Nonlin. Processes Geophys., 19, 69–80, 2012
www.nonlin-processes-geophys.net/19/69/2012/
doi:10.5194/npg-19-69-2012

Nonlinear Processes in Geophysics

Structural and statistical properties of the collocation technique for error characterization

S. Zwieback[1,*], K. Scipal[2], W. Dorigo[1], and W. Wagner[1]

[1]Vienna University of Technology, Institute of Photogrammetry and Remote Sensing, Vienna, Austria
[2]European Space Agency, Mission Science Division, Noordwijk, The Netherlands
[*]currently at: ETH Zürich, Institute of Environmental Engineering, Zürich, Switzerland

Correspondence to: S. Zwieback (zwieback@ifu.baug.ethz.ch)

Received: 18 October 2011 – Revised: 25 January 2012 – Accepted: 5 February 2012 – Published: 9 February 2012

Abstract. The validation of geophysical data sets (e.g. derived from models, exploration techniques or remote sensing) presents a formidable challenge as all products are inherently different and subject to errors. The collocation technique permits the retrieval of the error variances of different data sources without the need to specify one data set as a reference. In addition calibration constants can be determined to account for biases and different dynamic ranges. The method is frequently applied to the study and comparison of remote sensing, in-situ and modelled data, particularly in hydrology and oceanography. Previous studies have almost exclusively focussed on the validation of three data sources; in this paper it is shown how the technique generalizes to an arbitrary number of data sets. It turns out that only parts of the covariance structure can be resolved by the collocation technique, thus emphasizing the necessity of expert knowledge for the correct validation of geophysical products. Furthermore the bias and error variance of the estimators are derived with particular emphasis on the assumptions necessary for establishing those characteristics. Important properties of the method, such as the structural deficiencies, dependence of the accuracy on the number of measurements and the impact of violated assumptions, are illustrated by application to simulated data.

1 Introduction

Adequate knowledge of the error characteristics of different sensors, models, remote sensing products, etc. can be considered a prerequisite for their meaningful application in practice and scientific studies. It is, for example, necessary when assimilating satellite and in-situ observations with meteorological models (e.g. Munro et al., 2004), when combining data from different sources (e.g. Missaoui et al., 2011; Liu et al., 2011), and when analyzing such measurements or models as to their accuracy and range of validity (e.g. Stoffelen, 1998; Scipal et al., 2008). The validation of such products is intrinsically difficult due to the lack of knowledge of the "truth": the actual value of the parameter to be determined is never known with absolute certainty, and spatial as well as temporal mismatch often exert a confounding influence.

The collocation technique does not require the specification of a reference data set and is applicable when three or more data sources are available. It permits the estimation of the error variance of each sensor provided certain assumptions about the error structure are met. When applied to three data sources, it is called triple collocation and its popularity has grown considerably over the last decade. Most frequently it has been applied to remote sensing products in order to evaluate their error structure and compare them to models, in-situ and alternative remote sensing measurements.

The method was introduced by Stoffelen (1998) in order to study the error characteristics of wind vector data derived from a model, buoy measurements and scatterometer observations. Further oceanographic studies pertaining to wind speed, wave height or sea surface temperature measurements include Caires and Sterl (2003); Janssen et al. (2007); O'Carroll et al. (2008); Winterfeldt et al. (2010). Regarding land hydrological applications the comparison of soil moisture estimates from models, in-situ data and remote sensing products has become an active field of research (e.g. Scipal et al., 2008; Dorigo et al., 2011; Loew and Schlenz, 2011; Parinussa et al., 2011). The technique has, for example, also been applied to the study of evapotranspiration data by Miralles et al. (2011).

Apart from the estimation of the error structure, the determination of calibration constants is of vital practical importance as well in order to account for biases and different dynamic ranges of the products. This can be achieved within the collocation framework, as shown by Stoffelen (1998) and Muraleedharan et al. (2006). Alternative ways of estimating these calibration constants have also been proposed, e.g. the error-in-variables regression approach suggested by Scipal et al. (2008).

While the triple collocation technique has become a routine tool in calibration/validation studies of models and measured data, its statistical properties and sensitivity with respect to violated assumptions, e.g. the presence of correlations between different data sources, have not been analyzed in detail. This is one objective of this paper. The other one, which will be treated before that, is the analysis of the technique for a general number of data sources: which properties of measurement errors can be estimated and which cannot. These structural deficencies highlight the importance of expert knowledge and experience with the analyzed data sets for validation studies.

The notation and the error models related to the collocation technique are defined in Sect. 2. The triple collocation method, as applied in many previous studies, is introduced in Sect. 3. Subsequently – Sect. 4 – the inherent mathematical structure for an arbitrary number of data sources is analyzed and it is shown to what degree error covariances can be estimated. The key statistical properties of the estimators for the error characteristics and the calibration constants are derived and discussed in Sect. 5. These results are compared and applied to simulated data in Sect. 6 with particular emphasis on the determination of the uncertainties and the dependence on the number of samples.

2 Error models

The collocation method relies on a stochastic model in which the noise is additive. More specifically this paper treats three such models; the difference is due to the varying number of calibration constants included.

In general it is assumed that there are N sets of measurements with each set containing a measurement of each of the M data sources. In the classical triple collocation approach $M = 3$. In previous studies the different samples $n = 1 \ldots N$ referred to different epochs in time and we also adopt this view and terminology. The statistical properties and assumptions are not inherently related to this interpretation; the technique can thus be applied to non-temporal data as well. Angle brackets $\langle \cdot \rangle$ are adopted for expectation values.

2.1 Basic model

The basic model describes data sources that are mutually calibrated and only differ in an additive random error:

$$y_i^n = x^n + e_i^n \tag{1}$$

where y_i^n is the measurement number n by sensor i, x^n is the unknown variable and e_i^n is the corresponding error.

Note that in this paper x^n is treated as an unknown deterministic parameter and not as a random variable. An alternative view considers x to be random in principle: the analysis is conducted by conditioning on x^n, thus fixing their values as in the deterministic point of view. These different conceptions regarding x are equivalent and mirror the way the exogenous variables can be treated in regression analysis. The conclusions drawn from the collocation technique thus do not rely on any additional assumptions about the "truth" x except the validity of the model defined in Eq. (1).

2.2 Bias model

An extension of the basic model allows for the inclusion of an additive bias term α (e.g. Parrens et al., 2011). As the collocation technique makes no assumptions about the unknown parameter x, the zero point of one of the data sets has to be taken as reference; w.l.o.g. this is the set $m = 1$.

$$y_i^n = x^n + \alpha_i + e_i^n \tag{2}$$

Note that $\alpha_1 = 0$ and for $m = 1$ the bias model essentially collapses to the basic model.

2.3 Affine model

The affine model also accounts for a multiplicative bias or scale factor β (source). Following the reasoning for the bias model, $\beta_1 = 1$ and thus the first data set determines the zero point and the units to which the remaining data sets will refer after calibration (i.e. estimation of the calibration constants α and β).

$$y_i^n = \beta_i x^n + \alpha_i + e_i^n \tag{3}$$

2.4 Assumptions

The assumptions regarding the statistical characteristics of the error terms are crucial for the validity of the collocation technique. One important contribution of this paper is to show which assumptions are necessary for certain properties of the estimators to hold; the following will be referred to in the subsequent analyses:

Assumption 1 (Zero expectation) *the expected values of the error terms vanish, i.e.* $\langle e_i^n \rangle = 0$

Assumption 2 (Homoscedasticity) *the error (co)variances do not depend on time, i.e.* $\langle (e_i^n)^2 \rangle = \sigma_{ii}$, $\langle e_i^n e_j^n \rangle = \sigma_{ij}$
Assumption 2b *the time invariance also holds for the fourth moments, i.e.* $\langle (e_i^n)^4 \rangle = \gamma_i$

Assumption 3 (Zero crosscorrelation) *the correlations between different errors at the same epoch n are 0, i.e.* $\langle e_i^n e_j^n \rangle = 0, i \neq j$

Assumption 4 (Zero autocorrelation) *the correlations at different times vanish, i.e.* $\langle e_i^{n'} e_i^{n''} \rangle = 0, n' \neq n''$

Note that these assumptions about the errors pertain to the error model used for deriving and characterizing the estimators. Failure to meet one of them can in certain cases be circumvented by choosing an appropriate error model. If, for example, one of the error terms had a bias ($\langle e_i \rangle \neq 0$), the calibration constant α_i could account for this in the bias or the affine model.

