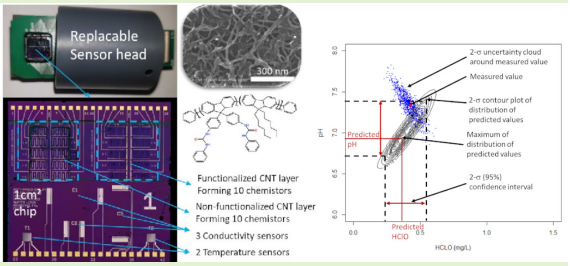
Uncertainty-Based Calibration Method for Environmental Sensors-Application to Chlorine and pH Monitoring With Carbon Nanotube Sensor Array

Guillaume Perrin, Berengere Lebental

Uncertainty-based calibration method for environmental sensors - application to chlorine and pH monitoring with carbon nanotube sensor array

Guillaume Perrin and Bérengère Lebental

***Abstract*— Societal demands in the field of air and water pollution monitoring require the ability to simultaneously detect a wide variety of chemical elements at very low concentrations in complex environments using compact and low-cost sensor devices. Although nanomaterial-based sensors have long been proposed as a solution to these exacting requirements, their detection accuracy is generally degraded in real-world conditions compared to laboratory conditions due to the effects of various interferents. To manage the related uncertainties and to associate a confidence to the estimations, it seems natural to formulate the calibration-estimation problem in a probabilistic framework. This probabilistic formulation, and its successful application to the monitoring of pH and active chlorine in drinking water, are the main contributions of this work. While other tested calibration methods only allowed the monitoring of active chlorine, our solution enables monitoring of both active chlorine and pH with uncertainties on par with the measurement uncertainties. Its success relies mainly on two adaptations of standard calibration methods: the consideration of measurement errors on reference calibration instruments (and not only on the sensor outputs as is more often done), and the introduction of two sources of model error, one accounting for the approximate character of the calibration model, the other for unmeasured - and possibly unknown - interferents in the calibration environment.**



***Index Terms*— Uncertainty quantification, sensor calibration, carbon nanotubes, drink water monitoring.**

I. Introduction

WATER (and air) quality has become a major public health issue [8]. An increasing number of cities are therefore interested in deploying technologies that will allow them to better monitor water pollution, and then limit it as much as possible. Traditionally, the concentrations of a reduced number of selected pollutants are measured in near real-time by a small number of high-accuracy monitoring stations located at strategic points. While these systems have a tremendous impact nowadays (for instance to manage city-wide pollution alerts or pollution-reduction policies), due to their distance from each other (because of their high cost), these stations do not allow to map contaminants in a localized way, nor to identify their sources. At the level of urban water networks, monitoring water quality at higher spatial resolution monitoring could for instance enable a faster detection of chemical contamination events in the network. This status strongly drives the development of compact and low-cost solutions that can be deployed discreetly in large numbers in urban area [16]. Nanomaterial-based sensors [33], [34] have long been proposed as a solution to this need for environmental monitoring, among which particularly carbon nanotubes [26]. It is generally thought that, due to their high surface over volume ratio, nanomaterials are more sensitive to chemicals than their bulk counterpart, and their selectivity can be enhanced through chemical engineering. In addition, they can be shaped into small devices (from cm^2 to μm^2), which offers the possibility to implement them in the context of environmental sensor networks or even further to develop chemical sensor arrays (often called electronic tongues or noses) enabling multiparameter sensing on chip [37]. The specifics of operation and sensitivity of such devices widely vary across the literature depending on the selected transduction mode and nanomaterial; in Section IV, we illustrates briefly a possible operation mode in the case of multiplexed carbon nanotube sensors chemistors. As a detrimental corollary, they often suffer from high noise, strong sensitivity to environmental perturbations (temperature, humidity, chemical

The authors thank Thomas Vezin for dataset preparation. 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.

G. Perrin and B. Lebental are with the Université Gustave Eiffel, COSYS, F-77454 Marne-la-Vallée, France (e-mails: guillaume.perrin@univ-eiffel.fr, berengere.lebental@univ-eiffel.fr)

interferents) and rapid ageing (characterized by drift).

Like any sensor, they must first be calibrated in a controlled environment as close as possible to the conditions encountered during deployment [1], [14]. The objective of this calibration phase is to identify the dependence between sensor response and pollutant concentrations. The identified model is then used during the deployment phase to determine pollutant concentrations based on the real-time sensor response.

Calibration is done most often through the introduction of a parametric model, whose complexity, i.e. the number of parameters on which the model depend, can strongly vary according to the applications [4], [6], [7], [19]. For nanomaterial-based sensors however, the definition of this model is a particularly challenging step, as they are likely to be sensitive to many more influence factors than traditional sensors. This is all the more true when considering sensor array approaches which are targeting multiple outputs. This difficulty is reflected in the increasing use of machine learning techniques for these systems [18], [23], [36], as they generally do not require the explanation of all the physical chemistry of the system via a mathematical model (see for instance [30] for a comparison between linear model-based and neural network based calibration). It should be noted that the very interesting generalization capacity of these techniques is most often based on the introduction of a very large number of parameters to be identified. Hence, proposing adapted regularization techniques to avoid overfitting is a central issue. Despite the use of these regularization techniques, in the case where only reduced volumes of data are available (which commonly happens in lab calibration activities), it often appears that only methods relying on a very small number of parameters are performing well. Among the techniques relying on few parameters, we focus here specifically on Gaussian process regression (GPR), as is done in [2], [17], [22], whose interest is also to associate a confidence level to the predictions it provides.

In the *small data* context, it is moreover essential to take into account all the uncertainties related to the experimental acquisitions of the data. This is particularly important for nanomaterial-based sensors, where significant measurement uncertainties are present in the sensor outputs, but also for environment monitoring, as the target pollutant concentrations are often close to the limit of detection of the instruments. Probabilistic or Bayesian approaches are therefore particularly attractive, as they offer a theoretical framework integrating all these uncertainties in a very natural way [3], [31]. Compared to other calibration approaches, their outputs are more informative: they provide not only the estimation of pollutant concentrations based on sensor outputs, but also the probability distribution of these concentrations, from which it is possible to extract one (or more) probable value and credibility intervals [9], [11], [15], [32]. These methods also offer the possibility to consider *a priori* expertise as well as a model error. Rarely introduced in the literature, the explicit integration of model error allows to account for incomplete knowledge of the sensor operating law.

Thus, the objective of this paper is to describe a Bayesian methodology for sensor calibration in a *small data* context which takes into account all the sources of uncertainty that can affect the results. The proposed method differs from the literature by its exhaustive treatment of experimental uncertainties as well as by the introduction of two terms for the model error. Indeed, while only uncertainties on sensor outputs are usually considered in sensor calibration phases, we will first show how to rigorously integrate into a GPR formalism the measurement uncertainties on all environmental quantities impacting sensor outputs. We will then show to what extent this GPR formalism can be extended to incorporate a model error that allows to represent not only the imperfect knowledge of the sensor calibration model, but also the potential presence of unknown influence factors. The significance, in terms of estimation accuracy, of these two contributions will first be illustrated on a test case based on simulated data, then on an experimental dataset generated by an array of carbon nanotube-based chemistors designed to monitor pH and active chlorine in drinking water.

The outline of this work is as follows: Section II presents the Bayesian formalism of sensor calibration and estimation we propose. Then, Section III highlights the interest of the proposed method on simulated data, while Section IV presents the application in water quality analysis. At last, Section V concludes the paper.

