Enhancing Multi-Hop Sensor Calibration with Uncertainty Estimates

Balz Maag
ETH Zurich
Zurich, Switzerland, CH-8092
Email: balz.maag@tik.ee.ethz.ch

Zimu Zhou
ETH Zurich
Zurich, Switzerland, CH-8092
Email: zzhou@tik.ee.ethz.ch

Lothar Thiele
ETH Zurich
Zurich, Switzerland, CH-8092
Email: thiele@ethz.ch

***Abstract*—Low-cost sensors, installed on mobile vehicles, provide a cost-effective way for fine-grained urban air pollution monitoring. However, frequent calibration is crucial for low-cost sensors to consistently deliver accurate measurements. Multi-hop calibration is a common practice to calibrate mobile sensor deployments, but is prone to severe error accumulation over hops. Prior research mitigates error accumulation by designing special calibration models, which only apply to linear models. In this paper, we propose an orthogonal approach by selecting reliable measurements for calibration at each hop. We analyze the impact of different data-induced uncertainties on calibration errors and devise a scheme to estimate these uncertainties of the calibrated outputs. We further propose an uncertainty-based metric for data filtering at each hop. We evaluate the effectiveness of our method in a real-world ozone sensor deployment. Experimental results show that our method works with both linear and non-linear calibration models and reduces calibration errors in multi-hop setups by up to 25% compared with existing techniques.**

1. Introduction

Advances in sensor technology have made it possible to monitor urban air pollution at high spatio-temporal resolution with low cost. Various small, cheap and low-power sensors have been successfully installed on vehicles [1]–[4] and measure concentrations of major air pollutants citywide. These measurements provide valuable air pollution information for quantitative studies and public services.

To ensure data consistency of these measurements, the sensors need periodical *calibration* after deployment. This is because the accuracy of low-cost sensors tend to degrade over time. Some studies report significant sensor drifts even after only one month of deployment [5]. By calibrating a low-cost sensor its measurements are transformed via a calibration model such that the calibrated measurements closely agree with measurements from a highly accurate reference [6]. Sensor calibration in a large-scale air pollution monitoring deployment is often difficult since typically there are only a few sparsely distributed references within a city.

Multi-hop calibration is an effective method to calibrate low-cost sensors in a mobile deployment with limited references [1]–[4], [7]. The idea is to exploit rendezvous between sensors, *i.e.*, situations when two or more sensors meet in time and space. During rendezvous, sensors are exposed to the same environment and sense the same phenomena, thus, creating a calibration opportunity. Whenever a low-cost sensor is in rendezvous with a reference sensor it can use the reference measurement for calibration. To further increase calibration opportunities, multi-hop calibration also exploits freshly calibrated sensors. That is, a freshly calibrated sensor is providing its calibrated measurements as a virtual reference to an uncalibrated sensor. This process can be repeated until all sensors are calibrated, forming a calibration graph over multiple hops. An example is illustrated in Fig. 1.

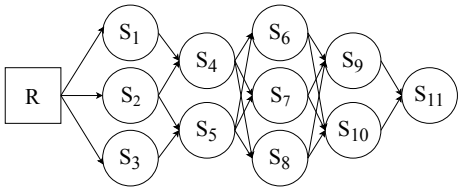


Figure 1. Example of a calibration graph consisting of one reference and 11 low-cost sensors. A reference sensor R is calibrating three low cost sensors $S_{\{1,2,3\}}$ in a first hop. These freshly calibrated sensors then provide their calibrated measurements as references to calibrate two additional sensors $S_{\{4,5\}}$. This procedure is continued until all 11 sensors are calibrated.

While multi-hop calibration allows to calibrate substantially more sensors in a mobile deployment, it usually encounters severe error accumulation over multiple hops [1], [3]. For multi-hop calibration to function, measurements of the virtual reference need to be as accurate as possible at every hop in order not to affect later calibration procedures. However, both the *data* and the *model* for calibration are not perfectly accurate in practice. Hence, calibration at each hop is never error-free. Consequently, the error of multi-hop calibration tends to accumulate over hops if not explicitly controlled. Prior approaches [1], [3] mitigate error accumulation by designing new calibration models. They only apply to linear calibration models whereas non-linear models are necessary to calibrate many other important air pollutants such as particulate matters [2], [4], [8]. Furthermore, previous studies take all measurements as input for calibration without differentiating their quality [9].

In this paper, we propose to reduce error accumulation in multi-hop calibration by considering data *uncertainties*.

Rather than designing error-resilient calibration models, we quantify the data-induced uncertainties and conduct explicit data filtering at each hop such that only reliable data are utilized for later calibration. To enable data filtering in multi-hop calibration, two problems need to be solved. *(i)* What uncertainties are introduced by sensor data into the calibration procedure and how to estimate them during data processing in the calibration pipeline? *(ii)* How to define a unified metric to quantify the uncertainties for data filtering at each hop? We address these problems and make the following contributions.

- We design a scheme to estimate two types of uncertainties: *epistemic* and *aleatoric*, which contribute to data-induced errors in sensor calibration. Our approach is agnostic to the underlying calibration models (linear or non-linear) and can be plugged into the standard calibration procedure. To the best of our knowledge, this is the first work to augment sensor calibration with uncertainty estimates.
- We comprehensively analyze the impact of different uncertainties on potential calibration errors and propose an uncertainty-based metric for data filtering in multi-hop calibration.
- We evaluate our method in real-world ozone (O3) sensor deployments. Our results show that we are able to reduce the calibration error in a multi-hop setup by up to 25% when using uncertainty-based data filtering compared to traditional calibration.

