J. Sens. Sens. Syst., 12, 45–60, 2023
https://doi.org/10.5194/jsss-12-45-2023

JSSS Journal of Sensors and Sensor Systems Open Access

Influence of measurement uncertainty on machine learning results demonstrated for a smart gas sensor

Tanja Dorst[1,2], **Tizian Schneider**[1,2], **Sascha Eichstädt**[3], **and Andreas Schütze**[1,2]

[1]ZeMA – Center for Mechatronics and Automation Technology gGmbH, Saarbrücken, Germany
[2]Lab for Measurement Technology, Department of Mechatronics, Saarland University, Saarbrücken, Germany
[3]Fachbereich 9.4, Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany

Correspondence: Tanja Dorst (t.dorst@lmt.uni-saarland.de)

Received: 1 August 2022 – Revised: 18 November 2022 – Accepted: 9 January 2023 – Published: 27 January 2023

Abstract. Humans spend most of their lives indoors, so indoor air quality (IAQ) plays a key role in human health. Thus, human health is seriously threatened by indoor air pollution, which leads to 3.8 $\times 10^6$ deaths annually, according to the World Health Organization (WHO). With the ongoing improvement in life quality, IAQ monitoring has become an important concern for researchers. However, in machine learning (ML), measurement uncertainty, which is critical in hazardous gas detection, is usually only estimated using cross-validation and is not directly addressed, and this will be the main focus of this paper. Gas concentration can be determined by using gas sensors in temperature-cycled operation (TCO) and ML on the measured logarithmic resistance of the sensor. This contribution focuses on formaldehyde as one of the most relevant carcinogenic gases indoors and on the sum of volatile organic compounds (VOCs), i.e., acetone, ethanol, formaldehyde, and toluene, measured in the data set as an indicator for IAQ. As gas concentrations are continuous quantities, regression must be used. Thus, a previously published uncertainty-aware automated ML toolbox (UA-AMLT) for classification is extended for regression by introducing an uncertainty-aware partial least squares regression (PLSR) algorithm. The uncertainty propagation of the UA-AMLT is based on the principles described in the *Guide to the Expression of Uncertainty in Measurement* (GUM) and its supplements. Two different use cases are considered for investigating the influence on ML results in this contribution, namely model training with raw data and with data that are manipulated by adding artificially generated white Gaussian or uniform noise to simulate increased data uncertainty, respectively. One of the benefits of this approach is to obtain a better understanding of where the overall system should be improved. This can be achieved by either improving the trained ML model or using a sensor with higher precision. Finally, an increase in robustness against random noise by training a model with noisy data is demonstrated.

1 Introduction

1.1 Indoor air quality and VOCs

As humans spend most of their lives indoors, the most significant environment for them is the indoor environment (Brasche and Bischof, 2005). For this reason, indoor air quality (IAQ) is of special importance as it plays a leading role with regard to the performance, well-being, and health of humans (Sundell, 2004; Asikainen et al., 2016). Volatile organic compounds (VOCs) are one of the main contributors to poor air quality, especially in indoor air, and can lead to serious health problems, e.g., leukemia, cancers, or tumors (Jones, 1999; Tsai, 2019). Nowadays, IAQ monitoring is mostly based on measurements of carbon dioxide (CO_2) emitted by humans as the primary indicator for poor indoor air, as CO_2 concentration is directly related to VOCs caused by human presence (Von Pettenkofer, 1858). However, this neglects the fact that not only humans emit VOCs but also their activities such as household cleaning, cooking, and smoking, as well as, for example, furniture, carpets, and even the building itself due to the building materials used (Spaul, 1994). To measure almost all types of VOCs in in-

Published by Copernicus Publications on behalf of the AMA Association for Sensor Technology.

door air, metal oxide semiconductor (MOS) gas sensors are widely used as they are low-cost, robust, and highly sensitive. To improve the limited selectivity of these sensors and enable the discrimination of specific pollutants, MOS gas sensors can be operated in dynamic modes, especially by using a temperature-cycled operation (TCO; Eicker, 1977; Lee and Reedy, 1999; Baur et al., 2015; Schütze and Sauerwald, 2020a; Baur et al., 2021). A TCO, especially in combination with modern microstructured gas sensors, yields extensive and rich response patterns that need to be interpreted using machine learning (ML) to extract the relevant information (Schütze and Sauerwald, 2020a).

For the data set used in this contribution, sensor responses of an SPG30 sensor (Sensirion AG, Stäfa, Switzerland) with four gas-sensitive layers in TCO were recorded (Rüffer et al., 2018). This contribution focuses on formaldehyde as an example of a highly relevant toxic gas and on the sum concentration of all VOCs (VOC_{sum}) in parts per billion (ppb) in the used data set, i.e., the sum of the concentrations of acetone, ethanol, formaldehyde, and toluene. VOC_{sum} should not be confused with the widely used total VOC (TVOC) value, as this is based on analytical measurements and takes into account only VOCs with medium volatility (Schütze and Sauerwald, 2020b). Formaldehyde (CH_2O) is one of the most toxic and carcinogenic gases in indoor air (Hauptmann et al., 2004; Zhang, 2018; NTP, 2021) and is released from a variety of sources. The most significant ones are pressed wood products, e.g., particle board and plywood paneling. The World Health Organization (WHO) set the guideline threshold for a 30 min average concentration to 0.1 mg m^{-3}, which corresponds to approximately 80.1 ppb for 760 mmHg and 20 °C (World Health Organization, 2010).

1.2 Automated ML toolbox

In recent years, an automated machine learning toolbox (AMLT) was developed and applied to different classification tasks (Schneider et al., 2017, 2018; Dorst et al., 2021). Its extension to an uncertainty-aware AMLT (UA-AMLT) for classification was presented in Dorst et al. (2022). As gas concentrations are continuous quantities, regression must be used, which is a supervised ML technique. In this contribution, the AMLT is therefore extended to be applicable for regression tasks and, furthermore, the corresponding uncertainty for the ML result is considered. The uncertainty propagation is based on the *Guide to the Expression of Uncertainty in Measurement* (GUM; BIPM et al., 2008a) and its Supplement 1 (BIPM et al., 2008b) and Supplement 2 (BIPM et al., 2011). These three documents establish general rules for evaluating and expressing measurement uncertainty. These rules and principles are applied in this contribution for estimating the uncertainty of an ML model prediction, thus extending the GUM approach to smart sensors.

To investigate the influence of measurement uncertainty on machine learning (ML) results, sensor raw data are manipulated by simulated additive white Gaussian noise. With these manipulated data sets, different ML models are determined based on feature extraction, feature selection followed by regression, and the influence of the Gaussian noise, which simulates increased sensor uncertainty in the ML results, is investigated. Gaussian (normally distributed) noise is a very good assumption for any process for which the central limit theorem holds. In addition, the influence of additive white uniform noise as a further noise model is investigated.

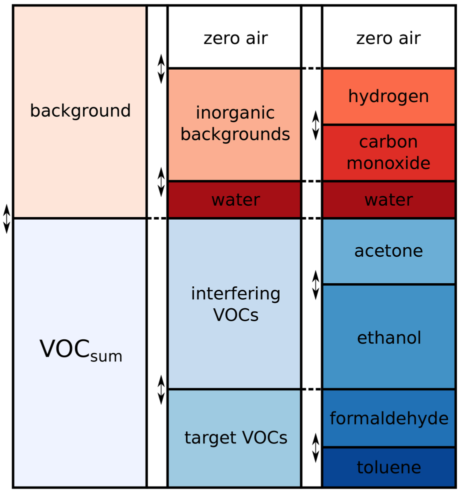

Figure 1. Gas composition for calibration consisting of random mixtures of VOCs (blue) and background gases (red; adapted from Baur et al., 2021).

2 Materials and methods

2.1 Data set

A data set published in Baur et al. (2021) is used to investigate the influence of measurement uncertainty on ML results. It consists of different calibration and field test measurements of gas mixtures with the MOS gas sensor SGP30 (Sensirion AG, 2020). The gas mixtures are composed of random mixtures of seven different gases that are relevant for indoor air quality. Various VOCs, i.e., acetone, ethanol, formaldehyde, and toluene, are used together with water vapor and inorganic background gases, i.e., hydrogen and carbon monoxide, as shown in Fig. 1. The gas concentrations are mixed using Latin hypercube sampling (LHS; McKay et al., 1979) to obtain unique gas mixtures (UGMs). In this contribution, only data from the initial calibration are used. The concentration ranges for all gases during the initial calibration are shown in Table 1.

The SGP30 sensor, with its four different gas-sensitive layers, is used in TCO to improve its selectivity, sensitivity, and stability (Schultealbert et al., 2018). As shown in Fig. 2, the temperature cycle consists of 10 steps at 400 °C, with a duration of 5 s each, followed by different low-temperature steps,

Table 1. Concentration ranges for all gases during the initial calibration phase (Amann et al., 2021b).

Substance	Minimum	Maximum
Humidity	25 % RH	70 % RH
Hydrogen	400 ppb	2000 ppb
Carbon monoxide	150 ppb	2000 ppb
Acetone	14 ppb	300 ppb
Ethanol	4 ppb	300 ppb
Formaldehyde	1 ppb	400 ppb
Toluene	4 ppb	300 ppb
$\mathrm{VOC_{sum}}$	300 ppb	1200 ppb

RH is the relative humidity.

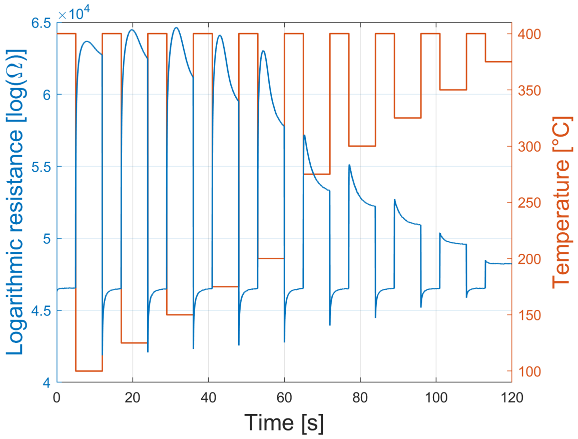

Figure 2. Logarithmic conductance of one sensor element (blue) and the temperature-cycled operation of the SGP30 (red).

with a duration of 7 s each. One single temperature cycle thus lasts 120 s and, due to the sampling rate of 20 Hz, consists of 2.400 measurement values for each gas-sensitive layer. The sensor output represents the logarithmic resistance shown for one cycle and one gas-sensitive layer in Fig. 2.

During the initial calibration phase, the SGP30 sensor is exposed to 500 UGMs for 10 temperature cycles (TCs) each. Due to the limited time response of the gas mixing apparatus (GMA) and synchronization problems between sensor and GMA, four TCs at the beginning and the last TC for each UGM are omitted so that only five TCs per UGM are evaluated. Furthermore, the first three UGMs are also not considered due to run-in effects. Thus, the data set comprises 2485 relevant cycles of 497 UGMs with stable gas concentrations from the initial calibration.

