Nonlinear calibration transfer based on hierarchical Bayesian models and Lagrange Multipliers: Error bounds of estimates via Monte Carlo - Markov Chain sampling

Felicia Seichter^a, Josef Vogt^{b,*}, Peter Radermacher^b, Boris Mizaikoff^{a,*}

^a*Institute of Analytical and Bioanalytical Chemistry, Ulm University, 89081 Ulm, Germany*

^b*Institute of Anesthesiologic Pathophysiology and Method Development, Ulm University Medical Center, 89081 Ulm, Germany*

Abstract

The calibration of analytical systems is time-consuming and the effort for daily calibration routines should therefore be minimized, while maintaining the analytical accuracy and precision. The 'calibration transfer' approach proposes to combine calibration data already recorded with actual calibrations measurements. However, this strategy was developed for the multivariate, linear analysis of spectroscopic data, and thus, cannot be applied to sensors with a single response channel and/or a non-linear relationship between signal and desired analytical concentration.

To fill this gap for a non-linear calibration equation, we assume that the coefficients for the equation, collected over several calibration runs, are normally distributed. Considering that coefficients of an actual calibration are a sample of this distribution, only a few standards are needed for a complete calibration data set.

The resulting calibration transfer approach is demonstrated for a fluorescence oxygen sensor and implemented as a hierarchical Bayesian model, combined with a Lagrange Multipliers technique and Monte-Carlo Markov-Chain sam-

*Corresponding author

Email addresses: josef.vogt@uni-ulm.de (Josef Vogt), boris.mizaikoff@uni-ulm.de (Boris Mizaikoff)

URL: www.uni-ulm.de/iabc (Boris Mizaikoff)

Preprint submitted to Analytica Chimica Acta

September 8, 2016

pling. The latter provides realistic estimates for coefficients and prediction together with accurate error bounds by simulating known measurement errors and system fluctuations. Performance criteria for validation and optimal selection of a reduced set of calibration samples were developed and lead to a setup which maintains the analytical performance of a full calibration. Strategies for a rapid determination of problems occurring in a daily calibration routine, are proposed, thereby opening the possibility of correcting the problem just in time.

Keywords: calibration transfer, Lagrange Multipliers, fluorescence, hierarchical models, oxygen sensor, Bayesian statistics

2010 MSC: 68N30, 65C05, 41A20, 6207, 3204, 32F25, 62J12, 32M02, 62P99, 65C40, 65D10, 70H03, 92F99

PACS: 02.70.Uu, 02.60.Ed, 07.07.Df, 42.81.Pa, 33.50.Dq, 02.50.Ng, 07.07.Df, 02.50.-r

1. Introduction

One key step when setting up an analysis system or developing a quantification technique for a particular analyte is designing an appropriate calibration strategy. For routine use, time efficiency and economic viability should be considered next to the analytical performance, as a substantial effort is required for determining an appropriate calibration model/function and maintaining it. Hence, one may be inclined to reuse a once established calibration model as long as possible in the day-to-day routine, but this may conflict with the demands on accuracy and precision. We therefore herein consider a technology for measuring a scalar signal that is supposed to be converted to a scalar analyte property, such as chemical concentration, using a calibration curve. One can define a theoretical calibration equation based on the physical measurement process; however, the physical properties could be affected by sensor imperfections, which are usually not captured by such a theoretical equation. If these imperfections cannot be eliminated by simple means and add to a nonlinear relationship between measurable signal and analyte quantity, one may resort to

empirical functions.

Empirical functions are defined via signal values at certain data or sampling points and allow for a smooth interpolation for intermediate values of the independent variable. Polynomial functions can cover any smooth curvature, as long as the polynomial order is high enough, but cannot efficiently approximate functions with one inflection point connecting two asymptotic lines, i.e. a response as shown for example in Figure 1. For a good fit to reference grid points, a higher order polynomial ($n > 10$) is necessary. But when interpolating between the grid points, the polynomial tends to oscillate around the expected curve. These effects can be ameliorated by regularization [1], yet, series of well-defined calibration samples are needed for an accurate approximation [2]. A rational function could better approximate the above mentioned curve profile with only few coefficients. However, for a statistically reliable determination and for an acceptable error performance, the number of calibration samples should at least exceed the number of calibration coefficients. This may require an effort one cannot spare in routine calibration scenarios. We consider the case where the general shape of the calibration curve does not distinctively change from one calibration to the next or may be comparable between different but identically constructed instruments. Nevertheless, it is assumed that a previously established calibration cannot be used at present because the difference between the actual sensor response and a previously measured one exceeds the actual machine precision in some range of the analyte concentration, thus leading to a systematic calibration error. As a historic calibration may not be used directly, it would be desirable to exploit at least the fact that all calibration curves share a similar appearance. This previously gathered information should therefore be smartly combined with the information obtained from just one or two actual calibration samples yielding an useful complete calibration data set with minimal additional effort.

Calibration transfer is a suitable approach to solve this problem [3] for a linear multivariate quantification. Corresponding strategies were primarily developed in the field of near infrared spectroscopy. The different strategies comprise a way

to remove the fluctuation of sensor status/difference between sensors etc. via multivariate transformation (i.e., standardization). Typically, complete actual spectra or some local features, collected for a certain reference condition, are transformed to match spectra found during the initial calibration, and then the calibration is applied to the transformed spectra[4] [3]. In addition, multivariate methods/preprocessing steps can be explored that are more robust against variance effects. The relative merits of such calibration transfer strategies are discussed in many publications; Noord as well as Feudale et al. give an excellent overview on the topic [4] [3]. While most publication refer to multivariate linear systems, there is no straight forward way to adapt or modify these calibration transfer schemes to a system that (i) only provides an univariate signal (i.e. like a chemical sensor) and/or (ii) is subject to a nonlinear response function. To the best of our knowledge, no corresponding strategies can be found in literature, and are thus the scope of this contribution.

As a solution, we propose an approach that is based on the assumption that calibration curves, collected at comparable conditions, form a family of curves whose similarity can be captured via appropriate statistics. Statistical concepts characterizing a group of lines collected under comparable conditions have evolved over time. In 1965, Rao[5] considered growth curves with random coefficients and provided equations to estimate the variability of the coefficients. Rosenberg [6] extended this approach in 1973 for curves collected at different subjects and described the inter-subject variability of curves with a multivariate normal distribution for the curve coefficients. Harville [7] developed formulas to assess coefficients for an individual curve and their covariance, if measurements for the individual curve next to group mean values and the group covariance were available. Laird [8] provided an efficient EM algorithm to estimate both the coefficients for individual curves and their group distribution. This class of approaches was termed 'random coefficients' model or regression, and has been summarized in a seminal textbook by Longford[9]. Herein, an improved approach to estimate the group-variance was presented. Examples how to adapt the random coefficient approach to multivariate problems were provided and

applied to curves with any shape that can be covered by polynomial functions
 in the independent variable. Also discussed is the gain in precision of the deter-
 mined polynomial coefficients, if a curve is treated as member of group of differ-
 ent curves, compared to the precision obtained if only actual measurements were
 used to define the polynomial. This gain in precision paves the way for a coeffi-
 cient based calibration transfer, as proposed in our present study. The random
 coefficient approach has been applied to a scalar NDIR (non-dispersive infrared)
 sensor signal that relies on two independent variables[2]. The corresponding re-
 sponse surface was described by a polynomial in the two independent variables.
 A principal component analysis on the variance of the polynomial coefficients,
 collected over several calibrations, allowed isolating a subspace comprising the
 potential values of the coefficients, driven by day-to-day variability. The dimen-
 sion of this subspace is much smaller than the number of unknown coefficients.
 Thus, confining the coefficients to this subspace dramatically reduced the num-
 ber of samples necessary for an actual calibration. Polynomials appear to be
 the method of choice since they allow the retracing of almost any shape of cali-
 bration curve. However, as discussed before, a 'regularization' or smoothing of
 the polynomial is necessary. The optimal tuning for such a smoothing is critical
 and we are not aware of any applications linked with random coefficient models.
 As we will demonstrate later, only a few coefficients are necessary to approx-
 imate curve such as described above with an rational function that, unfortu-
 nately, is nonlinear in the coefficients. However, solely for calibration curves
 that are linear in their coefficients, such as polynomials, one can easily derive
 formulas for the distribution of values predicted by the calibration curve, i.e.
 concentration values for a given signal measurement. Such a distribution con-
 siders the propagation of the uncertainty in the parameters defining the calibra-
 tion curve and in the measurements to predicted concentrations. Corresponding
 formulas are difficult to define for calibration equations that are inherently non-
 linear. Approximate solutions for nonlinear equations have been provided in
 [10][11], however, precise estimates for confidence ranges have to be assessed
 via bootstrapping or resampling methods[12], which in turn require an iteration

over different sets of curves, usually providing only approximate statistical features.