3 Basic triple collocation

The vast majority of applications of the collocation technique to study error characteristics have focussed on three different data sources as this is the minimum number needed in order to estimate the RMS error of each. An in-depth analysis is provided in Sect. 5; in this brief introduction only the key properties are stressed. To this end it is sufficient to look at the basic model.

3.1 Estimating the error variance

Applied to three sensors, the basic model postulates the following error structure:

$$y_1^n = x^n + e_1^n$$
$$y_2^n = x^n + e_3^n$$
$$y_3^n = x^n + e_3^n$$

By forming a difference between two simultaneous measurements, the parameter x^n vanishes such that for example

$$\left\langle \left(y_1^n - y_2^n\right)\left(y_1^n - y_3^n\right)\right\rangle = \left\langle e_1^n e_1^n - e_1^n e_2^n - e_1^n e_3^n + e_2^n e_3^n \right\rangle = \sigma_{11}$$

using assumptions 1, 2, and 3 and it thus seems natural to apply the following estimators

$$\hat{\sigma}_{11} = \frac{1}{N}\sum_{n=1}^{N}\left(y_1^n - y_2^n\right)\left(y_1^n - y_3^n\right) \tag{4}$$

$$\hat{\sigma}_{22} = \frac{1}{N}\sum_{n=1}^{N}\left(y_2^n - y_1^n\right)\left(y_2^n - y_3^n\right) \tag{5}$$

$$\hat{\sigma}_{33} = \frac{1}{N}\sum_{n=1}^{N}\left(y_3^n - y_1^n\right)\left(y_3^n - y_2^n\right) \tag{6}$$

These estimators are unbiased given assumptions 1, 2, and 3, as is shown in Sect. 5.1, where also their variance is given.

Analogously to Stoffelen (1998) it is also possible to postulate a fixed covariance between two data sets and modify these estimators accordingly. In the analysis of scatterometer-derived wind vectors, for instance, such correlations can arise due to the spatial scale mismatch of the sensors involved (Vogelzang et al., 2011). If, for example, the covariance $\langle e_2^n e_3^n \rangle = \sigma_{23}$, the expected value of $\frac{1}{N}\sum_{n=1}^{N}\left(y_1^n - y_2^n\right)\left(y_1^n - y_3^n\right)$ becomes $\sigma_{11} + \sigma_{23}$ so that $-\sigma_{23} + \frac{1}{N}\sum_{n=1}^{N}\left(y_1^n - y_2^n\right)\left(y_1^n - y_3^n\right)$ is an unbiased estimator for σ_{11}. The other formulae will have to changed in a similar fashion.

The subsequent discussions examine how this collocation technique can be generalized to more than three sensors and whether error covariances can be estimated as well. Furthermore the properties of this and related estimators are analyzed. Note that the term "collocation" is very general in nature; it can e.g. refer to a method for numerically solving differential equations or to the act of linking different data sets. The phrase "collocation technique", as used in this paper, encompasses generalizations of the triple collocation method for estimating the error variance, which do not require specification of a reference data set.

4 Structure of the collocation technique

This section is concerned with the general structure of the collocation technique; more specifically, the possible relaxation of assumption 3 (lack of cross-covariance) is studied for a general number of data sources. In the following the validity of the basic model and a general error covariance matrix Σ will be assumed; e.g. for $M = 3$:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

4.1 Brackets

As certain kinds of products are commonplace in the collocation method, we introduce the bracket notation.

$$[i,j;k,l]^n = (y_i^n - y_j^n)(y_k^n - y_l^n) = (e_i^n - e_j^n)(e_k^n - e_l^n) \tag{7}$$

Several properties of such brackets follow immediately from the definition in Eq. (7):

$$[i,j;i,j]^n = [i,j;i,k]^n + [j,i;j,k]^n \tag{8}$$
$$[i,j;i,k]^n = [i,j;i,l]^n - [i,j;k,l]^n \tag{9}$$

4.2 Sampling the error covariance matrix

The general bracket in Eq. (7) essentially samples the error matrix $\mathbf{E}^n$ with $e_{i,j}^n = e_i^n e_j^n$. As the expected value of $\mathbf{E}^n$ is the error covariance matrix (given assumption 1), this is of great importance to the collocation technique. Due to the distributivity of multiplication over addition we can, by averaging over multiple samples n, sample the error covariance matrix, provided assumption 2 is met. We can thus focus our discussion on the sampling of the error matrix.

Let us introduce a vectorization of the problem at hand. A bracket from Eq. (7), such as $[1,2;1,3]^n$ for $M=3$, can also be thought of as an inner product between the error matrix $\mathbf{E}^n$ and a symmetric bracket matrix $\mathbf{B}$:

$$[1,2;1,3]^n = \mathrm{Tr}\left(\mathbf{B}^T_{[1,2;1,3]} \cdot \mathbf{E}^n\right) \tag{10}$$

where Tr is the Trace operator, which defines an inner product (Cantrell, 2000). Note that this corresponds simply to a sum over the products of each corresponding pair of elements. The bracket matrix is given by

$$\mathbf{B}_{[1,2;1,3]} = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

In general we require these matrices to be symmetrical and that Eqs. (7) and (10) match.

It is clear from Eqs. (8) and (9) that there are dependencies between the brackets, i.e. that the corresponding **B**-matrices are linearly dependent. This also follows due the fact there are $\binom{M+1}{2}$ degrees of freedom in a $M \times M$ symmetric matrix.

The aim of Sect. 4.3 consists of identifying the structure of the possible information that can be gained from Eq. 10. Before this we will show that the brackets are sufficient for this task.

4.2.1 Sufficiency of brackets

A product of the form $(y_i - y_j)(y_k - y_l)$ can be generalized to one where on either side a linear combination of the measurements whose weights sum to 0 is allowed. This condition is necessary and sufficient for cancelling the unknown parameter.

All linear functionals fulfilling the above condition form an $M-1$ dimensional subspace of which the vectors $\boldsymbol{u}_m$, $m = 1 \ldots M-1$, form a basis: they are defined by

$$u_{m,l} = \delta_{m,l+1} - \delta_{m,l} \tag{11}$$

where δ is the Kronecker delta.

Such a general product can be written as

$$\left(\sum_{m=1}^{M} v_m y_m\right)\left(\sum_{m'=1}^{M} w_{m'} y_{m'}\right) \tag{12}$$

Consequently, the weights v_m and $w_{m'}$ can be expressed in the $\boldsymbol{u}_m$ basis as

$$v_m = \sum_{l=1}^{M-1} u_{l,m} p_l$$

$$w_{m'} = \sum_{l'=1}^{M-1} u_{l',m'} q_{l'}$$

so that Eq. (12) becomes

$$\left(\sum_{m=1}^{M}\sum_{l=1}^{M-1} u_{l,m} p_l y_m\right)\left(\sum_{m'=1}^{M}\sum_{l'=1}^{M-1} u_{l',m'} q_{l'} y_{m'}\right)$$
$$= \sum_{l=1}^{M-1}\sum_{l'=1}^{M-1} p_l q_{l'}\left[\sum_{m=1}^{M}\sum_{m'=1}^{M} u_{l,m} y_m u_{l',m'} y_{m'}\right]$$

where it turns out that the terms inside the square brackets are just brackets of the form $[l+1,l;l'+1,l']$, which shows that the most general product of Eq. (12) can be thought of as a linear combination of simple products as defined in Eq. (7). The brackets are thus perfectly general for our purposes.