2.2 Uncertainty-aware automated machine learning toolbox

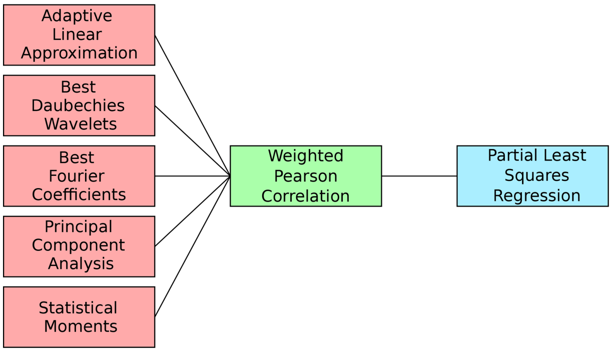
Adaptive Linear Approximation
Best Daubechies Wavelets
Best Fourier Coefficients
Principal Component Analysis
Statistical Moments
Weighted Pearson Correlation
Partial Least Squares Regression

Figure 3. Feature extraction (red), feature selection (green), and regression (blue) algorithms of the uncertainty-aware AMLT for regression tasks.

In general, regression is used for predicting a continuous quantity, whereas classification is used for predicting a discrete class label. As a basis for this publication, the AMLT for classification tasks (Schneider et al., 2017, 2018; Dorst et al., 2021) and its extended uncertainty-aware version (Dorst et al., 2022) are modified to also solve regression tasks. With the AMLT, feature extraction (FE) and feature selection (FS), as well as classification/regression and evaluation, are performed without expert knowledge and without a detailed physical model of the process to minimize model generation costs. Model training, in addition to application, can be carried out with the (uncertainty-aware) AMLT. Partial least squares regression (PLSR) as the de facto standard for quantification in the field of gas sensors (Wold et al., 2001; Gutierrez-Osuna, 2002) is used for regression tasks in the AMLT. Another well-known regression algorithm is principal component regression (PCR), which first performs the principal component analysis (PCA) as an unsupervised technique to obtain the principal components (PCs) and then uses these PCs to build the regression model. As a two-step model-building algorithm, the PCR makes interpreting the ML results harder in contrast to PLSR, which only has one step (Ergon, 2014). Using PCA leads to a relevant drawback of the PCR algorithm, as performing an unsupervised technique does not guarantee that the selected principal components for the regression model building are associated with the target. An advantage of PLSR is that it often has fewer components than PCR to achieve the same prediction level (De Jong, 1993a).

As shown in Fig. 3, five complementary FE algorithms are used within the AMLT, together with Pearson correlation for FS and PLSR, as the regression algorithm.

Adaptive linear approximation splits cycles into approximately linear segments, and for each segment, the mean value and slope are extracted as features from the time domain (Olszewski et al., 2001). The best Daubechies wavelets algorithm performs a wavelet transform using a Daubechies D4 wavelet (Daubechies, 1992) to extract 10 % of the wavelet coefficients with the highest average absolute value over all cycles as features from the time frequency domain. The best Fourier coefficients algorithm performs a Fourier transform, and 10 % of the amplitudes with the highest average absolute value over all cycles and their corresponding phases are extracted from frequency domain (Mörchen, 2003). Using principal component analysis, projections on the principal components are determined (Pearson, 1901; Jackson, 1991) and used as features from the time domain. Moreover, the statistical distribution of the measurement values also includes information in the time domain (Martin and Honarvar, 1995). Thus, the cycles are split into 10 approximately equally sized segments, and the four statistical moments (mean, standard deviation, skewness, and kurtosis) are extracted for each segment as features. These five FE algorithms and the Pearson correlation as FS lead to five different algorithm combinations, each benchmarked to choose the best one for the respective application. The best combination is determined by the smallest cross-validated root mean square error (RMSE), which is a measure for the differences between the predicted $\boldsymbol{y}_{\mathbf{pred}} \in \mathbb{R}^m$ and the observed target values $\boldsymbol{y}$ of the same dimension, i.e.,

$$\mathrm{RMSE}(\boldsymbol{y}_{\mathbf{pred}}, \boldsymbol{y}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_{\mathrm{pred}_i} - y_i)^2}\,. \tag{1}$$

The cross-validation (CV) scenario used is explained in Sect. 2.2. In general, different metrics can be used to describe the performance of a regression model; however, RMSE is one of the best interpretable error measures as it has the same unit as the prediction of the model and is also comparable to the (standard) measurement uncertainty used in describing data quality in measurement.

To use the UA-AMLT, a data matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$ for each sensor (or sensor layer) must be given, where m denotes the number of cycles of length n. In case of non-cyclic sensor data, data must be windowed to obtain the correct $m \times n$ format. Furthermore, there must be knowledge about the uncertainty matrix $\mathbf{U} \in \mathbb{R}^{m\times n}$, which assigns an uncertainty value u_{ij} to a measurement value d_{ij} $\forall\, i, j$. This means that correlation of errors at different time instants is neglected.

Uncertainty-aware feature extraction and selection

To perform FE, which mathematically describes the mapping $\mathbf{D} \longmapsto \mathbf{F_E}$, five complementary methods are used. In this step, one feature matrix $\mathbf{F_E} \in \mathbb{R}^{m\times k}$, $k \leq n$ is calculated for each of the FE methods. The uncertainty calculation is performed according to Dorst et al. (2022), so that, for every feature matrix $\mathbf{F_E}$, an uncertainty matrix $\mathbf{U_{F_E}}$ of the same dimension is calculated.

In the uncertainty-aware FS step, features are ranked according to their weighted Pearson correlation to the target value, i.e., in this contribution to the gas concentration. In weighted Pearson correlation, the reciprocals of the squared uncertainty values of the features are used as weights (Dorst et al., 2022). After ranking the features, a 10-fold stratified CV (Kohavi, 1995) is carried out for every possible number of features, and the minimum CV error is determined based on the optimal number of features $l \in \mathbb{N}$ found. From a mathematical point of view, FS is a mapping $\mathbf{F_E} \longmapsto \mathbf{F_S}$, with $\mathbf{F_S} \in \mathbb{R}^{m\times l}$, $l < k$ containing only the optimal number of the most relevant features according to weighted Pearson correlation. The corresponding uncertainty matrix is $\mathbf{U_{F_S}} \in \mathbb{R}^{m\times l}$.

2.3 Partial least squares regression

Let a predictor matrix $\mathbf{X} \in \mathbb{R}^{m\times l}$ and a responses matrix $\mathbf{Y} \in \mathbb{R}^{m\times s}$ be given. The basic algorithm for computing a PLSR of $\mathbf{Y}$ on $\mathbf{X}$ using n_{comp} PLSR components is developed in Wold et al. (1984). Performing PLSR means iteratively solving the following decompositions, such that the covariance between $\mathbf{X}$ and $\mathbf{Y}$ is maximized as follows:

$$\mathbf{X} = \mathbf{X_S} \cdot \mathbf{X_L}^\top + \mathbf{X}_{\mathrm{res}} \tag{2}$$

$$\mathbf{Y} = \mathbf{Y_S} \cdot \mathbf{Y_L}^\top + \mathbf{Y}_{\mathrm{res}}, \tag{3}$$

where $\mathbf{X_L} \in \mathbb{R}^{l\times n_{\mathrm{comp}}}$ and $\mathbf{Y_L} \in \mathbb{R}^{s\times n_{\mathrm{comp}}}$ denote the orthogonal loading matrices. $\mathbf{X_S} \in \mathbb{R}^{m\times n_{\mathrm{comp}}}$ and $\mathbf{Y_S} \in \mathbb{R}^{m\times n_{\mathrm{comp}}}$ are the predictor and response scores, respectively. The matrices $\mathbf{X}_{\mathrm{res}}$ and $\mathbf{Y}_{\mathrm{res}}$ are the residual terms for predictor and response, respectively, and are used as a start for the next iteration step.

In MATLAB®, the partial least squares regression (PLSR) is calculated using the SIMPLS (statistically inspired modification of the partial least squares) algorithm (De Jong, 1993b). The advantage of SIMPLS is that the regression coefficients are determined directly without inverse matrices or singular value decomposition. Assume that $\hat{\mathbf{X}} \in \mathbb{R}^{m\times(l+1)}$ denotes a matrix in which a vector of ones is prepended to $\mathbf{X}$ to compute coefficient estimates for a model with constant terms. With $\mathbb{1} \in \mathbb{R}^m$ denoting a vector containing only ones, it holds for the augmented matrix that $\hat{\mathbf{X}} = (\mathbb{1}\,|\,\mathbf{X}) \in \mathbb{R}^{m\times(l+1)}$. The SIMPLS algorithm involves the calculation of a weighted matrix $\mathbf{W} \in \mathbb{R}^{(l+1)\times n_{\mathrm{comp}}}$. For the SIMPLS algorithm, the following holds:

$$\mathbf{X_S} = \hat{\mathbf{X}} \cdot \mathbf{W} \quad \text{and} \tag{4}$$

$$\mathbf{Y} = \mathbf{X_S} \cdot \mathbf{Y_L}^\top. \tag{5}$$

Combining Eqs. (4) and (5) leads to the following:

$$\mathbf{Y} = \hat{\mathbf{X}} \cdot \mathbf{W} \cdot \mathbf{Y_L}^\top \tag{6}$$

$$= \hat{\mathbf{X}} \cdot \mathbf{B}, \tag{7}$$

where $\mathbf{B} \in \mathbb{R}^{(l+1)\times s}$ denotes the matrix containing intercept terms in the first row and PLSR coefficient estimates in the others (De Jong, 1993b).

Uncertainty-aware partial least squares regression

In this contribution, the target values $\boldsymbol{y} \in \mathbb{R}^m$ are only represented by one vector, which leads to the matrix $\mathbf{B}$ (see

Sect. 2.3) being also only a vector $\boldsymbol{\beta} \in \mathbb{R}^{l+1}$. The matrix of selected features is given by $\mathbf{F_S} \in \mathbb{R}^{m \times l}$. $\hat{\mathbf{F}}_\mathbf{S} = (\mathbb{1} \,|\, \mathbf{F_S}) \in \mathbb{R}^{m \times (l+1)}$ denotes the matrix where one column of ones at the beginning of $\mathbf{F_S}$ was added. For PLSR, the following holds:

$$\boldsymbol{y}_{\mathbf{pred}} = \hat{\mathbf{F}}_\mathbf{S} \cdot \boldsymbol{\beta}, \tag{8}$$

with $\boldsymbol{y}_{\mathbf{pred}} \in \mathbb{R}^m$ representing the predicted target values. The basis of the uncertainty values calculation for the prediction $\boldsymbol{y}_{\mathbf{pred}}$ are formulas given in Sect. 6.2 (“Propagation of uncertainty for explicit multivariate measurement models”) found in Supplement 2 of GUM (GUMS2; BIPM et al., 2011). This section of GUMS2 shows the covariance matrix calculation associated with an estimate of a multidimensional output quantity with the help of a sensitivity matrix using matrix–vector notation. This approach can be transferred to the propagation of uncertainty for PLSR. The first step is the transposing of Eq. (8), which leads to the following:

$$\boldsymbol{y}_{\mathbf{pred}}^\top = \boldsymbol{\beta}^\top \cdot \hat{\mathbf{F}}_\mathbf{S}^\top. \tag{9}$$