In the rest of this paper, we review related work in Sec. 2, explain calibration uncertainties in Sec. 3, introduce methods to estimate them in Sec. 4 and propose a data filtering metric in Sec. 5. We conduct simulations in Sec. 6 and real-world evaluations in Sec. 7 and conclude in Sec. 8.

2. Related Work

Our work is related to research on multi-hop calibration and measuring uncertainty.

2.1. Multi-Hop Calibration

Sensor calibration aims to improve the data quality of low-cost air pollution sensor deployments. We focus on multi-hop calibration, a popular approach to ensure data quality of mobile air pollution sensor deployments [1]–[4], [7]. We refer interested readers to [6] for a comprehensive review on air pollution sensor calibration.

Multi-hop calibration exploits rendezvous [1], [7] to recursively calibrate sensors using freshly calibrated sensors. However, multi-hop calibration suffers from error accumulation over multiple hops [1], [10]. Saukh *et al.* [1] prove that geometric mean regression can be used for offset and gain calibration without error accumulation. Maag *et al.* [3] formulate a constrained regression method to mitigate error accumulation for multi-hop sensor array calibration. Alternatively, Fu *et al.* [2] studied the impact of reference locations on the performance of multi-hop calibration and design a reference placement strategy.

Our work also aims to mitigate error accumulation in multi-hop calibration. Instead of redesigning calibration models [1], [3] or redistributing references [2], we control errors by allowing only reliable sensor measurements to propagate over hops. Our method applies to both linear [1], [3] and non-linear calibration models [4].

2.2. Measuring Uncertainty

Quantifying the expected performance of statistical models, for instance by estimating major uncertainty sources, is an important problem. Various works [11]–[13] differentiate between different uncertainties, the two most important ones being *epistemic* and *aleatoric*.

Epistemic uncertainty characterizes the knowledge of a model, *e.g.*, based on the data it has been trained on or if it features adequate modeling capabilities. Missing knowledge may lead to ignorant predictions. The most prominent method to estimate epistemic uncertainty is ensemble learning, where multiple models are trained to learn the same underlying function but on a different view of the dataset. These different views are often created by using bootstrapping [14] or bagging [15] and have been applied in general modeling frameworks [16], neural networks [17] or Bayesian learning [12].

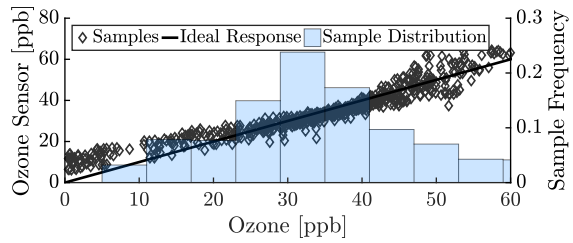
Aleatoric uncertainty, which captures the noise inherent in the measurements, can be modeled by auxiliary models [18] that learn the relationship between the input of a model and its corresponding expected error. Recent methods, especially tailored for complex machine learning methods, model the output as a probabilistic predictive distribution. Examples of such methods are Bayesian neural networks [19] or deep ensemble learners [20], [21].

Our work is inspired by all these works. In particular, we combine the concepts presented in [14], [17], [18] into one concise model and apply it in the context of sensor calibration, which is presented in detail in Sec. 4.

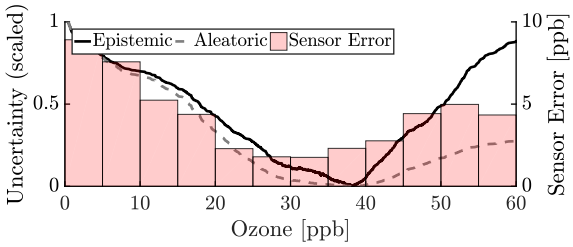
3. Uncertainties in Sensor Calibration

In this section we explain two types of uncertainties in the context of air pollution sensor calibration.

The general goal of sensor calibration is to find a calibration function cal (*i.e.*, calibration model) that transforms some raw sensor measurements X into a calibrated form $\hat{y} = cal(X)$. An optimal calibration model minimizes some norm between the calibrated measurements $\hat{y}$ and some reference or ground-truth measurements y. We use a real-world example of a low-cost ozone (O_3), MICS-OZ-47 [22], sensor which has been deployed next to a highly accurate governmental monitoring station. Ozone is an important atmospheric pollutant that contributes to respiratory symptoms when people engage in outdoor exercises and activities [23]. We take the state-of-the-art method [24] to calibrate the ozone sensor. Specifically, the ozone sensor is augmented with a temperature and humidity sensor (*i.e.*, a *sensor array*) and multiple least squares is applied to find a function cal that calibrates m measurements $X \in \mathbb{R}^{m \times 3}$ of the three sensors to the ozone reference $y \in \mathbb{R}^{m \times 1}$. The three sensors



(a)



(b)

Figure 2. Calibration results of an ozone sensor. Fig. 2a shows the calibrated measurements versus the actual ozone concentrations. These samples are also used to train the calibration model. The distribution across the ozone range is highlighted by the blue area. Fig. 2b shows the behaviour of the epistemic, see (2), and aleatoric uncertainty, see (3), as well as the sensor measurement error, *i.e.*, $|y - \hat{y}|$.

were placed inside a ventilated box and deployed next to the governmentally maintained ozone sensor in a suburban area in Switzerland [25] during 2 weeks in May, 2014.