4.3 Resolvable structure of the error covariance matrix

The previously defined sampling of the error matrix in Eq. (10) allows us to describe the quadratic estimators in terms of linear algebra. The key question is whether the complete error covariance matrix can be resolved by the brackets of Eq. (7). The answer is simple: No. Suppose you have 5 independent sensors (with the basic model defined in Eq. 1 applicable) and none of them has any measurement noise. All 5 will give the same result. On the other hand, suppose you have 5 sensors with equal error variance that are all perfectly correlated. Again, all 5 sensors will yield the same result and there is no way to tell the difference between the two cases without additional assumptions. It will now be shown that there is more structure that cannot be resolved by the collocation technique.

In terms of linear algebra a bracket corresponds to an inner product between the error matrix and the associated bracket matrix. The vector space and the important subspaces are denoted and defined as follows[1]:

> **Vector space $\boldsymbol{C^M}$** the vector space of all symmetric $M \times M$ matrices. Note that definiteness plays no role. Its dimension is $\binom{M+1}{2}$.
>
> **Vector space $\boldsymbol{B^M}$** the vector space spanned by the bracket matrices given there are M sensors. The previous discussion shows that (linear combinations of) the brackets can represent all possible products, i.e. they are perfectly general. B^M is a subspace of C^M
>
> **Vector space $\boldsymbol{K^M}$** the orthogonal complement of B^M (Lang, 1987), using the dot product of Eq. (10).

If B^M were of dimension $\binom{M+1}{2}$, we could sample the error matrix completely, i.e. we could reconstruct it. We will now show that its dimension is only $\binom{M}{2}$. We will first find two sets of independent vectors, one for B^M and one for K^M and then argue on dimensional grounds that these actually form a basis for their respective space.

[1] Actually it is not necessary to introduce the formalism of inner products, but rather to think of the brackets as a linear functional acting on the error matrix. The description adopted here, however, simplifies the analysis and notation cosiderably.

4.3.1 Vectors in B^M

Let us turn to the matrices corresponding to brackets of the form $[i,j;i,j]$ with $i<j$. There are clearly $\binom{M}{2}$ of these matrices $\mathbf{B}_{[i,j;i,j]}$. They are also independent because only $\mathbf{B}_{[i,j;i,j]}$ has a non-vanishing coefficient at positions (i,j) and (j,i).

4.3.2 Vectors in K^M

These vectors are by definition orthogonal to the brackets matrices $\mathbf{B}_{[i,j;k,l]}$. The elements (i,j) of one set of such vectors $\mathbf{A}_m$ are given by $\frac{1}{2}(\delta_{i,m}+\delta_{m,j})$, such that for $M=3$

$$\mathbf{A_1}=\frac{1}{2}\begin{pmatrix}2&1&1\\1&0&0\\1&0&0\end{pmatrix}\mathbf{A_2}=\frac{1}{2}\begin{pmatrix}0&1&0\\1&2&1\\0&1&0\end{pmatrix}\mathbf{A_3}=\frac{1}{2}\begin{pmatrix}0&0&1\\0&0&1\\1&1&2\end{pmatrix}$$

These vectors are independent because only $\mathbf{A}_i$ has a non-zero element at the (i,i) position.

The orthogonality with respect to B^M, although straightforward to show, is a bit tedious to establish because of the number of cases to consider. In the following the vector $\mathbf{A}_u$ will be dotted with all possible bracket matrices. As the brackets of the form $[i,i;i,i]$ vanish, we only have to look at those with 2, 3 or 4 distinct indices.

The first group of bracket matrices to consider consists of those $\mathbf{B}_{[i,j;i,j]}$ with $i\neq j$. There are two cases. Firstly, $u\neq i\neq j$. The dot product in Eq. (10) is clearly zero because $\mathbf{B}_{[i,j;i,j]}$ has only zero elements in row/column u. Secondly, we have $u=i$ (the $u=j$ case is analogous), in which case the (i,i) element exactly cancels the (i,j) and (j,i) elements.

The second group contains the matrices of the form $\mathbf{B}_{[i,j;i,k]}$ with $i\neq j\neq k$, $j<k$. There are three cases to consider. Firstly, if $u=i$ the (i,i) cancels with the (i,j), (j,i), (i,k), and (k,i) elements. Secondly, if $u=j$ ($u=k$ is analogous), the (i,j) and (j,i) elements cancel the (j,k) and (k,j) elements. Thirdly, when u is distinct from i, j, and k, the dot product vanishes because $\mathbf{B}_{[i,j;i,k]}$ has only zero elements in row/column u.

The third group consists of matrices of the form $\mathbf{B}_{[i,j;k,l]}$. When u is distinct from i, j, k, and l, we have orthogonality for the same reasons as above. When one of them, say $j=u$, then the (j,k) and (k,j) elements cancel the (j,l) and (l,j) elements.

4.3.3 Resolution and consequences

We have found $\binom{M}{2}$ independent vectors $\mathbf{B}_{[i,j;i,j]}$ in B^M and M vectors $\mathbf{A}_m$ in K^M. As orthogonality implies independence, we have found a set of $M+\binom{M}{2}=\binom{M+1}{2}$ independent vectors in C^M. We have thus found a basis for this space C^M (Lang, 1987) and consequently also a basis for B^M, whose dimension is thus established to be $\binom{M}{2}$. Covariance structure corresponding to its orthogonal complement K^M cannot be resolved, as the inner product yields 0.

Unfortunately it turns out there is an M dimensional subspace K^M which is invisible to the collocation technique. Furthermore the elements of this space are not particularly easy to interpret, the only exception being $\sum_{m=1}^{M}\mathbf{A}_m$, which corresponds to the case of perfectly correlated measurements described in Sect. 4.3. In practice it will be easier to postulate that certain covariances vanish. The standard triple collocation technique asserts exactly that, i.e. the three assumptions needed are exactly those that all correlations are 0.

5 Statistical analysis

In this section various estimators of elements of the error covariance matrix as well as the calibration constants will be analyzed; particular emphasis is placed on their expected values and variances. Each of the three models is discussed separately.

5.1 Basic model

Previous studies (e.g. Stoffelen, 1998; Scipal et al., 2008; Dorigo et al., 2011) predominantly applied the following estimator of the error variance (with $i\neq j\neq k$):

$$\begin{aligned}\hat{\sigma}_{ii}&=\frac{1}{N}\sum_{n=1}^{N}[i,j;i,k]^n \qquad (13)\\&=\frac{1}{N}\sum_{n=1}^{N}\left(e_i^n e_i^n+e_j^n e_k^n-e_i^n e_k^n-e_j^n e_i^n\right)\end{aligned}$$

By recourse to assumptions 1, 2 and 3 it follows that the estimator given by Eq. (13) is unbiased, i.e. $\hat{\sigma}_{ii}=\langle\sigma_{ii}\rangle$.

