, but the computation time increases substantially for large datasets.
- Concerning the response to different levels of input noise, we observe that the performance of GPR+IU gradually decreases as the noise increases, while the results associated with the NP method remain almost unchanged, suggesting minimal influence of measurement noise on NP.
- In addition, increasing the influence of the unobserved interferent on the sensors leads to a significant deterioration of the prediction results for both methods.
- Finally, increasing correlation rates between environmental variables poses an important challenge to calibration, leading to increased prediction errors and credibility intervals.

Overall, this study suggests that these two methods each have their own strengths, which should encourage us to try them both out for real-life test cases: NP is expected to be more robust in very noisy situations when a sufficient number of sensors is available, while GPR+IU is expected to cover less noisy situations with a lower number of sensors. This reinforces the conclusions reached in the real dataset: GPR+IU is overall slightly better than NP, as the number of independent sensors is very limited (only 2).

## 6. Conclusions

In this study, two calibration methods are proposed within a Bayesian framework. The aim of these methods is to incorporate various sources of uncertainty encountered by sensors in an uncontrolled environment. The first method is a non-parametric approach based on kernel density estimation, which offers simplicity of implementation, requires no prior assumptions about the data, and provides fast computation. The second approach is parametric and uses Gaussian process regression. Its main advantage lies is its ability to approximate complex and non-linear relationships between variables, in exchange for a higher computational cost. A further improvement of this second method is then introduced by considering uncertainties in the input data. This is particularly relevant when calibrating sensors in an open environment, where such uncertainties can significantly degrade the quality of sensor predictions. Two test cases are used to evaluate and compare these methods: a first one based on real data associated with a gas sensor array based on carbon nanotubes, and a second one based on simulated data. The main conclusion of these studies is that both proposed methods significantly reduce uncertainties and prediction errors compared to more standard approaches in the literature.

Several perspectives could be suggested for this work. Here we assume that the sensor outputs at time $t$ depend only on the environmental variables at time $t$. In reality, however, sensors can sometimes have long response times, so that their outputs depend on the environmental variables over a range of times. Another major difficulty is the problem of temporal drift, which can severely degrade the performance of sensors and calibration methods. A major improvement in these methods would be the ability to detect inaccuracies prediction due to response time and temporal drift and to automatically account for them in the sensor prediction.

## Appendix A. Computation of the mean and covariance of $(\boldsymbol{y}_{\text{train}}, \boldsymbol{y}_\star^{\text{mes}})$

Let $\boldsymbol{\mu}_{\text{train}}$ and $\boldsymbol{\mu}_\star$ be the expectations of $\boldsymbol{y}_{\text{train}} := ((\boldsymbol{y}_1^{\text{mes}})_1, \dots, (\boldsymbol{y}_n^{\text{mes}})_1, (\boldsymbol{y}_1^{\text{mes}})_2, \dots, (\boldsymbol{y}_n^{\text{mes}})_{d_y})$ and $\boldsymbol{y}_\star^{\text{mes}}$, let $\boldsymbol{C}_{\text{train}}$ and $\boldsymbol{C}_{\star\star}$ be their respective covariance matrices, and let $\boldsymbol{C}_\star$ be the cross-covariance matrix between $\boldsymbol{y}^{\text{train}}$ and $\boldsymbol{y}_\star^{\text{mes}}$. Using Eqs. (4.5) and introducing the notations $\boldsymbol{v}_i := (\boldsymbol{x}_i^{\text{mes}}, \boldsymbol{z}_i^{\text{mes}})$, $\boldsymbol{\epsilon}^v := (\boldsymbol{\epsilon}_x, \boldsymbol{\epsilon}_z)$ and $d_v := d_x + d_z$, we deduce, for each $1 \le j, k \le d_y$ and each $1 \le i, \ell \le n$,

$$(\boldsymbol{\mu}_{\text{train}})_{n(j-1)+i} := \mathbb{E}\left[(\boldsymbol{y}_i^{\text{mes}})_j\right] = \boldsymbol{h}_j(\boldsymbol{v}_i)\overline{\boldsymbol{\beta}}_j \tag{A.1}$$