We are interested in the distributions of the calibration errors across the sensing range (the bars in Fig. 2b). We observe for a calibrated sensor, the calibration errors vary at different sensor readings. The calibration error increases where the measurements to train the calibration model are limited or have high variation. For example, at low ozone concentrations below 10 ppb the training measurements are scarce. As a result the calibration model is overestimating the ozone concentration. Similarly, at concentrations above 45 ppb the measurements exhibit higher variance than at lower concentrations and, hence, the sensor error is also growing. We aim to explain the nonuniform calibration errors across the sensing range of a sensor using *uncertainties*.

Epistemic Uncertainty. This uncertainty captures the general ignorance of our calibration model and is typically caused by lack of knowledge, for instance by missing data or insufficient modeling power [11], [12]. It is in general difficult, if not impossible, to generate confident calibrated measurements in ranges where the calibration model is trained on little or no data at all. This is in particular true for complex models, such as non-linear methods, when they have to excessively extra- or interpolate. The same holds when a model, which lacks appropriate capabilities to capture a complex calibration function, is applied, *e.g.*, when a linear model is used to learn a non-linear function. However, epistemic uncertainty can typically be reduced by collecting and adding more data into the calibration process or applying more appropriate models with adequate capabilities for the problem. Epistemic uncertainty can point out potentially inaccurate sensor readings. In Fig. 2a, we can observe that most of the samples for training the calibration model are distributed between [20,40] ppb. Outside this range fewer samples have been used to train the calibration model. Consequently from 20 ppb to 0 ppb and above 40 ppb, both the calibration error and the epistemic uncertainty begin to grow, as shown in Fig. 2b. Note that above 40 ppb the epistemic uncertainty is equally high as for measurements below 20 ppb, however the sensor error is notably smaller. This additional error at low concentrations is due to a constant overestimation, which may be caused by missing modeling power. Note that there is in general no direct causation between sensor error and epistemic uncertainty. It is for instance still possible for a calibration model to learn an accurate calibration function even with lower amounts of samples. We defer the method to estimate the epistemic uncertainty to Sec. 4.

Aleatoric Uncertainty. This uncertainty captures the *natural noise*, *i.e.*, measurement error, in low-cost sensor measurements [11], [12], [18]. In Fig. 2a we can observe the scatter of the measurements is larger for higher concentrations. Especially between [45,60] ppb we can observe various outliers, which are at least 10 ppb from the ideal response. Potential reasons for this effect may be that the linear model is not capable to capture the underlying calibration function of the sensor or the uncalibrated measurements exhibit a generally higher noise in this high concentration region. As indicated in Fig. 2b, the aleatoric uncertainty increases in regions where the deviation of the measurements is increasing as well. Note that high variance of measurements in the training data does not necessarily imply high measurement error. A calibration model may still be able to perfectly capture the ideal underlying function between sensor and reference measurements, *e.g.*, if the noise is following a Gaussian distribution. Further, we can also observe an increased aleatoric uncertainty at lower concentrations. As we have discussed before, the calibration model is not able to perfectly capture the underlying function due to either missing data or modeling power and, thus, the aleatoric uncertainty also captures these high inaccuracies below 20 ppb. In Sec. 4 we lay out a method to learn the relationship between the sensor measurements and the aleatoric uncertainty.

Summary. Not all outputs of a calibrated sensor have the same accuracy due to the distributions of the raw sensor measurements to train the calibration model. This motivates us to associate auxiliary "uncertainties" to each output of a calibrated sensor to indicate the potential errors. We identify two types of uncertainties: *epistemic* and *aleatoric*. The former characterizes the uncertainty in output ranges where the measurements for training the calibration model are sparse or the model is lacking sufficient modeling capabilities, while the latter represents the uncertainty in output ranges where the calibrated measurements exhibit large inaccuracies. These two types of uncertainties provide extra information about the reliability of the outputs of a

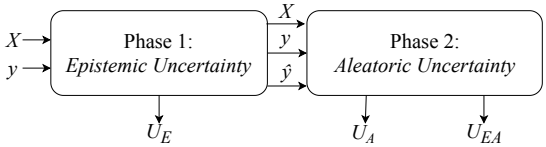


Figure 3. General overview of our model, which is divided into two separate phases. In phase 1 sensor calibration is performed and epistemic uncertainty U_E, see (2), information is retrieved. Phase 2 predicts the aleatoric uncertainty U_A, see (3), as well as the corresponding epistemic uncertainty U_{EA}, see (4), of this process.

calibrated sensor at different sensing ranges, which helps to filter unreliable outputs for further use (multi-hop calibration in our case). In the next two sections, we design methods to estimate the two types of uncertainties, which are shown in Fig. 2b, and integrate them into multi-hop sensor calibration for better accuracy.

4. Estimating Uncertainties

This section describes our scheme that can be plugged into any calibration model to output calibrated sensor measurements and the corresponding uncertainties.

4.1. General Overview

Given a dataset (X,y), where $X \in \mathbb{R}^{m\times k}$ are the uncalibrated sensor measurements that consist of k input feature vectors, *i.e.*, a sensor array consisting of k sensors such as the one in Sec. 3, and $y \in \mathbb{R}^{m\times 1}$ are reference sensor measurements, our scheme works in two phases.