II. Theoretical Framework

A. Notations and available information

Let t be the time, $\boldsymbol{x}(t) \in \mathbb{X} \subset \mathbb{R}^{d_x}$ be the vector gathering the concentrations of the d_x pollutants to monitor, $\boldsymbol{y}(t) \in \mathbb{Y} \subset \mathbb{R}^{d_y}$ be the vector gathering the d_y sensor outputs, and $\boldsymbol{z}(t) \in \mathbb{Z} \subset \mathbb{R}^{d_z}$ be the vector gathering the d_z environment characteristics that are likely to affect the sensor outputs (such as temperature or relative humidity [13], [35]). For simplicity reasons, we only focus on the monitoring of $\boldsymbol{x}$ at discrete times t_i, and we forget about the intermediate times. We therefore denote by $(\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{y}_i)$ the true value of $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$ at $t = t_i$. It is also assumed that the response time of the sensors is small compared to the time intervals between two measurements, as well as to the characteristic times of evolution of the pollutant concentrations. From a practical point of view, this amounts to assuming that each observation has waited long enough for the sensor response to stabilize, and that the value of $\boldsymbol{y}_i$ only depends on the value of $\boldsymbol{x}$ and $\boldsymbol{z}$ at time $t = t_i$ (and not at the previous times). While the observed values of $(\boldsymbol{z}_i, \boldsymbol{y}_i)$ are assumed to be available at all times, it is however important to notice that the measurements of $\boldsymbol{x}$ are only available in a finite (and often reduced) number of times noted $t = t_{i_k}$, with $1 \leq k \leq n$. Given this formalism, we generally call *sensor calibration* the step of identifying a relation between $\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{y}_i$ from the information gathered in

$$\mathcal{D}_n := \{(\boldsymbol{x}_{i_k}^{\text{obs}}, \boldsymbol{z}_{i_k}^{\text{obs}}, \boldsymbol{y}_{i_k}^{\text{obs}}),\ 1 \leq k \leq n\},$$

$$\boldsymbol{x}_i^{\text{obs}} := \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i^x,\ \boldsymbol{z}_i^{\text{obs}} := \boldsymbol{z}_i + \boldsymbol{\varepsilon}_i^z,\ \boldsymbol{y}_i^{\text{obs}} := \boldsymbol{y}_i + \boldsymbol{\varepsilon}_i^y, \quad (1)$$

where $(\boldsymbol{x}_{i_k}^{\text{obs}}, \boldsymbol{z}_{i_k}^{\text{obs}}, \boldsymbol{y}_{i_k}^{\text{obs}})$ is the measurement of $(\boldsymbol{x}_{i_k}, \boldsymbol{z}_{i_k}, \boldsymbol{y}_{i_k})$. And we call *estimation* the step of estimating the value of $\boldsymbol{x}_i$

from this learned relation and the observation $(\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}})$ of $(\boldsymbol{z}_i, \boldsymbol{y}_i)$ alone.

B. Specificity of the problem

Let $\boldsymbol{\varepsilon}_i^x, \boldsymbol{\varepsilon}_i^z, \boldsymbol{\varepsilon}_i^y$ be the measurement errors, such that:

$$\boldsymbol{x}_i^{\text{obs}} := \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i^x, \;\; \boldsymbol{z}_i^{\text{obs}} := \boldsymbol{z}_i + \boldsymbol{\varepsilon}_i^z, \;\; \boldsymbol{y}_i^{\text{obs}} := \boldsymbol{y}_i + \boldsymbol{\varepsilon}_i^y. \quad (2)$$

The relation between $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$ is moreover written under the form

$$\boldsymbol{y}_i = \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{z}_i) + \boldsymbol{\zeta}_i, \quad (3)$$

where $\boldsymbol{f}$ is an unknown function, and $\boldsymbol{\zeta}_i$ is a model error aggregating the potential effects on $\boldsymbol{y}_i$ of other quantities not listed in $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. Combining Eqs. (2) and (3), we obtain $n+1$ equations allowing to make a connection between $\boldsymbol{x}_i$, $\boldsymbol{z}_i^{\text{obs}}$, $\boldsymbol{y}_i^{\text{obs}}$, and the n triplets in $\mathcal{D}_n$:

$$\boldsymbol{y}_i^{\text{obs}} = \boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}} - \boldsymbol{\varepsilon}_i^z) + \boldsymbol{\varepsilon}_i^y + \boldsymbol{\zeta}_i, \quad (4)$$

$$\boldsymbol{y}_{i_k}^{\text{obs}} = \boldsymbol{f}(\boldsymbol{x}_{i_k}^{\text{obs}} - \boldsymbol{\varepsilon}_{i_k}^x, \boldsymbol{z}_{i_k}^{\text{obs}} - \boldsymbol{\varepsilon}_{i_k}^z) + \boldsymbol{\varepsilon}_{i_k}^y + \boldsymbol{\zeta}_{i_k}. \quad (5)$$

The specificity of the formalism presented in this work lies in the fact that for the targeted applications (in particular the calibration of chemical nanosensors), the measurement uncertainties on $\boldsymbol{x}$ and $\boldsymbol{z}$ are not negligible (because of experimental constraints), and that the knowledge of $\boldsymbol{x}$ and $\boldsymbol{z}$ does not allow to completely predict $\boldsymbol{y}$ (because of unknown or not-measured interferents), hence the very important role of the term $\boldsymbol{\zeta}_i$. The calibration and estimation steps are thus impacted by many sources of uncertainty: while the statistical properties of the measurement errors are generally provided, the choice of $(\boldsymbol{f}, \boldsymbol{\zeta}_i)$ from the data is often not trivial.

To account for the different sources of uncertainty affecting the calibration and estimation problems, a Bayesian framework is proposed [20]. It amounts to assuming that the value to be estimated, $\boldsymbol{x}_i$, the measurement errors, the model error $\boldsymbol{\zeta}_i$, but also the function $\boldsymbol{f}$ can be modeled by random quantities in order to account for their uncertain nature. Estimating the value of $\boldsymbol{x}_i$ in a Bayesian formalism requires first to introduce prior distributions for all these random quantities, and then to characterize as well as possible the statistical distribution of $\boldsymbol{x}_i$ given the information provided by $\boldsymbol{z}_i^{\text{obs}}$, $\boldsymbol{y}_i^{\text{obs}}$, and $\mathcal{D}_n$. Proposing an a priori for $\boldsymbol{x}_i$ is particularly interesting in a small data framework, as a priori knowledge often tends to prevent overfitting, to regularize the inversion problem, and to guide its solving to the most relevant areas.

C. Definition of the a priori models

For the sake of simplicity, the model errors and the measurement errors are assumed to be centered and Gaussian. There is very little loss in generality: a transformation of the inputs can be applied to make the errors Gaussian-like if the assumption is not verified. We note $\boldsymbol{C}_\zeta^{(i)}$, $\boldsymbol{C}_x^{(i)}$, $\boldsymbol{C}_z^{(i)}$ and $\boldsymbol{C}_y^{(i)}$ their respective covariance matrices. While the values of $\boldsymbol{C}_x^{(i)}$, $\boldsymbol{C}_z^{(i)}$ and $\boldsymbol{C}_y^{(i)}$ are usually provided, the covariance matrices of the model error are *a priori* unknown. A gaussian prior is also associated with function $\boldsymbol{f}$, which amounts to assuming that the sensor outputs can be seen as a particular realization of a vector-valued Gaussian process,

$$\boldsymbol{f} \sim \text{GP}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{C}(\boldsymbol{\theta})), \quad (6)$$

whose mean function, $\boldsymbol{\mu}$, and covariance function, $\boldsymbol{C}$, are assumed to be parameterized by a vector $\boldsymbol{\theta}$ to be determined from the data. There are several reasons for choosing to use the Gaussian process regression (GPR) formalism for modeling $\boldsymbol{f}$. First of all, we underline its flexibility and its very good properties for the approximation of functions defined on relatively low dimensional input spaces [28], [29]. Then, choosing $\boldsymbol{f}$ as Gaussian when all the other sources of uncertainty are also assumed to be Gaussian makes it possible to carry out in an analytical way a large part of the statistical treatments. Finally, we note that modeling $\boldsymbol{f}$ as a Gaussian process in the calibration-estimation process allows to integrate in a natural way the necessarily approximate character of the relation between $\boldsymbol{x}$, $\boldsymbol{z}$ and $\boldsymbol{y}$.