$$(\boldsymbol{\mu}_\star)_j := \mathbb{E}\left[(\boldsymbol{y}_\star^{\text{mes}})_j\right] = \boldsymbol{h}_j(\boldsymbol{v}_\star)\overline{\boldsymbol{\beta}}_j \tag{A.2}$$

$$(\boldsymbol{C}_{\text{train}})_{n(j-1)+i,n(k-1)+l} := \text{Cov}((\boldsymbol{y}_i^{\text{mes}})_j, (\boldsymbol{y}_l^{\text{mes}})_k) = \delta_{jk}A_{i\ell}^{(j)} + \delta_{i\ell}B_{jk} \tag{A.3}$$

$$(\boldsymbol{C}_\star)_{n(j-1)+i,k} := \text{Cov}((\boldsymbol{y}_i^{\text{mes}})_j, (\boldsymbol{y}_\star^{\text{mes}})_k) = \delta_{jk}A_{i\star}^{(j)} + \delta_{i\star}B_{jk} \tag{A.4}$$

$$(\boldsymbol{C}_{\star\star})_{j,k} := \text{Cov}((\boldsymbol{y}_\star^{\text{mes}})_j, (\boldsymbol{y}_\star^{\text{mes}})_k) = \delta_{jk}A_{\star\star}^{(j)} + B_{jk} \tag{A.5}$$

where:

$$\boldsymbol{A}^{(j)} := \boldsymbol{H}_j\boldsymbol{\Gamma}_j\boldsymbol{H}_j^T + \boldsymbol{K}_j + \boldsymbol{C}_y + \boldsymbol{\Sigma} + \boldsymbol{S}_j, \tag{A.6}$$

$$\boldsymbol{H}_j := \begin{bmatrix} \boldsymbol{h}_j(\boldsymbol{x}_1^{\text{mes}}; \boldsymbol{z}_1^{\text{mes}})^T \\ \vdots \\ \boldsymbol{h}_j(\boldsymbol{x}_n^{\text{mes}}; \boldsymbol{z}_n^{\text{mes}})^T \end{bmatrix}, \quad \boldsymbol{K}_j := \begin{bmatrix} k_v^{(j)}(\boldsymbol{v}_1, \boldsymbol{v}_1) & \cdots & k_v^{(j)}(\boldsymbol{v}_1, \boldsymbol{v}_n) \\ \vdots & \ddots & \vdots \\ k_v^{(j)}(\boldsymbol{v}_n, \boldsymbol{v}_1) & \cdots & k_v^{(j)}(\boldsymbol{v}_n, \boldsymbol{v}_n) \end{bmatrix}, \tag{A.7}$$

$$\boldsymbol{S}_j := \begin{bmatrix} s_1^{(j)} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & s_n^{(j)} \end{bmatrix}, \tag{A.8}$$

$$k_v^{(j)}(\boldsymbol{v}_i, \boldsymbol{v}_\ell) := k_x^{(j)}(\boldsymbol{x}_i^{\text{mes}}, \boldsymbol{x}_\ell^{\text{mes}}) k_z^{(j)}(\boldsymbol{z}_i^{\text{mes}}, \boldsymbol{z}_\ell^{\text{mes}}), \tag{A.9}$$

$$s_i^{(j)} := \sum_{p=1}^{d_v}\sum_{q=1}^{d_v} (\boldsymbol{C}_v)_{pq} \left( \text{Cov}\left( \frac{\partial \varepsilon_j^{\text{mod}}}{\partial (\boldsymbol{v}_i)_p}(\boldsymbol{v}_i), \frac{\partial \varepsilon_j^{\text{mod}}}{\partial (\boldsymbol{v}_i)_q}(\boldsymbol{v}_i) \right) + \frac{\partial \boldsymbol{h}_j}{\partial (\boldsymbol{v}_i)_p}(\boldsymbol{v}_i)^T \boldsymbol{\Gamma}_j \frac{\partial \boldsymbol{h}_j}{\partial (\boldsymbol{v}_i)_q}(\boldsymbol{v}_i) \right), \tag{A.10}$$