To use Sect. 6.2 of GUMS2, $\hat{\mathbf{F}}_\mathbf{S}$ and $\boldsymbol{\beta}$ must be transformed into vector and matrix, respectively. For the columns of $\hat{\mathbf{F}}_\mathbf{S}^\top$, the following holds:

$$\hat{\mathbf{F}}_\mathbf{S}^\top = \begin{pmatrix} & 1 & \dots & 1 \\ \boldsymbol{f}_{\boldsymbol{S_1}}^\top & \boldsymbol{f}_{\boldsymbol{S_2}}^\top & \dots & \boldsymbol{f}_{\boldsymbol{S_m}}^\top \end{pmatrix}, \tag{10}$$

where $\boldsymbol{f}_{\boldsymbol{S_i}}^\top \in \mathbb{R}^l$, $\forall i = 1, \dots, m$ denotes the selected features for the ith cycle. Thus, the matrix–vector representation is given by the following:

$$\tilde{\boldsymbol{F}}_{\boldsymbol{S}}^\top = \begin{pmatrix} 1 \\ \boldsymbol{f}_{\boldsymbol{S_1}}^\top \\ 1 \\ \boldsymbol{f}_{\boldsymbol{S_2}}^\top \\ \vdots \\ 1 \\ \boldsymbol{f}_{\boldsymbol{S_l}}^\top \end{pmatrix} \in \mathbb{R}^{(m \cdot (l+1)) \times 1}, \tag{11}$$

and $\tilde{\boldsymbol{\beta}}^\top \in \mathbb{R}^{m \times (m \cdot (l+1))}$, with

$$\tilde{\boldsymbol{\beta}}^\top = \begin{pmatrix} \boldsymbol{\beta}^\top & 0\dots0 & 0\dots0 & \dots & 0\dots0 \\ 0\dots0 & \boldsymbol{\beta}^\top & 0\dots0 & \dots & 0\dots0 \\ 0\dots0 & 0\dots0 & \boldsymbol{\beta}^\top & \dots & 0\dots0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0\dots0 & 0\dots0 & 0\dots0 & \dots & \boldsymbol{\beta}^\top \end{pmatrix}, \tag{12}$$

which leads to

$$\boldsymbol{y}_{\mathbf{pred}}^\top = \tilde{\boldsymbol{\beta}}^\top \cdot \tilde{\boldsymbol{F}}_{\boldsymbol{S}}^\top. \tag{13}$$

To propagate the uncertainty in the PLSR, the uncertainty matrix of the selected features $\mathbf{U}_{\mathbf{F_S}} \in \mathbb{R}^{m \times l}$ must be extended with a column associated with the first column of $\hat{\mathbf{F}}_\mathbf{S}$. Thus, it holds that $\hat{\mathbf{U}}_{\mathbf{F_S}} = \left(\mathbb{0} \,|\, \mathbf{U}_{\mathbf{F_S}}\right) \in \mathbb{R}^{m \times (l+1)}$. The transpose matrix $\hat{\mathbf{U}}_{\mathbf{F_S}}^\top$ is transferred to the diagonal matrix $\tilde{\mathbf{U}}_{\mathbf{F_S}}^\top \in \mathbb{R}^{(m \cdot (l+1)) \times (m \cdot (l+1))}$, where the rows of $\hat{\mathbf{U}}_{\mathbf{F_S}}$ are in the diagonal. Using Sect. 6.2.1.3 of BIPM et al. (2011) leads to the following:

$$\tilde{\mathbf{U}} = \tilde{\boldsymbol{\beta}}^\top \cdot \left(\tilde{\mathbf{U}}_{\mathbf{F_S}}^\top\right)^2 \cdot \left(\tilde{\boldsymbol{\beta}}^\top\right)^\top \tag{14}$$

$$= \tilde{\boldsymbol{\beta}}^\top \cdot \left(\tilde{\mathbf{U}}_{\mathbf{F_S}}^\top\right)^2 \cdot \tilde{\boldsymbol{\beta}}, \tag{15}$$

with $\tilde{\mathbf{U}} \in \mathbb{R}^{m \times m}$. To obtain the diagonal elements of $\tilde{\mathbf{U}}$, Eq. (15) can be simplified and retransformed to the following:

$$\mathbf{U}_{\mathrm{PLSR}}^\top = \left(\boldsymbol{\beta}^\top \circ \boldsymbol{\beta}^\top\right) \cdot \left(\hat{\mathbf{U}}_{\mathbf{F_S}}^\top \circ \hat{\mathbf{U}}_{\mathbf{F_S}}^\top\right) \tag{16}$$

$$\Leftrightarrow \quad \mathbf{U}_{\mathrm{PLSR}} = \left(\hat{\mathbf{U}}_{\mathbf{F_S}} \circ \hat{\mathbf{U}}_{\mathbf{F_S}}\right) \cdot (\boldsymbol{\beta} \circ \boldsymbol{\beta}) \tag{17}$$

$$= \hat{\mathbf{U}}_{\mathbf{F_S}}^{\circ 2} \cdot \boldsymbol{\beta}^{\circ 2}, \tag{18}$$

where $\circ$ denotes the Hadamard (element-wise) product (Horn, 1990). The uncertainty values associated with $\boldsymbol{y}_{\mathbf{pred}}$ can be calculated by the following:

$$\boldsymbol{U}_{\boldsymbol{y}_{\mathbf{pred}}} = \left(\left|\hat{\mathbf{U}}_{\mathbf{F_S}}^{\circ 2} \cdot \boldsymbol{\beta}^{\circ 2}\right|\right)^{\circ 1/2} \in \mathbb{R}^{m \times 1}, \tag{19}$$

where $|.|$ denotes the element-wise absolute value and $(.)^{\circ 1/2}$ the Hadamard (element-wise) square root (Reams, 1999).

3 Investigation of the influence of measurement uncertainty on ML results

To evaluate the influence of measurement uncertainty on ML results, the logarithmic resistance raw data of each sensor layer are modified by artificially generated additive white Gaussian noise of different signal-to-noise ratios (SNRs). This means that the logarithmic amplifier of the sensor is responsible for the noise. In general, the SNR is defined as the ratio of signal power to background noise power. SNR > 0 dB indicates that there is more signal than background noise. The maximum theoretical SNR in decibel (dB) for an analog-to-digital converter (ADC) can be determined, according to Bennett (1948), with the following:

$$\mathrm{SNR}(N) = 20 \cdot \log_{10}\left(2^N \cdot \sqrt{\frac{3}{2}}\right) \quad [\mathrm{dB}] \tag{20}$$

$$\approx 6.02 \cdot N + 1.76 \quad [\mathrm{dB}], \tag{21}$$

where N is the resolution of an ADC in bits. Thus, the maximum theoretical SNR for the 16 bit ADC of the SGP30 is approx. 98 dB. For this reason, only SNRs from 0 to 98 dB are considered in this publication. Figure 4 shows an example of raw and modified sensor data with different SNR values. The

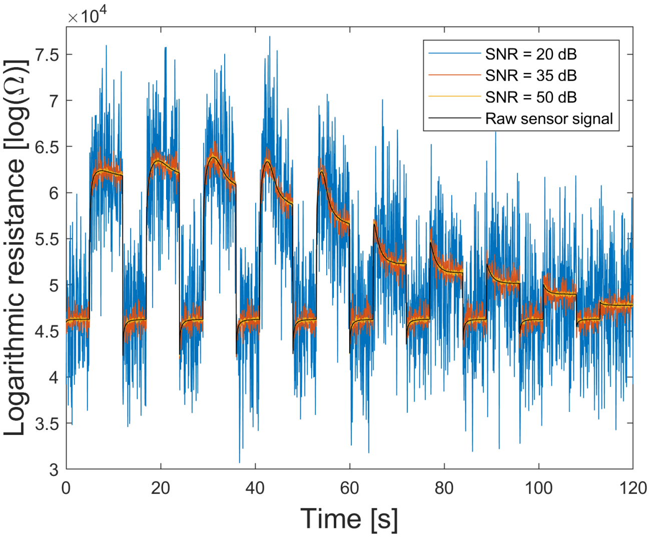

Figure 4. Raw (violet) and modified sensor signals with additive white Gaussian noise of different SNR values.

relation between the SNR and squared standard uncertainty σ^2 is given by the following:

$$\sigma^2 = 10^{\frac{\mathrm{SP}-\mathrm{SNR}}{10}}, \tag{22}$$

where the signal power (SP) is calculated by

$$\mathrm{SP} = 10 \cdot \log_{10}\left(\frac{\|\mathbf{A}\|_2^2}{m \cdot n}\right). \tag{23}$$

Here, $\mathbf{A} \in \mathbb{R}^{m \times n}$ denotes the data for one sensor. Thus, for example, 75 dB corresponds to $\sigma^2 = 91.71$, 80 dB to $\sigma^2 = 29.00$, and 98 dB to $\sigma^2 = 0.46$ for the first gas-sensitive layer of the SGP30. In practical applications, 98 dB is typically not reached because the measurement range of the ADC is larger than the range of actual measured values within the data set.

3.1 Application of AMLT

To investigate the influence of measurement uncertainty on machine learning results, the best FE algorithm must first be determined. To train, validate, and test a model, the data set is randomly split into 70 % training, 10 % validation, and 20 % test data by omitting complete UGMs in the training, validation, or test data set, respectively. This means that each of the 497 UGMs exists in either the training, validation, or test data but not in more than one at a time (see Fig. 5). Training the model is carried out by using the AMLT together with the training data and formaldehyde or $\mathrm{VOC_{sum}}$, respectively, as the target. The results obtained for $\mathrm{VOC_{sum}}$ as the target show the same trends and lead to the same conclusions as the results with formaldehyde as target and are therefore only shown in Sect. A2.

A 10-fold stratified CV is automatically performed in the AMLT to determine the best FE algorithm out of five

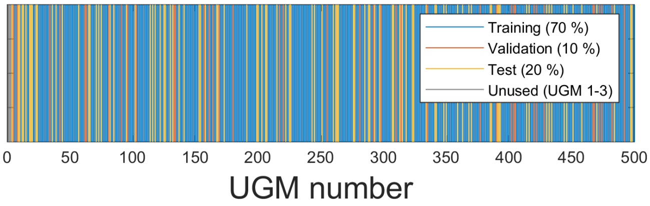

Figure 5. Randomized split of the UGMs into training, validation, and test data used in this contribution.

complementary FE methods. In contrast to the data split, which is carried out by omitting complete UGMs and used for performing group-based CV with validation data, the 10-fold stratified CV randomly omits individual TCs. The RMSE value resulting from the 10-fold CV is called the random CV error. To obtain quality information on the trained model, the differences between the predicted and the observed target values are measured using RMSE. The test RMSE (T − RMSE) results from applying the trained model to the test data. There can be a significant difference between a group-based CV error and T − RMSE, as the omitted UGMs are selected randomly. For each of the five algorithm combinations (see Fig. 3), a Monte Carlo different train, validation, and test data sets, is performed. The mean value and standard deviation are calculated for the three different errors resulting from using training, validation, and test data, with $n_{\mathrm{comp}} = 20$ in the PLSR algorithm. The reason for choosing $n_{\mathrm{comp}} = 20$ is given below. The results are shown in Fig. 6. Although principal component analysis (PCA) achieves the lowest random CV error mean value (14.7 ppb), with negligible variations for different splits and therefore seems to be the best FE algorithm, applying 10 % validation data will lead to a group-based CV error mean value of 24.7 ppb. This means that 10-fold stratified CV does not efficiently detect overfitting for this application as it does not omit complete UGMs and, thus, does not need to interpolate to different gas concentrations. A new, unpublished version of the AMLT already allows the user to define validation scenarios (random or group based). Here, adaptive linear approximation (ALA) as the second-best method with a random CV error mean value of 15.3 ppb is chosen for further investigations as there is no significant difference in the error mean values between omitting single TCOs (15.3 ppb; random validation) and complete UGMs (16.5 and 16.6 ppb; group-based validation). Thus, it is sufficient to evaluate the random CV error with 80 % training (including 10 % validation data) and 20 % test data split by omitting complete UGMs in the training or test data, respectively. Applying this trained model ($n_{\mathrm{comp}} = 20$; 80 % data used for model training) to the 20 % test data (see Fig. 5) leads to a shown in Fig. 7.