In order to make the input/output model identifiable, it is also assumed that the model errors are pair-wise independent, while having the same statistical properties,

$$\mathbb{E}\left[\boldsymbol{\zeta}_i \boldsymbol{\zeta}_j^T\right] = \delta_{ij} \boldsymbol{C}_\zeta, \quad (7)$$

where $\mathbb{E}[\cdot]$ is the mathematical expectation, and δ_{ij} is the Kronecker symbol equal to 1 if $i = j$ and 0 otherwise. It is also assumed that these model errors are statistically independent of the Gaussian process $\boldsymbol{f}$.

D. Bayesian inversion

Given values for $\boldsymbol{\theta}$ and $\boldsymbol{C}_\zeta$, predicting $\boldsymbol{x}_i$ amounts to searching the conditional distribution (also called posterior distribution) of $\boldsymbol{x}_i \mid \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$ (the vertical bar indicates the conditioning). To simplify the notations of the expressions to come, we abusively denote by $\pi(\boldsymbol{a})$ the probability density function (PDF) of any random vector $\boldsymbol{a}$ taking the value $\boldsymbol{a}$. Noting that $\boldsymbol{x}_i$ is *a priori* independent of $\mathcal{D}_n$ and $\boldsymbol{z}_i^{\text{obs}}$, the Bayes formula tells us that the law of $\boldsymbol{x}_i$ knowing $(\boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n)$ can be decomposed as:

$$\pi(\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n) \propto \pi(\boldsymbol{x}_i) \times \pi(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}}, \mathcal{D}_n), \quad (8)$$

where $\propto$ indicates a proportional relationship, $\pi(\boldsymbol{x}_i)$ is the *a priori* PDF of $\boldsymbol{x}_i$, which is assumed to be known from expert judgment, and $\pi(\boldsymbol{y}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n)$ is the likelihood function. For example, if the only *a priori* information on each component of $\boldsymbol{x}_i$ is a minimum and a maximum value, uniform laws on this admissible interval can be chosen as *a priori* PDF (see [20] for more details).

Concerning the likelihood function, it is however generally too difficult to work directly with the input/output models given by Eqs. (4) and (5), which are known as *error-in-variables* (ERI) models in the statistics [5], [21]. Indeed, as $\boldsymbol{\varepsilon}_{i_k}^x$, $\boldsymbol{\varepsilon}_{i_k}^z$ and $\boldsymbol{\varepsilon}_i^z$ are random, even if $\boldsymbol{f}$ is Gaussian, there is no reason for $\boldsymbol{y}_{i_k}^{\text{obs}}$ or

$\boldsymbol{y}_i^{\text{obs}}$ to be Gaussian (on the contrary, their probability distributions are generally not computable due to the compositions). As an extension of what is shown in [12], a possible way to circumvent this problem is to replace $\boldsymbol{f}(\boldsymbol{x}_{i_k}^{\text{obs}} - \boldsymbol{\varepsilon}_{i_k}^x, \boldsymbol{z}_{i_k}^{\text{obs}} - \boldsymbol{\varepsilon}_{i_k}^z)$ and $\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}} - \boldsymbol{\varepsilon}_i^z)$ by $\boldsymbol{g}(\boldsymbol{x}_{i_k}^{\text{obs}}, \boldsymbol{z}_{i_k}^{\text{obs}}; \boldsymbol{C}_x^{(i_k)}, \boldsymbol{C}_z^{(i_k)})$ and $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}}; \boldsymbol{0}, \boldsymbol{C}_z^{(i)})$ respectively, where $\boldsymbol{g}$ is a Gaussian process whose mean and covariance functions are the same than the ones of the (non-Gaussian) second order Taylor expansion of $\boldsymbol{f}$ around the available observations (see [25] for more details). Note that we are trying here to assess the noiseless value of $\boldsymbol{x}_i$. This explains the presence of a zero covariance matrix on $\boldsymbol{x}$ in the second expression of $\boldsymbol{g}$. Thanks to this approximation, the distribution of $(\boldsymbol{y}_i^{\text{obs}}, \boldsymbol{y}_{i_1}^{\text{obs}}, \ldots, \boldsymbol{y}_{i_n}^{\text{obs}})$ is this time Gaussian, and it is possible to derive explicitly the likelihood function $\pi(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}}, \mathcal{D}_n)$ by Gaussian conditioning. Looking at Eq. (8), now that we have explicit expressions for $\pi(\boldsymbol{y}_i^{\text{obs}} | \boldsymbol{x}_i, \boldsymbol{z}_i^{\text{obs}}, \mathcal{D}_n)$ and $\pi(\boldsymbol{x}_i)$, it is possible to evaluate, to an unknown multiplicative constant, the PDF of $\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$ in any point of $\mathbb{X}$. The fact that the multiplicative constant is not known complicates the manipulation of this PDF. Indeed, without this constant, the evaluation of the *a posteriori* PDF in a particular value $\boldsymbol{x}_i^\star$ of $\mathbb{X}$ does not make it possible to characterise the likelihood of $\boldsymbol{x}_i^\star$ in the absolute, but simply to specify its greater or lesser likelihood in relation to other values of $\boldsymbol{x}_i$ that were already evaluated. One way to circumvent this problem, which is the basis of the Markov Chain Monte Carlo (MCMC) methods, is to construct a Markov chain that has the desired distribution as its equilibrium distribution. Hence, by recording the states from the chain once it has converged, we can generate a set of m points $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(m)}$ that are approximately distributed according to the PDF of $\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$. In the following, standard Metropolis-Hastings algorithms (see [27] for more details) will be used to simulate Markov chains such that the stationary distributions of the chains coincide with the target distributions. Kernel-based methods (see [24] for more details) can finally be used to approximate the PDF of $\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$ from these points.

According to this formalism, we can distinguish two contributions to the estimation uncertainty: the uncertainty linked to the residual variance of the conditioned Gaussian process, which could potentially be reduced by increasing the number of observations of $(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y})$, and the model uncertainty linked to the nature of the approximation class (e.g. the choice of a Gaussian process to describe the input-output relationship, with specific forms for the mean and covariance functions). Without modification of the approximation class, this latter can hardly be reduced.

E. Offline versus online computational cost

The previous formalism requires the knowledge of $\boldsymbol{\theta}$ (to derive the mean and covariance functions of the Gaussian process $\boldsymbol{f}$) and $\boldsymbol{C}_\zeta$ (for the model errors). As the probability distribution of $(\boldsymbol{y}_{i_1}^{\text{obs}}, \ldots, \boldsymbol{y}_{i_n}^{\text{obs}})$ knowing $\boldsymbol{\theta}$ and $\boldsymbol{C}_\zeta$ is Gaussian, the likelihood function is explicit, so that we can estimate these quantities by their maximum likelihood estimators. We can thus distinguish two costs for the monitoring of the concentration of the pollutants of interest at time t_i. We call *offline cost* the computational cost associated with the processing of the measurements gathered in $\mathcal{D}_n$, i.e., the estimation of $\boldsymbol{\theta}$ and $\boldsymbol{C}_\zeta$, which we perform once and for all. On the contrary, we call *online cost* the computational cost associated with the generation of the $\boldsymbol{x}_i^{(1)}, \ldots, \boldsymbol{x}_i^{(m)}$ using sampling procedures for each t_i.

Remark: The results presented in the following sections are all from dedicated developments using the R software.

III. Application on Simulated Data

The objective of this analytical case (see Appendix A for details) is to allow, in a perfectly controlled environment, to validate the performances of the main features needed for the nanosensor test case presented in Section V, namely:

- the non-linear relationships between $\boldsymbol{x}$, $\boldsymbol{z}$, and $\boldsymbol{y}$,
- the presence of threshold phenomena (the sensitivities of the sensors may be very different at low or high concentration of pollutants, due to saturation effects for example),
- the presence of unmeasured interferents in the measurement environment,
- the presence of non-negligible measurement uncertainties on $\boldsymbol{y}$, but also on $\boldsymbol{x}$ and $\boldsymbol{z}$,
- form and coefficients of the calibration model relatively similar between the different sensors due to similar manufacturing processes.

For the sake of brevity, we restrict ourselves to the case where $d_x = d_z = d_y = 2$ and $n = 100$, but additional results are provided in Appendix B.