Phase 1. An ensemble of calibration models is trained to output, for an uncalibrated measurement $x_i \in X \in \mathbb{R}^{1\times k}$, a calibrated measurement $\hat{y_i} = cal(x_i) \in \mathbb{R}^{1\times 1}$ and a corresponding epistemic uncertainty value $U_E(\hat{y_i}) \in \mathbb{R}^{1\times 1}$. Any calibration model that performs a regression task can be applied to find cal. We use bootstrapping [14] to train our ensemble, *i.e.*, we artificially generate different training datasets to create diverse models in our ensemble.

Phase 2. A neural network ensemble is trained to learn the relationship between uncalibrated sensor measurements X and calibration errors $\varepsilon = (\hat{y} - y)^2 \in \mathbb{R}^{m\times 1}$ to quantify the aleatoric uncertainty $U_A(\hat{y}) \in \mathbb{R}^{1\times 1}$, also known as local error bars [18], for each calibrated measurement $\hat{y_i} \in \hat{y}$ of the sensor. This approach helps us to quantify the variance of our measurements around the ground-truth, *i.e.*, the calibration error as a function of the inputs $f(X) = \varepsilon$. Since the estimation of the aleatoric uncertainty is also subject to some potential error sources, we also estimate the epistemic uncertainty $U_{EA}(\hat{y_i}) \in \mathbb{R}^{1\times 1}$ of the aleatoric uncertainty estimation.

Fig. 3 illustrates the overall procedure of our scheme. We explain the two phases in detail below.

4.2. Estimating Epistemic Uncertainty

In the first phase we train multiple calibration models to output the calibrated sensor measurements, which then are used to generate epistemic uncertainty. To achieve this goal, we apply an ensemble of sensor calibration models via bootstrapping [14]. In particular, given the dataset (X,y), p different calibration models are trained. This is done via the following process:

1) Generate p bootstrapped datasets, each consisting of $(\overline{X}, \overline{y})$. This is achieved by the standard bootstrapping methodology, where each pair of bootstrapped samples $(\overline{X_i}, \overline{y_i})$ is randomly sampled with replacement from the input datasets (X,y). As a result, p different datasets with the same cardinality as the input dataset are artificially generated. According to the .632 bootstrapping rule [14], approximately 63.2% unique samples form one newly bootstrapped input dataset.
2) Train p calibration models using the p generated datasets. A typical calibration model minimizes the mean squared error $\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i^{(j)} - \overline{y}_i)^2$, where $\hat{y}^{(j)}$ is the output the j-th ($j \in \{\underline{1}, 2..., p\}$) model of the ensemble for a given $x_i \in \overline{X}$.

As a result, we get p different model outputs. Therefore, for a given $x_i \in X$ the final output is the mean over all p ensemble model outputs, *i.e.*,

$$\hat{y}_i = \frac{1}{p}\sum_{j=1}^{p}\hat{y}_i^{(j)} \tag{1}$$

Finally, the epistemic uncertainty is calculated by the standard deviation over all the p model outputs, *i.e.*,

$$U_E(\hat{y}_i) = \left(\frac{1}{p}\sum_{j=1}^{p}(\hat{y}_i^{(j)} - \hat{y}_i)^2\right)^{\frac{1}{2}}. \tag{2}$$

Since we use the mean over all outputs as final calibrated measurement, the epistemic uncertainty $U_E(\hat{y}_i)$ gives a notion of how much confidence we can have in this mean $\hat{y}_i$. The higher the disagreement of the p calibration models, the higher the epistemic uncertainty U_E, the less confidence we have in our calibrated measurement. If we add more knowledge, *e.g.*, in the form of data, to the calibration model we might be able to reduce this disagreement and consequently the epistemic uncertainty, see also Sec. 3.

4.3. Estimating Aleatoric Uncertainty

In the second phase the goal is the estimation of the aleatoric uncertainty inherent in the output of our ensemble of calibration models in phase 1. This is achieved by a second ensemble of models, which individually learn the relationship between the input vectors X and the calibration error $(\hat{y}_i - y_i)^2$. The calibration error measures the distance of our calibrated measurements to the ground-truth values and, therefore, the aleatoric uncertainty, see Sec. 3. Similarly to phase 1, this task could be performed by any type of regression technique. However in order to facilitate powerful modeling capabilities of non-linear functions we solely use non-linear neural networks. The following procedure performs the estimation of the aleatoric uncertainty.

1) Generate q new bootstrapped datasets $(\overline{X}, \overline{y})$, which are not identical to the ones as in phase 1.

2) For each dataset and each sample $x_i \in \overline{X}$ calculate the squared calibration error $\varepsilon_i = (\hat{y}_i - y_i)^2$, where $\hat{y}_i$ is the calibrated measurement from (1).
3) Train q models that approximate the function f with $f(X) = \varepsilon$.

In order to reduce the training efforts, we apply an optimized neural network ensemble structure [17]. Instead of q individual neural networks, one neural network with q outputs and a shared hidden structure is used. During each training run, the outputs are individually trained using the bootstrapped datasets and the hidden structure is updated in every training step for every output. The optimization functions for each output are also applying a L2-regularization [26] to assure smooth uncertainty estimations and to avoid over-fitting. Finally, to assure positive outputs (due to $\varepsilon \geq 0$), each output uses the soft-plus activation function, *i.e.*, $log(1 + e^x)$.