For $\mathrm{VOC_{sum}}$ as the target, the results are similar, and again, ALA is chosen as the best FE algorithm (random CV error mean value of 40.9 ppb) due to the overfitting of the trained model when using PCA (random CV error mean value of

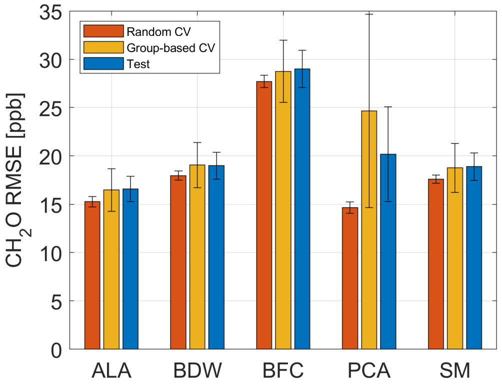

Figure 6. Random CV, group-based CV, and test RMSE of the five FE algorithms using Pearson as FS and PLSR with $n_{\mathrm{comp}} = 20$ for 100 trials with different randomized UGM splits and the formaldehyde concentration as target. ALA is the adaptive linear approximation, BDW is the best Daubechies wavelets, BFC is the best Fourier coefficients, PCA is the principal component analysis, and SM is the statistical moments.

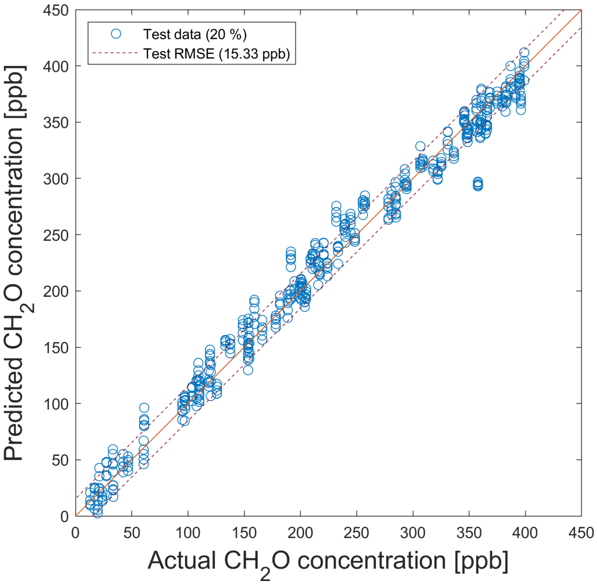

Figure 7. PLSR model for the quantification of formaldehyde for testing with test data from the data split shown in Fig. 5. Dashed lines indicate the RMSE of the test data (T − RMSE).

31.9 ppb; group-based CV error mean value of 61.3 ppb). Applying the model trained with the data split in Fig. 5 to the 20 % test data results in a T − RMSE of 46.1 ppb. The corresponding results are shown in Figs. A3 and A4.

To determine the optimal number of PLSR components, a Monte Carlo simulation (10 trials with different train and test data) was carried out, and the T − RMSE mean values of 10 trials, in addition to the corresponding standard deviations, were calculated. In Fig. 8, the T − RMSE value is plotted over the number of PLSR components for ALA as FE, Pearson as FS, and PLSR. For a small number of PLSR components, T − RMSE mean values have large standard deviations, for example, the standard deviation for $n_{\mathrm{comp}} = 1$ is $\sigma = 5.5$ ppb. If the number of PLSR components is greater than 10, then the standard deviations are in the range from 1.15 to 1.22 ppb; thus, the obtained models are highly reproducible. The lowest T − RMSE mean value is achieved for a high number of PLSR components (here 13.5 ppb is achieved with $n_{\mathrm{comp}} = 100$), but it is preferable to find a good tradeoff between the accuracy and computational cost, as a lower number of PLSR components reduces the computational effort. Therefore, the optimal number of PLSR components is determined using the elbow method (Thorndike, 1953) to ensure a stable model, with a T − RMSE of 15.3 ppb. The elbow point, i.e., the point after which no further significant change occurs, is determined by using the ALA algorithm. ALA automatically determines four segments as being the best segmentation of the T − RMSE curve (see Fig. 8). Thus, the optimal number of PLSR components is $n_{\mathrm{comp}} = 20$, as more components have no considerable influence on the T − RMSE, leading to higher computational cost and also increasing the risk of overfitting.

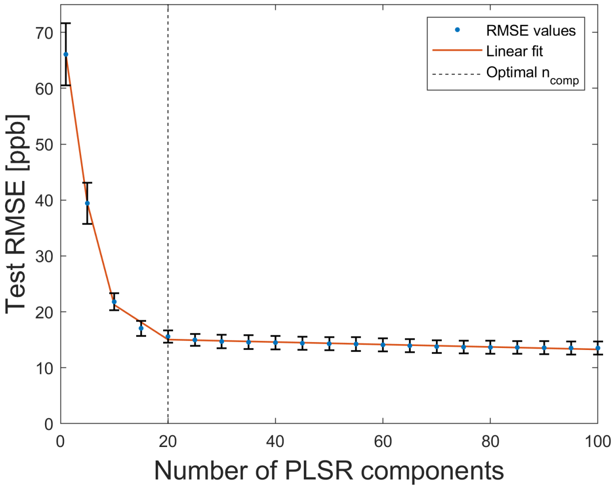

Figure 8. Elbow method applied to the T − RMSE curve for 10 trials. The optimal number of PLSR components is 20.

3.2 Influence of measurement uncertainty on ML results

In this contribution, two approaches for investigating the influence of the measurement uncertainty on machine learning results are considered, namely training a model with raw (see Sect. 3.2.1) and noisy (see Sect. 3.2.2) data, respectively. The trained models are used to predict data with varying noise levels between 0 and 98 dB in both use cases. Training a model with raw data means that the uncertainty associated with the raw data is propagated through the UA-AMLT, which saves on computational cost, as no retraining is necessary if the uncertainty changes. To validate the UA-AMLT, training with the noisy data of different SNRs is carried out and compared to the results of the uncertainty propagation approach. The number of PLSR components ($n_{\text{comp}} = 20$) determined with the elbow method is considered for formaldehyde and VOC_{sum} as the target. For formaldehyde, the number of PLSR components leading to the minimum $\text{T}-\text{RMSE}$ value, i.e., $n_{\text{comp}} = 100$, is also considered and compared to the results for

3.2.1 Model trained with raw data

The motivation for using raw data for training and noisy data for model application is the typical degradation of sensors over time (Jiang et al., 2006). To avoid a loss in sensor performance, periodical recalibration is typically required, which is often expensive and difficult or impossible to perform, as collecting sensors and sending them to the lab leads to the downtime of the IAQ monitoring system.

The test plus uncertainty RMSE ($\text{T}+\text{U}-\text{RMSE}$) is introduced as measure for the quality of the model considering the uncertainty values. This $\text{T}+\text{U}-\text{RMSE}$ value is the sum of the two RMSE values obtained by the test of the model ($\text{T}-\text{RMSE}$) and by propagating the measurement uncertainty through the toolbox ($\text{U}-\text{RMSE}$), respectively. It is calculated according to the following:

$$\boldsymbol{d_u} = \boldsymbol{y}_{\mathbf{pred}} + \boldsymbol{U}_{\boldsymbol{y}_{\mathbf{pred}}} \tag{24}$$

$$\boldsymbol{d_l} = \boldsymbol{y}_{\mathbf{pred}} - \boldsymbol{U}_{\boldsymbol{y}_{\mathbf{pred}}} \tag{25}$$

$$\text{RMSE}_{\text{T+U}} = \begin{cases} \text{RMSE}(\boldsymbol{d_u}, \boldsymbol{y}) & \boldsymbol{y}_{\mathbf{pred}} \geq \boldsymbol{y} \\ \text{RMSE}(\boldsymbol{d_l}, \boldsymbol{y}) & \text{otherwise} \end{cases}, \tag{26}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{y}_{\mathbf{pred}} \in \mathbb{R}^m$ denote the actual and the predicted target, respectively. $\mathbf{U}_{\boldsymbol{y}_{\mathbf{pred}}}$ contains the uncertainty values associated with the predicted target. Furthermore, noisy data RMSE (ND-RMSE) is used, which indicates the quality of the model when applying it to another simulated data set (2000 cycles) with the added white Gaussian noise (noisy data) of different SNRs.

First, it is of interest if the selected FE algorithm still performs well when applying the model trained with raw data on noisy test data. ALA was chosen as the best FE algorithm when applying a model trained with raw data on raw test data, as shown in Sect. 3.1. Applying the model on noisy test data leads to the $\text{T}+\text{U}-\text{RMSE}$ curves, as shown in Fig. 9. For SNR values greater than 65 dB, ALA achieves the smallest $\text{T}+\text{U}-\text{RMSE}$. In this range, the statistical moments (SM) also perform well, with the best Daubechies wavelets (BDW) achieving similar results for very high SNR $\leq$ 85 dB. The $\text{T}+\text{U}-\text{RMSE}$ difference between ALA (best algorithm) and SM is only 1.9 ppb for 98 dB. Between 50 and 65 dB, the smallest $\text{T}+\text{U}-\text{RMSE}$ is achieved using statistical moments. If SNR $\leq$ 50 dB, then PCA achieves the smallest $\text{T}+\text{U}-\text{RMSE}$. This means that PCA can compensate for noise in this range, but overfitting leads to higher error, as shown above for the raw data. This figure shows that the measurement uncertainty has a direct influence on the performance of the ML algorithm, and thus, different FE methods should be chosen for different SNR values.

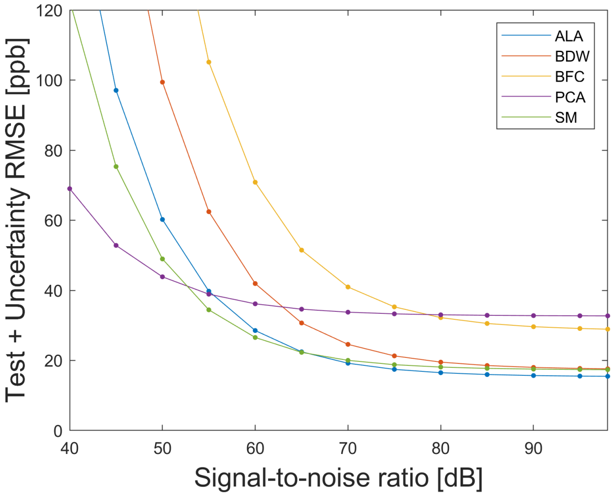

Figure 9. Test plus uncertainty RMSE ($\text{T}+\text{U}-\text{RMSE}$) curve for the five complementary FE algorithms, each in combination with the Pearson correlation for FS and PLSR.