Hence, we turn to the Bayesian equivalent to the random coefficient approach, the ‘hierarchical models’[13][14][15][16][17]. According to Gelman[13][14], Bayesian statistics is based on the explicit use of probabilities for quantifying uncertainty in inferences based on a statistical analysis. The first step in a Bayesian data analysis is setting up a full probability model or definition of a joint probability distribution for all observable and unobservable quantities in a problem. Applied to a calibration problem, this means defining a probability in which specific signals can be observed for samples with a given analyte composition. For such a definition, one needs a function, the calibration equation, which defines a functional relation between analyte in the sample and the signal, observed as response of the sensors. The shape of calibration curves is characterized by a set of coefficients $\mathbf{c}$, in the simplest case slope and intercept. In Bayesian methods, these coefficients are treated as stochastic instead of being an unknown deterministic quantity like in the more classical ‘frequentist’ framework [18]. As a stochastic quantity, the coefficients have a distribution, denoted as $P(\mathbf{c})$. For a full probability model, one also has to consider a measurement error – here we refer to it as ϵ . Let $\mathcal{D}$ denote the observed data values, in the present case the sensor response. One important distribution for the full probability model is the likelihood that $\mathcal{D}$ can be found for given set of calibration coefficients, and for given measurement error – it is denoted as $P(\mathcal{D}|\mathbf{c}, \epsilon)$. In a traditional approach, this likelihood would be maximized to obtain an estimate for the coefficients $\mathbf{c}$, which is usually solved by regression. In an Bayesian context however, the full probability model is defined as the joint probability distribution $P(\mathcal{D}|\mathbf{c}, \epsilon)P(\mathbf{c})$.

The next step in a Bayesian analysis is termed *conditioning on data* or estimating the conditional probability distribution of the calibration coefficients given the observed data. According to the Bayes’ Theorem[19][14][15], this conditional probability is is proportional to the just mentioned product, or:

$$P(\mathbf{c}|\mathcal{D}, \epsilon) \propto P(\mathcal{D}|\mathbf{c}, \epsilon)P(\mathbf{c}) \quad (1)$$

The distribution on the left side of equation (1), $P(\mathbf{c}|\mathcal{D}, \epsilon)$, is the posterior
 distribution. By obtaining a posterior distribution, interval estimates can be
 constructed and uncertainty information about estimates provided, since in
 Bayesian analysis, these intervals represent the most probable region of model
 parameters based on all available information[18]. In equation (1), the distri-
 bution $P(\mathbf{c})$ acts as a *prior* for the calibration coefficients. *Priors* reflects any
 prior knowledge of the coefficients and uncertainty achieved from earlier data.
 Most the time, so-called vague flat priors (with a uniform distribution) are used
 such that the posterior distribution is mainly affected by the data. As an ad-
 vantage of Bayesian analysis, information based on pre-existing knowledge can
 be implemented using priors[20].
 The key feature of the hierarchical approach is that the prior distribution $P(\mathbf{c})$
 is treated as an unknown, which, in the present case, has to be derived from
 data of the different calibration runs. It reflects the day to day distribution
 of the coefficients and allows implementing the concept that an actual calibra-
 tion data set belongs to a group of independently performed calibrations with
 well defined group mean values for their calibration coefficients and case-to-case
 variance. With Bayesian analysis, a descriptive probability model to observe
 data given some parameters can be linked with a likelihood function that re-
 flects our prior knowledge about the involved parameters. It allows calculating
 a posterior distribution for predicted quantities and with it the uncertainty in
 prediction. This posterior distribution reflects what we should believe about the
 value of the prediction, given the experimental data[20]. In a similar logic, a
 predictive distribution can be defined for estimating the analyte concentrations,
 which is based on new measurements and the determined coefficients. The cor-
 responding distribution for prediction is based on a prediction equation, derived
 by converting the calibration equation and the coefficient distribution given in
 equation (1) serves as a prior. The product of distributions used in the right
 side of equation 1 or that for the predictive distribution might be difficult to
 convert into an explicit or analytically given distribution of coefficients for non-
 linear calibration equations, not to mention the predictive distribution. At first

170 glance, this mirrors the situation described for the random coefficients models.
 However, as a big difference, such an analytical solution is not required. Most
 statistical Bayesian software packages simply 'draw a valid sample' from such a
 distribution product and reconstruct the resulting ('posterior') distribution from
 a large set of samples. Corresponding software packages(e.g. 'Winbugs'[21]) are
 175 freely available with tutorial examples of hierarchical regression models, such
 as the 'orange tree' example for nonlinear hierarchical models [10](Winbugs,
 example volumes II). Carlin[22, chapter 17] discusses implementations of a lin-
 ear hierarchical model and Gelman et al. [13] consider hierarchical models
 applied on functions that are nonlinear in the unknown coefficients. Given
 180 their versatility, Bayesian statistics are therefore slowly making their way into
 chemometrics[18][23][24].
 In this study, we have developed a Bayesian calibration transfer model imple-
 mented via the software package 'Stan'[25]. 'Stan' allows defining a statistical
 'model' for almost any aspect of Bayesian statistics. Based on such a model it
 185 derives the posterior probability distribution of the model parameters from a
 prior probability distribution and a likelihood function. It is possible to define
 different priors and probability distributions for every parameter to be deter-
 mined. A particular feature in Stan is that it is very effective at converg-
 ing from diffuse random initialization, though user-specified values can also be
 190 implemented[26][27]). Stan utilizes a compact and concise scripting language
 such that the algorithm developed for the specific situation is readily shared
 and reproduced among different users. Moreover, error propagation from mea-
 surements or from the uncertainty in the case-to-case variability towards the
 prediction for curve values can be easily obtained. This feature will be used to
 195 define performance criteria for validation, for optimization of the minimal set of
 calibration samples and to test the applicability of the approach. Furthermore,
 we will explore whether statistical properties of a group of sensors can be used
 for the efficient calibration of a new member (i.e., a new sensor) of the same
 product family. The resulting implementation fully exploits the potential for
 200 stability, accuracy and precision provided by the sensor technology, obtaining a

useful calibration with minimal measurement effort, yet accurate and realistic error bounds.

1.1. Example application: nonlinear calibration of an oxygen sensor

As practical implementation, the calibration transfer strategy was applied for the determination of the O₂ concentration using the luminescence (LS) quenching effect of a fluorophore immobilized in a commercially available, fiber-coupled, flow-through cell (FireStingO2, Pyro Science GmbH, Aachen, Germany).

Oxygen quenches the fluorescence of an excited dye immobilized in a sensor matrix, resulting in an optical signal that is reduced in fluorescence intensity, fluorescence lifetime, and shift of the phase angle in dependence on the oxygen concentration. Noise and signal fluctuations of the sensor increase for concentrations above 50 %, as less and less photons arrive at the detector and random detector noise gains prevalence. Therefore the manufacturer confines the high precision range to 0 - 50 % oxygen[28] and provides a factory calibration that favors simplicity and robustness rather than precision and accuracy. Fluorescence quenching is described by the Stern-Volmer equation [29] or, for a more realistic scenario, two quenching dyes can be considered [30] using two different constants K_{SV1} and K_{SV2} and relative contributions f_1 and f_2 to the combined quenching:

$$\frac{F}{F_0} = \frac{f_1}{1 + K_{SV1}[O_2]} + \frac{f_2}{1 + K_{SV2}[O_2]} \quad (2)$$

Here, F_0 is the incident fluorescence intensity, and F the fluorescence intensity after quenching.

Experimentally determined quenching values over the entire range of 0-100% oxygen reveal that the two dye version of equation (2) is necessary to describe the quenching response. Unfortunately, this equation cannot be converted to a function that is linear in the unknown coefficients, and hence, in the present case, has been replaced by the following empirical form:

$$\Delta\phi_t = \frac{c_1 x_t^2 + c_2 x_t + c_3}{x_t + c_4} + \epsilon_t \quad \text{or} \quad \Delta\phi_t = F(x_t, \mathbf{c}_t) + \epsilon_t \quad (3)$$

Figure 1: Raw value $\Delta\phi$ [°] vs. oxygen concentration [%]. Custom calibration using an empirical function (see equation (3)), fitted to a set of 11 calibration points spread over the range of 0 - 100 % at 10% intervals. Blue solid line: fit, dotted red line: 95 % confidence interval, black dots: experimentally obtained raw values (phase shift).