A. Performance metrics

In order to evaluate the relevance of the estimations, several indicators are calculated. According to Section II-D, for each value of i in

$$\mathcal{I}^{\text{test}} := \{1, \ldots, 2000\} \backslash \{i_1, \ldots, i_n\},$$

we calculate the PDF of $\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$, from which we extract mainly two types of information:

- the most likely value of $\boldsymbol{x}_i$, i.e. the value of $\boldsymbol{x}$ maximizing its PDF, which is denoted by $\boldsymbol{x}_i^{\text{MAP}}$, from which a normalized mean square error is calculated:

$$e^2 := \sum_{i \in \mathcal{I}^{\text{test}}} \left\| \boldsymbol{x}_i - \boldsymbol{x}_i^{\text{MAP}} \right\|^2 / \sum_{i \in \mathcal{I}^{\text{test}}} \left\| \boldsymbol{x}_i \right\|^2, \qquad (9)$$

- the 95% credible intervals for each component of $\boldsymbol{x}_i | \boldsymbol{z}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}, \mathcal{D}_n$. Two quantities of interest are then deduced from these credible intervals: the percentages $\%_1^{0.95}$ and $\%_2^{0.95}$ of true values of $(\boldsymbol{x}_i)_1$ and $(\boldsymbol{x}_i)_2$ belonging to these confidence intervals, and their lengths $L_1^{0.95}$ and $L_2^{0.95}$. Indeed, a monitoring method is considered relevant if the true value falls into the confidence intervals and if the confidence intervals have relevant dimensions, i.e. they are neither too large to remain informative, nor too small to keep enough chances to include the true values. The values of $L_1^{0.95}$ and $L_2^{0.95}$ should be compared to the length of the definition intervals of x_1 and x_2, which is 10 in each case.

B. Analysis of the results

To underline the importance of taking into account the measurement uncertainties on $\boldsymbol{x}$ and $\boldsymbol{z}$, as well as the introduction of a double model error, the performances of several calibration strategies of increasing complexity are now compared. It is important to note that all the methods compared in the following are based on a Bayesian framework. Only the parametric classes in which the input-output relationships for the sensors will be sought vary, as well as the addition or not of Model Error (ME) or the consideration of measurement Uncertainties on the Inputs (IU).

- SLR (Simple Linear Regression) is the case when we assume an input-output relationship of the form:

$$\begin{gathered}(\boldsymbol{y}_i)_j = b_0^j + \sum_{\ell=1}^{d_x} b_\ell^{x,j}(\boldsymbol{x}_i)_\ell + \sum_{\ell=1}^{d_z} b_\ell^{z,j}(\boldsymbol{z}_i)_\ell, \\ \boldsymbol{x}_i^{\text{obs}} = \boldsymbol{x}_i,\ \boldsymbol{z}_i^{\text{obs}} = \boldsymbol{z}_i,\ \boldsymbol{y}_i^{\text{obs}} = \boldsymbol{y}_i + \boldsymbol{\varepsilon}_i^y.\end{gathered} \tag{10}$$

  In that case, the uncertainties in the inputs are neglected, and a particularly simple relationship is postulated for the input-output relationship. Although this expression is very likely to be imperfect, no model error is introduced here. The uncertainty in the estimate of $\boldsymbol{x}_i$ then depends only on the output measurement uncertainties, as well as the uncertainties in the estimation of the coefficients $(b_0^j, b_\ell^{x,j}, b_\ell^{z,j})$. This configuration is expected to have difficulties in correctly estimating pollutant concentrations, but will serve as a reference, in the sense that it will take into account the least number of sources of uncertainty.
- SLR+ME (Simple Linear Regression + Model Error) is the case when the uncertainties on the inputs are again neglected and the input-output relationship is also defined by a simple linear regression, but considering this time an additive model uncertainty $\boldsymbol{\zeta}$ with components that are independent of each other.
- GLR+ME and GLR+ME+IU (Generalized Linear Regression + Model Error + Input Uncertainty) are the cases when the input-output relationship is modeled by a generalized linear regression, that is to say a polynomial function with optimized polynomial degrees, and when an additive model error term is considered. In order to evaluate the impact of considering the uncertainties on the inputs, they are accounted for in the model GLR+ME+IU (but not in GLR+ME).
- GPR+ME and GPR+ME+IU (Gaussian Process Regression + Model Error + Input Uncertainty) are the cases where the input-output relationship is this time modeled from a Gaussian process, as described in Section II-B. In both cases, a linear mean function is considered, the covariance functions are chosen in the class of Matern-5/2 covariance functions (see [29] for alternative classes of covariance functions), and the different sensor outputs are assumed to be statistically independent. As previsously, the input uncertainties are taken into account in GPR+ME+IU, but not in GPR+ME.

Table I compares the performance metrics of the different methods. We can first note the very low relevance of the SLR method, which, by neglecting the various sources of uncertainty, misleads us (high value of e^2) and leads us to overconfidence in areas that in reality have little chance of containing the true value (too small values for $L_1^{0.95}$ and $L_2^{0.95}$, and therefore values of $\%_1^{0.95}$ and $\%_2^{0.95}$ well below 0.95). Considering a model error obviously improves the results: the values of $\%_1^{0.95}$ and $\%_2^{0.95}$ are now close to (or even above) 0.95 for all approaches, even for SLR. However, for SLR, e^2 remains very high and the length of confidence intervals is too high compared to the range of variation of interest (not informative). By adding flexibility and non-linearity to the input-output relationship (by increasing the polynomial order with GLR, then switching to Gaussian process regression), the results improve: the values of e^2 decrease (the estimations are on average better centred on the true value), as do the lengths of the confidence intervals (the estimations become progressively more informative and exploitable). A significant improvement on these two aspects (reduction of e^2, $L_1^{0.95}$ and $L_2^{0.95}$) is finally observed when integrating the uncertainties on the model inputs. Indeed, neglecting them in the calibration phase forces the model to carry them over into the model error, which is carried into the estimation phase. This is no longer the case when they are integrated in the calibration phase, and the model error being reduced, the estimations are likely to be better adjusted.

Method	$e^2(\%)$	$\%_1^{0.95}$	$\%_2^{0.95}$	$L_1^{0.95}$	$L_2^{0.95}$
SLR	3.13	0.09	0.21	0.22	1.74
SLR+ME	3.16	0.97	0.93	8.22	3
GLR+ME	2.32	0.97	0.87	4.53	2.99
GLR+ME+IU	2.00	0.97	0.96	3.3	2.17
GPR+ME	0.36	0.98	1	2.31	2.26
GPR+ME+IU	0.26	0.93	0.97	1.26	1.22

TABLE I
COMPARATIVE PERFORMANCE FOR DIFFERENT CALIBRATION METHODS, FOR $\boldsymbol{n = 200, d_y = 2}$. CALIBRATION METHODS ARE DESCRIBED IN SECTION III-B AND PERFORMANCE METRICS IN SECTION III-A.

Remark: If the response of sensors with respect to the environmental variables in which they are placed is nonlinear, the interest of modeling the relationship between $\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}$ by nonlinear models is obvious. In a small data context, the gains provided by these non-linear models are nevertheless not always guaranteed. Indeed, these nonlinear models are most often based on the identification of a larger number of parameters, and it is not always possible to estimate them correctly (possible overfitting) when the number of observations is too small.

IV. APPLICATION TO PH AND CHLORINE MONITORING IN DRINKING WATER

A. Description of the sensor

In this section, the methodology is applied to estimate both active chlorine (HClO) concentration and pH in drinking water from a dataset generated in laboratory settings by a sensor node designed for installation into drinking water pipes (see Figure 1). The head of the sensor integrates a 1cm² sensor chip with 20 chemistors (e.g. resistive chemical sensors) based

Fig. 1. Two sensor nodes installed into a drinking water pipe and connected via USB cable to control computers.