The variance of the estimator follows from its definition

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_{ii})&=\left\langle\left(\hat{\sigma}_{ii}-\sigma_{ii}\right)^2\right\rangle=-\sigma_{ii}^2+\left\langle\hat{\sigma}_{ii}^2\right\rangle\\&=-\sigma_{ii}^2+\frac{1}{N^2}\sum_{n'=1}^{N}\sum_{n''=1}^{N}(e_i^{n'}-e_j^{n'})(e_i^{n'}-e_k^{n'})\\&\quad\cdot(e_i^{n''}-e_j^{n''})(e_i^{n''}-e_k^{n''})\\&=\frac{1}{N}\left(\gamma_i-\sigma_{ii}^2+\sigma_{ii}\sigma_{kk}+\sigma_{ii}\sigma_{jj}+\sigma_{jj}\sigma_{kk}\right) \qquad (14)\end{aligned}$$

where the last line is obtained by expanding the expression into sums of four error terms – the algebra mirrors the derivation of the variance of the classical variance estimator of a sample (e.g. Kenney and Keeping, 1956). Most of these terms cancel when assumptions 1, 2, 2b, 3 and 4 are invoked. The first of these terms t_1, for example, is:

$$\begin{aligned}t_1&=\sum_{n'=1}^{N}\sum_{n''=1}^{N}\langle(e_i^{n'})^2(e_i^{n''})^2\rangle\\&=N\langle(e_i^{n'})^4\rangle+N(N-1)\underbrace{\langle(e_i^{n'})^2(e_i^{n''})^2\rangle}_{n'\neq n''}\\&=N\gamma_i+N(N-1)\sigma_{ii}^2\end{aligned}$$

If the error e_i^n follows a Gaussian distribution, $\gamma_i = 3\sigma_{ii}^2$ and Eq. (14) simplifies to $\frac{1}{N}\left(2\sigma_{ii}^2 + \sigma_{ii}\sigma_{kk} + \sigma_{ii}\sigma_{jj} + \sigma_{jj}\sigma_{kk}\right)$ (Koopmans, 1995). In general the fourth-order moments will neither be known in advance nor estimated from the data. The Gaussian assumption thus provides simplified and approximate expressions for the error variances.

The derivation of the estimation variance leading to Eq. (14) relies on several assumptions – in particular assumption 4 (no autocorrelation), which does not affect the expected value given by Eq. (13). Alternatively a more general error model (including e.g. autocorrelation) could be used to compute the error variance. In practical cases such information is, however, very difficult to obtain; this is why Caires and Sterl (2003); Muraleedharan et al. (2006) suggested the bootstrap method for estimating the error variance from the data.

It is worthwhile to discuss the estimation of the RMSE (root mean square error), i.e. the square root of the error variance if assumption 1 holds. When taking the square root of $\hat{\sigma}_{ii}$ in Eq. (13) as an estimate of the RMSE – $\hat{r}_{ii} = \sqrt{\hat{\sigma}_{ii}}$ – the result will have a negative bias as

$$\mathrm{Var}\left(\hat{r}_{ii}\right) = \langle(\hat{r}_{ii})^2\rangle - \langle\hat{r}_{ii}\rangle^2 \geq 0 \Rightarrow \sqrt{\sigma_{ii}} \geq \langle\hat{r}_{ii}\rangle$$

This result (as well as the derivation) is exactly the same as for the usual sample variance/standard deviation (Kenney and Keeping, 1956) and could also have been derived by recourse to Jensen's inequality.

It is easy to determine whether the obtained covariance matrix is positive definite in case of a diagonal matrix: the diagonal elements must be greater than 0. Note that the collocation technique does not guarantee the retrieval of a valid (i.e. positive definite) covariance matrix.

5.2 Bias model

5.2.1 Calibration constants

In addition to the error structure, the bias terms α_i, $i \neq 1$ have to be estimated when the bias model is assumed. The latter can also be obtained from the differences between two measurements (Muraleedharan et al., 2006). The following estimator seems obvious:

$$\hat{\alpha}_i = \frac{1}{N}\sum_{n=1}^{N} y_i^n - y_1^n = \alpha_i + \frac{1}{N}\sum_{n=1}^{N} e_i^n - e_1^n \tag{15}$$

The unbiasedness follows immediately from the rightmost part by assumption 1.

Given assumptions 1, 2, 3, and 4, the estimator $\hat{\alpha}_i$ is the best linear (in the difference) unbiased estimator, as will now be shown. A general linear estimator is given by $\check{\alpha}_i = \sum_{n=1}^{N} w_n\left(y_i^n - y_1^n\right)$. From assumption 1 and the validity of the bias model, it follows that $\langle\check{\alpha}_i\rangle = \alpha_i \sum_{n=1}^{N} w_n$ such that the weights must sum to one for $\check{\alpha}_i$ to be unbiased. For such an unbiased estimator the variance under assumptions 1, 2, 3, and 4 is given by $\mathrm{Var}(\check{\alpha}_i) = (\sigma_{ii} + \sigma_{11})\sum_{n=1}^{N} w_n^2$, which, for $\hat{\alpha}_i$ from Eq. (15), evaluates to

$$\mathrm{Var}(\hat{\alpha}_i) = (\sigma_{ii} + \sigma_{11})\frac{1}{N} \tag{16}$$

This variance is the smallest value possible as

$$\begin{aligned}\mathrm{Var}(\check{\alpha}_i) - \mathrm{Var}(\hat{\alpha}_i) &= (\sigma_{ii} + \sigma_{11})\left[\left(\sum_{n=1}^{N} w_n^2\right) - \frac{1}{N}\right] \\ &= (\sigma_{ii} + \sigma_{11})\sum_{n=1}^{N}\left(w_n - \frac{1}{N}\right)^2 \\ &\geq 0\end{aligned}$$

where the second line follows from the unbiasedness condition.

5.2.2 Error terms

Let $[i'j';kl] = (y_i^n - \hat{\alpha}_i - y_j^n + \hat{\alpha}_j)(y_k^n - y_l^n)$ and similarly for various other brackets.

In order to derive an unbiased estimator of σ_{11}, we first look at $\sum_{n=1}^{N}\langle[1,j';1,k']^n\rangle$, where $j \neq k$. By recourse to assumptions 1, 2, 3, and 4, this evaluates to $(N-1)\sigma_{11}$. Note that assumption 4 (no autocorrelation) is invoked to derive this result because of the correlation between the individual error terms and the estimated calibration constants $\hat{\alpha}$. Thus we have an unbiased estimator for σ_{11}:

$$\hat{\sigma}_{11} = \frac{1}{N-1}\sum_{n=1}^{N}[1,j';1,k']^n \tag{17}$$

The denominator is $N-1$, as in the unbiased estimation of the population variance from the sample variance. The small bias of the naive estimator with N in the denominator is negligible for reasonably sized samples. Likewise, it can be shown that the following estimators for σ_{jj}, $i \neq 1$ are unbiased ($j \neq k \neq 1$):

$$\hat{\sigma}_{jj} = \frac{1}{N-1}\sum_{n=1}^{N}[j',1;j',k']^n \tag{18}$$

The evaluation of the variance of these two estimators turns out to be a veritable tour de force; it closely follows Kenney and Keeping (1956) and Eq. (14) but there are many more terms. For both Eq. (17) and Eq. (18) it follows that

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}_{uu}) = \frac{1}{N(N-1)^2}\big[&N^2(\gamma_u - 3\sigma_{uu}^2 + \sigma_{uu}\sigma_{vv} \\ &+ \sigma_{uu}\sigma_{ww} + \sigma_{vv}\sigma_{ww}) + N(-6\gamma_u \\ &+ 11\sigma_{uu}^2 - 4\sigma_{uu}\sigma_{vv} - 4\sigma_{uu}\sigma_{ww} \\ &- 4\sigma_{vv}\sigma_{ww}) + (13\gamma_u - 8\sigma_{uu}^2 + 11\sigma_{uu}\sigma_{vv} \\ &+ 11\sigma_{uu}\sigma_{ww} + 11\sigma_{vv}\sigma_{ww})\big]\end{aligned} \tag{19}$$

where $u = 1$, $v = j$, and $w = k$ for Eq. (17); and $u = j$, $v = 1$, and $w = k$ for Eq. (18). Remarkably the expressions in

Eqs. (17) and (18) for the variance are the same and also possess surprising permutational symmetries e.g. with respect to 1 and k in the second case, whereas the one between j and k in the first case is obvious. Comparing the variance given by Eq. (19) with the corresponding one of the basic model shown in Eq. (14), we see that the dominant terms (which diminish with $\frac{1}{N}$) are almost identical; the only difference is the coefficient of σ_{uu}^2 and it is due to the different denominators.