Here, x_t pertains to the independent variable, i.e., the oxygen concentration $[O_2]$, $\Delta\phi$ to the phase shift[31], the raw unit for the measured signal and ϵ_t refers to a randomly measured error, which will be considered in equation (1). As

Figure 1 indicates, this equation accurately reflects the signal-to-concentration relationship. Equation (3) defines the shape of the calibration curve via the coefficient set $c_1, \dots, c_4$, which is denoted as $\mathbf{c}_t$, and hence, the actual state of the sensor is captured by $\mathbf{c}_t$. Actual $\mathbf{c}_t$ values are determined by the calibration, and can thus be applied in a prediction equation. The latter can be obtained by solving equation (3) for the independent variable. The resulting prediction is denoted as y_t , for discrimination against the measurement, and is formally described as:

$$y_t = G(\Delta\phi_t, \mathbf{c}_t) \quad (4)$$

The calibration transfer will therefore be applied to determine the coefficients of $\mathbf{c}_t$.

Figure 1 shows a decreasing signal at higher oxygen concentrations, which is associated with an increased measurement error in this range. The typical

error distribution of a oxygen sensor for different oxygen concentration (here for several different days and sensors) is shown in Fig. 2 – the dependency on the oxygen concentration is clearly visible. This error is a physical phenomena caused by the naturally occurring intensity loss and resulting higher signal fluctuation, and a heterocedastic measurement error. Heteroscedasticity introduces a certain bias if not considered adequately. To avoid this one requirement of most classical chemometric techniques is a homoscedastic error distribution. One of the advantages of Stan is that each error included in the model in question can be characterized via an individual, even nonlinear, function, i.e. allowing the integration of heteroscedastic errors. As displayed in Fig. 3, the error can be represented by a function:

$$\sigma_i = a_0 + a_1[O_2]_i \quad (5)$$

It reflects a full recreation of the error function in the model. The Bayesian analysis obtained with the MCMC sampling of Stan allows to assess the propagation of measurement errors to the final error bounds for the oxygen concentration prediction. The latter should be very accurate and realistic.

2. Material and Methods

2.1. Experimental

All gas calibration samples were prepared from pure technical grade nitrogen and pure medical grade oxygen (both MTI IndustrieGase AG, Neu-Ulm, Germany) using a gas mixing pump (DIGAMIX 2M 301, H. Wösthoff Messtechnik GmbH, Bochum, Germany). The oxygen sensors were calibrated using 11 mixtures of oxygen and nitrogen spanning 0 – 100 % oxygen. Data evaluation was done using MATLAB 8.2 (R2013b, The Mathworks Inc, Naticks/MA, USA) and MatlabStan 2.7.0.0 [32] / Stan 2.8.0[25].

Six different sensors, named Sensor 1 to 6 (S1-S6) were studied over the period of one year. S1, S2, S5, and S6 represent an older generation of oxygen sensors, while S3 and S4 were purchased later. Functional properties of the sensors were

Figure 2: Typical error distribution of an oxygen sensor: oxygen concentration vs. standard deviation of all $\Delta\phi$ raw values of a measurement (measurement time ca. 1 min), each dot is the value of one day/sensor.

the same however. All sensor sets were measured in 3 to 46 replicates or calibration days, thus resulting in 122 different data sets. Sensor S3 and S4 are supposed be used under conditions where a calibration transfer is required for efficient and accurate calibration, and the calibrations for the other sensors were therefore considered the pool of previous measurements necessary to assess the day-to-day or sensor-to-sensor variability.

2.2. Numerical methods

The calibration transfer method is based on calibration equation (3), and the data set described above. Following major steps have to be executed, as described later in detail:

1. As the key step of the calibration transfer: define an approach, called 'Lagrange Multipliers', to combine mean values and the case-to-case covariance of calibration coefficients gained from a sufficient, large enough

calibration set pool with a minimal number of actual measurements to generate an actual calibration curve.

2. Define a hierarchical model, as outlined in Table 4, to use collected calibration sets to assess mean values and variance/covariance of coefficients across different sensors or different calibration sets of one specific sensor. Use the multivariate distribution of coefficients and the Lagrange Multipliers to determine actual calibration coefficients. Sample the corresponding predictions and generate parallel predictions from conventionally generated calibration curves.
3. Develop performance criteria for the calibration transfer and use them to determine the best combination of calibration samples for a calibration transfer by examining quality criteria.
4. Use these performance criteria to validate the calibration transfer model, and to evaluate its predictive capabilities via internal and external test samples.

2.2.1. Lagrange Multiplier approach

For the sake of simplicity, we assume that only one measurement with the index t is made to characterize the actual calibration. The independent variable of the calibration (i.e., concentration) shall be called x_t , while the dependent variable/sensor response shall be called $\Delta\phi_t$. We denote the set of actual calibration coefficients with $\mathbf{c}_t$. This vector $\mathbf{c}_t$ should satisfy equation (3) for a given independent variable x_t and for a measured response $\Delta\phi_t$. The key requirement is that this $\mathbf{c}_t$ should be a sample of a normal distribution around the mean coefficients $\bar{\mathbf{c}}$ and inter-day variance Σ . The $\mathbf{c}_t$ values that satisfy both criteria could be determined using a Lagrange Multiplier approach[33][34][35][36]. Here, the coefficients have to satisfy an equation that is linear in the unknown coefficients, but should be applied for the nonlinear calibration equation (3). Hence, this equation is rearranged to become linear in the calibration coefficient set $\mathbf{c}_t$.

310 This can be achieved for rational functions in the dependent variable as follows:

$$\begin{aligned}\Delta\phi_t(c_4 + x_t) &= c_1x_t^2 + c_2x_t + c_3 \\ \Delta\phi_tx_t &= c_1x_t^2 + c_2x_t + c_3 - \Delta\phi_tc_4 \\ \Delta\phi_tx_t &= (x_t^2, x_t, 1, -\Delta\phi_t)^T (c_1, c_2, c_3, c_4)\end{aligned}\quad (6)$$

The left side of the last row of equation (6) corresponds to a transformed response variable, which is noted in the following as $R_t = \Delta\phi_tx_t$. For the response variable R , one obtains the following linear equations:

$$R = B^T c_t; \quad \text{with} \quad B = \begin{pmatrix} x_t^2 & x_t & 1 & -\Delta\phi_t \end{pmatrix} \quad (7)$$

315 From the requirement that $\mathbf{c}_t$ should also be as close as possible to the mean coefficient set $\bar{\mathbf{c}}$, one can define the Lagrange-Norm

$$Q = \frac{1}{2}(c_t - \bar{c})^T \Sigma^{-1}(c_t - \bar{c}) + \lambda(B^T c_t - R) \quad (8)$$

Here, the first part pertains to the assumption that $\mathbf{c}_t$ is a sample of a normal distribution around $\bar{\mathbf{c}}$ and Σ , and the second part reflect the restraint that $\Delta\phi_t$ lies on the calibration curve. The Lagrange-Norm (8) is then minimized against $\mathbf{c}_t$ and λ . Minimizing by λ gives

$$\frac{\delta Q}{\delta \lambda} = 0 \quad , \text{ which gives } B^T c_t = R \quad (9)$$

320 Minimizing (8) by $\mathbf{c}_t$ results in

$$\begin{aligned}\frac{\delta Q}{\delta c_t} &= 0 \quad , \text{ which yields } \Sigma^{-1}(c_t - \bar{c}) + \lambda B = 0 \\ \text{or: } c_t &= \bar{c} - \lambda \Sigma B\end{aligned}\quad (10)$$

Insertion in (9) results in

$$\begin{aligned}B^T(\bar{c} - \lambda \Sigma B) &= R \\ -\lambda(B^T \Sigma B) &= R - B^T \bar{c} \\ \lambda &= (B^T \Sigma B)^{-1} B^T \bar{c} - (B^T \Sigma B)^{-1} R\end{aligned}\quad (11)$$