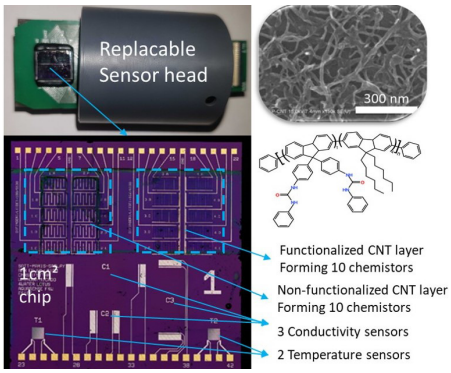


Fig. 2. The replaceable sensor head contains a chip with a 10×2 sensor array based on a random network of non-functionalized carbon-nanotubes (CNT) or CNT functionalized with the FF-UR polymer.

on random network of multi-walled carbon nanotubes (CNT) ink-jet printed on top of interdigitated electrodes. The CNT are either non-functionalized or polymer-functionalized (10 devices of each type, see Figure 2).

The focus of the present section is on the exploitation of the data generated by the CNT sensor array, so the fabrication and operation of the sensor node are detailed elsewhere [10]. Briefly, the multi-walled carbon nanotubes are dispersed in 1,2-dichlorobenzene. They are functionalized by mixing the CNT dispersion with a solution of the FF-UR polymer. CNT are batch-printed on the silicon chip using a Dimatix DMP inkjet printer. After annealing, to prevent CNT loss in water, the CNT devices are covered with a spin-coated PMMA layer which is then turned porous by non-solvent-induced phase separation process.

The resulting chip is then integrated into a sensor node as shown in Figure 1 and 2. For operation, each CNT device is activated sequentially for 1s at 5μA and the resulting voltage is measured. The voltage value used for further analyse is the averaged voltage for the last 100ms of the activation period.

In this mode of transduction, device resistance changes when certain chemicals enter close proximity to the CNT layer, which is fostered by the high surface area of CNT and their chemical affinity to these chemicals. The role of the polymer is to modify - for selected chemicals - the chemical affinity, the equilibrium distance and thus the resistance change.

B. Description of the dataset

A dataset from the 10×2 CNT sensor array was generated over four sets of lab experiments in drinking water at different pH, chlorine concentration and temperature. The chlorine and pH levels are controlled by successive additions of sodium hypochlorite (bleach), chlorhydric acid and sodium hydroxide to drinking water sampled from the tap. After each chemical addition, one waits until a steady state is reached by the chemistors (at least 15min) before proceeding to the next chemical addition. It is worth mentioning that this experimental plan results from a set of experimental constraints - total available time, stabilization duration after each chemical addition, rise of the pH after each bleach addition, range of chlorine and pH of interest - and thus is not optimized in terms of chlorine and pH estimation performance.

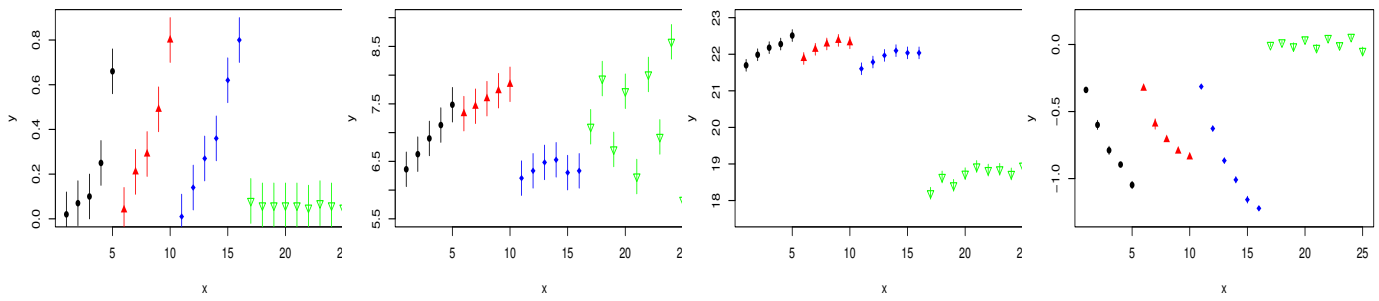
For each chemistor, the last resistance value of each step (considered most reliable) is extracted. The resistance values are normalized by subtracting the first resistance value of the dataset. Reference measurements are available for temperature, pH and free chlorine.

During the first three experiments, the free chlorine concentration is increased from 0 up to about 1mg/L over 5 to 6 points, at different pH ranging from 5.5 to 8.5. For the last experiment, only the pH varies significantly. While the temperatures are almost identical for the first three experiments, the temperature variation observed for the fourth experiment is large enough to have a non-negligible impact on the response of the sensors. These ranges of values for chlorine and pH were selected for their representativity in the field of drinking water. Indeed, drinking water network are usually operated at concentration of chlorine way below 1 mg/l, with target measurement accuracy required by the network operators below 0.05mg/L. As a consequence, from an applicative perspective, there is a strong interest in focusing on very low chlorine levels. Regarding pH, the typical values measured in drinking water networks range from 6 to 8..

It is important to mention that the sensors respond to active chlorine (hypochlorous acid HClO) and not to free chlorine, which is the sum of active chlorine and hypochlorite ions (ClO^-). The active chlorine concentration, noted c_{HClO} (in mg/L), can be calculated from the free chlorine mass concentration c_{Chl} (in mg/L) using the following formula:

$$c_{\mathrm{HClO}} = c_{\mathrm{Chl}} \left(1 + 0.98 \times 10^{\mathrm{pH}-\mathrm{pKa}(T)}\right)^{-1}, \qquad (11)$$

with T the temperature (in °C), pH the pH (unitless) and $\mathrm{pKa}(T) = 7.5 - 0.01 \times (T - 30)$. The standard deviation for pH is estimated at 0.15, for chlorine at 0.05mg/L and for temperature at 0.08°C. The uncertainty on HClO concentration is derived as the standard deviation of a set of HClO values calculated from 1000 sets of (chlorine, pH, temperature) values, themselves randomly generated using their known uncertainties. The standard deviations of the sensor outputs are calculated at each measurement point from an empirical estimate of the measurement noise related to the response fluctuations observed after stabilization. The average ratio between standard deviation and sensor response on all sensors is about 2%.



(a) Chlorine (mg/L) (b) pH (c) Temperature (°C) (d) One sensor response

Fig. 3. Graphical representation of the experimental dataset. The black points, the red upward triangles, the blue diamonds and the green downward triangles are respectively associated with experiments 1, 2, 3 and 4. For each measured value, the vertical lines correspond to ± 2 standard deviations.

The full dataset only contains 25 points, whose characteristics are shown in Figure 3. Recovering the notations of the former sections, we denote by $\boldsymbol{x}$ the active chlorine concentration and pH, $\boldsymbol{z}$ the temperature, and $\boldsymbol{y}$ the responses provided by the 20 CNT-based sensors. All measurement points are assumed to be statistically independent of each other.

1) Relevance of the dataset: The limited size of the dataset (due to the relatively long time needed to reach steady-state for each calibration step) clearly places the sensor calibration in a *small data* context, preventing the use of sophisticated machine learning approaches such as neural networks for the input-output relation. As is commonly practiced, the dataset was first exploited by a standard calibration approach - called simple linear regression (SLR) here - yielding good monitoring capability for HClO but no monitoring capability at all for pH. This was surprising as the sensors displayed comparable sensitivities (and differentiated between sensor types) to pH and HClO in deionized water. Further analysis of the calibration coefficients found by the SLR approach showed that in drinking water the sensitivity to pH was 5 to 10 times lower than to HClO. Henceforth, it turned out that the response to 1 pH unit variation by device was of the same range of magnitude than the response to 0.05mg/L variation in HClO, which was the uncertainty on HClO provided by the reference instrument. This strongly encouraged us to apply the formalism presented in Section II to finely integrate the different sources of uncertainty (with in particular the integration of the non-negligible measurement uncertainties on pH and HCLO, which is at the center of the proposed methodology).

