Atmos. Chem. Phys., 6, 4311–4320, 2006
www.atmos-chem-phys.net/6/4311/2006/


Atmospheric Chemistry and Physics

# Validation of remotely sensed profiles of atmospheric state variables: strategies and terminology

**T. von Clarmann**

Forschungszentrum Karlsruhe, Institut für Meteorologie und Klimaforschung, Karlsruhe, Germany

Received: 18 April 2006 – Published in Atmos. Chem. Phys. Discuss.: 20 June 2006
Revised: 16 August 2006 – Accepted: 15 September 2006 – Published: 25 September 2006

**Abstract.** This paper summarizes and classifies the various approaches to validation of remote measurements of atmospheric state variables, and tries to recommend a clear and unambiguous terminology. The following approaches have been identified: Intercomparison of individual profiles for accuracy validation; statistical comparison of matched pairs of measurements with respect to bias determination and precision validation; statistical intercomparison of randomly sampled measurements by two instruments, and comparison of a single measurement to an ensemble of measurements. Applicable statistics are shortly reviewed, and recipes for evaluation of the co-incidence error due to less than perfect coincidences are presented. An approach is suggested to quantitatively validate profile measurements when full covariance matrices are unavailable.

## 1 Introduction

Validation of a data product we understand means to confirm the predicted accuracy of the data product. After a series of self-consistence texts (e.g. Rodgers, 2000), the key element of validation is a statistical analysis of the differences between measurements of a new instrument to be validated, and of a reference instrument already validated. The purpose is to detect and remove any potential bias of the new measurement, and to verify that the estimated precision of the new measurement characterizes the measurements correctly.

Without any validated and bias-corrected reference measurement available, it may also be helpful to intercompare measurements by two or more non-validated instruments. This approach we call "cross validation". While this approach certainly is no validation in its rigorous sense, it still may help to better characterize the data products.

*Correspondence to:* T. von Clarmann
(thomas.clarmann@imk.fzk.de)

## 2 Terminology

Let $\boldsymbol{x}=(x_1, \cdots, x_N)^T$ be a vertical profile of an atmospheric state variable, sampled on a discrete vertical grid of $N$ altitude gridpoints, describing the true atmospheric state at the altitude resolution of the measurement to be validated. Let further $\hat{\boldsymbol{x}}=(\hat{x}_1, \cdots, \hat{x}_N)^T$ be a measurement of $\boldsymbol{x}$. The accuracy[1] $\boldsymbol{a}$ of the measurement $\hat{\boldsymbol{x}}$ we understand is the square root of the expectation value of the squared differences of the true quantities $x_n$ and their measurements $\hat{x}_n$:

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} \sqrt{\langle(\hat{x}_1 - x_1)^2\rangle} \\ \sqrt{\langle(\hat{x}_2 - x_2)^2\rangle} \\ \vdots \\ \sqrt{\langle(\hat{x}_N - x_N)^2\rangle} \end{pmatrix} \tag{1}$$

The bias $\boldsymbol{b}$ of a measurement is the expectation value of the deviation of the measured and the true quantity:

$$\boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix} = \langle \hat{\boldsymbol{x}} - \boldsymbol{x} \rangle. \tag{2}$$

Depending on the nature of the bias, it can also be multiplicative rather than additive and then is better reported as a relative quantity:

$$\boldsymbol{b}_{\text{mult.}} = \begin{pmatrix} b_{\text{mult.};1} \\ b_{\text{mult.};2} \\ \vdots \\ b_{\text{mult.};N} \end{pmatrix} = \langle \begin{pmatrix} \frac{\hat{x}_1}{x_1} - 1 \\ \frac{\hat{x}_2}{x_2} - 1 \\ \vdots \\ \frac{\hat{x}_N}{x_N} - 1 \end{pmatrix} \rangle \tag{3}$$

[1] In the statistics as well as remote sensing literature there are at least two different definitions of "accuracy". The definition in this paper is consistent with the one used by, e.g. Rodgers (2000), Haseloff and Hoffmann (1970) or Walther and Moore (2005), while it is in conflict with Bevington (1969), whose "accuracy" corresponds to the quantity which is called "bias" here.

Published by Copernicus GmbH on behalf of the European Geosciences Union.

The precision $\boldsymbol{p}$ of the measurement characterizes the reproduceability of the measurement,

$$\boldsymbol{p}=\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} \sqrt{\langle(\hat{x}_1-\langle\hat{x}_1\rangle)^2\rangle-\langle(x_1-\langle x_1\rangle)^2\rangle} \\ \sqrt{\langle(\hat{x}_2-\langle\hat{x}_2\rangle)^2\rangle-\langle(x_2-\langle x_2\rangle)^2\rangle} \\ \vdots \\ \sqrt{\langle(\hat{x}_N-\langle\hat{x}_N\rangle)^2\rangle-\langle(x_N-\langle x_N\rangle)^2\rangle} \end{pmatrix}, \quad (4)$$

where the second term under the square root corrects for the natural variability of $\boldsymbol{x}$. Accuracy, bias, and precision are related by

$$\begin{pmatrix} a_1^2 \\ a_2^2 \\ \vdots \\ a_N^2 \end{pmatrix} = \begin{pmatrix} b_1^2 \\ b_2^2 \\ \vdots \\ b_N^2 \end{pmatrix} + \begin{pmatrix} p_1^2 \\ p_2^2 \\ \vdots \\ p_N^2 \end{pmatrix} \quad (5)$$

Usually, remotely sensed data are provided along with careful data characterization, which includes estimates of the random error covariance matrix $\mathbf{S}_{\text{random}}$ and the systematic error covariance matrix $\mathbf{S}_{\text{sys}}$. In the case of remote measurements, these error estimates typically are the linear mapping of known uncertainties (measurement noise, model parameter uncertainties etc.) onto the retrieved quantities $\hat{x}_n$ (Rodgers, 1990). This means, that these variances and covariances are ex ante estimates, i.e., they do not rely on any statistical comparison with external data. If the measurement includes a priori information, also the smoothing error, characterized by the covariance matrix $\mathbf{S}_{\text{smooth}}$, representing the mapping of the difference between the a priori assumption and the true state onto the measurement, contributes to the total error budget (Rodgers, 2000), and we get for the total error covariance matrix $\mathbf{S}_{\text{total}}$:

$$\mathbf{S}_{\text{total}} = \mathbf{S}_{\text{s+r}} + \mathbf{S}_{\text{smooth}} = \mathbf{S}_{\text{sys}} + \mathbf{S}_{\text{random}} + \mathbf{S}_{\text{smooth}}, \quad (6)$$

where $\mathbf{S}_{\text{s+r}}$ is the sum of $\mathbf{S}_{\text{random}}$ and $\mathbf{S}_{\text{sys}}$. The diagonal elements of these matrices are the related variances $\sigma^2_{\text{total};n}=s_{\text{total};n,n}$, $\sigma^2_{\text{sys};n}=s_{\text{sys};n,n}$, $\sigma^2_{\text{random};n}=s_{\text{random};n,n}$, and $\sigma^2_{\text{smooth};n}=s_{\text{smooth};n,n}$, respectively. The smoothing error can be composed of components constant with time ($\sigma^2_{\text{smooth,sys};n}$) and components randomly varying with time ($\sigma^2_{\text{smooth,random};n}$). Validation then means to confirm the ex ante error estimates by verification that for all $n$ from 1 to $N$

$$a_n^2 = \sigma^2_{\text{s+r};n} + \sigma^2_{\text{smooth};n} = \sigma^2_{\text{total};,n} \quad (7)$$

$$b_n^2 = \sigma^2_{\text{sys};n} + \sigma^2_{\text{smooth,sys};n} \quad (8)$$

$$p_n^2 = \sigma^2_{\text{random};n} + \sigma^2_{\text{smooth,random};n}. \quad (9)$$

This is not as easy as it might seem, because the true atmospheric state $\boldsymbol{x}$, which is needed to evaluate precision, bias, and accuracy with the formalism outlined above is not available. Instead, we use independent reference measurements, which allow to infer ex post estimates of bias, precision, and accuracy. A useful strategy in validation is to first search for a possible bias, to quantify the bias in order to allow its correction, and to finally validate the estimated precision. Optimally, also the causes of the bias will be understood and removed. The scope of this paper, however, is restricted to the detection and quantification of the bias, and the validation of the precision estimates.