$$B_{jk} := \overline{\boldsymbol{\beta}}_j^T \left( \sum_{p=1}^{d_v}\sum_{q=1}^{d_v} (\boldsymbol{C}_v)_{pq} \frac{\partial \boldsymbol{h}_j}{\partial (\boldsymbol{v}_i)_p}(\boldsymbol{v}_i) \frac{\partial \boldsymbol{h}_k}{\partial (\boldsymbol{v}_i)_q}(\boldsymbol{v}_i)^T \right) \overline{\boldsymbol{\beta}}_k. \tag{A.11}$$

$A_{i\star}^{(j)}$ and $A_{\star\star}^{(j)}$ have the same expressions as $A_{i\ell}^{(j)}$, with the difference that the uncertainties on $\boldsymbol{x}_\star$ do not appear anymore, and that in this case $\boldsymbol{\epsilon}^v := (\boldsymbol{0}, \boldsymbol{\epsilon}_z)$.

## Appendix B. Parametric analyses based on simulated data

Here, several studies are carried out by modifying a parameter that was used to generate the data. For each analysis, only one parameter is modified; the other parameters are chosen according to an identical reference characterized by $d_y = 3$, $N = 200$, noise level=5%, $\alpha_u = 0$, $\rho = 0.5$. The figures in the different tables correspond to averages over 10 tests for greater consistency of results.

### B.1. Influence of the number of sensors

Table B.1. Increasing the sensor number $d_y$.

| Method | $d_y$ | $R_1^2$ | $R_2^2$ | $\text{MAE}_1$ | $\text{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|---|
| NP | 2 | 0.89 | 0.87 | 0.33 | 0.39 | 2.03 | 2.08 | 97.16 | 95.04 |
| | 3 | 0.92 | 0.88 | 0.28 | 0.38 | 1.68 | 1.98 | 96.92 | 94.36 |
| | 5 | 0.97 | 0.87 | 0.16 | 0.39 | 0.92 | 1.91 | 96.24 | 93.40 |
| | 7 | 0.97 | 0.87 | 0.17 | 0.41 | 0.89 | 1.84 | 95.40 | 92.32 |
| | 9 | 0.97 | 0.93 | 0.17 | 0.30 | 0.90 | 1.30 | 95.00 | 91.60 |
| | 10 | 0.97 | 0.94 | 0.17 | 0.28 | 0.90 | 1.20 | 94.88 | 90.64 |
| GPR+IU | 2 | 0.97 | 0.89 | 0.17 | 0.38 | 1.11 | 1.89 | 98.10 | 94.40 |
| | 3 | 0.98 | 0.89 | 0.16 | 0.38 | 0.88 | 1.78 | 96.05 | 92.20 |
| | 5 | 0.99 | 0.89 | 0.11 | 0.37 | 0.61 | 1.54 | 96.05 | 88.85 |
| | 7 | 0.99 | 0.90 | 0.10 | 0.35 | 0.54 | 1.34 | 96.45 | 86.30 |
| | 9 | 0.99 | 0.94 | 0.09 | 0.29 | 0.46 | 1.01 | 96.25 | 82.50 |
| | 10 | 0.99 | 0.94 | 0.08 | 0.28 | 0.43 | 0.98 | 95.50 | 83.25 |

Firstly, a study was carried out to compare the calibration according to the number of sensors used from $d_y = 2$ to $d_y = 10$. The results are presented in Table B.1. An analysis of the results of the two methods shows an improvement as the number of sensors increases, both in terms of accuracy indicators and credibility intervals. However, there is a limit to the improvement of the GPR+IU method, which stabilises at around nine sensors, unlike the NP method. It is worth noting that the credibility interval for $x_2$ decreases significantly even at ten sensors. Another observation is that as the number of sensors increases, both methods seem to converge to identical results especially for

the second variable. Consequently, the NP method would require more sensors than GPR+IU, and subsequently more information, to achieve convergence. In addition, the percentage of points of the credibility intervals seems to decrease as the number of sensors increases for the GPR+IU method, indicating an improvement for variable 1 (it becomes closer to 95%) but a deterioration for variable 2 (this time away from the expected 95%). This could be due to the chosen plug-in approach, which fixes the model errors right after estimation, without associating uncertainties with these estimates.