Figure 10a shows $\text{T}-\text{RMSE}$, $\text{T}+\text{U}-\text{RMSE}$, and ND-RMSE values for a model trained on raw data for SNR $\geq$ 40 dB (approx. maximum theoretical SNR for a 6 bit ADC, according to Eq. 21), using the data split shown in Fig. 5 with ALA as FE, the Pearson correlation for FS, and PLSR. The $\text{T}-\text{RMSE}$ values in Fig. 10 are constant because the model was trained with one specific raw data split (see Fig. 5). For large SNR values, it can be assumed that the added white Gaussian noise is smaller than the SNR of the raw data and, therefore, has no significant influence, as is indeed observed. The $\text{T}+\text{U}$ and ND errors show a similar increase with reduced SNR for both models with 20 and 100 PLSR components, with the ND error being slightly lower than the $\text{T}+\text{U}$ error. This indicates that the model uncertainty estimated by propagating the error through the toolbox, i.e., the $\text{T}+\text{U}$ error, overestimates the true model uncertainty slightly but still provides valuable insight into the sensitivity of the ML model to noisy data. To obtain an accurate model for predicting formaldehyde concentrations with $n_{\text{comp}} = 20$, the SNR of the data set should not fall below 70 dB, as, for this SNR, the $\text{T}+\text{U}-\text{RMSE}$ is approx. 19.2 ppb, which is an acceptable uncertainty for determining the formaldehyde concentration with a threshold limit value (TLV) of 81 ppb. An SNR of 45 dB for $n_{\text{comp}} = 20$ and of 50 dB for $n_{\text{comp}} = 100$ results in $\text{T}+\text{U}-\text{RMSE}$ and ND-RMSE values of approx. 80 ppb, i.e., similar to the TLV, which means that the sensor system would no longer be useful for estimating the formaldehyde concentration. For SNR < 80 dB, the model based on 20 PLSR components is more robust against noise than a model with 100 PLSR components, yielding lower RMSE values. In contrast, for SNR $\geq$ 80 dB, a higher number of

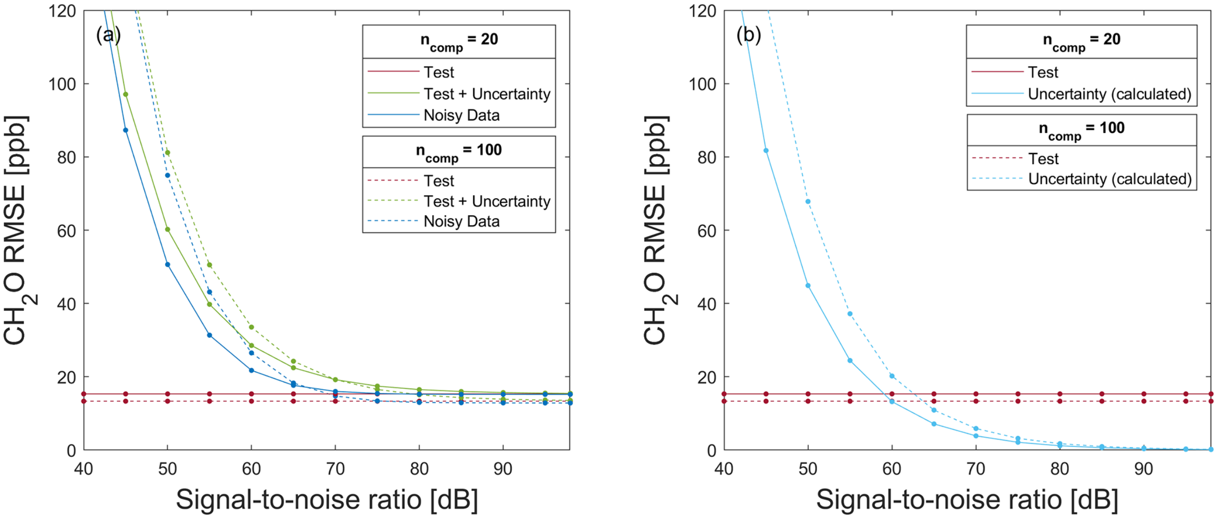

Figure 10. RMSE for testing a model trained with 80 % raw data for formaldehyde prediction on **(a)** 20 % test data without (red) and with associated uncertainty values (green), in addition to the application of the model on a noisy data set (blue) and **(b)** 20 % raw test data (red), and the calculated uncertainty RMSE (light blue) resulting from the difference in T + U − RMSE and T − RMSE for a different number of PLSR components.

PLSR components performs slightly better, i.e., an improvement in the RMSE values can be achieved with more PLSR components if the noise level in the data is very low. In the case of an SNR value of 98 dB, the T − RMSE value is 15.33 and 13.34 ppb for $n_{\mathrm{comp}} = 20$ and $n_{\mathrm{comp}} = 100$, respectively.

Figure 10b shows the T − RMSE and the U − RMSE. U − RMSE is calculated as the difference between T + U − RMSE and T − RMSE (see Fig. 10a). This figure shows that the influence of the trained model (expressed by T − RMSE) on T + U − RMSE is constant, while the influence of the measurement uncertainty (expressed by U − RMSE) decreases steadily with increasing SNR. For $n_{\mathrm{comp}} = 20$ ($n_{\mathrm{comp}} = 100$), U − RMSE is smaller than T − RMSE when SNR is greater than 60 dB (65 dB).

The results for the additive white uniform noise and formaldehyde as target are nearly the same as for the additive white Gaussian noise (see Fig. A8a). Similar results for $\mathrm{VOC}_{\mathrm{sum}}$ as the target are shown in Fig. A5a.

To demonstrate the effect of the noise on test data, PLSR models trained with raw data ($n_{\mathrm{comp}} = 20$) for the quantification of formaldehyde and $\mathrm{VOC}_{\mathrm{sum}}$ are shown in Figs. A1 and A6 for the two different SNR values, respectively.

3.2.2 Model training with noisy data

The second use case occurs when using low-performance sensors or sensor systems that provide significant noisy data or where the electronics/ADCs add significant noise. For the investigation of the influence of measurement uncertainty on regression results, ALA as FE and Pearson correlation as FS are used together with PLSR. Formaldehyde as the target is discussed here, as $\mathrm{VOC}_{\mathrm{sum}}$ leads to similar results, which are shown in Appendix A2. Only results for white Gaussian noise are shown here, as the results for white uniform noise are similar (see Appendix A3).

Figure 11a shows T − RMSE, T + U − RMSE, and ND-RMSE values for a model trained on noisy data for SNR ≥ 40 dB, using the data split shown in Fig. 5. Compared to Fig. 10a, the T + U − RMSE is significantly smaller for the model trained with noisy data, i.e., the ML model can suppress noise if it is contained in the training data. For example, for SNR = 50 dB and 20 PLSR components, the T + U − RMSE is 60.22 ppb when training with raw data, while it is only 34.3 ppb when training with noisy data. In general, a model can be made more noise resistant by adding additive white Gaussian noise to the training data. Comparing Figs. 10a and 11a, note that the regression results are similar for noisy and raw data for SNR ≥ 80 dB, thus indicating again that the noise level of the raw data is approx. 80 dB. The same holds for T + U − RMSE when SNR ≥ 80 dB. Of course, there is no need to train the model with noisy data with an added noise level lower than the noise of the original data. As already observed for the model trained with raw data, the RMSE values can be reduced by using more components, but here, this observation holds for all SNR levels, as the noise is contained in the training data, so there is no overfitting. This means that training a model with raw data once is sufficient, and no new model must be trained with noisy data. The associated measurement uncertainty values must only be used in the application of the model, which saves much computational cost. Figure 11b shows that, for T + U − RMSE values, the contribution resulting from the measurement uncertainty is always lower than the contribution from the test of the model. This means that the noise is already trained in

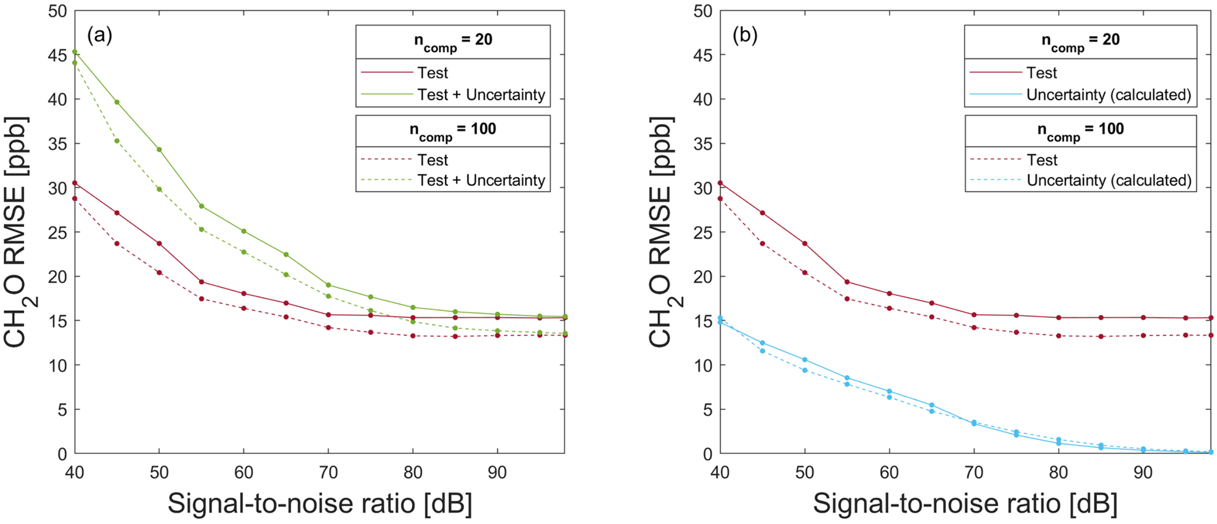

Figure 11. RMSE for testing a model trained with 80 % noisy data for formaldehyde prediction on **(a)** 20 % test data without (red) and with associated uncertainty values (green) and **(b)** 20 % raw test data (red) and the calculated uncertainty RMSE (light blue) resulting from the difference in T + U − RMSE and T − RMSE for a different number of PLSR components. Note the magnification compared to Fig. 10.

the model, and improving the used sensor would significantly improve the ML results.

For white uniform noise, similar results are shown in Fig. A8b.

In case of VOC_{sum} as the target, the results are similar, despite the fact that the RMSE values are higher than for formaldehyde, as shown in Fig. A5b. No significant difference between the RMSE values when training with raw and noisy data, respectively, is observed for SNR values higher than 70 dB.

To demonstrate the effect of noise on test data, PLSR models trained with noisy data ($n_{comp} = 20$) for the quantification of formaldehyde and VOC_{sum} are shown in Figs. A2 and A7, for two different SNR values, respectively.

4 Conclusion and outlook

In this contribution, the uncertainty-aware AMLT for classification tasks presented in Dorst et al. (2022) was first extended for solving regression problems. In accordance with the GUM, an analytical method for uncertainty propagation of PLSR was implemented. The code for this UA-AMLT for classification and regression tasks was published on GitHub (https://github.com/ZeMA-gGmbH/LMT-UA-ML-Toolbox, last access: 18 January 2023). For different SNR levels, the UA-AMLT automatically selects the best ML algorithm based on the overall test plus the uncertainty RMSE.