## 3 Comparison of co-incident measurements

### 3.1 General aspects

Let $\hat{\boldsymbol{x}}_{\text{val}}$ and $\hat{\boldsymbol{x}}_{\text{ref}}$ be two vertical profiles of the same quantity, measured by instruments "val" (instrument to be validated) and "ref" (independent reference instrument), respectively. The profiles and related diagnostic data have to be represented on a common grid, which usually implies regridding of one or both profiles (Calisesi et al., 2005). Further, if the measurements include a priori information, both profiles have to be transformed to the same a priori profile, and the smoothing error of the difference, $\mathbf{S}_{\text{smooth,diff}}$, has to be estimated (Rodgers and Connor, 2003). This smoothing error difference can be minimized or nullified using a method proposed by Ridolfi et al. (2006a), or it can be restricted to sub-scale differences by transformation of the data to a dedicated representation (von Clarmann and Grabowski, 2006).

Rodgers and Connor (2003) suggest to quantify profile intercomparison by application of a $\chi^2$ test.

$$\chi^2 = (\hat{\boldsymbol{x}}_{\text{val}} - \hat{\boldsymbol{x}}_{\text{ref}})^T \mathbf{S}^{-1}_{\text{diff}} (\hat{\boldsymbol{x}}_{\text{val}} - \hat{\boldsymbol{x}}_{\text{ref}}), \quad (10)$$

where $\mathbf{S}_{\text{diff}}$ is the ex ante estimate of the error covariance matrix of the difference $\hat{\boldsymbol{x}}_{\text{val}}-\hat{\boldsymbol{x}}_{\text{ref}}$. The actual value of $\chi^2$ allows to conclude if the differences $\hat{\boldsymbol{x}}_{\text{val}}-\hat{\boldsymbol{x}}_{\text{ref}}$ are consistent with the ex ante estimates of the uncertainties of the difference, represented by its covariance matrix $\mathbf{S}_{\text{diff}}$, or if there is a significant inconsistency. The integral of the $\chi^2$ probability density function from the actual $\chi^2$ to infinity yields the probability $P_{\text{acc}}(\chi^2)$ that the actual $\chi^2$ value occurs accidently, i.e. that the differences can be explained by the error estimates; the integral from zero to the actual $\chi^2$ yields the probability of a substantial difference. In particular, underestimation of the variances of the differences would lead to a $\chi^2$ value larger than its expectation value, which, in the case of a regular $\mathbf{S}_{\text{diff}}$, is the number of comparison pairs. If the probability of a substantial difference is above a certain threshold, e.g. 95%, then the difference is significant at 5% confidence level and there is statistical evidence that there is something wrong with the data or related error estimates. If this probability is below 95%, then the falsification has failed because the difference is not significant at the chosen (5%) confidence level.

Usually, more than one pair of co-incident profiles is available, and Eq. (10) can be applied to a larger ensemble of comparison pairs. For a large ensemble of $K$ independent

intercomparison pairs, the $\chi^2$ test can either be performed for $L$ subsets of data or as one single $\chi^2$ test involving the $\chi^2$ probability density function of $K$ degrees of freedom. The division into subsets allows to check if the $\chi^2$ values found follow the expected $\chi^2$ distribution. E.g., Migliorini et al. (2004) have detected suspicious ozone profiles in their comparison ensemble by comparison of the expected and the found $\chi^2$ distribution. The probability of disagreement for the complete ensemble, $P_{\rm dis}(L, \chi^2_{\rm max})=1-P_{\rm acc}(L, \chi^2_{\rm max})$, can be estimated according to the multiplication axiom,

$$P_{\rm dis}(L, \chi^2_{\rm max}) \le P_{\rm dis}(1, \chi^2_{\rm max})^L, \tag{11}$$

where $\chi^2_{\rm max}$ is the largest $\chi^2$ value found in the ensemble of comparison subsets, and $P_{\rm dis}(1, \chi^2_{\rm max})$ is the related probability of substantial disagreement in the subset where $\chi^2_{\rm max}$ is found. The drawback of this approach is that it is not sufficient in a sense that the probability estimate is based on the maximum $\chi^2$ value only and thus does not use all available information. This implies that this test is not very robust because it is very sensitive to outliers. For determination of $P_{\rm acc}(\chi^2)$ the single $\chi^2$ test involving the complete ensemble of comparison pairs is superior because of its inherent sufficiency. The safest is to combine both approaches. Discrepancies can then point at non-representative outliers in the comparison ensemble.

Large probabilities of substantial differences can have three different causes: 1. The ex ante error estimates may have been too optimistic. 2. The initial ensemble size was chosen too small and not representative. In this case, a larger comparison ensemble may help to achieve a larger $P_{\rm acc}(\chi^2)$. If, however, the initial sample was representative, even larger $\chi^2$ values will most probably occur in a larger ensemble, and $P_{\rm acc}(\chi^2)$ will not improve. It is, of course, important to work with pre-defined random samples and not to adjust the sample or the sample size to the maximum $P_{\rm acc}(\chi^2)$. 3. Large $\chi^2$ can also be associated with a particular subset of the sample which can be characterized by some objective criterion. Migliorini et al. (2004), e.g. have found problems in $O_3$ data from spectra suspected to be cloud contaminated. In such a case it may be appropriate to define a kind of data filter and to validate only the subset of the data which passes the filter. There are, however, two traps in this approach: First, the filter should not use the quantity to be validated itself as a filter criterion. Second, the new analysis system, of which the newly defined filter is a part, has to be validated using an independent comparison ensemble. When the original sample is used, it will always be possible to tune the data filter such that good agreement between the intercompared data is achieved.

While, in the case of good ex ante error estimates, we certainly can do better then just reject the hypothesis of inconsistency at 5% confidence level, which still allows a probability $P_{\rm dis}(L, \chi^2_{\rm max})$ of up to 95%, we cannot prove equivalence of $\hat{\boldsymbol{x}}_{\rm val}$ and $\hat{\boldsymbol{x}}_{\rm ref}$ at a confidence level better than 50%, because

$$\lim_{L\to\infty} P_{\rm acc}(L, \langle\chi^2_{\rm max}\rangle) = 0.5. \tag{12}$$

Knowledge of all error terms contributing to $\mathbf{S}_{\rm diff}$ is crucial to allow a meaningful estimate of the significance of the differences $\hat{\boldsymbol{x}}_{\rm val}-\hat{\boldsymbol{x}}_{\rm ref}$. The ex ante estimate of the covariance matrix of the difference with elements $s_{{\rm diff};m,n}$ is usually calculated as

$$\mathbf{S}_{\rm diff} = \mathbf{S}_{\rm s+r,val} + \mathbf{S}_{\rm s+r,ref} + \mathbf{S}_{\rm coinc.} + \mathbf{S}_{\rm smooth,diff}, \tag{13}$$

where $\mathbf{S}_{\rm coinc.}$ describes the spatial and temporal co-incidence error in terms of variances and co-variances, which are important to be quantified and considered in the case of less than perfect co-incidences of the two measurements (see Sect. 3.2). If both the validation and the reference measurement have a common error source, this introduces correlations. This applies e.g. when the same or correlated temperature profiles or spectroscopic data are used to derive both $\hat{\boldsymbol{x}}_{\rm val}$ and $\hat{\boldsymbol{x}}_{\rm ref}$. In the case of such correlations, $\mathbf{S}_{\rm diff}$ can be evaluated as

$$\mathbf{S}_{\rm diff} = (\mathbf{I}, -\mathbf{I}) \begin{pmatrix} \mathbf{S}_{\rm s+r,val}, \mathbf{C}_{\rm s+r,val,ref} \\ \mathbf{C}^T_{\rm s+r,val,ref}, \mathbf{S}_{\rm s+r,ref} \end{pmatrix} (\mathbf{I}, -\mathbf{I})^T + \mathbf{S}_{\rm coinc.} + \mathbf{S}_{\rm smooth,diff}, \tag{14}$$

where $\mathbf{I}$ is $N\times N$ unity and where matrix $\mathbf{C}_{\rm s+r}$ contains the related covariance elements $r_{{\rm s+r,val,ref};m,n}\sigma_{{\rm s+r,val},m}\sigma_{{\rm s+r,ref};n}$ between the new measurement "val" and the reference measurements "ref", where $r_{{\rm s+r,val,ref};m,n}$ is the correlation coefficient of the combined systematic and random errors of the validation measurement at altitude $m$ and the reference measurement at altitude $n$. Comparison of two individual profiles does not allow to distinguish between precision and bias validation.