Modifying the functions has the effect of completely changing the results. It should be noted that if one of the sensors has a very linear and simple relationship with a variable, fewer sensors may be needed to achieve convergence for that variable. Careful consideration should also be given to the selection of the most appropriate sensors.

### B.2. **Influence of the number of points in the training set**

Table B.2. Increasing the size $N$ of the training set.

| Method | $N$ | $R_1^2$ | $R_2^2$ | $\mathrm{MAE}_1$ | $\mathrm{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|---|
| NP | 50 | 0.86 | 0.82 | 0.38 | 0.47 | 1.66 | 1.82 | 90.67 | 87.42 |
| | 100 | 0.89 | 0.84 | 0.33 | 0.43 | 1.65 | 1.94 | 93.07 | 91.64 |
| | 200 | 0.93 | 0.86 | 0.28 | 0.40 | 1.60 | 1.96 | 95.51 | 93.42 |
| | 500 | 0.94 | 0.88 | 0.25 | 0.37 | 1.56 | 1.97 | 97.60 | 96.09 |
| | 700 | 0.95 | 0.89 | 0.23 | 0.35 | 1.57 | 1.99 | 98.49 | 96.67 |
| | 1000 | 0.96 | 0.90 | 0.22 | 0.34 | 1.56 | 1.99 | 99.07 | 97.78 |
| GPR+IU | 50 | 0.93 | 0.77 | 0.28 | 0.56 | 1.28 | 2.39 | 92.84 | 90.98 |
| | 100 | 0.96 | 0.84 | 0.19 | 0.44 | 1.00 | 2.10 | 94.49 | 93.11 |
| | 200 | 0.98 | 0.87 | 0.15 | 0.39 | 0.85 | 1.81 | 95.82 | 93.96 |
| | 500 | 0.98 | 0.91 | 0.13 | 0.31 | 0.75 | 1.58 | 96.84 | 95.42 |
| | 700 | 0.99 | 0.91 | 0.12 | 0.30 | 0.72 | 1.59 | 96.58 | 96.62 |
| | 1000 | 0.99 | 0.92 | 0.11 | 0.29 | 0.69 | 1.53 | 97.29 | 96.09 |

With a sufficient amount of data, it may be interesting to see what happens when the size of the training set is increased. The results are shown in Table B.2 for $N = 50$ to $N = 1000$. The first thing to notice is that the calibration seems to work relatively well even with few data. NP seems to work as well as GPR+IU and even better for the first variable. In the case of a small amount of data, it may be wiser to work with this method, which is also faster especially considering that GPR+IU may require more data than NP to correctly estimate the model parameters. More importantly, both methods show an improvement in terms of prediction error and credibility interval length with more training data. The GPR+IU method appears to have a superior performance. However, when $N = 1000$, the computation time is significantly longer. Another observation is that, regardless of the approach used, enlarging the training set boosts the percentage of belonging to the credibility intervals, despite the fact that the intervals become smaller. Consequently, the predictions become more rational and reliable.

### B.3. **Influence of the input noise**

With regards to experimental data following sensory deployment, access is limited only to the measured data. Understanding the results can be difficult in the presence of significant noise. Using simulated data, it is possible to compare these two methods and their responses to different percentages of noise on the variables. The results of this study are presented on Table B.3. The noise is added by a

Table B.3. Increasing the input noise (in % of the standard deviation of the inputs).