This yields the factor λ , which can be inserted into Equation (10) to obtain the desired $\mathbf{c}_t$. If more calibration samples are used, λ and R are vectors and $\mathbf{B}$ is a matrix. In the following, we subsume the equations above to a 'Lagrange function':

$$\mathbf{c}_{LG,t} = LGF(\bar{\mathbf{c}}, \Sigma, \Delta\phi_{limited}) \quad (12)$$

Hierarchical model

'Stan' is an open-Source Bayesian programming language that allows defining a statistical model with a structure that is outlined in Figure 3 and exemplified for the case of a nonlinear calibration based on Equation (3). The data set $\mathcal{D}$ is defined in the *Data Block*, at the top of the figure. Unknown parameters, which have to be determined from data, are defined in the **parameter** block - in the present example the coefficients $\mathbf{c}_t$ and parameters to describe a signal dependent measurement error. The software samples these parameters from a prior distribution, which can be specified in the model block. Otherwise a default, non-informative prior is used. In our case, we choose flat priors in a specified range (uniform distribution), derived from expected values gained from earlier calibration and measurements (i.e. the default initialization in Stan) as initialization values. In the **model** block, the actual, underlying statistical model is defined, and reflects the probability $P(\mathcal{D}|\mathbf{c})$, as defined in equation 1. First, a measurement error for each calibration level is defined. Then the latter is used to assess the probability that a measurement can be linked to the calibration curve for each replicate. This results in a combined probability of finding the data given the parameters to consider. Note that $P(\mathcal{D}|\mathbf{c})$, as used in equation 1, is realized in a series of normal distributions $\mathcal{N}(F(x_i, \mathbf{c}_t), \sigma_i)$ for individual measurements, where the mean value is calculated from $F(x_i, \mathbf{c}_t)$, which is a nonlinear function in the coefficients $\mathbf{c}_t$. A heteroscedastic measurement error σ_i defines the variance of this distribution. The lines in the *Model Block* demonstrate a key feature of Stan: Instead of using an analytical expression for a complex conditional probability, a sampling statement is defined, which may be composed of different elements. During a simulation run, 'Stan' picks

or samples a set of parameters, calculates the probabilities for this sample and checks whether the combined probability is large enough to be consistent with the measured data and the proposed distributions. If they are consistent, the parameter set is accepted as a sample of the underlying, unknown parameter distribution; otherwise, it is rejected. Strategies for rejection/acceptance, and for picking up an appropriate parameter set, follow the principles of Monte-Carlo Markov-Chain (MCMC) sampling[37, 38, 22]. Typically, a sequence of 2,000 or more parameter sets is sampled, and MCMC-update and rejection criteria are designed to ensure that each set is a sample of the underlying distribution. During sampling for each valid set of parameters, the **generated quantities block** is called. Here, the actual set of parameters can be used for prediction: values on the calibration curve at different levels can be assessed for each set of parameters. O₂ concentration values can be predicted based on measured signals for different levels and replicates. About 50 replicate actual measurements were used for each level, which vary only in their current measurement error range. The resulting predictions depend and fluctuate with the actual parameter set and the measurement error and are collected in a sampling chain comparable to the sampled parameters. The distributions of parameters and derived values, such as predicted concentrations, in the sampled chains reflect the underlying 'true' distributions. Therefore, these sequences can be used to reconstruct the underlying distributions and can be used to calculate statistical parameters including average, standard deviation and 5 % and 95 % quantiles. In the present case, it would be very difficult to devise an analytical expression for the distribution of the predicted O₂ concentration or $y_{i,j}$ values for the non-linear prediction equation, given the uncertainty in the average day-to-day coefficient values and given the measurement error. The simple sampling statement in this block provides adequate results. A complete 'Stan' script to execute this example can be found in the Appendix as well as the supplementary material. 'Stan' uses a very efficient MCMC variant, i.e., the Hamiltonian Monte Carlo / No-U-Turn Sampler[39] and can be set up on all major operation systems[25]. Interfaces to most major statistical software packages are available, including,

at the moment R, MATLAB[32], Python, Julia and Stata. Since the underlying 'Stan' programming script remains the same for each platform and interface, an effortless transfer between the different software platforms and therefore, rapid and facile exchange of finished scripts is enabled. A full description and manual of the 'Stan' programming language can be found in the supporting information of the software [40, 25, 26, 27].

The case-to-case variability of the calibration coefficients is described by a mean coefficient set $\bar{\mathbf{c}}$ and covariance matrix Σ , which form a multivariate normal distribution and should be determined from several measured calibration sets. The determination scheme outlined in Figure 3 is extended by expanding the data set to cover several calibration sets, and $\bar{\mathbf{c}}$ and Σ are added to the parameter list. This means that on the upper level of the sampling, the software draws a sample of $\bar{\mathbf{c}}$ and Σ from a default, non informative prior distribution. These values are used to define the probability $P(\mathbf{c})$, which is used in the model block as prior distribution. The model block defines statements to sample from hierarchical product $P(\mathcal{D}|\mathbf{c})P(\mathbf{c})$ used in equation 1. For sampling, an additional higher-level loop is put around the two loops for replicate measurements and calibration levels. The higher-level loop pertains to different calibration sets. Here an actual coefficient set $\mathbf{c}_k$ is sampled from $P(\mathbf{c})$ and these coefficients are used in the inner loop to sample from the probabilities to find specific measurement values. This sampling is repeated for each of the different calibration sets. In the inner loop, we also use a limited set of measurements for each calibration set, which is denoted $\Delta\phi_{limited}$ to determine coefficients based on the Lagrange-Multiplier approach using Equation (12). The sampling across different levels yields the parameters for the case-to-case variability of the coefficients, and the other parameters for the inner loops. As the uncertainties in the coefficients due to measurement errors are covered in the inner loop, they do not affect the estimates for the case-to-case variability. In the **generated quantities** block of the expanded script, the Lagrange-derived coefficients and the coefficients from the hierarchical model are both used to obtain predicted concentrations, which are denoted with $\mathbf{z}$ for the Lagrange-Model. The chain

Data block
data {structure of the data input}
Measurements $\Delta\phi_{i,j}$ at i -different level, each with j replicates
transformed data { }
calculate mean values, i.e., $\Delta\bar{\phi}_i$, the mean signal at concentration level i
Parameter block
parameters {to be determined via sampling}
Parameters, in the present case:
$\mathbf{c}_t$, the calibration coefficients
parameters a_0, a_1 to describe a random measurement error
of the measured signal
generated parameters { }
Transformation and definition of further parameters
Model block
model {Definition of model and sampling}
loop over all concentration levels i
$\sigma_i = a_0 + a_1 x_i$
loop over all replicates measurements j
$\Delta\phi_{i,j} \sim \mathcal{N}(F(x_i, \mathbf{c}_t), \sigma_i)^\dagger$
Generated quantities block
generated quantities {Predicted data}
loop over all replicates measurements j
$y_{i,j} = G(\Delta\phi_{i,j}, \mathbf{c}_t)$

Figure 3: General setup of a Stan program

$^\dagger a \sim \mathcal{N}(b, c)$ reads: a is normal distributed with mean value b and standard deviation c .

Data block
<code>data {structure of the data input}</code> Measurements $\Delta\phi_{k,i,j}$ for k -different calibration sets, at i -different levels, each with j replicates
Parameter block
<code>parameters {to be determined via sampling}</code> i.e., for the regression example, in addition: $\bar{\mathbf{c}}$, average calibration coefficients Σ , their case to case covariance
Model block
<code>model {extension for hierarchical model:}</code> loop over calibration runs k $\mathbf{c}_k \sim \mathcal{MN}(\bar{\mathbf{c}}, \Sigma)^\ddagger$ loop over all replicates measurements j loops as in table 3, except $\mathbf{c}_t$ is replaced by $\mathbf{c}_k$
Generated quantities block
<code>generated quantities{ }</code> loop over calibration runs k $\mathbf{c}_{LG,t} = LGF(\bar{\mathbf{c}}, \Sigma, \phi_{\text{limited}})$ loop over all concentration levels i loop over all replicates measurements j $y_{k,i,j} = G(\phi_{k,i,j}, \mathbf{c}_t)$ $z_{k,i,j} = G(\phi_{k,i,j}, \mathbf{c}_{LG,t})$

Figure 4: Extension to a hierarchical model

$^\ddagger \mathbf{a} \sim \mathcal{MN}(\mathbf{b}, \Sigma)$ reads: vector $\mathbf{a}$ is multivariate normal distributed with mean vector $\mathbf{b}$ and covariance Σ .

of sampled predictions is then used to derive a quality index for the full model, and for the reduced Lagrange-model.