### 3.2 Determination of co-incidence error in time and space

Usually, only profiles are selected for comparison which meet a certain co-incidence criterion in time and space or any other adequate co-ordinates $d$ like solar zenith angle, potential vorticity, equivalent latitude etc. The actual difference $\Delta d$ in this quantity is the mismatch, and the maximum allowed mismatch is the co-incidence criterion $\Delta d_{\rm max}$.

Variability of most atmospheric state variables is composed of a functional term and a random term. The abundance of a certain species, for example, may have a typical latitudinal dependence or a typical diurnal variation, which are superimposed by random fluctuations caused by the actual small-scale atmospheric situation. Whenever applicable, the functional term should be corrected first by a correction function $\boldsymbol{M}$, which can either be some appropriate parametrization, or alternatively a tabulated data set. With $d_{\rm val}$ and $d_{\rm ref}$ being the co-ordinates of the validation and

reference measurements, respectively, the uncorrected reference measurement $\hat{\boldsymbol{x}}_{\mathrm{ref,uncorrected}}$ is corrected as

$$\hat{\boldsymbol{x}}_{\mathrm{ref}} = \hat{\boldsymbol{x}}_{\mathrm{ref,uncorrected}} + \boldsymbol{M}(d_{\mathrm{val}}) - \boldsymbol{M}(d_{\mathrm{ref}}) \quad (15)$$

and only the residual random part of the co-incidence error with respect to the corrected reference measurement $\hat{\boldsymbol{x}}_{\mathrm{ref}}$ should be characterized by the covariance matrix $\mathbf{S}_{\mathrm{coinc.}}$. Otherwise the co-incidence error may not follow a Gaussian distribution, and errors due to systematic sampling differences in $d$ may inadvertently be treated as random co-incidence errors. This may happen, e.g., if the abundance of an atmospheric constituent which is characterized by a strong diurnal change is observed by two instruments at instrument-specific local times. An example of application of a correction function $\boldsymbol{M}$ is found in Ridolfi et al. (2006b)[2] who use ECMWF (European Centre for Medium-Range Weather Forecasts) temperature analyses to estimate the component of the differences between MIPAS and radiosonde temperatures which are explained by mismatch in space and time. A similar approach was chosen by Cortesi et al. (2006)[3] for ozone.

To quantify the residual co-incidence error caused by finer structures in $d$ than those accounted for by the correction function $\boldsymbol{M}$, a sufficiently fine resolved typical reference data set $\hat{\boldsymbol{x}}_{\mathrm{r}}$ of state variable $\boldsymbol{x}(d)$ is needed. Let the reference data set contain $K(\Delta d)$ independent pairs of data points separated by the mismatch $\Delta d{=}d_{\mathrm{val}}{-}d_{\mathrm{ref}}$. Then, the co-incidence error $\mathbf{S}_{\mathrm{coinc.}}$ can be evaluated as a function of $\Delta d$ as

$$s_{\mathrm{coinc.};m,n}(\Delta d) = \frac{\sum_{k=1}^{K(\Delta d)} \left(\Delta\hat{x}_{\mathrm{r};m}(\Delta d)\right)_k \left(\Delta\hat{x}_{\mathrm{r};n}(\Delta d)\right)_k}{K} - s_{\mathrm{err,diff};m,n} \quad (16)$$

where

$$\left(\Delta\hat{x}_{\mathrm{r};m}(\Delta d)\right)_k = \left(\hat{x}_{\mathrm{r};m}(d) - \hat{x}_{\mathrm{r};m}(d+\Delta d) - M_m(d) + M_m(d+\Delta d)\right)_k \quad (17)$$

and

$$\left(\Delta\hat{x}_{\mathrm{r};n}(\Delta d)\right)_k = \left(\hat{x}_{\mathrm{r};n}(d) - \hat{x}_{\mathrm{r};n}(d+\Delta d) - M_n(d) + M_n(d+\Delta d)\right)_k; \quad (18)$$

$m$ and $n$ identify the profile gridpoints, and

$$s_{\mathrm{err,diff};m,n} = 2s_{\mathrm{err};m,n}, \quad (19)$$

where $s_{\mathrm{err};m,n}$ is an element of the random error profile covariance matrix $\mathbf{S}_{\mathrm{err}}$ of the reference data set of the state variable $\hat{x}_r$. The factor of two accounts for error propagation through the calculation of the difference. $\mathbf{S}_{\mathrm{err}}$, which is assumed constant and uncorrelated with geolocation, cannot be obtained from the scatter of the reference sample, because the latter contains the natural variability we are trying to isolate. Instead, a true ex ante estimate is needed, e.g. by error propagation calculation, sensitivity studies or similar means. The $M$ terms account for the difference already explained by the functional mismatch correction.

In order to get $K(\Delta d)$ large enough for meaningful statistics, binning of $\mathbf{S}_{\mathrm{coinc.}}$ is recommended, i.e. evaluation of $\mathbf{S}_{\mathrm{coinc.}}([\Delta d_1, \Delta d_2])$ for all mismatches in a range from $\Delta d_1$ to $\Delta d_2$, where $\mathbf{S}_{\mathrm{coinc.}}$ is sufficiently linear in $\Delta d$. If such a bin $[\Delta d_1, \Delta d_2]$ covers the entire co-incidence criterion, i.e. $\Delta d_1{=}0$ and $\Delta d_2$ equals the co-incidence criterion, it is no longer necessary to care about the $\Delta d$-dependence of $\mathbf{S}_{\mathrm{coinc.}}$ but the mean co-incidence error $\overline{\mathbf{S}_{\mathrm{coinc.}}}{\approx}\mathbf{S}_{\mathrm{coinc.}}(\overline{\Delta d})$ can be used for the entire ensemble of co-incidences.

Meteorological analyses, satellite measurements or modeled atmospheric fields can be used as reference data sets to evaluate the co-incidence error on a larger scale. It is important to carefully assess any possible reduction of the horizontal variability in these datasets through application of background or a priori knowledge in the sense of variational data assimilation (e.g. Ide et al., 1997) or optimal estimation retrievals (Rodgers, 1976), respectively. For determination of small-scale temporal fluctuations stationary in situ measurements or ground-based remote sensing measurements are better suited, while for small-scale spatial fluctuations aircraft measurements are the first choice.

Multi-dimensional co-incidence can be assessed component-wise by evaluation of Eq. (16) for each co-incidence direction (e.g. latitude, longitude and time) and summing up the respective co-incidence error covariance matrices. In the case where the variation of the state variable under assessment is correlated between two of these dimensions, the summation has to be replaced by the following scheme:

$$\mathbf{S}_{\mathrm{coinc.}} = (\mathbf{I}, \mathbf{I}) \begin{pmatrix} \mathbf{S}_{\mathrm{coinc.};1}, \mathbf{C}_{\mathrm{coinc.};1,2} \\ \mathbf{C}^T_{\mathrm{coinc.};1,2}, \mathbf{S}_{\mathrm{coinc.};2} \end{pmatrix} (\mathbf{I}, \mathbf{I})^T, \quad (20)$$

where the subscripts of the covariance matrices $\mathbf{S}_{\mathrm{coinc.};l}$ and the cross-dimension covariances $\mathbf{C}_{\mathrm{coinc.};k,l}$ denote the dimensions along which the variabilities are analyzed. Such correlations may apply, e.g., to the mixing ratio of an inert trace gas the abundance of which is ruled by transport processes. The existence of a prevailing direction of wind in combination with a prevailing gradient in the field of the state variable then introduces such correlations.

Another option to handle co-incidence errors in $L$ dimensions is to define a norm of the following type which transforms the multi-dimensional mismatch

[2] Ridolfi, M., Blum, U., Carli, B., et al.: Geophysical Validation of temperature retrieved from MIPAS/ENVISAT atmospheric Limb-emission measurements, Atmos. Chem. Phys. Discuss., in preparation, 2006b.