| Method | Noise (%) | $R_1^2$ | $R_2^2$ | $\text{MAE}_1$ | $\text{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|---|
| NP | 1 | 0.92 | 0.88 | 0.29 | 0.38 | 1.63 | 2.01 | 95.32 | 95.08 |
| | 2 | 0.93 | 0.88 | 0.29 | 0.38 | 1.64 | 1.99 | 95.84 | 94.92 |
| | 5 | 0.92 | 0.88 | 0.29 | 0.38 | 1.64 | 2.01 | 95.40 | 95.32 |
| | 10 | 0.92 | 0.88 | 0.29 | 0.39 | 1.62 | 1.99 | 95.44 | 95.24 |
| | 20 | 0.92 | 0.88 | 0.29 | 0.39 | 1.62 | 1.99 | 95.24 | 94.88 |
| GPR+IU | 1 | 0.98 | 0.88 | 0.15 | 0.38 | 0.84 | 1.83 | 95.40 | 94.60 |
| | 2 | 0.98 | 0.88 | 0.15 | 0.37 | 0.84 | 1.83 | 95.72 | 94.16 |
| | 5 | 0.98 | 0.88 | 0.15 | 0.38 | 0.87 | 1.84 | 96.08 | 94.44 |
| | 10 | 0.97 | 0.88 | 0.18 | 0.39 | 1.05 | 1.99 | 97.16 | 96.24 |
| | 20 | 0.96 | 0.87 | 0.22 | 0.42 | 1.30 | 2.34 | 98.24 | 96.76 |

percentage ranging from 1% to 20% of the standard deviation of the inputs. The same study could have been carried out by varying the noise on environmental variables or sensor outputs. One might assume that the more noisy the data, the more difficult the calibration. This appears to be the case for the GPR+IU method, which shows a gradual decline in performance as the noise level increases and the credibility interval widens. Nevertheless, the percentages are still justified. This indicates that even though we are less certain and precise in our predictions due to measurement uncertainties, the target value can still be found within the interval.

Examining the results for the NP technique, they appear almost identical, suggesting a negligible influence of measurement noise. This can be rationalised by the exclusion of noise from the method, which solely relies on data. In addition, it is noteworthy that as the noise increases, the outcomes of the GPR+IU approach converge with those of the NP approach.

However, in this case we are only adjusting the amount of noise. In real-world scenarios, the sensors face various challenges at the same time, which may lead to different results.

### B.4. **Influence of the unobserved interferent**

Table B.4. Increasing the influence of the unobserved interferent ($\alpha_u$)

| Method | $\alpha_u$ | $R_1^2$ | $R_2^2$ | $\text{MAE}_1$ | $\text{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|---|
| NP | 0 | 0.90 | 0.85 | 0.32 | 0.43 | 1.71 | 2.13 | 95.32 | 94.32 |
| | 0.01 | 0.90 | 0.85 | 0.31 | 0.43 | 1.70 | 2.13 | 95.44 | 94.52 |
| | 0.1 | 0.89 | 0.85 | 0.32 | 0.43 | 1.74 | 2.10 | 95.72 | 93.88 |
| | 1 | 0.67 | 0.85 | 0.56 | 0.43 | 2.74 | 2.16 | 94.20 | 94.80 |
| GPR+IU | 0 | 0.97 | 0.86 | 0.17 | 0.40 | 0.97 | 2.00 | 96.80 | 94.08 |
| | 0.01 | 0.96 | 0.86 | 0.19 | 0.41 | 0.98 | 1.97 | 94.48 | 93.72 |
| | 0.1 | 0.96 | 0.87 | 0.20 | 0.39 | 0.94 | 1.95 | 92.52 | 94.48 |
| | 1 | 0.67 | 0.72 | 0.57 | 0.61 | 2.11 | 3.04 | 84.80 | 95.56 |

Here we want to quantify the effect of environmental variables that affect the sensor but are unknown, $\alpha_u$ being the importance of that variable $u$. In Table B.4, we show the results for $\alpha_u$ from 0 to 1. For both approaches, it is difficult to draw conclusions when $\alpha_u \in \{0, 0.01, 0.1\}$, as there is sometimes a small decrease in the results and sometimes a small increase. Such changes are certainly an outcome of fluctuations in the results of the methods, but they are not significant. Nevertheless, $\alpha = 1$ shows a visible deterioration in both the prediction errors and the lengths of the credibility

intervals. The results for the first pollutant are similar for both methods. For the second pollutant, NP is superior to GPR+IU. Therefore, in cases of significant model error, NP may be a more attractive option due to its faster processing time.