The influence of measurement uncertainty on machine learning results is investigated in depth with two use cases, namely model training with raw and noisy data generated by adding white Gaussian noise. For both use cases, the analysis shows where the measurement system must be improved to achieve better ML results. In general, there are two distinct possibilities, i.e., improving either the ML model or the used sensor. In case of an RMSE resulting from measurement uncertainty tending towards zero, an improvement of the ML model is suggested. In the range where U − RMSE is already very small (see Fig. 10b), a better ML model should be obtained as optimizing the sensor, including the data acquisition electronics, will only lead to even lower U − RMSE values close to zero, which does not significantly impact the overall T + U − RMSE. In contrast to that, in ranges where U − RMSE is higher, minimizing this RMSE by optimizing the physical sensor system should be the objective. To reduce the T − RMSE resulting from the ML model, using a better model would be necessary, as this can significantly influence the ML results. A better model can be achieved, for example, by using a higher number of PLSR components, as shown in this contribution, or by using deep learning, which can also improve the T − RMSE (Robin et al., 2021).

Finally, it is shown that increased robustness of the machine learning model can be achieved by adding white Gaussian noise to the raw training data.

In future work, the influence of different types of colored noise on ML results can be investigated, as this contribution has addressed only different additive white noise models. Therefore, the correlation must be considered within the uncertainty propagation, and this is only possible for the feature extractors. Furthermore, the difference between noise produced by the data acquisition electronics, especially the logarithmic amplifier as simulated in this contribution, and noise produced by the sensor could be investigated. To simulate sensor noise or electronic noise before the logarithmic amplifier noise, the noise must already be added to the inverse logarithmic of the logarithmic resistance raw data.

Appendix A: Additional figures

A1 Formaldehyde as target

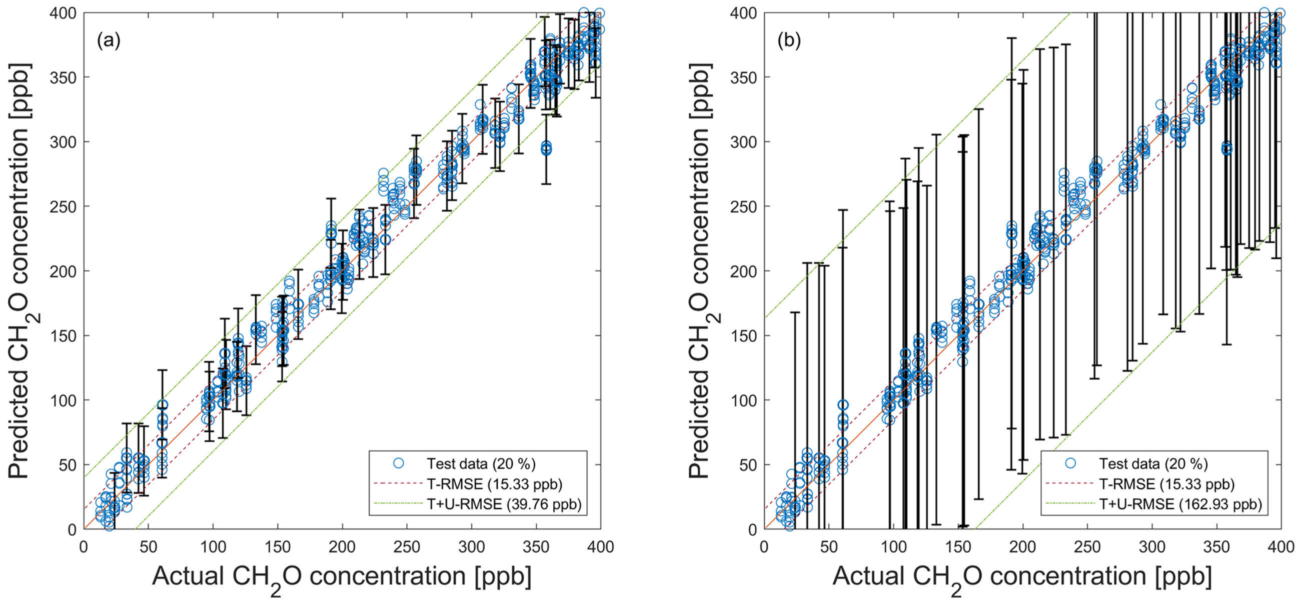

Figure A1. PLSR model (trained with raw data; $n_{\text{comp}} = 20$) applied to test data (see Fig. 5) for the quantification of formaldehyde and the propagated uncertainty. **(a)** SNR = 55 dB. **(b)** SNR = 40 dB. Dashed red and green lines indicate the test RMSE (T − RMSE) and the test plus uncertainty RMSE (T + U − RMSE) based on test data, respectively. For better visibility, error bars are only shown for every 10th prediction.

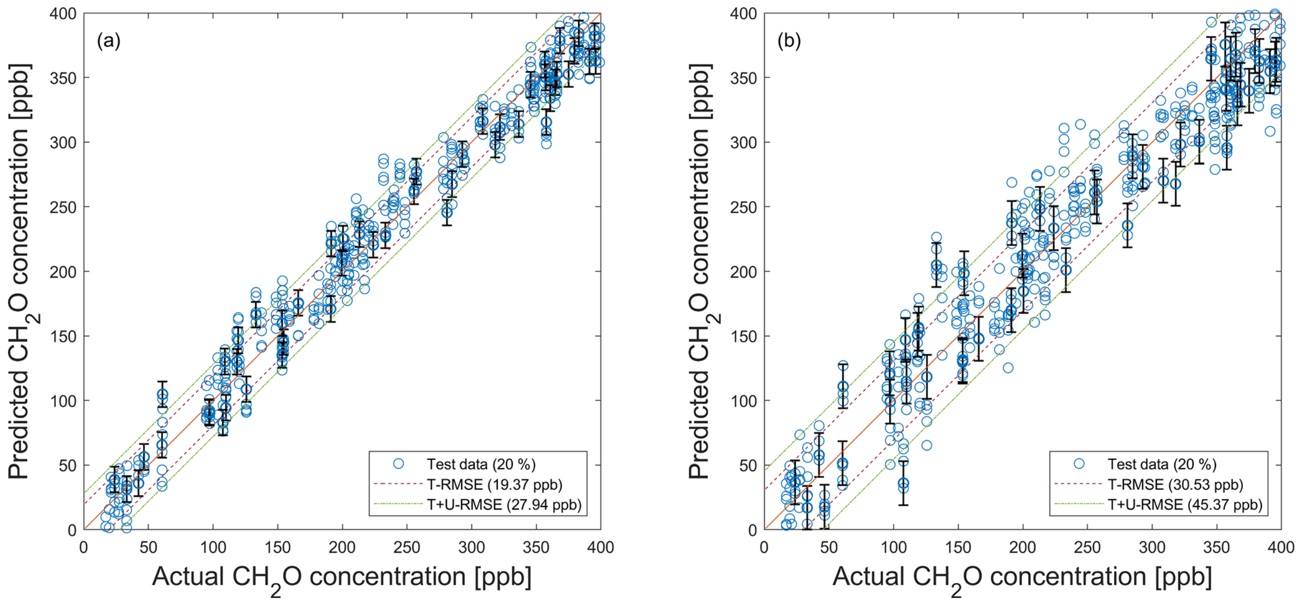

Figure A2. PLSR model (trained with noisy data; $n_{\text{comp}} = 20$) applied to noisy test data (see Fig. 5) for the quantification of formaldehyde using their associated standard uncertainty. **(a)** SNR = 55 dB. **(b)** SNR = 40 dB. Dashed red and green lines indicate the test RMSE (T − RMSE) and the test plus uncertainty RMSE (T + U − RMSE) based on test data, respectively. For better visibility, error bars are only shown for every 10th prediction.

A2 VOC_{sum} as target

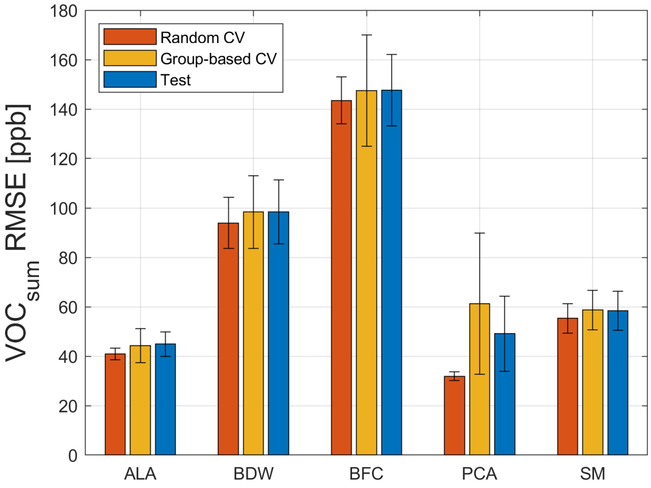

Figure A3. Random CV, group-based CV, and test RMSE of the five FE algorithms, using Pearson as FS and PLSR with $n_{comp} = 20$ for 100 trials with different data splits and the VOC_{sum} concentration as target. ALA is the adaptive linear approximation, BDW is the best Daubechies wavelets, BFC is the best Fourier coefficients, PCA is the principal component analysis, and SM is the statistical moments.

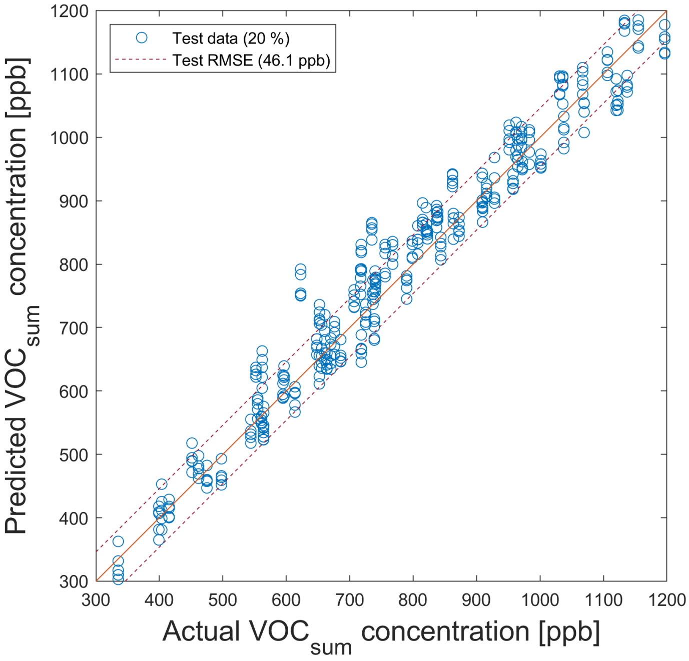

Figure A4. PLSR model for the quantification of VOC_{sum} for testing with test data from the data split shown in Fig. 5. Dashed lines indicate the RMSE of test data (T − RMSE).