[3] Cortesi, U., Blom, C., Blumenstock, Th., et al.: Co-ordinated validation activity and quality assessment of MIPAS-ENVISAT Ozone data, Atmos. Chem. Phys. Discuss., in preparation, 2006.

$\Delta\boldsymbol{d}=(\Delta d_1,\ldots,\Delta d_L)$ to a scalar mismatch distance $\Delta d$:

$$\Delta d = \sqrt{\sum_1^L (w_l \Delta d_l)^2}, \tag{21}$$

where $w_l$ are weighting factors reflecting the expected variability of the state variable with the respective direction $l$. Steck et al. (2006)[4], e.g., have used

$$\Delta d = \sqrt{\Delta_{\rm long}^2 + \Delta_{\rm lat}^2 + (\Delta_{\rm t} v_{\rm w})^2} \tag{22}$$

where $\Delta_{\rm long}$ and $\Delta_{\rm lat}$ are longitudinal and latitudinal mismatch distances, $\Delta_{\rm t}$ is the mismatch in time, $v_{\rm w}$ is the typical windspeed. This particular norm holds for analysis of transport-dominated abundances of trace species without prevailing gradients and wind directions.

### 3.3 Smoothing

Additional complication arises if the measurements to be compared characterize air parcels of non-zero extension in the direction of $d$. In this case the smoothing error in direction of $d$ and the co-incidence error can no longer be treated as independent. Smoothing error and co-incidence error are errors of the same nature, since both characterize differences caused by the fact that two instruments observe different parts of the atmosphere. The pure co-incidence error describes the error component caused by the variation of the atmospheric state over a distance $\Delta d$ between the two disjoint air parcels observed by the two instruments to be compared. Contrary to that, the smoothing error difference quantifies the error component caused by the different weights of different parts within one air parcel observed by two instruments. Both error terms can be estimated by one formalism which includes both the smoothing error application and the co-incidence error application as well as the case of partly overlapping air parcels.

Here we first discuss the quite general case that for both the reference measurement and the measurement to be validated smoothing in all three spatial directions has to be considered, and where the observed air parcels may or may not overlap. Inclusion of additional dimensions (e.g. time) is straightforward and will not explicitly be discussed here. Later, some convenient simplifications will be mentioned.

In a first step, we store all relevant elements of the 3-dimensional (3-D) fine-resolved field of the atmospheric state variable under assessment in a vector $\boldsymbol{x}_{3-\rm D}=(x_{3-{\rm D};1},\ldots,x_{3-{\rm D};L})^T$. Those elements are considered relevant, whose entries in the rows of either of the 3-D averaging kernel matrices of the size $L\times L$, $\mathbf{A}_{3-\rm D,ref}$ and $\mathbf{A}_{3-\rm D,val}$, are non-zero, i.e., elements of the 3-D field which are seen by at least one of both instruments. When reshaping the 3-D field to the vector $\boldsymbol{x}_{3-\rm D}$, the ordering is arbitrary but unambiguous and we use the notation $l(i,j,n)$ for the $l$-th element of the vector which represents the element of the $n$-th altitude, and the $i$-th and $j$-th geolocation in the original 3-D field. Following the concept of Rodgers (2000) but disregarding noise and other measurement errors, the atmospheric state as seen by the instrument to be validated and the reference instrument are

$$\hat{\boldsymbol{x}}_{3-\rm D,val} = \mathbf{A}_{3-\rm D,val}\boldsymbol{x}_{3-\rm D} + (\mathbf{I}-\mathbf{A}_{3-\rm D,val})\boldsymbol{x}_{3-\rm D,a} \tag{23}$$

and

$$\hat{\boldsymbol{x}}_{3-\rm D,ref} = \mathbf{A}_{3-\rm D,ref}\boldsymbol{x}_{3-\rm D} + (\mathbf{I}-\mathbf{A}_{3-\rm D,ref})\boldsymbol{x}_{3-\rm D,a}, \tag{24}$$

respectively, where $\mathbf{I}$ is $L\times L$ unity and where $\boldsymbol{x}_{3-\rm D,a}$ is the common a priori 3-D information which may be included in the measurements. The difference $\Delta\hat{\boldsymbol{x}}_{3-\rm D}$ of the measurements caused both by spatial mismatch and different smoothing characteristics then is

$$\Delta\hat{\boldsymbol{x}}_{3-\rm D} = (\mathbf{A}_{3-\rm D,val}-\mathbf{A}_{3-\rm D,ref})(\boldsymbol{x}_{3-\rm D}-\boldsymbol{x}_{3-\rm D,a}), \tag{25}$$

where the smoothing component is accounted for by the deviation of each of the relevant sub-matrices of $\mathbf{A}_{3-\rm D}$ from unity, while the co-incidence component caused by $\Delta d$ is accounted for by the different placement of the relevant sub-matrices in the full $\mathbf{A}_{3-\rm D}$ matrix. The difference profile $\Delta\hat{\boldsymbol{x}}$ at the nominal geolocation $(i,j)$ then has the elements

$$\Delta\hat{x}_n = \Delta\hat{x}_{3-{\rm D};l(i,j;n)} \tag{26}$$

This way to calculate the differences can be applied to the correction scheme suggested in Eq. (15) where the difference $\boldsymbol{M}(d_{\rm val})-\boldsymbol{M}(d_{\rm ref})$ can be replaced by the $\Delta\hat{x}$ values from Eq. (26) applied to the $\Delta\hat{\boldsymbol{x}}_{3-\rm D}$ field from Eq. (25), where $\boldsymbol{x}_{3-\rm D}$ is generated by the model $\boldsymbol{M}$. Further, Eqs. (25–26) can be used to calculate the $\Delta\hat{x}(\Delta d)$ terms in Eq. (16) which is used for a statistical estimate of the residual co-incidence error after a possible correction. In the latter application, the elements $s_{{\rm err};m,n}$ of the covariance matrix $\mathbf{S}_{\rm err}$ describing the part of the observed differences of pairs of profiles caused by errors in the 3-D dataset itself rather than its variability in $d$ at a given geolocation $(i,j)$ are calculated as

$$s_{\rm err,diff;m,n} = \left((\mathbf{A}_{3-\rm D,val}-\mathbf{A}_{3-\rm D,ref})\mathbf{S}_{3-\rm D}(\mathbf{A}_{3-\rm D,val}-\mathbf{A}_{3-\rm D,ref})^T\right)_{l(i,j;m),l(i,j;n)}, \tag{27}$$

where $\mathbf{S}_{3-\rm D}$ is the $L\times L$ covariance matrix representing the uncertainties of $\hat{\boldsymbol{x}}_{3-\rm D}$. Equation (16) can then be extended to describe the combined smoothing and co-incidence error:

$$(\mathbf{S}_{\rm coinc}+\mathbf{S}_{\rm smooth})_{m,n} =$$

[4]Steck, T., Blumenstock, T., Clarmann, T., Glatthor, N., Grabowski, U., Hase, F., Hochschild, G., Höpfner, M., Kellmann, S., Kiefer, M., Kopp, G., Linden, A., Milz, M., Oelhaf, H., Stiller, G. P., Wetzel, G., Zhang, G., Fischer, H., Funke, B., Wand, D. Y., Gathen, P., Hansen, G., Stebel, K., Kyrö, E., Allaart, M., Redondas Marrero, A., Remsberg, E., Russell III, J., Steinbrecht, W., Yela, M., and Raffalski, U.: Validation of ozone measurements from MIPAS-Envisat, in preparation, 2006.

$$\frac{1}{K}\Bigg(\sum_{k=1}^{K}\Big((\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})(\boldsymbol{x}_{3-\mathrm{D,r}}-\boldsymbol{M}_{3-\mathrm{D}})\Big)_{l(i(k),j(k),m)} \quad (28)$$
$$\times\Big((\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})(\boldsymbol{x}_{3-\mathrm{D,r}}-\boldsymbol{M}_{3-\mathrm{D}})\Big)_{l(i(k),j(k),n)}\Bigg)$$
$$-(\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})\mathbf{S}_{3-\mathrm{D}}(\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})^{T},$$

where $l((i(k),\, j(k),\, n)$ denotes the position of the point of the 3 dimensional field in the vector, and where $i(k)$, $j(k)$, and $m$ or $n$ denote the horizontal nominal co-ordinates and the altitude of the $k$-th sample pair used to evaluate the combined co-incidence and smoothing error covariance matrix. $\boldsymbol{M}_{3-\mathrm{D}}$ is the 3-dimensional correction model tabulated in a vector according to the predefined reshaping convention $(\boldsymbol{M}_{3-\mathrm{D}})_{k(i,j,n)}=M_n(d(i,\, j))$.