### B.5. **Influence of the correlation between inputs.**

Table B.5. Increasing the $\rho$ coefficient characterizing the correlation between inputs.

| Method | $\rho$ | $R_1^2$ | $R_2^2$ | $\mathrm{MAE}_1$ | $\mathrm{MAE}_2$ | $\mathcal{L}_1^{95\%}$ | $\mathcal{L}_2^{95\%}$ | $\mathcal{I}_1^{95\%}$ | $\mathcal{I}_2^{95\%}$ |
|---|---|---|---|---|---|---|---|---|---|
| NP | 0 | 0.91 | 0.67 | 0.26 | 0.46 | 1.48 | 2.25 | 96.75 | 94.40 |
| | 0.5 | 0.91 | 0.80 | 0.31 | 0.47 | 1.79 | 2.33 | 96.40 | 94.45 |
| | 0.9 | 0.90 | 0.90 | 0.54 | 0.54 | 2.85 | 2.57 | 95.80 | 94.90 |
| GPR+IU | 0 | 0.98 | 0.74 | 0.14 | 0.39 | 0.74 | 1.88 | 96.65 | 93.35 |
| | 0.5 | 0.97 | 0.83 | 0.17 | 0.42 | 0.96 | 1.97 | 96.05 | 93.65 |
| | 0.9 | 0.95 | 0.91 | 0.39 | 0.51 | 2.07 | 2.35 | 95.85 | 93.90 |

In a real-world setting, environmental variables such as pollutants, temperature, and humidity display a high degree of correlation. The present study, whose results are reported in Table B.5, examines the effect of increasing the $\rho$ correlation rate on calibration results. Specifically, three rates were examined: a zero rate, representing uncorrelated variables; an average rate of 0.5, reflecting variables with low correlation; and a high rate of 0.9 for variables with a high correlation. As the correlation increases, the calibration becomes increasingly challenging due to the rise in prediction error and a significant increase in the credibility intervals. It is also worth noting that despite the increased error on the second variable, the coefficient of determination also increases. We might have expected it to decrease.

## References

[1] Moritz Berger, Christian Schott, and Oliver Paul. Bayesian Sensor Calibration. *IEEE Sensors Journal*, 22(20):19384–19399, 2022.

[2] Gookbin Cho, Sawsen Azzouzi, Gaël Zucchi, and Bérengère Lebental. Electrical and Electrochemical Sensors Based on Carbon Nanotubes for the Monitoring of Chemicals in Water – A Review. *Sensors*, 22(1): article no. 218, 2022.

[3] Council of the European Union European Parliament. Directive 2008/50/EC of the European Parliament. The council of 21 may 2008 on ambient air quality, and cleaner air for europe. *Official Journal of the European Union*, 2008.

[4] Atefeh Daemi, Yousef Alipouri, and Biao Huang. Identification of robust Gaussian Process Regression with noisy input using EM algorithm. *Chemometr. Intell. Lab. Syst.*, 191:1–11, 2019.

[5] Shivani Dhall, B. R. Mehta, A. K. Tyagi, and Kapil Sood. A review on environmental gas sensors: Materials and technologies. *Sensors International*, 2: article no. 100116, 2021.

[6] Tarn Duong, Arianna Cowling, Inge Koch, and M. P. Wand. Feature significance for multivariate kernel density estimation. *Comput. Stat. Data Anal.*, 52(9):4225–4242, 2008.

[7] Philip Erickson, Michael Cline, Nishith Tirpankar, and Tom Henderson. Gaussian processes for multi-sensor environmental monitoring. In *2015 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 208–213. IEEE, 2015.

[8] Maurizio Filippone and Guido Sanguinetti. Approximate inference of the bandwidth in multivariate kernel density estimation. *Comput. Stat. Data Anal.*, 55(12):3104–3122, 2011.