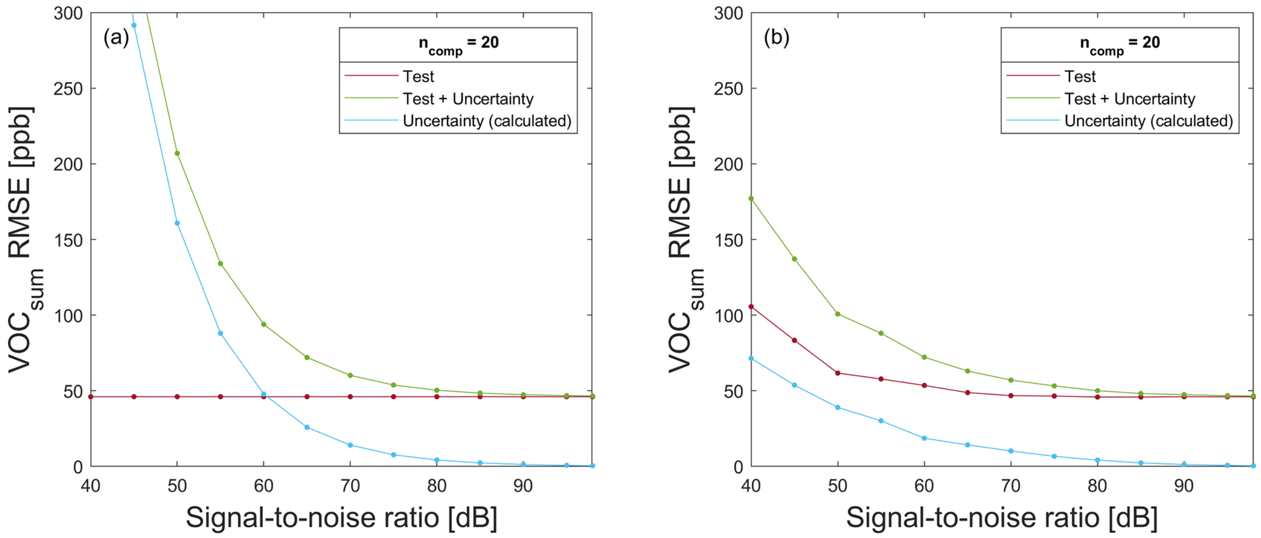

Figure A5. RMSE for testing a model trained with 80 % **(a)** raw data and **(b)** noisy data for $\mathrm{VOC_{sum}}$ prediction on 20 % test data without (red) and with associated uncertainty values (green), in addition to the calculated uncertainty RMSE (blue) resulting from the difference in $\mathrm{T+U-RMSE}$ and $\mathrm{T-RMSE}$.

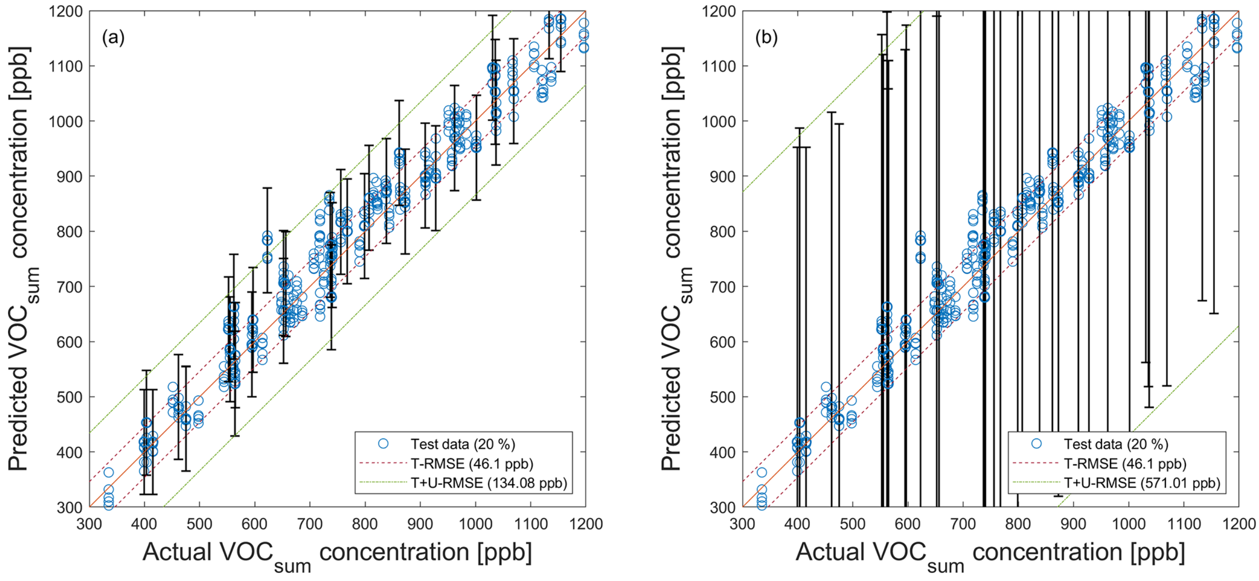

Figure A6. PLSR model (trained with raw data; $n_{\mathrm{comp}} = 20$) applied to test data (see Fig. 5) for the quantification of $\mathrm{VOC_{sum}}$ and propagated uncertainty. **(a)** $\mathrm{SNR} = 55\,\mathrm{dB}$. **(b)** $\mathrm{SNR} = 40\,\mathrm{dB}$. Dashed red and green lines indicate the test RMSE ($\mathrm{T-RMSE}$) and the test plus uncertainty RMSE ($\mathrm{T+U-RMSE}$) based on test data, respectively. For better visibility, error bars are only shown for every 10th prediction.

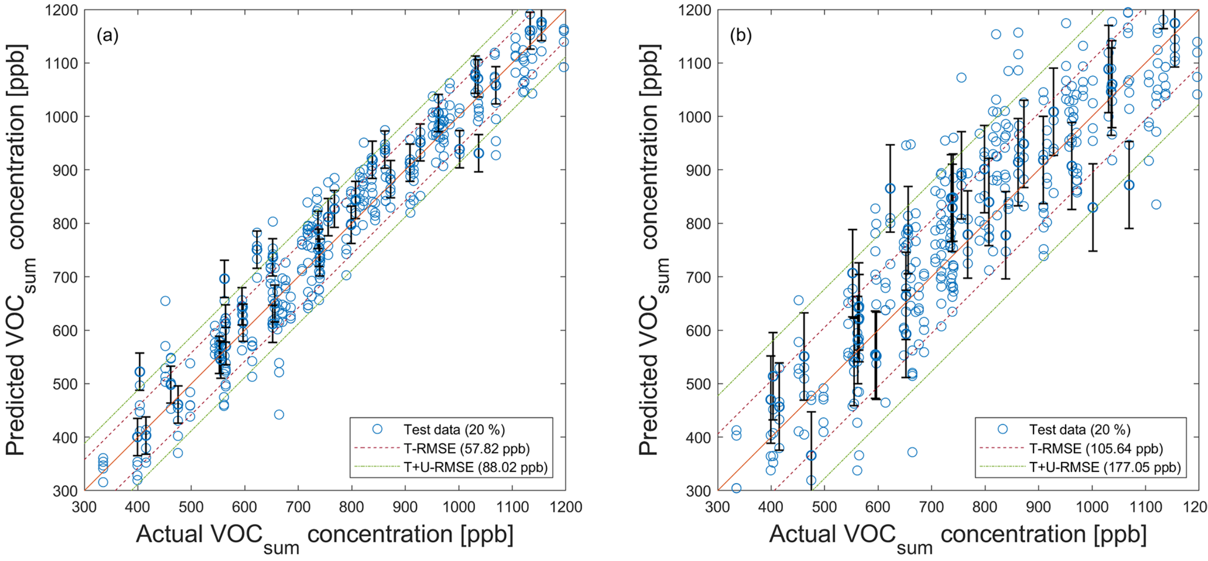

Figure A7. PLSR model (trained with noisy data; $n_{\text{comp}} = 20$) applied to noisy test data (see Fig. 5) for the quantification of VOC$_{\text{sum}}$ using their associated standard uncertainty. **(a)** SNR = 55 dB. **(b)** SNR = 40 dB. Dashed red and green lines indicate the test RMSE (T − RMSE) and the test plus uncertainty RMSE (T + U − RMSE) based on test data, respectively. For better visibility, error bars are only shown for every 10th prediction.

A3 Additive white uniform noise and formaldehyde as target

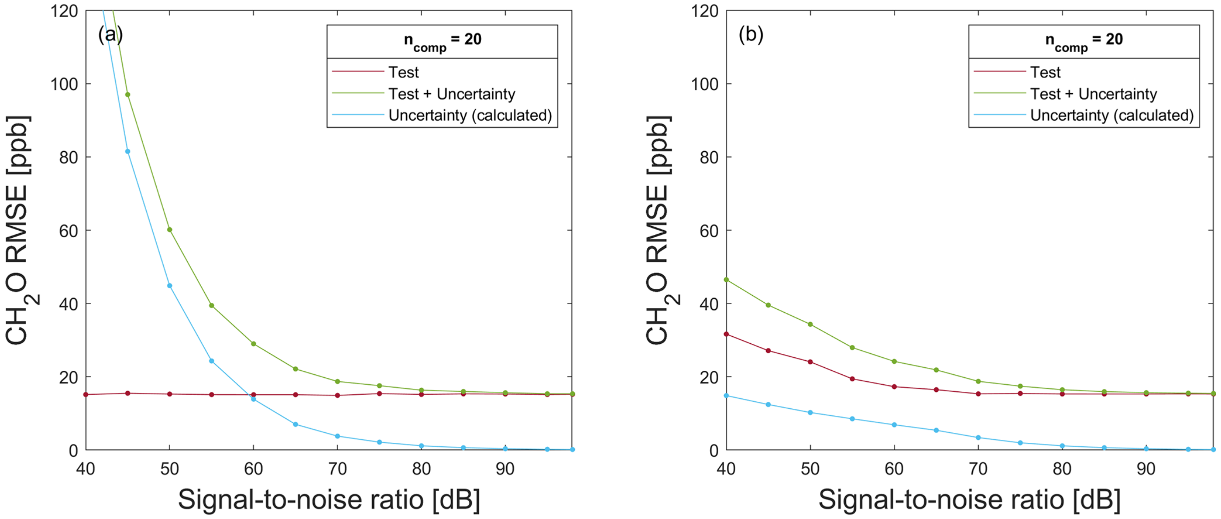

Figure A8. RMSE for testing of a model trained with 80 % **(a)** raw data and **(b)** noisy data (added white uniform noise to raw data) for a formaldehyde prediction on 20 % test data without (red) and with associated uncertainty values (green), in addition to the calculated uncertainty RMSE (blue) resulting from the difference in T + U − RMSE and T − RMSE.

Code and data availability. The paper uses data obtained from different calibration and field test measurements of gas mixtures with a MOS gas sensor. The data set is available on Zenodo https://doi.org/10.5281/zenodo.4593853 (Amann et al., 2021a).

The uncertainty-aware AMLT (Dorst et al., 2022; https://doi.org/10.1515/teme-2022-0042) includes all the code for data analysis associated with the current submission and is available at https://github.com/ZeMA-gGmbH/LMT-UA-ML-Toolbox (last access: 15 April 2022).

Author contributions. TD carried out the analysis, visualized the results, and wrote the original draft of the paper. TS, SE, and AS contributed with substantial revisions.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The uncertainty-aware automated ML toolbox was developed within the project 17IND12 Met4FoF from the EMPIR program co-financed by the Participating States and from the European Union's Horizon 2020 Research and Innovation program. We acknowledge support by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) and Saarland University within the "Open Access Publication Funding" program.