Similar to the ensemble in phase 1, the final output, in this case the estimated aleatoric uncertainty U_A for a given $x_i \in X$, is the mean over all q ensemble model outputs, *i.e.*,

$$U_A(\hat{y}_i) = \hat{\varepsilon}_i = \frac{1}{q}\sum_{j=1}^{q} \hat{\varepsilon}_i^{(j)}, \tag{3}$$

where $\hat{\varepsilon}_i^{(j)}$ is the j-th ($j \in \{1, 2, ..., q\}$) output of the neural network.

Finally, we also estimate the epistemic uncertainty of the aleatoric uncertainty estimation, given by,

$$U_{EA}(\hat{y}_i) = \left(\frac{1}{q}\sum_{j=1}^{q}(\hat{\varepsilon}_i^{(j)} - \hat{\varepsilon}_i)^2\right)^{\frac{1}{2}}. \tag{4}$$

This process allows us to model the calibration error, or the variance of our calibrated measurements, as a function of the input and consequently of a single calibrated measurement. Similarly to U_E, which we calculate in phase 1, the epistemic uncertainty U_{EA} serves as a measure of confidence in our aleatoric uncertainty estimation.

5. Integrating Uncertainties In Calibration

In this section, we first interpret different situations we may experience when applying our scheme and then propose a data filtering metric for sensor calibration.

5.1. Interpretation: Epistemic versus Aleatoric

Before explaining different situations, we first discuss the relationship between the two epistemic uncertainties U_E and U_{EA}. Both U_E and U_{EA} capture the epistemic behavior of the two ensembles we use. However, U_E captures the uncertainty we have in our calibration model and U_{EA} in our calibration error estimation. Basically, both metrics decrease as the number of samples, *i.e.*, the knowledge we feed into our model, increases, because the ensembles converge to a common output. For simplicity and due to the same basic interpretation we treat U_E and U_{EA} as similar in the following description of the different situations but differentiate them during data filtering (Sec. 5.2).

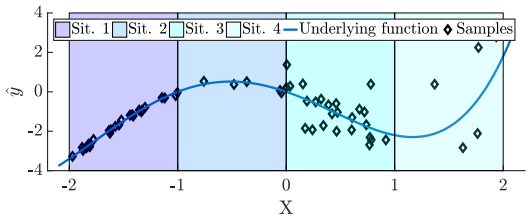


Figure 4. Four different uncertainty situations regarding epistemic U_E/U_{EA} versus aletoric U_A. From left ro right, situation 1) low—low, 2) high—low, 3) low—high, 4) high—high.

Situation 1: Low U_E/U_{EA}—Low U_A. In a situation where we have low epistemic and low aleatoric uncertainty at the same time we have high confidence in our calibrated measurement. This ideal case is shown in situation 1 in Fig. 4. The low epistemic uncertainty U_E and U_{EA} suggests that all the individual models in the two ensembles agree on their outputs and, thus, have been trained with a sufficient amount of data and appropriate modeling power. A low aleatoric uncertainty U_A at the same time, also points out that the calibration error of the measurements during training is low.

Situation 2: High U_E/U_{EA}—Low U_A. The next situation, see also situation 2 in Fig. 4, appears when we have only a few samples, therefore higher epistemic uncertainty, but these are not affected by large noise, thus low U_A. Although the low aleatoric uncertainty points out that the calibration model is able to approximate an accurate calibration function, there is only little proof. In order to gain more confidence the calibration model should be trained with more samples and reevaluated, especially if the model has to extra- or interpolate large areas of the input space given by X.

Situation 3: Low U_E/U_{EA}—High U_A. In the third situation the calibrated measurements show a large variation and, thus, also lead to a large calibration error. This situation may appear if the uncalibrated input measurements are affected by noise or the calibration model is not able to find an optimal calibration function, for instance due to missing features or insufficient complexity of the model.

Situation 4: High U_E/U_{EA}—High U_A. This is the most undesirable situation (situation 4 in Fig. 4). It occurs if the calibration model has been trained on only a few measurements and the corresponding outputs are far from the true measurements. In this situation we have no confidence in our calibration model due to missing knowledge and a potentially poorly calibrated model due to the observation of high calibration errors during training.

Summary. All situations except the ideal case with overall low uncertainty point out potential problems with the calibration model. In particular we face the potential of inaccurate calibrated measurements. In order to further process the calibrated measurements for calibration in subsequent hops we need to pinpoint confident measurements by exploiting our uncertainty estimates. We show how to integrate these uncertainties into multi-hop sensor calibration below.

5.2. Uncertainty based Data Filtering

Given a single uncalibrated sensor, which has been in rendezvous with one or multiple freshly calibrated sensors, we can collect a dataset $(X, \hat{y}, U_E(\hat{y}), U_{EA}(\hat{y}), U_A(\hat{y}))$ with uncalibrated sensor measurements X, virtual references $\hat{y}$ and the corresponding uncertainties $U_E(\hat{y}), U_{EA}(\hat{y})$ and $U_A(\hat{y})$ for each sample. Because the calibrated measurements may not be identical with the ground-truth measurements, we need to make sure that we remove potentially inaccurate samples before we perform the calibration for the uncalibrated sensor. In order to do this, we apply a simple heuristic to filter samples with general high uncertainty. In particular we treat a calibrated sample $\hat{y}_i$ as a confident sample if the following rule applies:

$$(U_E(\hat{y}_i) \leq p_E) \wedge (U_A(\hat{y}_i) \leq p_A) \wedge (U_{EA}(\hat{y}_i) \leq p_{EA}), \quad (5)$$

where $p_{\{E,A,EA\}}$ are thresholds. We set these thresholds by percentiles over all the available values for each uncertainty metric of the dataset, *e.g.*, p_E is set to the 90-th percentile of all $U_E(\hat{y})$ in the dataset. We will investigate the effect of this threshold in Sec. 7. As we have seen in Sec. 5.1, any situation with a high uncertainty metric is highlighting a potentially inaccurate calibrated measurement. The heuristic given in (5) is therefore removing any measurement that exhibits at least one of the three uncertainties with a too large value according to the set threshold.