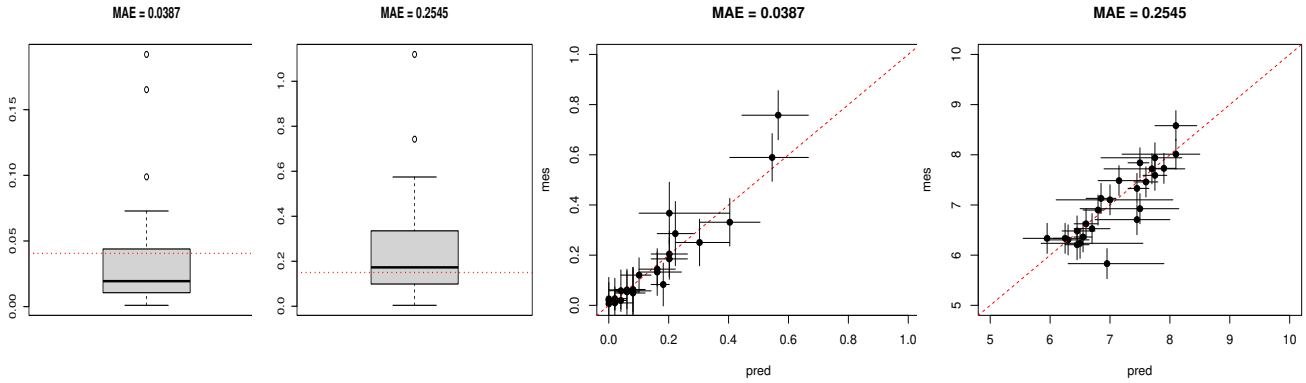
2) Performance analysis: In a first step, due to the very low number of observation points, the monitoring capabilities of the model are analyzed through a leave-one-out (LOO) approach, i.e. we estimate the value of each measurement from all other measurements except the one to be estimated. By applying the formalism presented in Section II, we then have 25 estimations of pH and active chlorine, in the form of 25 *a posteriori* PDFs. In the same way as for the analytical example, we can extract the most likely estimation, noted $\boldsymbol{x}_i^{\text{MAP}}$, as well as confidence intervals (with confidence level 95% by default). Most likely estimations and provided measurements can therefore be compared using the mean absolute error (MAE), such that for $1 \leq k \leq 2$:

Method	MAE_1(HClO)	MAE_2(pH)
SLR	0.056	1.36
SLR+ME	0.054	1.75
GLR+ME	0.064	0.872
GLR+ME+IU	0.068	1.057
GPR+ME	0.054	1.75
GPR+ME+IU	**0.039**	**0.254**

TABLE II
COMPARISON OF THE MAE VALUES OBTAINED FOR DIFFERENT CALIBRATION METHODS IN THE LOO APPROACH.

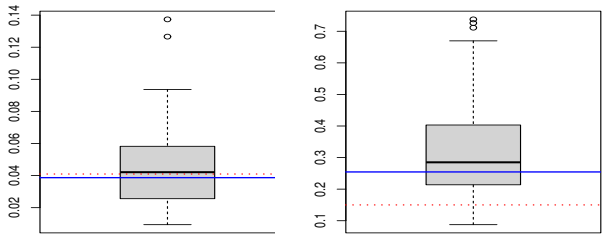
$$\text{MAE}_k = \frac{1}{25}\sum_{i=1}^{25} \Delta_{i,k}, \;\; \Delta_{i,k} = |(\boldsymbol{x}_i^{\text{MAP}})_k - (\boldsymbol{x}_i^{\text{mes}})_k|. \quad (12)$$

The obtained MAE values are summarized in Table II, for the six calibration methods that were described in Section III. We thus notice that only the GPR+ME+IU method, i.e. the one based on a Gaussian process regression and a correct integration of the input uncertainties, is able to monitor the pH in a satisfactory manner (this is to say with a reasonable MAE). Focusing on this method only, Figures 4-a,b then show, in the form of boxplots, the dispersion of the errors $\Delta_{i,k}$, and compare them to the standard deviation associated with the measurements (in red dotted line). It can then be noted on these figures that this particular method is specially efficient in monitoring HClO, the median error being even much lower than the average uncertainty of the measurements, and quite interesting for the pH, where the median error is very close to this average uncertainty. The consistency of the results in terms of uncertainties is then evaluated graphically in Figures 4-c,d, where the uncertainties on both the estimations and the measurements can be visualized and compared. In these graphs, the more the points are aligned with the first bisectrix, the lower the MAE. And an intersection of the vertical and horizontal lines with this bisectrix indicates an overlap of the 95% confidence intervals associated with the measurements and the estimations, which we expect to find most often if



(a) Chlorine (mg/L) (b) pH (c) Chlorine (mg/L) (d) pH

Fig. 4. Assessment of the monitoring capabilities of the proposed GPR+ME+IU method in the LOO approach. Subfigures (a) and (b) show the dispersion of the differences $\Delta_{i,k}$ between the most likely estimations and the provided measurements for the active chlorine and the pH. The red dotted lines are the mean standard deviations associated with the measurements. In subfigures (c) and (d), each point represents a couple $((x_i^{\text{MAP}})_k, (x_i^{\text{mes}})_k)$, when the horizontal and vertical lines correspond to the 95% confidence intervals for the estimations and measurements respectively.



(a) Chlorine (mg/L) (b) pH

Fig. 5. Dispersion of MAEs when trying to monitor $\mathbf{20\%}$ of the points randomly chosen using the $\mathbf{80\%}$ remaining points. The red dotted lines are the mean standard deviations associated with the measurements, and the blue solid lines give the MAEs obtained in the LOO case.

the model is correctly calibrated. These figures again show the very good monitoring capacity of the proposed model on these data: in coherence with the MAE obtained, the estimation uncertainties are of the same order of magnitude as the measurement uncertainties, and the only badly estimated data are the extreme points (the smallest pH and the highest concentration of HClO).

Finally, it is true that the leave-one-out approach is not truly representative of a sensor calibration approach, as in practice, calibration is learnt on initial experiments then applied on the following data. Considering the small size of the dataset and the strong differences between experiments (they do not repeat each other in terms of pH or temperature covered by the experiment), it was not possible to validate exactly this calibration process. Instead, we elected to randomly select 80% of the dataset for training (calibration) and the remainder for testing (estimation). A set of 100 different datasplits was tested. Very promising results were obtained in that second configuration compared to the LOO case, as it is shown in Figure 5. MAEs of 0.044 and 0.32 were observed respectively for HClO and pH, i.e. values only slightly higher than those obtained for the LOO (by respectively 15% and 26%). And the variability in these MAE values can easily be explained by the small size of the dataset: in some datasplits, part of the pH or HClO range of values available in the full dataset may not be covered by the calibration dataset.

V. Conclusions and Prospects

Deploying low-cost sensors in an open environment is a difficult task, both in terms of sensor calibration and exploitation of the results they provide. Indeed, the sensitivities observed in laboratories with respect to chemical quantities of interest can be degraded by other uncontrollable environmental variables, such as temperature or humidity, and nothing assures us that the sensors will not also react to other unidentified substances. In this context, it is essential to integrate a potential model error, and to seek to integrate the other sources of uncertainties as well as possible, at the level of the outputs of the model but also at the level of its inputs. This work thus proposes a Bayesian framework making it possible to meet these expectations, in particular in a small data context, i.e. for which the number of input-output pairs is relatively low for the sensors calibration. The relevance of this method is demonstrated not only on an analytical case, but also on an experimental dataset describing the monitoring of active chlorine and pH in drinking water using a CNT-based sensor array integrated into a low-cost sensor node that can be deployed in drinking water pipes. In the analytical case, Gaussian Process Regression with Input Uncertainties and Model Error leads to a reduction of the error metric (RMS error) by a factor of 12 compared to Simple linear regression (standard calibration), while in the experimental case, the reduction of the error (MAE) is by 40% on active chlorine and by a factor 5 on pH. One of the perspectives of this work is to use this calibration framework more extensively on this dataset (and on similar datasets generated by the other sensor nodes from the same

fabrication batch) as a tool to understand the root cause of sensor responses (for instance, role of the polymer in the sensitivity, role of the sensor multiplixing). Another aspect is to address the time response of the devices. Indeed, all of this work is nevertheless placed in a static framework, i.e. in which the estimation of chemical quantities at a time t is based solely on the responses of the sensors at this same time, and the sensor responses are considered to be in a steady state. Generalizing the proposed approach in a more dynamic framework, for which the estimation can also integrate measurements at previous times, is a direction of improvement that seems very valuable to reduce calibration duration and increase the size of calibration datasets.