5.3 Affine model

The analysis of the affine model is vastly more difficult than before. It is impossible to draw conclusions along the lines of the basic and the affine model because of the multiplicative nature of the β term.

First the determination of the scale factor β is analyzed. Following Muraleedharan et al. (2006), we first introduce the differenced measurements y'

$$y_i^{'n} = y_i^n - \frac{1}{N}\sum_{n'=1}^{N} y_i^{n'} = y_i^n - \langle y_i \rangle_e$$

the right most part of which defines the empirical average $\langle\cdot\rangle_e$. Similarly, let $x^{'n} = x^n - \langle x\rangle_e$ and $e_i^{'n} = e_i^n - \langle e_i\rangle_e$. The empirical covariance between two differenced measurements in the affine model is given by

$$\left\langle y_1' y_i' \right\rangle_e = \beta_i \langle x'^2\rangle_e + \beta_i \langle x' e_1'\rangle_e + \langle x' e_i'\rangle_e + \langle e_i' e_1'\rangle_e$$
$$\left\langle y_i' y_j' \right\rangle_e = \beta_i\beta_j \langle x'^2\rangle_e + \beta_i \langle x' e_j'\rangle_e + \beta_j \langle x' e_i'\rangle_e + \langle e_i' e_j'\rangle_e$$

The expected values of these two terms are given by (assumptions 1 and 3):

$$\left\langle \langle y_1' y_i'\rangle_e \right\rangle = \beta_i \langle x'^2\rangle_e$$
$$\left\langle \left\langle y_i' y_j'\right\rangle_e \right\rangle = \beta_i\beta_j \langle x'^2\rangle_e$$

where – as explained in Sect. 2.1 – the average is taken with respect to the error terms whereas x is treated as a deterministic parameter. This suggests the following estimator ($i \neq j \neq 1$) (Caires and Sterl, 2003):

$$\hat{\beta}_j = \frac{\left\langle y_i' y_j' \right\rangle_e}{\left\langle y_1' y_i' \right\rangle_e} \tag{20}$$

Due to it being a quotient of two dependent random variables, the statistical properties cannot be derived in the same way as those of α in the bias model.

Assuming β_i known for a moment, it follows from the definition of the affine model that

$$y_i^n - \beta_i y_1^n = \alpha_i - \beta_i e_1^n + e_i^n \tag{21}$$

which suggests the following estimator

$$\hat{\alpha}_i = \frac{1}{N}\sum_{n=1}^{N} y_i^n - \beta_i y_1^n \tag{22}$$

which is actually unbiased if assumption 1 holds. If, however, only an estimate of β_i is available, the complex dependencies between this estimate and the error terms render a straightforward analysis impossible.

It was noted in Sect. 1 that different ways of obtaining estimates of the calibration constants have been proposed, e.g. an iterative scheme based on error-in-variables regression (Scipal et al., 2008) or the linear re-scaling by Miralles et al. (2010); Hain et al. (2011). Alternatively, the calibration could also have been determined in a previous study or based on a completely different method not connected to the collocation technique. The advantage of using the simple estimator for α in the bias model, which essentially just matches the first moments, lies in the conceptual simplicity and the ease with which analytical properties of the estimators can be derived. As the latter breaks down in the presence of the scale factor β, the relative merits and drawbacks of the different estimators of the calibration constants remain an open topic of research.

5.4 Basic model with known correlations

In this section the estimators are adapted so that known (or postulated) correlations between different data sets can be taken into account. An example was already given in Sect. 3, where one non-vanishing covariance was assumed. Following the analysis in this section, it is immediately obvious that the following estimators are unbiased:

$$\hat{\sigma}_{ii} = \sigma_{ij} + \sigma_{ik} - \sigma_{jk} + \frac{1}{N}\sum_{n=1}^{N} [i,j;i,k]^n \tag{23}$$

The computation of the variance can be repeated along the lines of Eq. (14); however, terms such as $\sum_{n'=1}^{N}\sum_{n''=1}^{N}\langle e_j^{n'} e_k^{n'} e_j^{n''} e_k^{n''}\rangle$ are encountered. This one yields σ_{jk}^2 if $n' \neq n''$ by assumption 4 but additional assumptions about the higher order structure are required in order to evaluate it for $n' = n''$. For simplicity's sake it will be assumed that the errors at time n are samples from a multivariate normal distribution, for which we have that (Koopmans, 1995)

$$\langle v_i v_j v_k v_l\rangle = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk} \tag{24}$$

where $v_i \ldots v_l$ are elements of a zero mean multivariate normal distribution and σ_{ij} denotes the (i,j) element of its covariance matrix.

After collecting the terms, the following expressions for the variances are obtained $i \neq j \neq k$:

$$\begin{aligned}\mathrm{Var}\left(\hat{\sigma}_{ii}\right) = \frac{1}{N}\big(&2\sigma_{ii}^2 - 4\sigma_{ii}\sigma_{ij} - 4\sigma_{ii}\sigma_{ik} + \sigma_{ii}\sigma_{jj} \\ &+ 2\sigma_{ii}\sigma_{jk} + \sigma_{ii}\sigma_{kk} + \sigma_{ij}^2 + 6\sigma_{ij}\sigma_{ik} \\ &- 2\sigma_{ij}\sigma_{jk} - 2\sigma_{ij}\sigma_{kk} + \sigma_{ik}^2 - 2\sigma_{ik}\sigma_{jj} \\ &- 2\sigma_{ik}\sigma_{jk} + \sigma_{jj}\sigma_{kk} + \sigma_{jk}^2\big)\end{aligned} \tag{25}$$

which relies on assumptions 1, 2, 2b, 3, 4 and the normality requirement for evaluating the 4th order moments. It is also consistent with Eq. (14) and possesses permutational symmetry with respect to j and k.

5.5 Basic model with one covariance estimated

The last constellation of error terms to be analyzed is the basic model when correlations are estimated. There are two scenarios of interest: (i) all other covariances besides the one estimated vanish and (ii) there are additional non-zero covariances.

5.5.1 Correct covariance estimated

For the remainder of this part, assumption 3 will be generalized: all error covariances apart from σ_{ij} vanish. Among the possible estimators of σ_{ij}, those of the following form are particularly amenable to analysis ($i \neq j \neq k \neq l$):

$$\begin{aligned}\hat{\sigma}_{ij} &= \frac{1}{N}\sum_{n=1}^{N}[i,k;j,l]^n \qquad (26)\\ &= \frac{1}{N}\sum_{n=1}^{N}\left(e_i^n e_j^n - e_i^n e_l^n - e_k^n e_j^n + e_k^n e_l^n\right)\end{aligned}$$

where the unbiasedness follows directly from assumption 1, 2, and 3 (modified).

In order to compute the variance of this estimator, we will proceed as in the previous subsection Sect. 5.4: the error is assumed to follow a multivariate normal distribution and thus Eq. (24) applies. The usual expansion and collection of terms yields

$$\mathrm{Var}(\hat{\sigma}_{ij}) = \frac{1}{N}\left(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj} + \sigma_{ii}\sigma_{ll} + \sigma_{jj}\sigma_{kk} + \sigma_{kk}\sigma_{ll}\right) \quad (27)$$

where also assumption 4 is invoked.