6. Simulations

In this section, we conduct simulations to validate the effectiveness of our method to estimate the uncertainties.

6.1. Setup

We evaluate our scheme using artificial data samples $(x, y(x))$, with $x \sim U(0, 2\pi)$ and $y(x) \sim \mu(x) + N(0, \sigma(x))$, where

$$\mu(x) = \sin\left(5 \cdot \frac{x}{2}\right) \cdot \sin\left(3 \cdot \frac{x}{2}\right)$$
$$\sigma(x) = \frac{3}{20} + \frac{1}{4} \cdot \sin(4 \cdot x) + 2 \cdot \cos\left(\frac{6}{5} \cdot x\right)$$

The dataset is sampled 2000 times and samples, where $x \in \left[\frac{\pi}{2}, \frac{3 \cdot \pi}{2}\right]$, are used to train our models in phase 1 of the scheme. Because the underlying function $\mu(x)$ is non-linear, we use an ensemble of non-linear neural networks to approximate the function. The neural network uses the same optimized structure as the one to predict the aleatoric uncertainty, *i.e.*, one shared hidden structure and p outputs. Specifically, we apply for each of the two networks one shared hidden layer with 16 neurons with *tanh*-activation functions and $p = q = 20$ ensemble outputs.

6.2. Result

Fig. 5 shows the simulation results. First of all, in Fig. 5a we observe that within the training area, the approximate function $\hat{y}(x)$ fits the true function $\mu(x)$. Outside the training area the approximation is not fitting at all. This result is not surprising because the model has to extrapolate. Accordingly, the epistemic uncertainties U_E and U_{EA} (see Fig. 5b) capture this behavior. Within the training area the uncertainty is small and grows outside the training area. Further, we also observe a difference between the two uncertainties U_E and U_{EA}. As mentioned in Sec. 4, both capture the same behavior use however different underlying models in the respective ensembles and, consequently, output different epistemic uncertainty values.

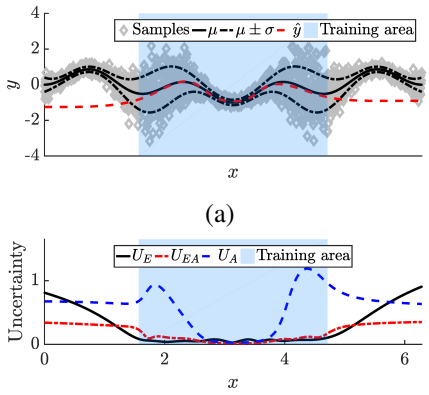


Figure 5. Performance on artifical data. Fig. 5a shows that our model is able to perform typical regression tasks. In Fig. 5b we see the typical behavior of the epistemic uncertainty $U_{\{E,EA\}}$, which starts to grow outside the training area, and that we are able to estimate the the noise variance in form of the aleatoric uncertainty U_A.

Secondly, the samples used for training follow a input dependent normal distribution. Especially at the borders of the training area the variation is significantly higher than in the center. The aleatoric uncertainty U_A is highlighting this effect, see Fig. 5b. It follows in fact the variance of the noise distribution $\sigma(x)$. This result is expected, because the optimization function of the aleatoric estimator is set to exactly estimate this variance, see (3).

Overall, we can conclude that our scheme is able to *i)* perform typical regressions tasks while also estimating *ii)* epistemic (U_E and U_{EA}) and *iii)* aleatoric (U_A) uncertainty.

7. Evaluations

In this section, we apply our scheme and evaluate its performance on real-world sensor data.

7.1. Setup

Dataset. The dataset consists of measurements from 11 different low-cost metaloxide ozone sensor prototypes. Each sensor is also paired with a temperature sensor, which is used to tackle the gas sensors dependencies on environmental conditions. The sensors are placed next to the same governmental ozone sensor as the one described in Sec. 3 that serves as reference. We placed all sensors in the same ventilated box and collected with a sampling interval of 10 min approximately 4000 samples of each sensor during the month of October, 2014.

Test setup. We artificially build a multi-hop calibration setup where a subset of the sensors is calibrated by the reference sensor, which in return are used to calibrate another subset of sensors, and so on until all sensors are calibrated. We use five different calibration graph structures, for example the one in Fig. 1, with varying number of parent nodes, *i.e.*, references, (1 to 4) and number of hops (1 to 5). Each pair of sensor nodes collects 200 samples, *i.e.*, we assume they are 200 times in rendezvous. To simulate a setup with diverse sensors, in particular diverse uncertainties, we randomly choose different measurement distributions, for instance see Fig. 2a, for each sensor by varying their overall range of collected measurements during rendezvous. Each sensor covers between $[25, 60]\%$ of the total ozone range during the measurement period. The effective measurement ranges are randomly sampled in every experiment. The performance of each calibrated sensor is evaluated on 200 separate testing samples using the normalized root-mean-squared-error $\text{NRMSE} = \frac{\text{RMS}(y-\hat{y})}{\text{RMS}(y)}$, where y is the actual ozone concentration and $\hat{y}$ the calibrated measurements of our sensors. Overall, we perform 200 experiments with different graph setups, arrangement of the sensors in the graph and measurement distributions in each experiment.