415 2.2.2. Quality index for a full calibration

One important parameter to assess the quality of a calibration is the root mean square error of prediction (RMSEP). It is defined for each calibration level as:

$$RMSEP_{level} = \sqrt{\frac{\sum_i^n (y_i - x_i)^2}{n}} \quad (13)$$

where y_i is the concentration predicted from the calibration and x_i the real concentration of sample i. The lower the RMSEP, the better the predictions of the calibration model fit the expected values. The average RMSEP ($\overline{RMSEP}$) over all runs, days and levels is tentatively a useful indicator for the overall calibration performance of a model. However, herein random measurement errors and prediction errors increase with increasing oxygen concentration levels, and a plain average of RMSEPs over the different levels can be dominated by the errors at higher levels. Optimizing such an average RMSEP can result in minimization of the error at high oxygen concentration levels. For an objective quality criterion, the errors at each level have to be weighted with a factor. To develop such a factor, one has to consider that the RMSEP of the Lagrange fit cannot be better than the RMSEP of the single or hierarchical fit, since the underlying calibration samples determine the general quality of the fit. The aim of the calibration transfer is to maintain possibly close the predictive capabilities of a full calibration while reducing the calibration effort. To quantify this with a scalar value, a relative error variable $\Delta RMSEP_i$ is defined for each level i using the RMSEP of the Lagrange fit ($RMSEP_{LG_i}$) and the RMSEP of the hierarchical fit ($RMSEP_{HM_i}$):

$$\Delta RMSEP_i = \frac{RMSEP_{LG_i} - RMSEP_{HM_i}}{RMSEP_{HM_i}} \quad (14)$$

The sum of the $\Delta RMSEP_i$ of one Lagrange fit over all levels is now considered as a measure of success of the calibration transfer over the entire calibration

range, and is defined as

$$\Sigma\Delta RMSEP = \sum_{i=1}^{11} (\Delta RMSEP_i) \quad (15)$$

$\Sigma\Delta RMSEP$ is then calculated for each calibration run. A low $\Sigma\Delta RMSEP$ is an indicator that the reduced calibration transfer fit comes close to the predictive abilities of a full calibration, i.e., that the calibration transfer was successful. If calibration sample values of a day deviate from the usual expected accuracy, resulting in a 'bad' calibration, the values for $RMSEP_{LG}$ and $RMSEP_{HM}$ will be equally worse and therefore cancel each other out. Hence, $\Sigma\Delta RMSEP$ cannot detect such a situation. In order to account for this problem, $\Sigma\Delta RMSEP2$ is defined. Instead of using separate $RMSEP_{HM_i}$ calculated for each calibration day as weighting factor, the average $\overline{RMSEP_{HM_i}}$ of all fits in the calibration model is used:

$$\Delta RMSEP_{i2} = \frac{RMSEP_{LG_i} - \overline{RMSEP_{HM_i}}}{\overline{RMSEP_{HM_i}}} \quad (16)$$

$$\Sigma\Delta RMSEP2 = \sum_{i=1}^{11} (\Delta RMSEP_{i2}) \quad (17)$$

This value allows detecting calibrations that provide less accuracy than expected compared to the usual calibrations.

2.2.3. Quality index for an actual calibration transfer

For a typical calibration based on multiple samples, problems with an individual sample, either during calibration sample preparation or caused by unusual sensor fluctuation, may be identified, if the corresponding measurement points are off the calibration curve. If the number of measurements is smaller than the number of calibration coefficients, and if the coefficients for the calibration curve are determined by Lagrange Multipliers, then the measured points always lie on the calibration curve and the lack of fit criterion fails. For the same reason, the criterion based on $\Sigma\Delta RMSEP$ cannot be used. Erroneous measurements might drive the determined location of $\mathbf{c}_t$ away from the mean

value $\bar{\mathbf{c}}$, and thus, may have the same effect as a sensor drifting from its previously determined mean value. In both cases, the key assumption of the present calibration transfer scheme is challenged, i.e., that the coefficient values to be determined ($\mathbf{c}_t$) are part of a normal distribution around the mean values $\bar{\mathbf{c}}$ and covariance matrix Σ . With increasing distance from the mean value, probability decreases that $\mathbf{c}_t$ is a sample of the multi-normal distribution around $\bar{\mathbf{c}}$ and Σ . This probability is defined as:

$$P(c_t | \bar{\mathbf{c}}) \sim \text{Multinormal}(\bar{\mathbf{c}}, \Sigma) \quad (18)$$

A probability $P(c_t | \bar{\mathbf{c}})$ of a current calibration that has changed compared to those obtained by the model calibration sets indicates that a set of $\mathbf{c}_t$ is far from $\bar{\mathbf{c}}$, and that the calibration transfer criteria may not be fulfilled due to problems either caused by sensor drift or deficient measurements. Actual values for $\mathbf{c}_t$ and its corresponding $P(c_t | \bar{\mathbf{c}})$ depend on the measurements and the multivariate distribution of the coefficients. All these determinants can be assessed only with limited precision and therefore widen the confidence range for the probability that $\mathbf{c}_t$ assumes a given value. This confidence range can be assessed by the MCMC simulation, and we propose that the width of the confidence range of $P(c_t | \bar{\mathbf{c}})$ may be a quality parameter for proper calibration transfer.

Another indicator may be obtained using an additional reference or standard sample with known concentration or independent value. Let $\Delta\phi_{real}$ be its measurement, $\sigma\Delta\phi$ its variance and $\Delta\phi_{pred}$ its value predicted by the calibration curve. Then the probability that the predicted value $\Delta\phi_{pred}$ is within the normal distribution of the corresponding measured value is determined as:

$$P(\Delta\phi_{pred} | \Delta\phi_{real}) \sim \text{Normal}(\Delta\phi_{real}, \sigma\Delta\phi) \quad (19)$$

Values of $P(\Delta\phi_{pred} | \Delta\phi_{real})$ of an actual calibration that have changed compared to the values gained during model building indicate that the control point lies off the calibration curve, indicating problems with the calibration transfer. In addition, equation (17) can also be calculated for the reduced calibration set used with a calibration transfer. Thus, three criteria for the applicability

Figure 5: Available data sets and their division in training and test sets

of the calibration transfer are available: $P(c_t | \bar{c})$, $P(\Delta\phi_{pred} | \Delta\phi_{real})$ and $\Delta RMSEP_i2$, calculated for the reduced sample set. While building the calibration model, values for these criteria can be assessed from the data that were used to build the calibration model. Corresponding values can be collected over the MCMC runs performed with model building. This provides some reference values and distributions for the quality criteria describing proper calibration transfer. For comparison, MCMC sampling distributions can be collected for the criteria calculated from two or three samples measured for the actual calibration transfer. If the new distribution of a criterion differs extensively from those of the calibration model, then the performance of an actual calibration transfer may be compromised. This helps with quickly deciding during routine work whether a calibration is having problems and if it is possible to correct it in real-time.

2.2.4. Validation of the calibration transfer

Three different strategies for validating the calibration transfer were performed, based on the training and test sets shown in Figure 5. For each validation strategy, a specific training set was validated against two different data

sets, i.e., one test set consisted of selected calibration runs removed from the full set and kept aside for validation. Two versions were explored: in (i) a test set of calibrations, collected after a certain time, was validated against a training set collected before to assess sensor drift, and (ii) training and test sets were randomly selected from the same set of calibrations, for external validation. In the later strategy, the calibration runs collected for one sensor were validated against calibration runs collected for another sensor, to explore a calibration transfer from known sensors to unknown sensors. In detail, the corresponding validations steps are:

1. Validation Set 1: consisting of all sets of S1, S2, S3, S4
As test set 1A: selected sets removed from the full set of calibration runs
As test set 1B: set consisting of all calibration runs of S5 and S6
2. Validation Set 2: set consisting of all runs of S3
As test set 2A: selected sets removed from the full set of calibration runs
As test set 2B: set consisting of all calibration runs of S4
3. Validation Set 3: set consisting of all calibration runs of S3 and S4
As test set 3: selected sets removed from the full set of calibration runs

In a full calibration, relying only on actual calibration samples, 11 different levels or samples were used. From these 11 calibration levels, a limited set is combined with statistical information on the coefficients for a calibration curve to define an actual calibration. Four calibration levels were readily available: 0 %, 20 %, 30 % and 100 % oxygen. For brevity, in the present study only these four levels have been considered for the calibration transfer.