If the $\mathbf{S}_{\mathrm{err}}$ matrices vary between the data pairs over which the summation in Eqs. (16) or (28) runs, this has to be taken into account by appropriate weighting of the terms.

If an ex ante estimate of the covariance matrix $\mathbf{S}_{3-\mathrm{D,a}}$ which describes the variability and correlations of $\boldsymbol{x}_{3-\mathrm{D}}$ is avaliable, the formalism proposed by Rodgers and Connor (2003) for estimation of the smoothing error difference can be directly applied to estimate the combined effect of smoothing error difference and co-incidence error, $\mathbf{S}_{\mathrm{coinc.}}+\mathbf{S}_{\mathrm{smooth.}}$, at a given location $(i,\, j)$, without evaluation of a reference data set $\hat{\boldsymbol{x}}_{\mathrm{r}}$:

$$s_{\mathrm{coinc.};m,n}+s_{\mathrm{smooth.};m,n}= \quad (29)$$
$$\Big((\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})^{T}\mathbf{S}_{3-\mathrm{D,a}}(\mathbf{A}_{3-\mathrm{D,val}}-\mathbf{A}_{3-\mathrm{D,ref}})\Big)_{l(i,j;m),l(i,j;n)}$$

For particular cases the relationship of smoothing error and co-incidence error can be simplified due to the nature of measurements to be compared. Further, the rigorous approach often is not possible since the needed data are unavailable. For example, $\hat{\boldsymbol{x}}_{\mathrm{r}}$ may not be available as sufficiently high resolved 3-D field but only as 1- or 2-dimensional cross-section. Further, the $\mathbf{A}_{3-\mathrm{D}}$ averaging kernel matrix may not be provided along with the data product. Sometimes the rigorous approach is inappropriate, because the given data sets are not accurate enough to justify the related effort. In this case, the errors introduced by the approximations may no longer be significant.

Obvious simplifications are: The instruments may sound non-overlapping air parcels; this is reflected by disjoint non-zero entries in the 3-D averaging kernel matrices and, as a consequence, correlation terms between the air parcels in Eq. (27) turn zero. The reference measurement can be a point measurement, e.g. an in situ measurement, where no averaging kernel has to be considered; in this case, the reference data set $\hat{\boldsymbol{x}}_{\mathrm{r}}$ can first be smoothed by application of the matrix of vertical averaging kernels and the assessment of the horizontal smoothing and co-incidence error then requires only the horizontal components of the 3-D averaging kernel matrix. Further, $\mathbf{S}_{3-D}$ may have some diagonal structure, e.g. because the errors are assumed uncorrelated in geolocations. Often the cross-line-of-sight component of the 3-D averaging kernel matrix is negligible. Various retrieval schemes use $\boldsymbol{x}_{3-\mathrm{D,a}}=0$.

The entries of $\mathbf{A}_{3-\mathrm{D}}$ related to vertical smoothing are contained in the profile averaging kernel matrix, which often is available as part of the diagnostics of the data sets under assessment. For the elements related to horizontal smoothing the situation is different: For nadir sounders and the cross-line-of sight component of limb sounders, related averaging components just are identical to the field of view function, possibly widened by the horizontal smearing caused by the motion of the instrument while the measurement is made. If the field of view is narrower than the horizontal spacing of gridpoints in the data set $\hat{\boldsymbol{x}}_{\mathrm{r}}$, this component can be ignored. The horizontal along-line-of-sight component can be obtained e.g. from perturbational analysis or analytically from 2-D radiative transfer modelling and retrieval tools (see, e.g. Steck et al., 2005; Carlotti et al., 2001). If the horizontal components of $\mathbf{A}_{3-\mathrm{D}}$ are not available, they can be approximated by $\mathbf{RA}$, where $\mathbf{A}$ is the vertical profile averaging kernel matrix, and $\mathbf{R}$ the $I\times N$ dimensional ray-tracing operator, which maps altitudes $z_1,\cdots,z_n$ to along-track distances $d_1,\cdots,d_I$ according to the observation geometry. Elements of $\mathbf{A}$ representing contributions from below the tangent altitude are assigned to the tangent point geolocation. This approximation, however, neglects both the mapping of any horizontal smoothing error onto the retrieved profile, and the asymmetry of the horizontal averaging kernel around the tangent point of a limb viewing measurement. This approach has been chosen by Ridolfi et al. (2006b)[2] and Cortesi et al. (2006)[3] to account for the horizontal smoothing of MIPAS in the co-incidence correction. These authors have used vertically smoothed ECMWF fields as correction model in Eq. (15) and have considered the horizontal smoothing by the formalism of Eqs. (25–26), where $\mathbf{A}_{3-\mathrm{D,ref}}$ was assumed unity and where $\boldsymbol{x}_{3-\mathrm{D,a}}$ was zero.

## 4 Bias determination

To determine the bias between two measurement systems, a statistical ensemble of measurements is needed. This ensemble can either be composed of $K$ matching pairs of measurements or random samples of $K$ and $L$ measurements of each measuring system, respectively.

### 4.1 Statistical bias determination with matching pairs of measurements

The mean difference between measurements to be validated and co-incident reference measurements can be compared with its statistical uncertainty in order to determine any bias between the measurement to be validated and the reference measurement and its significance. With $K$ pairs of co-incident measurements available, the bias $\boldsymbol{b}_{\mathrm{diff}}$ between these measurements is estimated as (here and henceforth we use

the $\breve{}$ symbol to denote ex post estimates based on a statistical comparison with reference data)

$$\breve{\boldsymbol{b}}_{\text{diff}} = \frac{\sum_{k=1}^{K}(\hat{\boldsymbol{x}}_{\text{val};k} - \hat{\boldsymbol{x}}_{\text{ref};k})}{K} \tag{30}$$

The statistical uncertainty of the bias is characterized by the related covariance matrix $\breve{\mathbf{S}}_{\text{bias}}$, the elements of which are estimated as

$$\begin{aligned}\breve{s}_{\text{bias};m,n} = &\frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m,k} - \overline{\hat{x}}_{\text{val};m})(\hat{x}_{\text{val};n,k} - \overline{\hat{x}}_{\text{val};n})}{K(K-1)} + \\ &\frac{\sum_{k=1}^{K}(\hat{x}_{\text{ref};m,k} - \overline{\hat{x}}_{\text{ref};m})(\hat{x}_{\text{ref};n,k} - \overline{\hat{x}}_{\text{ref};n})}{K(K-1)} - \\ &\frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m,k} - \overline{\hat{x}}_{\text{val};m})(\hat{x}_{\text{ref};n,k} - \overline{\hat{x}}_{\text{ref};n})}{K(K-1)} - \\ &\frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};n,k} - \overline{\hat{x}}_{\text{val};n})(\hat{x}_{\text{ref};m,k} - \overline{\hat{x}}_{\text{ref};m})}{K(K-1)} \\ = &\frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m,k} - \hat{x}_{\text{ref};m,k} - \breve{b}_{\text{diff};m})(\hat{x}_{\text{val};n,k} - \hat{x}_{\text{ref};n,k} - \breve{b}_{\text{diff};n})}{K(K-1)},\end{aligned} \tag{31}$$

where

$$\overline{\hat{x}} = \frac{\sum_{k=1}^{K}\hat{x}_k}{K}. \tag{32}$$

This assessment does not need any error estimates of $\hat{\boldsymbol{x}}_{\text{val}}$ or $\hat{\boldsymbol{x}}_{\text{ref}}$.