REFERENCES

[1] David M. Broday and The Citi-Sense Project Collaborators. Wireless distributed environmental sensor networks for air pollution measurement - the promise and the current reality. *Sensors*, 17(10), 2017.

[2] Tao Chen and Bo Wang. Bayesian variable selection for gaussian process regression: Application to chemometric calibration of spectrometers. *Neurocomputing*, 73(13):2718–2726, 2010. Pattern Recognition in Bioinformatics Advances in Neural Control.

[3] José María Cordero, Rafael Borge, and Adolfo Narros. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sensors and Actuators B: Chemical*, 267:245–254, 2018.

[4] Houxin Cui, Ling Zhang, Wanxin Li, Ziyang Yuan, Mengxian Wu, Chunying Wang, Jingjin Ma, and Yi Li. A new calibration system for low-cost sensor network in air pollution monitoring. *Atmospheric Pollution Research*, 12(5):101049, 2021.

[5] Atefeh Daemi, Yousef Alipouri, and Biao Huang. Identification of robust gaussian process regression with noisy input using em algorithm. *Chemometrics and Intelligent Laboratory Systems*, 191:1–11, 2019.

[6] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, and G. Di Francia. Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches. *Sensors and Actuators B: Chemical*, 255:1191–1210, 2018.

[7] Florentin Delaine, Bérengère Lebental, and Hervé Rivano. ¡italic¿in situ¡/italic¿ calibration algorithms for environmental sensor networks: A review. *IEEE Sensors Journal*, 19(15):5968–5978, 2019.

[8] Directive 2008/50/EC of the European Parliament. The council of 21 may 2008 on ambient air quality, and cleaner air for europe, 2008.

[9] Leonardo Tomazeli Duarte, Christian Jutten, and SaÏd Moussaoui. A bayesian nonlinear source separation method for smart ion-selective electrode arrays. *IEEE Sensors Journal*, 9(12):1763–1771, 2009.

[10] Bérengère Lebental et al. Lotus project, deliverable d2.2 first version lotus prototype public summary, 2022.

[11] Zongyu Geng, Feng Yang, Xi Chen, and Nianqiang Wu. Gaussian process based modeling and experimental design for sensor calibration in drifting environments. *Sensors and Actuators B: Chemical*, 216:321–331, 2015.

[12] Agathe Girard and Roderick Murray-smith. Learning a gaussian process model with uncertain inputs, 2003.

[13] R. Jayaratne, X. Liu, P. Thai, M. Dunbabin, and L. Morawska. The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog. *Atmospheric Measurement Techniques*, 11(8):4883–4890, 2018.

[14] M. Kamionka, P. Breuil, and C. Pijolat. Calibration of a multivariate gas sensing device for atmospheric pollution measurement. *Sensors and Actuators B: Chemical*, 118(1):323–327, 2006. Eurosensors XIX.

[15] Shima Khatibisepehr, Biao Huang, and Swanand Khare. Design of inferential sensors in the process industry: A review of bayesian methods. *Journal of Process Control*, 23(10):1575–1596, 2013.

[16] P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter. The rise of low-cost sensing for managing air pollution in cities. *Environ International*, 75:199–205, 2015.

[17] Sungyeop Lee and Jangbom Chai. An enhanced prediction model for the on-line monitoring of the sensors using the gaussian process regression. *Journal of Mechanical Science and Technology*, 33(5):2249–2257, 2019.

[18] B. Liu, Q. Zhao, Y. Jin, J. Shen, and C. Li. Application of combined model of stepwise regression analysis and artificial neural network in data calibration of miniature air quality detector. *Scientific reports*, 11(1):3247, 2021.

[19] Balz Maag, Zimu Zhou, and Lothar Thiele. A survey on sensor calibration in air pollution monitoring deployments. *IEEE Internet of Things Journal*, 5(6):4857–4870, 2018.

[20] J. M . Marin and C. P. Robert. *Bayesian core*. Springer-Verlag, New York, 2007.

[21] 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, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[22] Javier G. Monroy, Achim J. Lilienthal, Jose-Luis Blanco, Javier Gonzalez-Jimenez, and Marco Trincavelli. Probabilistic gas quantification with mox sensors in open sampling systems—a gaussian process approach. *Sensors and Actuators B: Chemical*, 188:298–312, 2013.

[23] Corrado Di Natale, Fabrizio A.M. Davide, Arnaldo D'Amico, Wolfgang Göpel, and Udo Weimar. Sensor arrays calibration with enhanced neural networks. *Sensors and Actuators B: Chemical*, 19(1):654–657, 1994.

[24] 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.

[25] V. S. Pugachev. *Theory of random functions and its application to control problems*. Pergamon Press, 1967.

[26] Pengfei Qi, Ophir Vermesh, Mihai Grecu, Ali Javey, Qian Wang, Hongjie Dai, Shu Peng, and K. J. Cho. Toward large arrays of multiplex functionalized carbon nanotube sensors for highly sensitive and selective molecular detection. *Nano Letters*, 3(3):347–351, 2003.

[27] R. T. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo method*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2008.

[28] J. Sacks, W. Welch, T. Mitchell, and H. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–435, 1989.

[29] T. J. Santner, B.J. Williams, and W.I. Notz. *The design and analysis of computer experiments*. Springer, New York, 2003.

[30] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air. In *SENSORS, 2014 IEEE*, pages 21–24, 2014.

[31] Georgi Tancev and Federico Grasso Toro. Variational bayesian calibration of low-cost gas sensor systems in air quality monitoring. *Measurement: Sensors*, 19:100365, 2022.

[32] Julia Tsitron, Cortney R. Kreller, Praveen K. Sekhar, Rangachary Mukundan, Fernando H. Garzon, Eric L. Brosha, and Alexandre V. Morozov. Bayesian decoding of the ammonia response of a zirconia-based mixed-potential sensor in the presence of hydrocarbon interference. *Sensors and Actuators B: Chemical*, 192:283–293, 2014.

[33] Deepika Tyagi, Huide Wang, Weichun Huang, Lanping Hu, Yanfeng Tang, Zhinan Guo, Zhengbiao Ouyang, and Han Zhang. Recent advances in two-dimensional-material-based sensing technology toward health and environmental monitoring applications. *Nanoscale*, 12:3535–3559, 2020.

[34] Joseph Wang. Nanomaterial-based electrochemical biosensors. *Analyst*, 130:421–426, 2005.

[35] Peng Wei, Zhi Ning, Sheng Ye, Li Sun, Fenhuan Yang, Ka Chun Wong, Dane Westerdahl, and Peter K. K. Louie. Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring. *Sensors*, 18(2), 2018.

[36] Jie Yu. A bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers & Chemical Engineering*, 41:134–144, 2012.

[37] G. Zucchi, B. Lebental, L. Loisel, S. Ramachandran, A. Flores Gutierrez, X. Wang, M. Godumala, and L. Bodelot. Chemical sensors based on carbon nanotubes functionalised by conjugated polymers for analysis in aqueous medium, U.S. Patent Application No. 16/604,423.