3. Results and Discussion

3.1. Determination of optimum calibration model and set points

To define a calibration model, one has to select the training set used to define the case-to-case variability and select the test set for later validation. Possible combinations are outlined in Figure 5. However, at this point of time, no test

sets were removed yet, and all data sets are in the training set. Before validation,
one has to determine the best number and combination of calibration set points
used for the Lagrange Multiplier calculation. The 15 possible combinations of
using up to four set points from four possible calibration levels were investigated.
Both $\overline{RMSEP}$ and $\Sigma\Delta RMSEP$ were averaged over all levels and days and
were used as comparison points. Here, the criteria for "best" was a compromise
between the minimum number of set points to minimize calibration efforts, while
still maintaining the error levels acceptable.

O ₂ conc. [%]	$\overline{RMSEP}$			$\Sigma\Delta RMSEP$		
	set 1	set 2	set 3	set 1	set 2	set 3
0	1.76	2.11	44.00	41.24	104.34	11203
20	44.75	1.66	46.20	2148	527.22	8278
30	1.32	1.76	0.68	77.38	928.86	112.74
100	1.24	1.20	6.38	100.67	6104	219.25
0, 100	1.17	1.01	0.87	28.00	24.18	18.00
0, 20	1.31	0.89	0.65	15.14	10.56	5.58
0, 30	1.29	0.72	0.61	13.82	6.13	4.50
20, 100	0.73	0.95	0.79	63.60	131.59	75.78
30, 100	0.72	0.90	0.74	66.13	252.22	106.65
20, 30	1.52	0.76	0.70	69.72	105.85	75.33
0, 20, 100	0.72	0.87	0.75	5.93	11.10	8.99
0, 30, 100	0.71	0.80	0.69	5.58	9.53	7.45
0, 20, 30	1.45	0.93	0.73	15.37	8.91	7.09
20, 30, 100	0.77	0.78	0.66	63.14	158.30	95.36
0, 20, 30, 100	0.76	0.70	0.63	7.02	8.97	7.59

Table 1: Comparison of possible combinations of training sets and calibration set points. The row in bold face font indicates the best choice for the training model, which was applied subsequent calculations

From all possible models, the version using training set 3 consisting of S3

	$\overline{RMSEP}$	$\Sigma\Delta RMSEP$
Mean	0.6070	4.4881
Std.dev.	0.0006	0.0223

Table 2: Repeatability of MCMC sampling

and S4 and two calibration set points (i.e., 0 and 30 % oxygen), clearly shows the best performance. The combination of these two set points establishes the best compromise between calibration stability and availability of calibration gas. Concurrently, Table 1 shows that, as previously mentioned, the $\overline{RMSEP}$ is vulnerable to larger error variances of higher oxygen levels, which limits its applicability as indicator for the performance of the calibration transfer. In contrast, $\Sigma\Delta RMSEP$ indicates reliably that the calibration transfer was successful, since only the deviation from the error expected for the given data is considered. This parameter is therefore more sensitive to a particular deficient calibration transfer vs. $\overline{RMSEP}$, in which extreme values may cancel each other out. To show the reproducibility of the MCMC sampling, the optimum model was repeated five times. The results are summarized in Table 2 and demonstrate excellent reproducibility of the MCMC sampling.

3.2. Model validation

The calibration transfer model was validated to explore the performance of the transfer in the case of new (i.e., "unknown") calibrations, but also to evaluate its robustness, if different calibration sets are removed. 25 % of the calibration samples used before were randomly removed before the model building process and were used as test sets. This external validation was repeated 10 times. Three different calibration set models were compared using the optimum number and combination of calibration set points shown in Table 1. Table 3 indicates that the values for $\overline{RMSEP}$ and $\Sigma\Delta RMSEP$ of the training and test sets of the three models are quite stable and comparable when randomly extracting sets from calibration runs for validation. Comparing the averages over the 10 repetitions of each validation run with the obtained values summarized in Table

	# training set /	$\overline{RMSEP}$		$\Sigma\Delta RMSEP$	
	# test set	training	test	training	test
Mean	1/1A	1.12	1.52	11.78	13.74
Std.dev.		0.09	0.16	1.11	0.66
Mean	2/2A	0.85	0.87	9.23	10.60
Std.dev.		0.08	0.15	1.77	2.46
Mean	3/3	0.61	0.64	4.43	5.69
Std.dev.		0.03	0.05	0.32	1.03
Mean	3/2A	0.60	0.65	4.88	5.12
Std.dev.		0.03	0.11	0.68	0.57

Table 3: External model validation using random subsets cross validation

1, both calibration and validation are quite close and no significant deterioration of the predictive abilities can be detected, which is confirmed by the associated standard deviations. Only the model based on calibration set 2 (only S3) shows larger standard deviations and deviation from the full set model. It appears that the calibration robustness of this model is reduced vs. the other models, if calibration runs are removed.

3.3. Calibration of sensor drift

For the calibration transfer, it is assumed that coefficients randomly fluctuate around mean values. A systematic drift of the coefficients over time would certainly challenge the applicability of the transfer model. It is therefore important to determine whether the calibration transfer is stable over a period of a few months without requiring recalibration. As a test, several calibration sets of S3 recorded during a later period were removed, and used as a validation set. When testing them against calibration model 2 and 3, the results shown in the upper part of Table 4 were achieved. The $\Sigma\Delta RMSEP$ value is quite stable for calibration and validation, and comparable to the values achieved for the random validation. Consequently, the sensor drift for this example case appears negligible.

# training	# test	$\overline{RMSEP}$		$\Sigma\Delta RMSEP$	
set	set	training	test	training	test
Sensor drift over time					
2	2A	0.96	0.75	8.57	11.13
3	2A	0.64	0.53	4.71	4.50
Transfer across different sensors					
S1-S4(1)	S5/S6(1B)	1.28	1.47	13.75	21.33
S3(2)	S4(2B)	0.70	0.59	4.86	6.07
S4	S3	0.50	1.31	3.80	22.92
S3/S4(3)	S1/S2/S5/S6	0.61	1.87	3.97	21.87

Table 4: External validation of sensor drift over time and transfer across different sensors

3.4. Calibration transfer for new unknown sensors

As derived from real-world analytical scenarios, it is frequently desirable to apply an established calibration transfer model to systems/sensors that have not been included in the previous calibration process; in the best case, these sensors may be used without a full calibration routine executed for the new sensor. The lower part of table 4 presents several combinations for calibration transfer. While calibration model 2, validated with test model 2B, shows promising results and nearly no deterioration in calibration performance, the other models reveal increased deviations. This indicates that in some instances, there may be some loss in calibration accuracy, if an unknown sensor is calibrated via the transfer routine without inclusion in the calibration model. Hence, whether the calibration transfer is successful for an unknown and non-calibrated sensor needs to be assessed for each individual case.

3.5. Quality control of a calibration

We have proposed to use $P(c_t | \bar{c})$, $P(\Delta\phi_{pred} | \Delta\phi_{real})$ and $\Sigma\Delta RMSEP_2$ as defined in equations 17 to 19 serving as criterion parameters for the quality of a calibration transfer. Plots of $P(c_t | \bar{c})$ vs. $P(\Delta\phi_{pred} | \Delta\phi_{real})$ and $P(c_t | \bar{c})$

Figure 6: Normal Calibration: a) $P(c_t | \bar{c})$ vs. $P(\Delta\phi_{pred} | \Delta\phi_{real})$ for full calibration sets, b) $P(c_t | \bar{c})$ vs. $\Sigma\Delta RMSEP_2$ for full calibration sets, c) $P(c_t | \bar{c})$ vs. $P(\Delta\phi_{pred} | \Delta\phi_{real})$ for reduced calibration sets, d) $P(c_t | \bar{c})$ vs. $\Sigma\Delta RMSEP_2$ for reduced calibration sets.