The consistence of $\breve{b}_n$ and the ex ante estimate of the systematic error $\sigma_{\text{sys},n}$ of the retrieved state parameter $\hat{x}_n$, or of $\breve{b}_{\text{diff};n}$ and $\sigma_{\text{diff,sys},n}$, respectively, can easily be checked (see, e.g. Ridolfi et al., 2006b[2], for application to MIPAS temperature validation, or Cortesi et al., 2006[3], for ozone validation). Evaluation of the significance of the bias then requires $\chi^2$ statistics, where

$$\langle \chi^2_{\text{bias}} \rangle = \langle \breve{\boldsymbol{b}}^T_{\text{diff}} \breve{\mathbf{S}}^{-1}_{\text{bias}} \breve{\boldsymbol{b}}_{\text{diff}} \rangle = N. \tag{33}$$

The consistency of the bias with the ex ante estimates of the systematic error components can also be evaluated by a $\chi^2$ test:

$$\langle \chi^2_{\text{bias}} \rangle = \langle \breve{\boldsymbol{b}}^T_{\text{diff}} (\mathbf{S}_{\text{diff,sys}} + \breve{\mathbf{S}}_{\text{bias}})^{-1} \breve{\boldsymbol{b}}_{\text{diff}} \rangle = N. \tag{34}$$

Covariance matrix $\mathbf{S}_{\text{diff,sys}}$ includes the systematic components of $\mathbf{S}_{\text{diff}}$:

$$\mathbf{S}_{\text{diff,sys}} = \mathbf{S}_{\text{val,sys}} + \mathbf{S}_{\text{ref,sys}} + \mathbf{S}_{\text{smooth,sys}} \tag{35}$$

Obviously, neither the root mean squares difference of profiles obtained from two measurement systems nor $1/\sqrt{K}$ of the root mean squares difference are a measure of the significance of the bias.

If ex ante precision estimates of differences

$$\mathbf{S}_{\text{diff,random}} = \mathbf{S}_{\text{val,random}} + \mathbf{S}_{\text{ref,random}} + \mathbf{S}_{\text{coinc.}} + \mathbf{S}_{\text{smooth,random}} \tag{36}$$

are available and the uncertainties are known to vary within the sample, the measurements can be weighted accordingly to determine the weighted bias $\breve{\boldsymbol{b}}_{\widetilde{\text{diff}}}$:

$$\breve{\boldsymbol{b}}_{\widetilde{\text{diff}}} = \left( \sum_{k=1}^{K} \mathbf{S}^{-1}_{\text{diff,random};k} \right)^{-1} \left( \sum_{k=1}^{K} \mathbf{S}^{-1}_{\text{diff,random};k} (\hat{\boldsymbol{x}}_{\text{val};k} - \hat{\boldsymbol{x}}_{\text{ref};k}) \right) \tag{37}$$

The bias uncertainty in terms of covariance matrix then is

$$\mathbf{S}_{\text{bias}} = \left( \sum_{k=1}^{K} \mathbf{S}^{-1}_{\text{diff,random};k} \right)^{-1}, \tag{38}$$

which is an ex ante estimate. Thus, the use of $\breve{\boldsymbol{b}}_{\widetilde{\text{diff}}}$ and $\mathbf{S}_{\text{bias}}$ as determined from Eqs. (38) and (37) for precision validation (see Sect. 5) is not fully conclusive, because it depends on typically unvalidated ex ante precision estimates.

The most probable estimate of the multiplicative bias from a given sample of $K$ measurement pairs $\hat{x}_{\text{val};n,k}$ and $\hat{x}_{\text{ref};n,k}$, each affected by an additive random error of constant expectation is

$$\breve{b}_{\text{diff,mult.};n} = \frac{\breve{b}_{\text{diff};n}}{\frac{\sum_{k=1}^{K}\hat{x}_{\text{ref};n,k}}{K}} = \frac{\breve{b}_{\text{diff};n}}{\overline{\hat{x}}_{\text{ref};n}} = \frac{\sum_{k=1}^{K}\hat{x}_{\text{val};n,k}}{\sum_{k=1}^{K}\hat{x}_{\text{ref};n,k}} - 1. \tag{39}$$

This estimator gives larger weight to the ratios determined from large $\hat{x}_{\text{val};n,k}$ and $\hat{x}_{\text{ref};n,k}$, because their ratio is less affected by the measurement error. The covariance matrix $\breve{\mathbf{S}}_{\text{bias,mult.}}$ has the elements

$$\begin{aligned}\breve{s}_{\text{bias,mult.};m,n} = &\frac{\breve{r}_{\text{val};m,n}\breve{\sigma}_{\overline{\text{val}};m}\breve{\sigma}_{\overline{\text{val}};n}}{\bar{x}_{\text{ref};m}\bar{x}_{\text{ref};n}} \\ &+ \frac{\bar{x}_{\text{val};m}\bar{x}_{\text{val};n}\breve{r}_{\text{ref};m,n}\breve{\sigma}_{\overline{\text{ref}};m}\breve{\sigma}_{\overline{\text{ref}};n}}{\bar{x}^2_{\text{ref};m}\bar{x}^2_{\text{ref};n}} \\ &- \frac{\bar{x}_{\text{val};m}\breve{\sigma}_{\overline{\text{val}};n}\breve{\sigma}_{\overline{\text{ref}};m}\breve{r}_{n,m}}{\bar{x}^2_{\text{ref};m}\bar{x}_{\text{ref};n}} \\ &- \frac{\bar{x}_{\text{val};n}\breve{\sigma}_{\overline{\text{val}};m}\breve{\sigma}_{\overline{\text{ref}};n}\breve{r}_{m,n}}{\bar{x}^2_{\text{ref};n}\bar{x}_{\text{ref};m}},\end{aligned} \tag{40}$$

where

$$\breve{\sigma}_{\overline{\text{val}};m} = \frac{1}{\sqrt{K}}\breve{\sigma}_{\text{val};m} = \sqrt{\frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m,k} - \overline{\hat{x}}_{\text{val};m})^2}{K(K-1)}}, \tag{41}$$

$$\breve{\sigma}_{\overline{\text{ref}};m} = \frac{1}{\sqrt{K}}\breve{\sigma}_{\text{ref};m} = \sqrt{\frac{\sum_{k=1}^{K}(\hat{x}_{\text{ref};m,k} - \overline{\hat{x}}_{\text{ref};m})^2}{K(K-1)}}, \tag{42}$$

$$\breve{r}_{\text{val};m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\text{val};m,k} - \overline{\hat{x}}_{\text{val};m})(\hat{x}_{\text{val};n,k} - \overline{\hat{x}}_{\text{val};n})}{(K-1)\breve{\sigma}_{\text{val};m}\breve{\sigma}_{\text{val};n}}, \tag{43}$$

$$\breve{r}_{\text{ref};m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\text{ref};m,k} - \overline{\hat{x}}_{\text{ref};m})(\hat{x}_{\text{ref};n,k} - \overline{\hat{x}}_{\text{ref};n})}{(K-1)\breve{\sigma}_{\text{ref};m}\breve{\sigma}_{\text{ref};n}}, \tag{44}$$

and

$$\breve{r}_{m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k} - \bar{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{ref};n,k} - \bar{\hat{x}}_{\mathrm{ref};n})}{(K-1)\breve{\sigma}_{\mathrm{val};m}\breve{\sigma}_{\mathrm{ref};n}}. \quad (45)$$

This gives for the variances

$$\begin{aligned}\breve{\sigma}^2_{\mathrm{bias,mult.};n} &= \breve{s}_{\mathrm{bias,mult.};n,n} \\ &= \frac{1}{\bar{\hat{x}}^4_{\mathrm{ref};n}}(\breve{\sigma}^2_{\overline{\mathrm{val}};n}\bar{\hat{x}}^2_{\mathrm{ref};n} + \breve{\sigma}^2_{\overline{\mathrm{ref}};n}\hat{x}^2_{\mathrm{val};n} \\ &\quad -2\breve{r}_{n,n}\breve{\sigma}_{\overline{\mathrm{val}};n}\breve{\sigma}_{\overline{\mathrm{ref}};n}\bar{\hat{x}}_{\mathrm{ref},n}\bar{\hat{x}}_{\mathrm{val};n}), \end{aligned} \quad (46)$$

The simplified expression

$$\breve{\bar{\sigma}}_{\mathrm{b,rel};n} = \frac{\breve{\sigma}_{\mathrm{bias};n}}{\bar{\hat{x}}_{\mathrm{ref},n}} \quad (47)$$

ignores the uncertainty of $\bar{\hat{x}}_{\mathrm{ref},n}$.