Financial support. This research has been supported by the European Metrology Programme for Innovation and Research (Met4FoF (grant agreement no. 17IND12)) and the European Union's Horizon 2020 Research and Innovation program.

Review statement. This paper was edited by Sebastian Wood and reviewed by two anonymous referees.

References

Amann, J., Baur, T., and Schultealbert, C.: Measuring Hydrogen in Indoor Air with a Selective Metal Oxide Semiconductor Sensor: Dataset, Zenodo [data set], https://doi.org/10.5281/zenodo.4593853, 2021a.

Amann, J., Baur, T., Schultealbert, C., and Schütze, A.: Bewertung der Innenraumluftqualität über VOC-Messungen mit Halbleitergassensoren - Kalibrierung, Feldtest, Validierung, tm - Tech. Mess., 88, S89–S94, https://doi.org/10.1515/teme-2021-0058, 2021b.

Asikainen, A., Carrer, P., Kephalopoulos, S., Fernandes, E. d. O., Wargocki, P., and Hänninen, O.: Reducing burden of disease from residential indoor air exposures in Europe (HEALTHVENT project), Environ. Health, 15, S35, https://doi.org/10.1186/s12940-016-0101-8, 2016.

Baur, T., Schütze, A., and Sauerwald, T.: Optimierung des temperaturzyklischen Betriebs von Halbleitergassensoren, tm - Tech. Mess., 82, 187–195, https://doi.org/10.1515/teme-2014-0007, 2015.

Baur, T., Amann, J., Schultealbert, C., and Schütze, A.: Field Study of Metal Oxide Semiconductor Gas Sensors in Temperature Cycled Operation for Selective VOC Monitoring in Indoor Air, Atmosphere, 12, 647, https://doi.org/10.3390/atmos12050647, 2021.

Bennett, W. R.: Spectra of quantized signals, Bell Syst. Tech. J., 27, 446–472, https://doi.org/10.1002/j.1538-7305.1948.tb01340.x, 1948.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML: JCGM 100: Evaluation of measurement data – Guide to the expression of uncertainty in measurement, https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6 (last access: 18 January 2023), 2008a.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML: JCGM 101: Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement" – Propagation of distributions using a Monte Carlo method, https://www.bipm.org/documents/20126/2071204/JCGM_101_2008_E.pdf/325dcaad-c15a-407c-1105-8b7f322d651c (last access: 18 January 2023), 2008b.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML: JCGM 102: Evaluation of measurement data – Supplement 2 to the "Guide to the expression of uncertainty in measurement" – Extension to any number of output quantities, https://www.bipm.org/documents/20126/2071204/JCGM_102_2011_E.pdf/6a3281aa-1397-d703-d7a1-a8d58c9bf2a5 (last access: 18 January 2023), 2011.

Brasche, S. and Bischof, W.: Daily time spent indoors in German homes – Baseline data for the assessment of indoor exposure of German occupants, Int. J. Hyg. Envir. Heal., 208, 247–253, https://doi.org/10.1016/j.ijheh.2005.03.003, 2005.

Daubechies, I.: Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, https://doi.org/10.1137/1.9781611970104, 1992.

De Jong, S.: PLS fits closer than PCR, J. Chemometr., 7, 551–557, https://doi.org/10.1002/cem.1180070608, 1993a.

De Jong, S.: SIMPLS: An alternative approach to partial least squares regression, Chemometr. Intell. Lab., 18, 251–263, https://doi.org/10.1016/0169-7439(93)85002-X, 1993b.

Dorst, T., Robin, Y., Schneider, T., and Schütze, A.: Automated ML Toolbox for Cyclic Sensor Data, MSMM 2021 – Mathematical and Statistical Methods for Metrology 2021, 149–150, http://www.msmm2021.polito.it/content/download/245/1127/file/MSMM2021_Booklet_c.pdf (last access: 18 January 2023), 2021.

Dorst, T., Schneider, T., Eichstädt, S., and Schütze, A.: Uncertainty-aware automated machine learning toolbox, tm - Tech. Mess., in press, https://doi.org/10.1515/teme-2022-0042, 2022 (code available at: https://github.com/ZeMA-gGmbH/LMT-UA-ML-Toolbox, last access: 18 January 2023).

Eicker, H.: Method and apparatus for determining the concentration of one gaseous component in a mixture of gases, US patent US4012692A, http://www.google.tl/patents/US4012692 (last access: 18 January 2023), 1977.

Ergon, R.: Principal component regression (PCR) and partial least squares regression (PLSR), John Wiley & Sons, Ltd, chap. 8, 121–142, https://doi.org/10.1002/9781118434635.ch08, 2014.

Gutierrez-Osuna, R.: Pattern analysis for machine olfaction: a review, IEEE Sens. J., 2, 189–202, https://doi.org/10.1109/JSEN.2002.800688, 2002.

Hauptmann, M., Lubin, J. H., Stewart, P. A., Hayes, R. B., and Blair, A.: Mortality from solid cancers among workers in formaldehyde industries, Am. J. Epidemiol., 159, 1117–1130, https://doi.org/10.1093/aje/kwh174, 2004.

Horn, R. A.: The Hadamard product, in: Matrix theory and applications, edited by: Johnson, C. R., Proc. Sym. Ap., 40, 87–169, https://doi.org/10.1090/psapm/040/1059485, 1990.

Jackson, J. E.: A Use's Guide to Principal Components, John Wiley & Sons, Inc., https://doi.org/10.1002/0471725331, 1991.

Jiang, L., Djurdjanovic, D., Ni, J., and Lee, J.: Sensor Degradation Detection in Linear Systems, in: Engineering Asset Management, edited by: Mathew, J., Kennedy, J., Ma, L., Tan, A., and Anderson, D., Springer London, London, 1252–1260, https://doi.org/10.1007/978-1-84628-814-2_138, 2006.

Jones, A. P.: Indoor air quality and health, Atmos. Environ., 33, 4535–4564, https://doi.org/10.1016/S1352-2310(99)00272-1, 1999.

Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, in: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 20–25 August 1995, IJCAI'95, 2, 1137–1143, 1995.

Lee, A. P. and Reedy, B. J.: Temperature modulation in semiconductor gas sensing, Sensor. Actuat. B-Chem., 60, 35–42, https://doi.org/10.1016/S0925-4005(99)00241-5, 1999.

Martin, H. R. and Honarvar, F.: Application of statistical moments to bearing failure detection, Appl. Acoust., 44, 67–77, https://doi.org/10.1016/0003-682X(94)P4420-B, 1995.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21, 239–245, https://doi.org/10.2307/1268522, 1979.

Mörchen, F.: Time series feature extraction for data mining using DWT and DFT, Department of Mathematics and Computer Science, University of Marburg, Germany, Technical Report, 33, 1–31, 2003.

NTP (National Toxicology Program): Report on Carcinogens, 15th edn., https://doi.org/10.22427/NTP-OTHER-1003, 2021.

Olszewski, R. T., Maxion, R. A., and Siewiorek, D. P.: Generalized feature extraction for structural pattern recognition in time-series data, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, https://www.cs.cmu.edu/~bobski/pubs/tr01108-twosided.pdf (last access: 18 January 2023), 2001.

Pearson, K.: LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559–572, https://doi.org/10.1080/14786440109462720, 1901.

Reams, R.: Hadamard inverses, square roots and products of almost semidefinite matrices, Linear Algebra Appl., 288, 35–43, https://doi.org/10.1016/S0024-3795(98)10162-3, 1999.

Robin, Y., Amann, J., Baur, T., Goodarzi, P., Schultealbert, C., Schneider, T., and Schütze, A.: High-Performance VOC Quantification for IAQ Monitoring Using Advanced Sensor Systems and Deep Learning, Atmosphere, 12, 1487, https://doi.org/10.3390/atmos12111487, 2021.

Rüffer, D., Hoehne, F., and Bühler, J.: New Digital Metal-Oxide (MOx) Sensor Platform, Sensors, 18, 1052, https://doi.org/10.3390/s18041052, 2018.

Schneider, T., Helwig, N., and Schütze, A.: Automatic feature extraction and selection for classification of cyclical time series data, tm - Tech. Mess., 84, 198–206, https://doi.org/10.1515/teme-2016-0072, 2017.

Schneider, T., Helwig, N., and Schütze, A.: Industrial condition monitoring with smart sensors using automated feature extraction and selection, Meas. Sci. Technol., 29, 094002, https://doi.org/10.1088/1361-6501/aad1d4, 2018.

Schultealbert, C., Baur, T., Schütze, A., and Sauerwald, T.: Facile Quantification and Identification Techniques for Reducing Gases over a Wide Concentration Range Using a MOS Sensor in Temperature-Cycled Operation, Sensors, 18, 744, https://doi.org/10.3390/s18030744, 2018.

Schütze, A. and Sauerwald, T.: Dynamic operation of semiconductor sensors, in: Semiconductor Gas Sensors, 2nd edn., edited by: Jaaniso, R. and Tan, O. K., Woodhead Publishing Series in Electronic and Optical Materials, Woodhead Publishing, 385–412, https://doi.org/10.1016/B978-0-08-102559-8.00012-4, 2020a.

Schütze, A. and Sauerwald, T.: Indoor air quality monitoring, in: Advanced Nanomaterials for Inexpensive Gas Microsensors, edited by: Llobet, E., Micro and Nano Technologies, Elsevier, 209–234, https://doi.org/10.1016/B978-0-12-814827-3.00011-6, 2020b.

Sensirion AG: Datasheet SGP30, https://sensirion.com/media/documents/984E0DD5/61644B8B/Sensirion_Gas_Sensors_Datasheet_SGP30.pdf (last access: 18 January 2023), 2020.

Spaul, W. A.: Building-related factors to consider in indoor air quality evaluations, J. Allergy Clin. Immun., 94, 385–389, 1994.

Sundell, J.: On the history of indoor air quality and health, Indoor air, 14, 51–58, 2004.

Thorndike, R. L.: Who belongs in the family?, Psychometrika, 18, 267–276, https://doi.org/10.1007/BF02289263, 1953.

Tsai, W.-T.: An overview of health hazards of volatile organic compounds regulated as indoor air pollutants, Rev. Environ. Health, 34, 81–89, https://doi.org/10.1515/reveh-2018-0046, 2019.

Von Pettenkofer, M.: Über den Luftwechsel in Wohngebäuden, Cotta, München, https://opacplus.bsb-muenchen.de/title/BV013009721 (last access: 18 January 2023), 1858.

Wold, S., Albano, C., Dunn, W. J., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjöström, M.: Multivariate Data Analysis in Chemistry, in: Chemometrics: Mathematics and Statistics in Chemistry, edited by: Kowalski, B. R., Springer, Dordrecht, Netherlands, 17–95, https://doi.org/10.1007/978-94-017-1026-8_2, 1984.

Wold, S., Sjöström, M., and Eriksson, L.: PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab., 58, 109–130, https://doi.org/10.1016/S0169-7439(01)00155-1, 2001.

World Health Organization (WHO): WHO guidelines for indoor air quality: selected pollutants, WHO Regional Office for Europe, Copenhagen, Vol. 9, ISBN: 978-9-2890-0213-4, 2010.

Zhang, L.: Formaldehyde, Issues in Toxicology, The Royal Society of Chemistry, https://doi.org/10.1039/9781788010269, 2018.