Data filtering. In order to show the impact of our data filtering, we investigate the effect of different filter thresholds $p_{\{E,A,EA\}}$, see (5). As described in Sec. 5.2, the thresholds are defined by the percentile values of each individual uncertainty of a calibration dataset. We use 6 different levels at $\{100, 90, 80, 70, 60, 50\}\%$, where 100% leads to no measurements filtered. For simplicity we use the same percentage-level for all three uncertainties.

Calibration models. We apply our scheme on three different calibration models, linear SCAN [3], non-linear regression trees [4] and non-linear neural networks [4], [8]. SCAN [3] is a linear regression model specifically developed to reduce error accumulation in multi-hop calibration. Regression trees [4] and neural networks [4], [8] have mainly been used used in one-hop calibration, *i.e.*, calibration to perfect references, and allow the modeling of more complex calibration functions. Although these methods might suffer from additional error accumulation over multiple hops, their complex modeling capabilities can be helpful to calibrate graphs with small overall number of hops. Similar to Sec. 6, the neural network that learns the aleatoric uncertainty uses one hidden layer with 16 neurons with tanh-activation functions and 20 ensemble members. The same network structure is also applied when we apply neural networks to perform the calibration in phase 1.

7.2. Results

Error over multiple hops. Fig. 6 shows the average calibration error of the sensors when calibrated at different hops in the graph. The different bars correspond to different percentile levels of the data filtering. We can observe that data filtering is able to reduce the calibration error. Especially at hops 3 to 5, using only samples with high confidence performs notably better than using all samples for calibration. While the relative improvement in terms of calibration error compared to no data filtering at hops 1 and 2 is at most 5%, it is in average 11% at hop 3 and 25% at hop 4 and 5. We also observe that the best performing percentile level is 70% at hops 3 and larger. Using lower percentiles results in a loss of accuracy, because too many measurements have been removed. In future work, it will be important to enhance our scheme with capabilities to find the optimal thresholds. Note that the two other calibration models, regression trees and neural networks, do not perform equally well compared to SCAN. This can be observed in Fig. 7, where we show the lowest calibration error of all threshold values at each individual hop for the three methods. SCAN achieves an upto 69% lower error than regression trees and 63% than neural networks. The data filtering is decreasing the error in average only by 3% for regression trees and neural networks. The error accumulation over multiple hops has in fact a big impact on the overall accuracy and begins to dominate at later hops. This is mainly because the methods are not developed for multi-hop calibration with large number of hops. However they can be powerful in setups with small hop numbers, which is presented in the following.

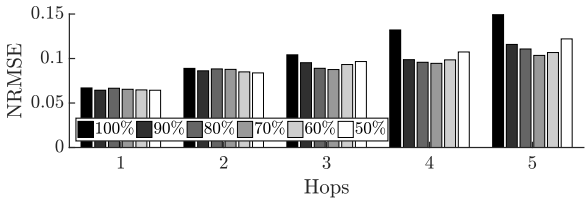


Figure 6. Calibration error of SCAN over multiple hops at different threshold percentile values. The higher the number of hops, the more important it is to reduce the impact of potentially inaccurate measurements.

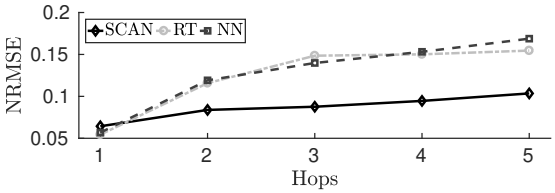
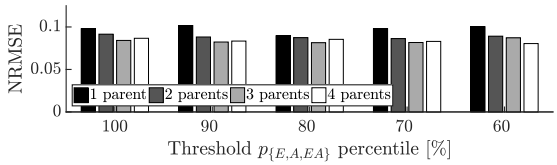
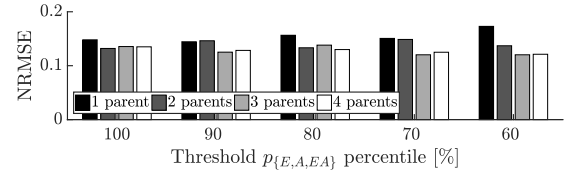


Figure 7. Lowest calibration error over multiple hops of the three methods SCAN, Regression Trees (RT) and neural networks (NN) over all threshold values. SCAN clearly outperforms the two other methods with an up to 69% lower error.

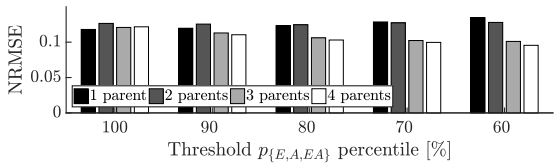
Error versus number of parents. Fig. 8 shows the average calibration error of our sensors at different filtering thresholds with different number of hops. We use a graph structure where these sensors are all 2 hops away from the reference, meaning they are calibrated by 1 to 4 parents that have been calibrated by the reference. For all three calibration models, SCAN in Fig. 8a, regression trees in Fig. 8b and neural networks in Fig. 8c we can observe a similar trend. The more parents are used to calibrate a sensor, the more effective is our uncertainty based data filtering. This result is expected due to the fact that more parents with different uncertainties at different concentrations can potentially induce high levels of noise in the calibration dataset. By only keeping the most confident samples from the different parents we are able to



(a) SCAN



(b) Regression Trees



(c) Neural Networks

Figure 8. Calibration error at different threshold levels for different number of parent nodes. Filtering samples with high uncertainty becomes more important with more parents for all three methods.

generate an improved dataset that reduces the calibration error. In fact, we are able to achieve the lowest calibration error for all three calibration models by using 4 parents for calibration in combination with our data filtering.