vs. $\Sigma\Delta RMSEP_2$ for both full calibration and reduced Lagrange calibration transfer set are shown in Figure 6. They include the results of all 2,000 MCMC chains calculated for 35 different calibration runs and show the distributions of the quality criteria. For the reduced calibration set, the two calibration samples at 0% and 30% O₂ concentration, which were used to assess the coefficient, plus one control sample at 20% O₂ were applied. The lower panels in Figure 6, pertaining to the reduced set, show a pattern comparable to the upper panels, which reflect the fit quality obtained with the 11 measurements of the full calibration set. This indicates that the three measurements of the reduced set were in fact sufficient to assess the fit quality of an actual calibration, and similar enough to serve as a 'rapid warning parameter' (i.e., cross-check) for failed calibrations. The values for the quality criteria cluster around $P(c_t | \bar{c})$ of 8 and $P(\Delta\phi_{pred} | \Delta\phi_{real})$ were around 2 with two exceptions. A closer look at the corresponding individual calibration curves indicates that they were still acceptable in terms of calibration performance, despite the fact that they do not have

the high quality of the other calibration data sets. With the quality criteria, it is desired that two different failures should be detected. Firstly, an accidental measurement of the wrong test gas, resulting in a wrong $\Delta\phi$ value and secondly, an accidental calibration measurement before reaching equilibrium of the test gas in the instrument after switching from one concentration level to the next, thus causing a larger than usual standard deviation. To explore the sensitivity of the quality criteria against possible failures, the raw calibration data of one calibration run was artificially altered to simulate first, a case of a deficient measurement of one calibration sample, by increasing all $\Delta\phi$ values at 0% O₂ by several units from the expected value of 50. Secondly, the case of a measurement prior reaching equilibrium was simulated by lowering the first two $\Delta\phi$ values at 0% O₂ by several units, thus resulting in a much higher standard deviation than usually expected. The third investigated case was a combination of both. All cases herein are modeled after the most common experimentally observed calibration errors for such sensors and thus serve as excellent model-cases for problems occurring during real-world routine calibration situations. Evidently, such problems need to be recognized and remedied during routine calibration protocols. The results of these effects are shown in the three panels in Figure 7.

Here, a distinct difference of the altered set (blue) compared to the others is immediately evident. The parameter used as quality control criteria show a significantly worsened parameter value and larger spread compared to those achieved by a not disturbed calibration data set. Recalling the definition of probabilities, it shows that the $\mathbf{c}_t$ values calculated from the 'wrong' calibration set are considerably more instable and vary significantly more compared to a correctly calibrated data set, which in turn impacts on the predictive performance of that calibration.

The probabilities $P(c_t | \bar{c})$ and $P(\Delta\phi_{pred} | \Delta\phi_{real})$, the calibration error representation $\Sigma\Delta RMSEP_2$, as well as their graphical representations appear to be excellent indicators for rapid recognition of problems during a daily calibration transfer routine. Therefore, these results allow for a rapid and immediate

Figure 7: Set 1 modified: first column: $P(c_t | \bar{c})$ vs. $P(\Delta\phi_{pred} | \Delta\phi_{real})$ for reduced calibration sets, second column: $P(c_t | \bar{c})$ vs. $\Sigma\Delta RMSEP_2$ for reduced calibration sets. First row: wrong test gas measured (wrong $\Delta\phi$ value), second row: too fast start of measurement (large standard deviation), third row: both cases combined (worst case scenario).

correction of potential problems. Only in the case of persisting problems, a full
recalibration of the sensor in question has to be considered.

4. Conclusion and outlook

We have developed a Bayesian calibration transfer algorithm using Lagrange Multipliers and a hierarchical model that is implemented within a multi-platform software package. Using Monte-Carlo Markov-Chain sampling, an accurate
statistic estimation of error bounds concerning any aspect of the statistical model pertaining to calibration transfer is ensured. This includes estimates for the coefficients of a nonlinear calibration equation, determination of the raw measurement error as a function of the signal intensity, estimates for the variability of the calibration coefficients across different calibration runs, and –
derived from these parameters, quantitative estimates for the prediction error associated with the calibration transfer. This provides the basis for defining quality criteria for a full calibration and for a calibration transfer, which can be used to detect problems with daily calibration routine for correcting problems just in time. With these tools, the calibration transfer approach developed
herein could be optimized to exploit the potential for stability, accuracy and precision provided by the applied sensor technology for obtaining valid calibration models with minimal experimental effort.

A key concept of the present approach is to represent a nonlinear response curve by a few coefficients and perform statistics on these coefficients. Generically,
this concept may be applied also to more complex, e.g., spectroscopic data. With a principal component analysis (PCA) performed on optical spectra, the actual shape of the spectrum may be represented by a few 'scores'. The latter may be nonlinear functions dependent on the composition of the sample, which may be completely characterized by a few coefficients. Using such a conversion,
the calibration transfer concept developed herein may be applied on spectra with score values that are nonlinear, e.g., in the concentration of the sample components.

While the present study only the implementation of the calibration transfer algorithm for univariate oxygen sensors is shown, this calibration strategy was also developed for the multivariate quantification of carbon dioxide via an infrared gas sensor combining a Fourier transform infrared spectrometer (FTIR) with a highly miniaturized substrate-integrated hollow waveguide (iHWG) gas cell [41][42][43][44][45][46] for correcting the influence of oxygen on the carbon dioxide signal. As both sensor systems – fluorescence for O₂ and infrared for CO₂ – are calibrated using the same calibration gases (i.e., mixtures of O₂ and CO₂ in nitrogen) and within the same procedure, the prerequisites for applying the developed calibration transfer protocols are immediately evident for minimizing the calibration efforts in daily routine work.

This combined sensors system is currently operated during routine studies at the Mouse Intensive Care Unit (MICU) of the Institute of Anesthesiologic Pathophysiology and Method Development at Ulm University, and serves as an on-line monitoring system for quantifying oxygen and carbon dioxide in exhaled breath of a ventilated mice[47][48][49]. As will be shown during future studies, the combination of the calibration transfer algorithms discussed herein with such higher-order sensing concepts will therefore facilitate the integration of these sensing technologies into the ventilation system of the MICU, in particular since the daily calibration efforts will be minimized.

5. Acknowledgement

The authors would like to thank the Federal State of Baden-Württemberg (Landesgraduiertenförderung) for partial financial support. This work was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory (LLNL) under Contract DE-AC52-07NA27344. This project was partially funded under LLNL sub-contract Nos. B598643 and B60301. The Boehringer Ingelheim Ulm University BioCenter (BIU) is acknowledged for financial support.

6. Appendix

```
functions{
  real prediction(vector coeffs,real O2)
  // this is the calibration function
  {
    return((coeffs[1] *O2 *O2 + coeffs[2] * O2 + coeffs[3])/(O2 + coeffs[4]));
  }
  real O2prediction(vector coeffs,real dphi)
  // the inverse calibration, converts measurements to concentration
  {
    real a;
    real b;
    real c;
    real squareroot;

    a<-coeffs[1];
    b<-coeffs[2]-dphi;
    c<-coeffs[3]-dphi*coeffs[4];
    squareroot<-sqrt(b*b-4*a*c);
    return((-b-squareroot)/(2*a)*100);
  }
  real RMSEP(vector y, vector yhat, int N)
  // root mean squared error of prediction
  {
    real summed;

    summed<-squared_distance(y,yhat);
    return(sqrt(summed/N));
  }
}

data
{
  int<lower=0> levels; //number of calibration levels/data points
  vector[levels] O2;
  vector[levels] dphi;
}

transformed data
{
  vector[levels] O2_trans;
  O2_trans<-O2/100;
}

parameters
```

```

{
  vector[4] Coeffs;
  real<lower=0> sigma;  // the measurement error is also determined
}

755 model
{
  real pred;

760 for (n in 1:levels)
  {
    pred<-prediction(Coeffs,O2_trans[n]);
    dphi[n]~normal(pred,sigma);
  }
765 }

generated quantities
{
  vector[levels] O2_prediction;
  real RMSEP_calibration;
770 for (n in 1:levels)
  {
    O2_prediction[n]<-O2prediction(Coeffs,dphi[n]);
  }
775 RMSEP_calibration<-RMSEP(O2,O2_prediction,levels);
}

```

7. References

- [1] T. Hastie, R. Tibshirani, J. H. Friedman, The Elements of Statistical Learning, Springer, 2003.
- 780 [2] J. A. Vogt, W. Fabinski, J. Kappler, H. Fischer, M. Georgieff, Response surface calibration of 13CO₂-NDIR offset values: A ‘random coefficients’ approach, Chemometrics and Intelligent Laboratory Systems 107 (2) (2011) 377–383.
- [3] R. N. Feudale, N. A. Woody, H. Tan, A. J. Myles, S. D. Brown, J. Ferré, 785 Transfer of multivariate calibration models: a review, Chemom. Intell. Lab. Syst. 64 (2) (2002) 181–192.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

[4] O. E. de Noord, Multivariate calibration standardization, *Chemometrics and Intelligent Laboratory Systems* 25 (2) (1994) 85–97.