5.5.2 Incorrect covariance estimated

We will now look at the consequences of additional cross-covariances on the estimator of Eq. (26); i.e. when these correlations are not properly accounted for. The estimator of Eq. (26) is generally biased in such cases:

$$\begin{aligned}\langle\hat{\sigma}_{ij}\rangle &= \frac{1}{N}\sum_{n=1}^{N}\left\langle e_i^n e_j^n - e_i^n e_l^n - e_k^n e_j^n + e_k^n e_l^n\right\rangle\\ &= \sigma_{ij} - \sigma_{il} - \sigma_{kj} + \sigma_{kl} \qquad (28)\end{aligned}$$

6 Simulation

In this section the previously gained insight is applied and compared to simulated data; this allows us to study the impact of violated assumptions on the retrieved results. As the emphasis of this paper rests on the generalization of the collocation technique to $M > 3$ and the treatment of cross-correlations, the estimation of calibration constants is forgone and only the basic model of Eq. (1) considered.

At each epoch n the noise terms e_i^n are sampled from a zero mean Gaussian distribution with specified covariance matrix Σ, the numerical values of which will be given in the relevant subsections. These noise terms at different epochs are independent.

The time series of the parameter x^n is simulated as well even though the results of the collocation technique are unaffected due to the inherent differencing, e.g. Eq. (13). It is generated by independently drawing from a uniform distribution (lower limit: 0, upper limit: 10) at each epoch n and subsequently smoothing this result with a 5 element boxcar filter.

A particularly important aspect of the simulation study is the analysis of the results as the number of samples N grows. In this case the new samples are not merely appended to the previous data but the entire sample is re-drawn.

Section 6.1 demonstrates the dependence of the results and their accuracy on the number of samples N available. The influence of cross-correlations on the collocation method is studied in Sect. 6.2. Section 6.3 deals with the possibility of estimating cross-covariances in quadruple collocation, i.e. when $M = 4$.

6.1 Triple collocation

In this first study we will look at three sensors, the (unitless) error covariance matrix $\Sigma_{\mathbf{a}}$ of which is taken to be

$$\Sigma_{\mathbf{a}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

This is thus a standard triple collocation analysis where all assumptions (1, 2, 2b, 3, and 4) are met. Figure 1 shows an exemplary time series generated by the approach described in the previous section.

The error variances can be estimated by recourse to Eq. (13), and more specifically Eqs. (4)–(6). The variance can be computed using Eq. (14) – the simplification for Gaussian noise applies. The results for two sensors as a function of the number of samples N is displayed in Fig. 2; the lines indicate the ± 2SE range, where the standard error SE is the square root of the estimator variance, Eq. (14).

The estimator variance given by Eq. (14) drops off as N^{-1}, which corresponds to a line with slope -1 in a log-log plot; the different multiplicative factors get mapped to different intercepts. This is illustrated for the same two sensors in Fig. 3. The data values are empirically estimated variances: for each N 50 time series are generated and the variance of the estimated $\hat{\sigma}_{ii}$ plotted.

These results about the uncertainty in the estimates allow us to address a question of great importance: the number

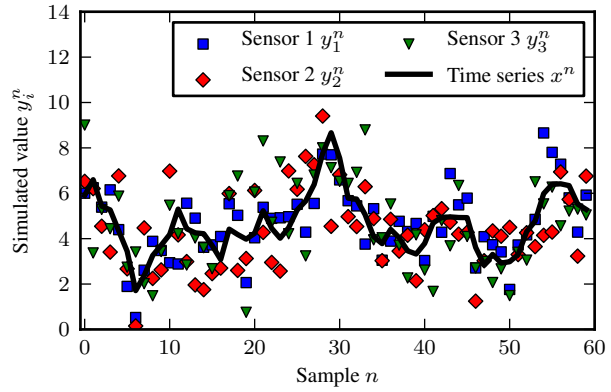

Fig. 1. Exemplary time series with $N = 60$, generated by the approach described in Sect. 6.

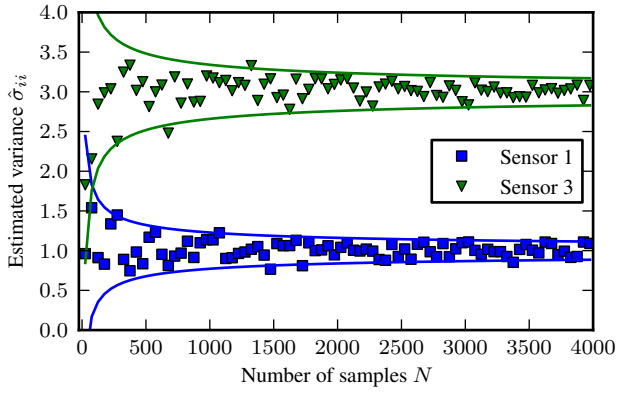

Fig. 2. Estimated variance $\hat{\sigma}_{ii}$ as function of the number of samples N. The solid lines indicate the ± 2SE range (around the actual value), as determined from Eq. (14).

of samples N needed to achieve reliable results. The validity of the variance formula 8 depends on assumptions 1, 2, 2a, 3, and 4. The autocorrelation assumption is particularly problematic in time series studies. Nevertheless, if these assumptions hold and the noise can be modelled as a normal distribution, the variance was shown to be $\frac{1}{N}\left(2\sigma_{ii}^2 + \sigma_{ii}\sigma_{kk} + \sigma_{ii}\sigma_{jj} + \sigma_{jj}\sigma_{kk}\right)$. If all error variances are similar in size this can be approximated by

$$\mathrm{Var}\left(\hat{\sigma}_{ii}\right) \approx \frac{5}{N}\sigma_{ii}^2 \Rightarrow s = \sqrt{\frac{5}{N}}\sigma_{ii} \tag{29}$$

whereas otherwise, we can take sensor i to be the one with the largest error variance and interpret this formula as a conservative bound. In Eq. (29) the simplified standard error s is just the square root of the approximate variance. In practice one is often not particularly interested in the absolute standard error (s), but in the standard error relative to the quantity of interest (σ_{ii}), which has the great advantage of being a dimensionless quantity. In our case this relative error $r = \frac{s}{\sigma_{ii}} = \sqrt{\frac{5}{N}}$. The frequently touted advice (Scipal et al., 2008; Dorigo et al., 2011) of needing at least $N = 100$ samples corresponds to $r = 0.22$; if we want a relative uncertainty of 10 %, we need 500 samples, provided all the previously mentioned assumptions hold. Note that positive autocorrelation (which is the one most commonly encountered) generally increases the standard error and thus results in overly optimistic estimates of the uncertainty, whereas the simplification required to obtain Eq. (29) leads to conservative values of r.

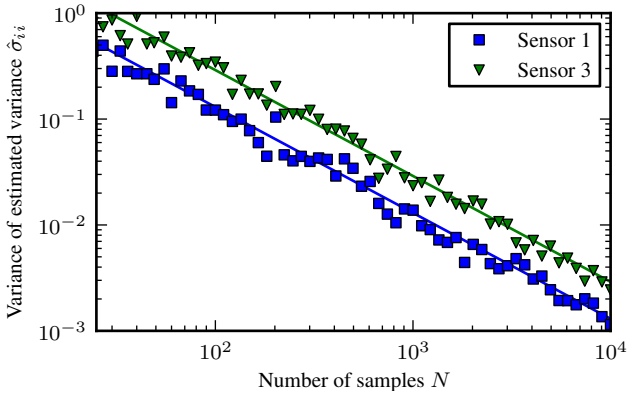

Fig. 3. Convergence of the variance of the estimated variance as the number of samples N grows. The markers indicate the sample variance obtained by running 50 simulations for each N. The solid lines are the theoretical values given by Eq. (14).