APPENDIX

A. Definition of the analytical test case

To generate the simulated dataset, the sensor model is expressed as a non-linear relationship between sensor outputs $\boldsymbol{y}$, environmental variables $\boldsymbol{z}$, pollutant concentrations of interest $\boldsymbol{x}$, plus a quantity w that can be seen as another environmental variable or pollutant concentration that has an uncontrolled

influence on the sensor response. For $d_x = d_z = 2$, this model is written as:

$$\begin{aligned} y_j(t) &= \beta_j^1 + \beta_j^2 \log(x_1(t) + 0.25) \\ &\quad + \beta_j^3 (x_2(t)/2 + (2x_2(t) - 9)1_{x_2(t)>4.5}) \\ &\quad + \beta_j^4 z_1(t) + \beta_j^5 \exp(\beta_j^6 (z_2(t) - \beta_j^7)) + \beta_j^8 w(t). \end{aligned}$$

Here, $1_{x_2(t)>4.5}$ is equal to one if $x_2(t) > 4.5$ and to 0 otherwise, which allows to model a threshold law between y_j and x_2. For each $1 \leq j \leq d_y$ (values of d_y between 2 and 10 will be considered in the following), the coefficients $(\beta_j^1, \ldots, \beta_j^8)$ are chosen as independent realizations of the Gaussian vector:

$$\boldsymbol{\xi} \sim \mathcal{N}\left(\begin{matrix} (8, 4, -3.5, 0.1, 2, 0.1, 60, 0.02), \\ \text{diag}\begin{pmatrix} 4, 9, 9, 0.16, 4 \times 10^{-4}, \\ 10^{-4}, 4, 9 \times 10^{-6} \end{pmatrix} \end{matrix} \right), \qquad (13)$$

where for a vector $\boldsymbol{a}$, $\text{diag}(\boldsymbol{a})$ is the diagonal matrix constructed from the elements of $\boldsymbol{a}$. The numerical values in $\boldsymbol{\xi}$ have been carefully chosen to establish a hierarchy among the inputs in terms of sensitivity to the sensors. The different sensors are thus on average not very sensitive to w and z_1, behave almost the same way with respect to z_2, while presenting on average important sensitivities with respect to x_1 and x_2. The sensor outputs are all positively correlated with the input variables, except for x_2, where a negative correlation is imposed. This singular behavior should be accompanied by an easier estimation of x_2 than x_1. To illustrate on a water quality monitoring application, there would be in this case data from d_y sensors with the same operating law but difference sets of coefficients $(\beta_j^1, \ldots, \beta_j^8)$. They would target the parameters x_1 and x_2 (for instance pH and chlorine levels) while the parameters z_1 and z_2 (for instance temperature and salinity) would be provided by other sensing means. The parameter w would be an additional parameter of influence to the sensors, but not known either by external means or by the sensors (for instance the concentration in lead).

To generate the data, we draw 2000 values of $(\boldsymbol{x}, \boldsymbol{z}, w)$ uniformly and independently in $[0, 10]^2 \times [10, 20] \times [20, 90] \times [0, 10]$ (these being the ranges of variation of the target and influential parameters). These points are noted $(\boldsymbol{x}_i, \boldsymbol{z}_i, w_i)_{i=1}^{2000}$, and we denote by $(\boldsymbol{y}_i)_{i=1}^{2000}$ the associated values of the sensor outputs. Non-negligible noise is then added to the input and output observations to get closer to the experimental conditions:

$$\boldsymbol{x}_i^{\text{obs}} = \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i^x, \quad \boldsymbol{z}_i^{\text{obs}} = \boldsymbol{z}_i + \boldsymbol{\varepsilon}_i^z, \quad \boldsymbol{y}_i^{\text{obs}} = \boldsymbol{y}_i + \boldsymbol{\varepsilon}_i^y,$$

where $\boldsymbol{\varepsilon}_i^x$, $\boldsymbol{\varepsilon}_i^z$ and $\boldsymbol{\varepsilon}_i^y$ are independent realizations of three independent centered Gaussian random vectors, with respective covariance matrices $0.04 \times \boldsymbol{I}_2$, $\text{diag}(0.01, 0.5)$, and $0.04 \times \boldsymbol{I}_{d_y}$, where $\boldsymbol{I}_p$ denotes the $(p \times p)$-dimensional identity matrix.

Among these 2000 noisy observation points, n triplets noted $\{(\boldsymbol{x}_{i_k}^{\text{obs}}, \boldsymbol{z}_{i_k}^{\text{obs}}, \boldsymbol{y}_{i_k}^{\text{obs}}),\ 1 \leq k \leq n\}$ are randomly chosen to define the training set, and the other triplets define the test set. We recall that for the test set, only $\boldsymbol{z}_i^{\text{obs}}$ and $\boldsymbol{y}_i^{\text{obs}}$ are available for the identification of $\boldsymbol{x}_i$.

(d_y, n)	$e^2(\%)$	$L_1^{0.95}$	$L_2^{0.95}$
$(\mathbf{2}, 200)$	0.26	1.26	1.22
$(\mathbf{4}, 200)$	0.15	0.84	0.55
$(\mathbf{6}, 200)$	0.14	0.8	0.51
$(\mathbf{8}, 200)$	0.18	0.74	0.5
$(\mathbf{10}, 200)$	0.18	0.73	0.5
$(5, \mathbf{50})$	0.45	1.08	0.62
$(5, \mathbf{100})$	0.20	0.98	0.66
$(5, \mathbf{200})$	0.15	0.81	0.53
$(5, \mathbf{300})$	0.11	0.76	0.51
$(5, \mathbf{400})$	0.08	0.65	0.44
$(5, \mathbf{500})$	0.08	0.65	0.39

TABLE III
INFLUENCE OF THE VALUES OF $\boldsymbol{d_y}$ AND $\boldsymbol{n}$ ON THE ESTIMATIONS FOR THE GPR+ME+IU CALIBRATION METHOD.

B. Additional results on the analytical case

Using the same notations as in Section III and focusing only on the GPR+ME+IU method, it is possible from Table III to quantify the impact of an increase in the number of sensors, d_y, or an increase in the size of the training set, n. Unsurprisingly, by increasing n, the monitoring results improve, but there is a limit to improvements. This is due to the unknown environmental factor w and also the various uncertainties. If it was possible to observe w, or to reduce the experimental uncertainties, no doubt the results would be improved. The convergence of the results with respect to the increase of the number of sensors is more delicate to analyze: by adding observations, the estimation uncertainties are reduced (the values of $L_1^{0.95}$ and $L_2^{0.95}$ decrease), but this does not necessarily translate into a better centering of the estimations on the true value (the value of e^2 is not completely monotonous). Knowing the quantity w and reducing the uncertainties decrease the quantity of data needed to reach the best performances and improves the performances for a given quantity of data. The performances also reach an optimum with the number of sensors according to the e^2 metrics. This optimum number of sensors increases with the quantity of available calibration observations. The interpretation is that the uncertain information of the too many sensors becomes contradictory and the limited number of data (compared to the number of sensors) limits the capability to identify the appropriate laws for the random variables; this transfers into the rise of the model error.

G. Perrin graduated from Ecole Polytechnique and Ecole Nationale des Ponts et Chaussées in 2010. He received the Ph.D. degree in mechanics from University Paris-Est in 2013, and the habilitation to direct research in mathematics from the University Paris-Sud in 2019. From 2013 to 2020, he was a research scientist at CEA (French atomic agency), and since 2020, he has been a research director at Gustave Eiffel University, in the IMSE laboratory of the COSYS department. His research interests include statistical learning, inverse problem solving in statistics and verification and validation approaches for the design and guarantee of systems under uncertainty. His work has been valorized in the framework of several research projects mainly concerning the validation of AI-based systems, the proposal of innovative methods for monitoring, maintaining and optimizing the performance of dynamic mechanical systems.

B. Lebental graduated from Ecole Polytechnique's Engineering Program in 2006, and from Ecole Polytechnique's Physics and Nanotechnology Master of Sciences Programs in 2007. She obtained her Ph.D. in Civil Engineering from Université Paris-Est, France, in 2010, on the topic of Carbon nanotubes sensors for monitoring construction materials. Since 2010, she is researcher at Université Gustave Eiffel. Her research focuses on nano-enabled mechanical and chemical sensors for smart cities. Her research ranges from nanosensor design and fabrication to field deployment and data exploitation. She has coordinated several collaborative research projects on sensors for smart cities, such as Sense-City (http://sense-city.ifsttar.fr/) or Proteus (http://www.proteus-sensor.eu/). She is cofounder of the startup Altaroad (https://www.altaroad.com/en/) which provides treacability solutions for materials in the construction industry.