[5] C. R. Rao, The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves, *Biometrika* 52 (3-4) 790 (1965) 447–458.

[6] B. Rosenberg, Linear regression with randomly dispersed parameters, *Biometrika* 60 (1) (1973) 65–72.

[7] D. Harville, Extension of the Gauss-Markov Theorem to Include the Esti-
795 mation of Random Effects, *Ann. Statist.* 4 (2) (1976) 384–395.

[8] N. M. Laird, J. H. Ware, Random-effects models for longitudinal data, *Biometrics* 38 (4) (1982) 963–974.

[9] N. T. Longford, *Random Coefficient Models*, Clarendon Press, Oxford, 1994.

800 [10] D. M. B. Mary J. Lindstrom, Nonlinear Mixed Effects Models for Repeated Measures Data, *Biometrics* 46 (3) (1990) 673–687.

[11] M. Davidian, D. M. Giltinan, Analysis of repeated measurement data using the nonlinear mixed effects model, *Chemometrics and Intelligent Laboratory Systems* 20 (1) (1993) 1–24.

805 [12] B. Efron, R. Tibshirani, *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy*, Statistical Science.

[13] A. Gelman, J. Hill, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, 1st Edition, Cambridge University Press, Cam-
810 bridge, New York, 2006.

[14] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, *Bayesian Data Analysis*, Third Edition, Chapman & Hall/CRC Texts in Statistical Science, CRC Press, 2013.

- [15] J. Kruschke, Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Elsevier Science, 2014.
- [16] D. Draper, Bayesian modeling, inference and prediction (2011).
- [17] S. Brooks, A. Gelman, G. Jones, X. Meng, Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, 2011.
- [18] H. Chen, B. R. Bakshi, P. K. Goel, Toward Bayesian chemometrics—A tutorial on some recent advances, *Analytica Chimica Acta* 602 (1) (2007) 1–16.
- [19] P. M. Lee, Bayesian Statistics: An Introduction, Wiley, 2012.
- [20] T. Sorensen, S. Vasishth, Bayesian linear mixed models using Stan : A tutorial for psychologists , linguists , and cognitive scientists, *arXiv.org* (2015) 1–30 [arXiv:arXiv:1506.06201v1](#).
- [21] D. J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, WinBUGS; A Bayesian Modelling Framework: Concepts, Structure, and Extensibility, *Statistics and Computing* 10 (4) (2000) 325–337.
- [22] Gilks, Richardson, Spiegelhalter (Eds.), Markov chain Monte Carlo Methods in Practice, Chapman and Hall, 1995.
- [23] N. Armstrong, D. B. Hibbert, An introduction to Bayesian methods for analyzing chemistry data: Part 1: An introduction to Bayesian theory and methods, *Chemometrics and Intelligent Laboratory Systems* 97 (2) (2009) 194–210.
- [24] D. B. Hibbert, N. Armstrong, An introduction to Bayesian methods for analyzing chemistry data: Part II: A review of applications of Bayesian methods in chemistry, *Chemometrics and Intelligent Laboratory Systems* 97 (2) (2009) 211–220.

- 840 [25] Stan: A C++ Library for Probability and Sampling, Version 2.8.0 (2015).
URL <http://mc-stan.org/>
- [26] A. Gelman, D. Lee, Stan : A probabilistic programming language for Bayesian inference and optimization, Journal of Statistical Software in press.
- 845 [27] A. Gelman, D. Lee, J. Guo, Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization, Journal of Educational and Behavioral Statistics 40 (5) (2015) 530–543. doi:10.3102/1076998615606113.
- [28] P. S. GmbH, Flow-Through Cells with Oxygen Sensor (2016).
850 URL <http://www.pyro-science.com/flow-through-cells-with-optical-oxygen-probe.html>
- [29] O. S. Wolfbeis, Luminescent sensing and imaging of oxygen: Fierce competition to the Clark electrode, BioEssays 37 (8) (2015) 921–928.
- [30] E. R. Carraway, J. N. Demas, B. A. DeGraff, J. R. Bacon, Photophysics and
855 photochemistry of oxygen sensors based on luminescent transition-metal complexes, Analytical Chemistry 63 (4) (1991) 337–342.
- [31] J. R. Lakowicz, Principles of Fluorescence Spectroscopy, Springer US, 2007.
- [32] B. Lau, MatlabStan: the MATLAB interface to Stan (2015).
URL <http://mc-stan.org/matlab-stan.html>
- 860 [33] D. P. Bertsekas, Nonlinear programming, Athena Scientific optimization and computation series, Athena Scientific, 1999.
- [34] D. P. Bertsekas, W. Rheinboldt, Constrained Optimization and Lagrange Multiplier Methods, Computer science and applied mathematics, Elsevier Science, 2014.
- 865 [35] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

- [36] D. Barber, Bayesian Reasoning and Machine Learning, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.
- [37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of State Calculations by Fast Computing Machines, J Chem Phys 21 (1953) 1087–1092.
- [38] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika (1970) 97–109.
- [39] M. D. Hoffman, A. Gelman, The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, Journal of Machine Learning Research 15 (2014) 1593–1623.
- [40] Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0 (2015).
URL <http://mc-stan.org/>
- [41] Petrucci, J. F. da Silveira, Wilk, A., Cardoso, A. A., Mizaikoff, B., Online Analysis of H₂S and SO₂ via Advanced Mid-Infrared Gas Sensors., Analytical Chemistry (Washington, DC, United States) 87 (19) (2015) 9605–9611.
- [42] Perez-Guaita, D., Kokoric, V., Wilk, A., Garrigues, S., Mizaikoff, B., Towards the determination of isoprene in human breath using substrate-integrated hollow waveguide mid-infrared sensors., Journal of Breath Research 8 (2) (2014) 026003/1–026003/8.
- [43] Rohwedder, J. J. R., Pasquini, C., Fortes, P. R., Raimundo, I. M. Jr, Wilk, A., Mizaikoff, B., iHWG-nNIR: a miniaturised near-infrared gas sensor based on substrate-integrated hollow waveguides coupled to a micro-NIR-spectrophotometer, Analyst (Cambridge, United Kingdom) 139 (14) (2014) 3572–3576.
- [44] Petrucci, J. F. da Silveira, Fortes, P. R., Kokoric, V., Wilk, A., Raimundo, I. M., Cardoso, A. A., Mizaikoff, B., Monitoring of hydrogen sulfide via

substrate-integrated hollow waveguide mid-infrared sensors in real-time,
Analyst (Cambridge, United Kingdom) 139 (1) (2014) 198–203.

[45] Wilk, A., Carter, J. C., Chrisp, M., Manuel, A. M., Mirkarimi, P., Alameda, J. B., Mizaikoff, B., Substrate-integrated hollow waveguides: a new level of integration in mid-infrared gas sensing., Analytical Chemistry (Washington, DC, United States) 85 (23) (2013) 11205–11210.

[46] J. C. Carter, M. P. Chrisp, A. M. Manuel, B. Mizaikoff, A. Wilk, S.-S. Kim, Substrate-integrated hollow waveguide sensors., U.S. Pat. Appl. Publ.

[47] Seichter, F., Wilk, A., Woerle, K., Kim, S.-S., Vogt, J. A., Wachter, U., Radermacher, P., Mizaikoff, B., Multivariate determination of $^{13}\text{CO}_2/^{12}\text{CO}_2$ ratios in exhaled mouse breath with mid-infrared hollow waveguide gas sensors., Analytical and Bioanalytical Chemistry 405 (14) (2013) 4945–4951.

[48] Woerle, K., Seichter, F., Wilk, A., Armacost, C., Day, T., Godejohann, M., Wachter, U., Vogt, J., Radermacher, P., Mizaikoff, B., Breath Analysis with Broadly Tunable Quantum Cascade Lasers., Analytical Chemistry (Washington, DC, United States) 85 (5) (2013) 2697–2702.

[49] Wilk, A., Seichter, F., Kim, S.-S., Tuetuencue, E., Mizaikoff, B., Vogt, J. A., Wachter, U., Radermacher, P., Toward the quantification of the $^{13}\text{CO}_2/^{12}\text{CO}_2$ ratio in exhaled mouse breath with mid-infrared hollow waveguide gas sensors., Analytical and Bioanalytical Chemistry 402 (1) (2012) 397–404.