8. Conclusion

Uncertainty in sensor calibration is an omnipresent phenomenon. In this work, we present a scheme to estimate two major uncertainties in typical regression tasks, epistemic and aleatoric uncertainty. We apply our approach to improve the calibration of low-cost ozone sensors in a multi-hop setup, an important practice to maintain high data quality over time. By estimating uncertainties of our sensor measurements and using this gained information in an uncertainty-based data filtering, we are able to reduce the calibration error in the multi-hop setup by upto 25% compared to existing approaches.

Acknowledgements

This work was funded by the Swiss National Science Foundation (SNSF) under the FLAG-ERA CONVERGENCE project.

References

[1] O. Saukh, D. Hasenfratz, and L. Thiele, "Reducing multi-hop calibration errors in large-scale mobile sensor networks," in *Proc. ACM IPSN*, 2015, pp. 274–285.

[2] K. Fu, W. Ren, and W. Dong, "Multihop calibration for mobile sensing: K-hop calibratability and reference sensor deployment," in *Proc. IEEE INFOCOM*, 2017, pp. 1–9.

[3] B. Maag, Z. Zhou, O. Saukh, and L. Thiele, "SCAN: Multi-hop calibration for mobile sensor arrays," *Proc. ACM IMWUT*, vol. 1, no. 2, pp. 19:1–19:21, 2017.

[4] Y. Lin, W. Dong, and Y. Chen, "Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement," *Proc. ACM IMWUT*, vol. 2, no. 1, pp. 18:1–18:18, 2018.

[5] M. Mueller, J. Meyer, and C. Hueglin, "Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of zurich," *Atmospheric Measurement Techniques*, vol. 10, no. 10, pp. 3783–3799, 2017.

[6] B. Maag, Z. Zhou, and L. Thiele, "A survey on sensor calibration in air pollution monitoring deployments," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4857–4870, 2018.

[7] Y. Xiang, L. S. Bai, R. Pledrahita, R. P. Dick, Q. Lv, M. Hannigan, and L. Shang, "Collaborative calibration and sensor placement for mobile sensor networks," in *Proc. ACM IPSN*, 2012, pp. 73–83.

[8] Y. Cheng, X. Li, Z. Li, S. Jiang, Y. Li, J. Jia, and X. Jiang, "AirCloud: A cloud-based air-quality monitoring system for everyone," in *Proc. ACM SenSys*, 2014, pp. 251–265.

[9] O. Saukh, D. Hasenfratz, C. Walser, and L. Thiele, "On rendezvous in mobile sensing networks," in *Real-World Wireless Sensor Networks*. Springer, 2014, pp. 29–42.

[10] F. Kizel, Y. Etzion, R. Shafran-Nathan, I. Levy, B. Fishbain, A. Bartonova, and D. M. Broday, "Node-to-node field calibration of wireless distributed air pollution sensor network," *Environmental Pollution*, vol. 233, pp. 900–909, 2018.

[11] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105 – 112, 2009, risk Acceptance and Risk Communication.

[12] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Proc. NIPS*, 2017, pp. 5574–5584.

[13] W. D. Penny and S. J. Roberts, "Neural network predictions with error bars," *Dept. of Electrical and Electronic Engineering*, 1997.

[14] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.

[15] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] W. S. Parker, "Ensemble modeling, uncertainty and robust predictions," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 4, no. 3, pp. 213–223, 2013.

[17] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped DQN," in *Proc. NIPS*, 2016, pp. 4026–4034.

[18] D. A. Nix and A. S. Weigend, "Learning local error bars for nonlinear regression," in *Proc. NIPS*, 1995, pp. 489–496.

[19] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NIPS*, 2017, pp. 6402–6413.

[21] S. Yao, Y. Zhao, H. Shao, A. Zhang, C. Zhang, S. Li, and T. Abdelzaher, "RDeepSense: Reliable deep mobile computing models with uncertainty estimations," *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, pp. 173:1–173:26, 2018.

[22] SGX Sensortech, "MiCS-OZ-47 ozone sensor (datasheet)," https://www.sgxsensortech.com/, 2014.

[23] M. Lippmann, "Health effects of ozone. A critical review," *Japca*, vol. 39, no. 5, pp. 672–695, 1989.

[24] B. Maag, O. Saukh, D. Hasenfratz, and L. Thiele, "Pre-deployment testing, augmentation and calibration of cross-sensitive sensors," in *Proc. ACM/IEEE EWSN*, 2016, pp. 169–180.

[25] J. J. Li, B. Faltings, O. Saukh, D. Hasenfratz, and J. Beutel, "Sensing the air we breathe – the OpenSense Zurich dataset," in *AAAI Conference on Artificial Intelligence*, 2012.

[26] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in *Proc. NIPS*, 1991, pp. 875–882.