Only if the expected errors of the differences are proportional to reference values, the mean relative deviation of a state parameter $x_n$ at altitude gridpoint $n$ is calculated as

$$\breve{b}_{\mathrm{diff},\widetilde{\mathrm{mult.}};n} = \frac{\sum_{k=1}^{K}\frac{\hat{x}_{\mathrm{val};n,k}-\hat{x}_{\mathrm{ref};n,k}}{\hat{x}_{\mathrm{ref};n,k}}}{K}. \quad (48)$$

is a better estimate of the multiplicative bias, because in this case the uncertainty of each of the ratios is equal, requiring equal weight of each ratio in the average. The elements of its covariance matrix are calculated as

$$\breve{s}_{\mathrm{bias},\widetilde{\mathrm{mult.}};m,n} = \frac{\sum_{k=1}^{K}(\frac{\hat{x}_{\mathrm{val};m,k}-\hat{x}_{\mathrm{ref};m,k}}{\hat{x}_{\mathrm{ref};m,k}} - \breve{b}_{\mathrm{diff},\widetilde{\mathrm{mult.}};m})(\frac{\hat{x}_{\mathrm{val};n,k}-\hat{x}_{\mathrm{ref};n,k}}{\hat{x}_{\mathrm{ref};n,k}} - \breve{b}_{\mathrm{diff},\widetilde{\mathrm{mult.}};n})}{K(K-1)}, \quad (49)$$

### 4.2 Bias determination by statistical comparison of random samples

It is not necessary to use matched pairs for validation. Random samples are sufficient but any sampling artefacts have to be carefully excluded. A parametrization as suggested in Sect. 3.2, Eq. (15) may help to reduce systematic sampling errors.

When two instruments provide large but independent, i.e. unmatched, random samples of measurements, the bias can be determined as the difference of respective mean values:

$$\breve{\boldsymbol{b}}_{\mathrm{diff}} = \frac{\sum_{k=1}^{K}\hat{\boldsymbol{x}}_{\mathrm{val};k}}{K} - \frac{\sum_{l=1}^{L}\hat{\boldsymbol{x}}_{\mathrm{ref};l}}{L} = \bar{\hat{\boldsymbol{x}}}_{\mathrm{val}} - \bar{\hat{\boldsymbol{x}}}_{\mathrm{ref}}, \quad (50)$$

where $K$ and $L$ are the respective sample sizes. The respective covariance matrix has the elements

$$\begin{aligned}\breve{s}_{\mathrm{bias};m,n} &= \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k}-\bar{\hat{x}}_{\mathrm{val};m,k})(\hat{x}_{\mathrm{val},n,k}-\bar{\hat{x}}_{\mathrm{val};n,k})}{K(K-1)} \\ &\quad + \frac{\sum_{l=1}^{L}(\hat{x}_{\mathrm{ref},m,l} - \bar{\hat{x}}_{\mathrm{ref};m,l})(\hat{x}_{\mathrm{ref},n,l} - \bar{\hat{x}}_{\mathrm{ref};n,l})}{L(L-1)}. \end{aligned} \quad (51)$$

Obviously, any non-randomness of the samples can cause an apparent bias or hide an existing bias.

## 5 Precision validation

### 5.1 Precision determination with matching pairs of measurements

The expectation value of the root mean squares difference of a pair of measurements of the same atmospheric state is the accuracy of the difference. In terms of variances and covariances of a profile of differences, this means that

$$\langle(\hat{x}_{\mathrm{val};m} - \hat{x}_{\mathrm{ref};m})(\hat{x}_{\mathrm{val};n} - \hat{x}_{\mathrm{ref};n})\rangle = s_{\mathrm{diff};m,n}. \quad (52)$$

For accuracy validation, this can be rewritten in terms of $\chi^2$ statistics, where the actual $\chi^2$ is evaluated from a sample of size $K$ of profiles with $N$ altitude gridpoints each:

$$\langle\chi^2\rangle = \langle\sum_{k=1}^{K}(\hat{\boldsymbol{x}}_{\mathrm{val};k}-\hat{\boldsymbol{x}}_{\mathrm{ref};k})^T\mathbf{S}^{-1}_{\mathrm{diff}}(\hat{\boldsymbol{x}}_{\mathrm{val};k}-\hat{\boldsymbol{x}}_{\mathrm{ref};k})\rangle = K\times N \quad (53)$$

If, however, there is a bias $\boldsymbol{b}_{\mathrm{diff}}$ between the measurement systems, this should be evaluated in a preceding step (see Sect. 4, Eq. 30) and removed in order to validate the precision of the measurement rather than the accuracy. This leads to the following $\chi^2$ statistics:

$$\begin{aligned}\langle\chi^2\rangle &= \langle\sum_{k=1}^{K}(\hat{\boldsymbol{x}}_{\mathrm{val};k}-\hat{\boldsymbol{x}}_{\mathrm{ref};k}-\breve{\boldsymbol{b}}_{\mathrm{diff}})^T\mathbf{S}^{-1}_{\mathrm{diff,random}} \\ &\quad (\hat{\boldsymbol{x}}_{\mathrm{val};k}-\hat{\boldsymbol{x}}_{\mathrm{ref};k}-\breve{\boldsymbol{b}}_{\mathrm{diff}})\rangle \\ &= (K-1)N \end{aligned} \quad (54)$$

$\mathbf{S}_{\mathrm{diff,random}}$ is the random component of $\mathbf{S}_{\mathrm{diff}}$ according to Eq. (14) and $\breve{\boldsymbol{b}}_{\mathrm{diff}}$ has been estimated from the same sample (see, e.g. Ridolfi et al., 2006b[2], for application to MIPAS temperature validation, or Cortesi et al., 2006[3], for ozone validation).

While Eq. (53) can be evaluated for single profiles ($K = 1$), Eq. (54) needs a sample of profiles ($K>1$) in order to distinguish between precision and bias, unless an altitude-independent bias $\breve{\bar{\boldsymbol{b}}}_{\mathrm{diff}}=(\breve{\bar{b}}_{\mathrm{diff}},\ldots,\breve{\bar{b}}_{\mathrm{diff}})^T$ is assumed, where

$$\breve{\bar{b}}_{\mathrm{diff}} = \frac{\sum_{n=1}^{N}(\hat{x}_{\mathrm{val};n} - \hat{x}_{\mathrm{ref};n})}{N}. \quad (55)$$

Equation (54) then reads

$$\begin{aligned}\langle\chi^2\rangle &= \langle(\hat{\boldsymbol{x}}_{\mathrm{val}}-\hat{\boldsymbol{x}}_{\mathrm{ref}}-\breve{\bar{\boldsymbol{b}}}_{\mathrm{diff}})^T\mathbf{S}^{-1}_{\mathrm{diff,random}}(\hat{\boldsymbol{x}}_{\mathrm{val}}-\hat{\boldsymbol{x}}_{\mathrm{ref}}-\breve{\bar{\boldsymbol{b}}}_{\mathrm{diff}})\rangle \\ &= N-1. \end{aligned} \quad (56)$$