6.2 Triple collocation: crosscorrelation

In order to study the influence of a violation of assumption 3, we now take the covariance matrix to be

$$\Sigma_{\mathbf{b}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

which corresponds to a correlation $\rho_{23} = 0.33$. When Eqs. (4)–(6) are adopted for estimating the error variances, the expected value of each of them is 2, i.e. all estimates are wrong. This behaviour is illustrated using the simulated data in Fig. 4. If, on the other hand, the applicable set of equations Eq. (23) is adopted, the correct results are obtained, as is made evident in Fig. 5. Note that the covariance σ_{23} has to be known and that the normality assumption is invoked for computing the standard errors.

6.3 Quadruple collocation

The possibility of estimating error covariances will be demonstrated by applying the collocation technique to a

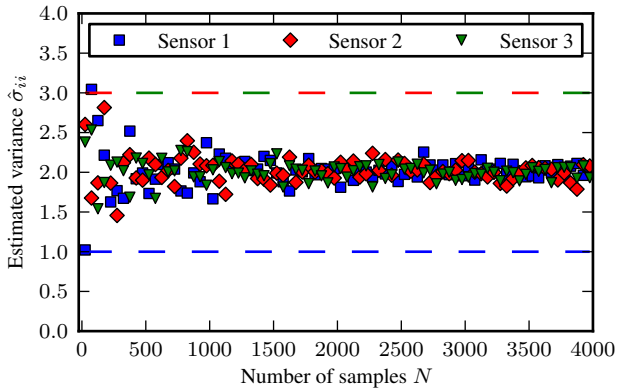

Fig. 4. Estimated variance of the three sensors computed with Eqs. (4)–(6) in the presence of correlations. The dashed lines indicate the correct values.

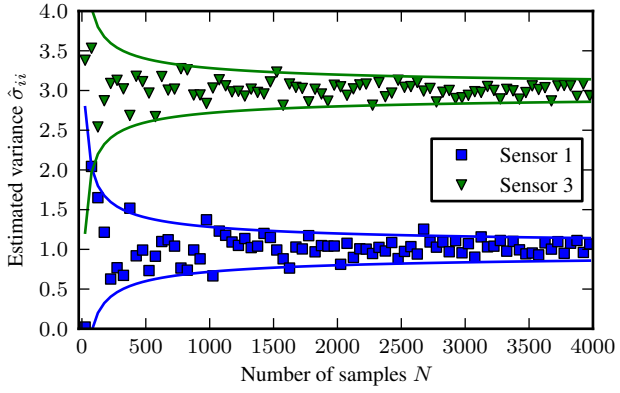

Fig. 5. Estimated variance computed with the correct formula Eq. (23), where the correct covariance σ_{23} is known. The solid lines indicate the ± 2SE range around the actual value.

scenario consisting of four data sets with a covariance matrix $\Sigma_{\mathbf{c}}$

$$\Sigma_{\mathbf{c}} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{5}{2} \end{pmatrix}$$

Only one covariance is estimated as this facilitates the test of the positive-definiteness of the obtained covariance matrix. Figure 6 illustrates the results of the quadruple collocation technique in the presence of one non-zero covariance (σ_{12}) and when the estimators are chosen accordingly: $\hat{\sigma}_{11} = \frac{1}{N}\sum_{n=1}^{N}[1,3;1,4]^n$, $\hat{\sigma}_{22} = \frac{1}{N}\sum_{n=1}^{N}[2,3;2,4]^n$, $\hat{\sigma}_{12} = \frac{1}{N}\sum_{n=1}^{N}[1,3;2,4]^n$.

Figure 7 shows the results of the application of the following estimators to the simulated data:

- ${}_a\hat{\sigma}_{22} = \frac{1}{N}\sum_{n=1}^{N}[2,3;2,4]^n$
- ${}_b\hat{\sigma}_{22} = \frac{1}{N}\sum_{n=1}^{N}[2,1;2,3]^n$

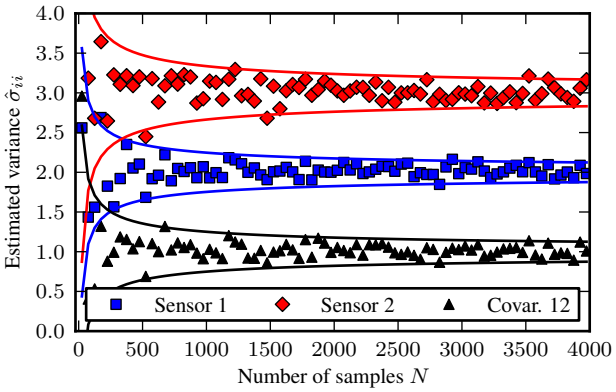

Fig. 6. Estimated variances $\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$ and covariance $\hat{\sigma}_{12}$ when σ_{12} is the only non-zero covariance. The solid lines indicate the ± 2SE range (around the actual value), as determined from Eq. (14) and (27).

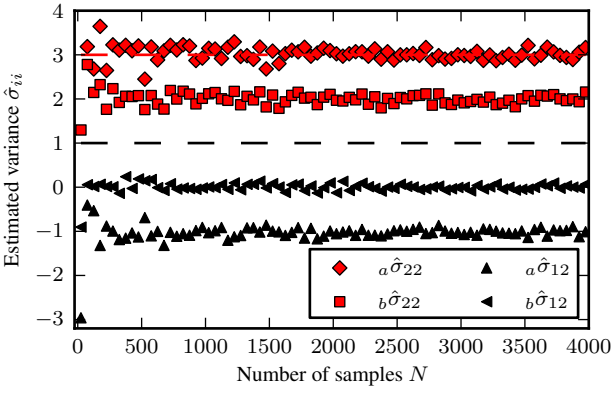

Fig. 7. Estimated variances $\hat{\sigma}_{33}$ and covariance $\hat{\sigma}_{23}$ when σ_{12} is the only non-zero covariance ($\Sigma_{\mathbf{c}}$). The bias between the two estimators for both the variance and the covariance is due to the covariance not taken into account. The dashed lines indicate the correct values.

- ${}_a\hat{\sigma}_{23} = \frac{1}{N}\sum_{n=1}^{N}[2,4;3,1]^n$
- ${}_b\hat{\sigma}_{23} = \frac{1}{N}\sum_{n=1}^{N}[2,1;3,4]^n$

where the left indices a and b denote two different estimators of the same parameter: the expected values of both estimators a and b would clearly be identical, if there were no non-zero covariance terms that were not taken into account. From a practical point of view this offers the possibility of detecting such covariances. However, in light of the results obtained in Sect. 4.3, the collocation technique can only resolve one part of the covariance matrix: the one in the vector space $\mathbf{B}^M$. The detection and estimation of the covariance structure thus has to rely on additional assumptions (apart from the validity of the additive error models and assumptions 1 and 2) and these will likely be determined by expert knowledge and experience with the respective data sources.

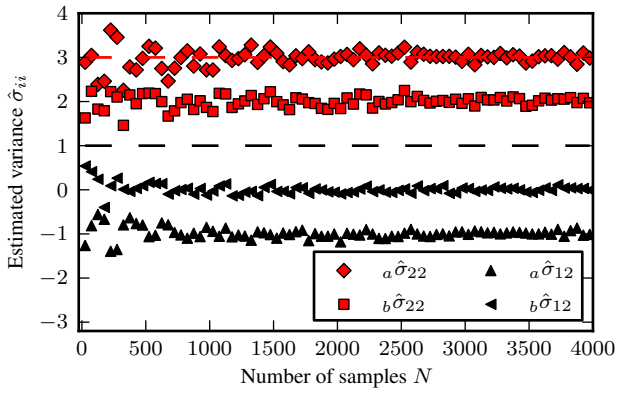

Fig. 8. Estimated variances $\hat{\sigma}_{33}$ and covariance $\hat{\sigma}_{23}$ when σ_{12} is the only non-zero covariance ($\mathbf{\Sigma_d}$). The bias between the two estimators for both the variance and the covariance is due to the covariance not taken into account. The dashed lines indicate the correct values. Note that this example only differs from the one in Fig. 7 in the noise: this illustrates the structural deficiencies of the collocation technique.