[9] F. J. Kelly and J. C. Fussell. Air pollution and public health: emerging hazards and improved understanding of risk. *Environ. Geochem. Health*, 37:631–649, 2015.

[10] Simon Mak and V. Roshan Joseph. Minimax designs using clustering. *J. Comput. Graph. Stat.*, 27(1):166–178, 2018.

[11] Andrew Mchutchon and Carl Rasmussen. Gaussian Process Training with Input Noise. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1341–1349. Curran Associates, Inc., 2011.

[12] Ali Mokhtari, Maryam Ghodrat, Pooya Javadpoor Langroodi, and Azadeh Shahrian. Wind speed sensor calibration in thermal power plant using Bayesian inference. *Case Studies in Thermal Engineering*, 19: article no. 100621, 2020.

[13] Mahendra Pal, Yodit Ayele, Angesom Hadush, Sumitra Panigrahi, and Vijay Jadhav. Public Health Hazards Due to Unsafe Drinking Water. *Air and Water Borne Diseases*, 7, 2018.

[14] G. Perrin and C. Cannamela. A repulsion-based method for the definition and the enrichment of opotimized space filling designs in constrained input spaces. *J. SFdS*, 158(1):37–67, 2017.

[15] G. Perrin and C. Durantin. Taking into account input uncertainties in the Bayesian calibration of time-consuming simulators. *J. SFdS*, 160(2):24–46, 2019.

[16] G. Perrin, C. Soize, and N. Ouhbi. Data-driven kernel representations for sampling with an unknown block dependence structure under correlation constraints. *Journal of Computational Statistics and Data Analysis*, 119:139–154, 2018.

[17] Guillaume Perrin and Bérengère Lebental. Uncertainty-Based Calibration Method for Environmental Sensors—Application to Chlorine and pH Monitoring With Carbon Nanotube Sensor Array. *IEEE Sensors Journal*, 23(5):5146–5155, 2023.

[18] Michael J. Pyrcz and Clayton V. Deutsch. Transforming data to a gaussian distribution. In Jared Deutsch, editor, *Geostatistics Lessons*. R Core Team, 2018. Retrieved from `https://geostatisticslessons.com/lessons/normalscore`.

[19] R. T. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Springer Series in Statistics. John Wiley & Sons, 2008.

[20] Thomas J. Santner, Brian J. Williams, and William I. Notz. *The design and analysis of computer experiments*. Springer Series in Statistics. Springer, 2003.

[21] Gustavo R. Taira, Adriano G. Leal, Alessandro S. Santos, and Song W. Park. Bayesian Neural Network-Based Calibration for Urban Air Quality Sensors. In Ludovic Montastruc and Stephane Negny, editors, *32nd European Symposium on Computer Aided Process Engineering*, volume 51 of *Computer Aided Chemical Engineering*, pages 1549–1554. Elsevier, 2022.

[22] Georgi Tancev and Federico Grasso Toro. Variational Bayesian calibration of low-cost gas sensor systems in air quality monitoring. *Measurement: Sensors*, 19: article no. 100365, 2022.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 58(1):267–288, 1989.

[24] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, 2009.

[25] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[26] An Wang, Yuki Machida, Priyanka deSouza, Simone Mora, Tiffany Duhl, Neelakshi Hudda, John L. Durant, Fábio Duarte, and Carlo Ratti. Leveraging machine learning algorithms to advance low-cost air sensor calibration in stationary and mobile settings. *Atmospheric Environment*, 301: article no. 119692, 2023.

[27] Naomi Zimmerman, Albert Presto, Sriniwasa Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis Robinson, Allen Robinson, and Subramanian Ramachandran. Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance. *Atmospheric Measurement Techniques Discussions*, pages 1–36, 2017.

[28] N. Zougab, S. Adjabi, and C. C. Kokonendji. Bayesian estimation of adaptive bandwidth matrices in multivariate kernel density estimation. *Comput. Stat. Data Anal.*, 75:28–38, 2014.