### 5.2 Precision validation by comparison of random samples

The scatter of a sample of measurements is composed of both the measurement random error (characterized by covariance matrices $\mathbf{S}_{\mathrm{random,val}}$ or $\mathbf{S}_{\mathrm{random,ref}}$, respectively) and the natural variability (characterized by its covariance matrix $\mathbf{S}_{\mathrm{nat}}$). The natural variability of two randomly sampled data sets,

however, is the same, regardless if we observe the atmosphere with the one or the other instrument. Thus, we have to verify

$$\begin{aligned}\breve{\mathbf{S}}_{\mathrm{val,nat}} &= \breve{\mathbf{S}}_{\mathrm{val,sample}} - \mathbf{S}_{\mathrm{val,random}} \\ &= \breve{\mathbf{S}}_{\mathrm{ref,sample}} - \mathbf{S}_{\mathrm{ref,random}} \\ &= \breve{\mathbf{S}}_{\mathrm{ref,nat}},\end{aligned} \tag{57}$$

where the elements of $\breve{\mathbf{S}}_{\mathrm{val;sample}}$ are

$$\breve{s}_{\mathrm{val,sample};m,n} = \frac{\sum_{k=1}^{K}(\hat{x}_{\mathrm{val};m,k} - \overline{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{val};n,k} - \overline{\hat{x}}_{\mathrm{val};n})}{K-1} \tag{58}$$

and where the elements of $\breve{\mathbf{S}}_{\mathrm{ref;sample}}$ are

$$\breve{s}_{\mathrm{ref,sample};m,n} = \frac{\sum_{l=1}^{L}(\hat{x}_{\mathrm{ref};m,l} - \overline{\hat{x}}_{\mathrm{ref};m})(\hat{x}_{\mathrm{ref};n,l} - \overline{\hat{x}}_{\mathrm{ref};n})}{L-1}. \tag{59}$$

Testing if related variances are equal can be performed with the $F$-test (see, e.g., Press et al., 1989). Care must be taken that the samples are really random samples of the same population. Further, this strategy discussed here is particularly sensitive to an artificial reduction of the variability of one of the measurement data sets through the use of retrieval schemes involving Bayesian statistics, where each single profile is pushed towards some a priori information (see, e.g. Rodgers, 2000, for application of Bayesian statistics to retrieval theory). Beyond this, the result of the $F$-test is particularly sensitive to deviations of the actual distributions from normal distributions. While the approach proposed here is valid theoretically, these inherent traps make it difficult to use, and the author is not aware of any actual application to atmospheric measurements.

## 6 Comparison of a single measurement with a random sample of measurements

If only a single profile measurement is available which does not co-incide with any of the measurements to be validated, it can be checked if this single profile measurement belongs to the distribution defined sample of size $K$ of the measurements to be validated. The applicable $\chi^2$ test then uses

$$\begin{aligned}\chi^2 = (\overline{\hat{\boldsymbol{x}}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}})^T (\breve{\mathbf{S}}_{\mathrm{val,ensemble}} + \mathbf{S}_{\mathrm{val,sys}} + \mathbf{S}_{\mathrm{ref,s+r}} \\ + \mathbf{S}_{\mathrm{smooth,diff}})^{-1} (\overline{\hat{\boldsymbol{x}}}_{\mathrm{val}} - \hat{\boldsymbol{x}}_{\mathrm{ref}})\end{aligned} \tag{60}$$

where $\mathbf{S}_{\mathrm{smooth,diff}}$ characterizes the applicable smoothing error difference, and $\breve{\mathbf{S}}_{\mathrm{val,ensemble}}$ is the ex post ensemble covariance matrix of the measurements to be validated. Its elements are calculated as

$$\breve{s}_{\mathrm{val,ensemble};m,n} = \sum_{k=1}^{K} \frac{(\hat{x}_{\mathrm{val};m,k} - \overline{\hat{x}}_{\mathrm{val};m})(\hat{x}_{\mathrm{val};n,k} - \overline{\hat{x}}_{\mathrm{val};n})}{K-1}. \tag{61}$$

Again, considerations as outlined in the context of Eq. (14) may apply.

## 7 What if full retrieval error covariance matrices are not available?

Without ex ante estimates of the profile covariances available, we cannot draw any quantitative conclusion on the reliability of the retrieved profiles in the sense of $\chi^2$ statistics. Often, however, after debiasing, there are at least no horizontal error correlations to be considered. Then, state variables can be compared and $\chi^2$ statistics can be set up for a large ensemble of size K of scalar measurements to be validated $\hat{x}_{\mathrm{val},n,k}$ and reference measurements $\hat{x}_{\mathrm{ref},n,k}$ at a single selected altitude $z(n)$. This corresponds to "map validation" instead of "profile validation". All formulation discussed in this paper then is applied to the particular case where $N$=1. $\chi^2$-testing in this application leads to a valid conclusion on the reliability of a measurement $\hat{x}_{\mathrm{val},n}$ at the selected altitude $z(n)$. Of course, this procedure can be performed for all altitudes of interest independently. We consider a profile measurement system validated if we can validate the values at each altitude. If, after debiasing, correlations in the time domain can be excluded, the rationale outlined above also can be applied to time series validation. Ridolfi et al. (2006b)[2] have combined the map validation and time series validation approach by statistically analyzing differences between MIPAS temperatures and radiosonde temperatures from two stations measured at various times. The statistical analysis was peformed for altitude bins defined such that each MIPAS limb scan (i.e. each profile) was represented only once in each bin, justifying to disregard any error correlations in altitude.

## 8 Conclusions

Recipes and terminology for statistical validation of a profile measurement system have been suggested which cover both bias and precision validation and which are applicable to both matched pairs of co-incident measurements and random samples of measurements. Further, a recipe has been suggested to validate profile measurements in a statistical rigorous way even if their full profile covariance matrices are not available. While in real life it will not always be possible to apply these approaches at full rigorosity, validation scientists certainly will find workarounds and simplifications. It is hoped that this paper at least supports better communication in the validation community by suggesting a more or less consistent terminology. Further, ad hoc validation approaches may serve their purpose better, once clarified which rigorous approach they are meant to replace.

*Acknowledgements.* The author would like to thank S. Ceccherini, K.-H. Fricke, S. Mikuteit, M. Ridolfi, G. Stiller, and the reviewers for helpful comments.

Edited by: P. Hartogh

## References


Bevington, B. R.: Data reduction and error analysis for physical sciences, MacGraw-Hill Book Company, New York, 1969.

Calisesi, Y., Soebijanta, V. T., and van Oss, R.: Regridding of remote soundings: Formulation and application to ozone profile comparison, J. Geophys. Res., 110, D23306, doi:10.1029/2005JD006122, 2005.

Carlotti, M., Dinelli, B. M., Raspollini, P., and Ridolfi, M.: Geofit approach to the analysis of limb-scanning satellite measurements, Appl. Opt., 40, 1872–1885, 2001.

Haseloff, O. W. and Hoffmann, H.-J.: Kleines Lehrbuch der Statistik, de Gruyter, Berlin, 1970.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified Notation for Data Assimilation: Operational, Sequential and Variational, J. Meteorol. Soc. Japan, 75, 1B, 1997.

Migliorini, S., Piccolo, C., and Rodgers, C. D.: Intercomparison of direct and indirect measurements: Michelson Interferometer for Passive Atmospheric Sounding (MIPAS) versus sonde ozone profiles, J. Geophys. Res., 109, D19316, doi:10.1029/2004JD004988, 2004.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.: Numerical Recipes, Cambridge University Press, Cambridge, 1989.

Ridolfi, M., Ceccherini, S., and Carli, B.: Optimal interpolation method for intercomparison of atmospheric measurements, Opt. Lett., 31, 855–857, 2006.

Rodgers, C. D.: Retrieval of Atmospheric Temperature and Composition From Remote Measurements of Thermal Radiation, Rev. Geophys. Space Phys., 14, 609–624, 1976.

Rodgers, C. D.: Characterization and error analysis of profiles retrieved from remote sounding measurements, J. Geophys. Res., 95, 5587–5595, 1990.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice, vol. 2 of Series on Atmospheric, Oceanic and Planetary Physics, edited by: Taylor, F. W., World Scientific, 2000.

Rodgers, C. D. and Connor, B. J.: Intercomparison of remote sounding instruments, J. Geophys. Res., 108, 4116, doi:10.1029/2002JD002299, 2003.

Steck, T., Höpfner, M., von Clarmann, T., and Grabowski, U.: Tomographic retrieval of atmospheric parameters from infrared limb emission observations, Appl. Opt., 44, 3291–3301, 2005.

von Clarmann, T. and Grabowski, U.: Elimination of hidden a priori information from remotely sensed profile data, Atmos. Chem. Phys. Discuss., 6, 6723–6751, 2006, http://www.atmos-chem-phys-discuss.net/6/6723/2006/.

Walther, B. A. and Moore, J. L.: The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance, Ecography, 28, 815–829, 2005.