As an example consider the same estimators applied to the covariance matrix $\mathbf{\Sigma_d}$:

$$\mathbf{\Sigma_d} = \begin{pmatrix} 2 & 1 & \frac{1}{2} & 0 \\ 1 & 3 & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{5}{2} \end{pmatrix}$$

where the difference to $\mathbf{\Sigma_c}$ is exactly one of those unresolvable elements of the subspace $\mathbf{K}$. The results are plotted in Fig. 8 and apart from the fluctuation due to the noise there is no perceptible difference.

7 Conclusions

The collocation technique has become a widely applied tool for analyzing the error structure of different data sources. Previous studies have mostly concentrated on the comparison of three data sets as this is the minimum number necessary to determine the error variance of each. Despite its popularity the method, as well as its theoretical properties, has not been scrutinized in detail – in this paper we demonstrate which part of the error covariance matrix can be resolved for an arbitrary number of data sources. The second contribution is formed by a detailed analysis of the statistical properties of the various estimators, such as the bias and the standard error.

Particular emphasis is placed on the assumptions necessary to obtain those results. These are notoriously hard to check when working with real data and further investigations are necessary in order to establish suitable tests and plots; this pertains to the possible presence of autocorrelation, time dependence of the error structure and calibration constants, etc. The structural deficiencies of the collocation technique (as derived in Sect. 4) exert an additional confounding influence as they imply that certain correlation structures are simply not resolvable by the method. Expert knowledge about the sensors and models of interest thus remains a necessity for the correct application of the collocation method.

Several simulation studies reveal the consequences of violated assumptions and also serve as empirical confirmation of the results obtained in Sect. 5. The dependence of the accuracy on the number of samples is analyzed in detail and with respect to all the different simulation scenarios.

Acknowledgements. The authors acknowledge the support of the Austrian Space Application Programme (ASAP) through the GMSM project and of the European Space Agency (ESA) through the Climate Change Initiative (CCI) Soil Moisture project. The authors would like to thank the reviewers for their comments and suggestions.

Edited by: D. Maraun
Reviewed by: A. Stoffelen and another anonymous referee

References

Caires, S. and Sterl, A.: Validation of ocean wind and wave data using triple collocation, J. Geophys. Res., 108, 3098–3114, 2003.

Cantrell, C. D.: Modern Mathematical Methods for Physicists and Engineers, Cambridge University Press, 2000.

Dorigo, W. A., Scipal, K., Parinussa, R. M., Liu, Y. Y., Wagner, W., de Jeu, R. A. M., and Naeimi, V.: Error characterisation of global active and passive microwave soil moisture datasets, Hydrol. Earth Syst. Sci., 14, 2605–2616, doi:10.5194/hess-14-2605-2010, 2010.

Hain, C., Crow, W., Mecikalski, J., Anderson, M., and Holmes, T.: An intercomparison of available soil moisture estimates from thermal infrared and passive microwave remote sensing and land surface modeling, J. Geophys. Res., 116, D15107, doi:10.1029/2011JD015633, 2011.

Janssen, P., Abdallah, S., Hersbach, H., and Bidlot, J.-R.: Error Estimation of Buoy, Satellit, and Model Wave Height Data, J. Atmos. Ocean. Technol., 24, 1665–1677, 2007.

Kenney, J. and Keeping, E.: Mathematic of Statistics Part Two, Van Nostrand, 2nd Edn., 1956.

Koopmans, L.: The spectral analysis of time series, Academic Press, 1995.

Lang, S.: Linear Algebra, Springer, 1987.

Liu, Y. Y., Parinussa, R. M., Dorigo, W. A., De Jeu, R. A. M., Wagner, W., van Dijk, A. I. J. M., McCabe, M. F., and Evans, J. P.: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrol. Earth Syst. Sci., 15, 425–436, doi:10.5194/hess-15-425-2011, 2011.

Loew, A. and Schlenz, F.: A dynamic approach for evaluating coarse scale satellite soil moisture products, Hydrol. Earth Syst. Sci., 15, 75–90, doi:10.5194/hess-15-75-2011, 2011.

Miralles, D., Crow, W., and Cosh, M.: Estimating Spatial Sampling Errors in Coarse-Scale Soil Moisture Estimates Derived from Point-Scale Observations, J. Hydrometeor., 11, 1423–1429, 2010.

Miralles, D. G., De Jeu, R. A. M., Gash, J. H., Holmes, T. R. H., and Dolman, A. J.: Magnitude and variability of land evaporation and its components at the global scale, Hydrol. Earth Syst. Sci., 15, 967–981, doi:10.5194/hess-15-967-2011, 2011.

Missaoui, O., Frigui, H., and Gader, P.: Land-Mine Detection With Ground-Penetrating Radar Using Multistream Discrete Hidden Markov Models, IEEE Trans. Geosci. Remote Sens., 49, 2080–2099, 2011.

Munro, R., Köpken, C., Kelly, G., Thépaut, J., and Saunders, R.: Assimilation of Meteosat radiance data within the 4D-Var system at ECMWF: Data quality monitoring, bias correction and single-cycle experiments, Q. J. R. Meteorol. Soc., 130, 2293–2313, 2004.

Muraleedharan, G., Rao, A., Mourani, S., and Mahapatra, D.: Analysis of Triple Collocation Method for validation of model predicted significant wave height data, J. Ind. Geophys. Union, 10, 79–84, 2006.

O'Carroll, A., Eyre, J., and Saunders, R.: Three-Way Error Analysis between AATSR, AMSR-E, and In Situ Sea Surface Temperature Observations, J. Atmos. O., 25, 1197–1207, 2008.

Parinussa, R., Meesters, A., Liu, Y., Dorigo, W., Wagner, W., and de Jeu, R.: Error Estimates for Near-Real-Time Satellite Soil Moisture as Derived From the Land Parameter Retrieval Model, IEEE Geosci. Remote Sens. Lett., 8, 779–783, 2011.

Parrens, M., Zakharova, E., Lafont, S., Calvet, J.-C., Kerr, Y., Wagner, W., and Wigneron, J.-P.: Comparing soil moisture retrievals from SMOS and ASCAT over France, Hydrol. Earth Syst. Sci. Discuss., 8, 8565–8607, doi:10.5194/hessd-8-8565-2011, 2011.

Scipal, K., Holmes, T., de Jeu, R., Naeimi, V., and Wagner, W.: A possible solution for the problem of estimating the error structure of global soil moisture data sets, Geophys. Res. Lett., 35, L24403, doi:10.1029/2008GL035599, 2008.

Stoffelen, A.: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, J. Geophys. Res., 103, 7755–7766, 1998.

Vogelzang, J., Stoffelen, A., Verhoef, A., and Figa-Saldaña, J.: On the quality of high-resolution scatterometer winds, J. Geophys. Res., 116, C10033, doi:10.1029/2010JC006640, 2011.

Winterfeldt, J., Andersson, A., Klepp, C., Bakan, S., and Weisse, R.: Comparison of HOAPS, QuikSCAT, and Buoy Wind Speed in the Eastern North Atlantic and the North Sea, IEEE Trans. Geosci. Remote Sens., 48, 338–348, 2010.