Equivalence of data fusion and simultaneous retrieval

Simone Ceccherini, Bruno Carli and Piera Raspollini

Istituto di Fisica Applicata "Nello Carrara" del Consiglio Nazionale delle Ricerche, Via Madonna del Piano 10, 50019 Sesto Fiorentino (Firenze), Italy

INTRODUCTION

When two or more instruments sound the same portion of atmosphere and observe the same species either in different spectral regions or with different geometries, two strategies are possible for the retrieval of the best vertical profile estimate that exploits all the available information. First, we can use all the observations acquired by the different instruments as inputs of a single retrieval algorithm that produces a single profile. We refer to this approach as the *simultaneous retrieval*. Second, we can use the observations of the different instruments to retrieve from each one an independent vertical profile and then use a posteriori an algorithm that combines the different profiles and determines a single estimate. We refer to this approach as the *data fusion*. The simultaneous retrieval is of difficult implementation but provides the best estimate. The implementation of data fusion is more easy but usually it is not expected to have the same performance as the simultaneous retrieval.

We present herewith a new data fusion method for atmospheric vertical profiles retrieved by remote sensing measurements. It uses an algorithm that takes into account both the retrieval errors of the single profiles, described by the covariance matrices (CM), and the sensitivity of the retrieved profiles to the true profile, described by the averaging kernel matrices (AKM).

We compare the results of the new data fusion method with those of the simultaneous retrieval using the measurements of the MIPAS (Michelson Interferometer for Passive Atmospheric Sounding) instrument onboard the ENVISAT satellite. The observations of a MIPAS limb sounding sequence are divided into two complementary sets and two profiles are independently retrieved from the two sets of observations. The two profiles are fused with the data fusion algorithm and the result is compared with the profile retrieved using simultaneously all the observations of the sequence. The performance of the new data fusion method is also compared with those of weighted and arithmetic means.

THE NEW DATA FUSION METHOD

We suppose to have N independent simultaneous measurements of the vertical profile of an atmospheric species referred to a specific geolocation. Performing the retrieval of the N measurements we obtain N vectors $\hat{\mathbf{x}}_i$ (i=1, 2, …, N) that provide independent estimates of the profile on a common vertical grid. These vectors are characterized by the CMs $\mathbf{S}_i$ and the AKMs $\mathbf{A}_i$.

If we expand at the first order the relationship that exists between retrieved profile and true profile, exploiting the definition of the AKM, we obtain the following equation :

$$\hat{\mathbf{x}}_i = \mathbf{x}_{ai} + \mathbf{A}_i\left(\mathbf{x}_{true} - \mathbf{x}_{ai}\right) + \boldsymbol{\sigma}_i, \qquad (1)$$

where we have indicated with $\mathbf{x}_{ai}$ the a priori profile used in the i-th retrieval, with $\mathbf{x}_{true}$ the true profile and with $\boldsymbol{\sigma}_i$ the error on the vertical profile obtained propagating the errors of the observations through the retrieval process. Rearranging Eq. (1) we obtain:

$$\hat{\mathbf{x}}_i - \left(\mathbf{I} - \mathbf{A}_i\right)\mathbf{x}_{ai} = \mathbf{A}_i\mathbf{x}_{true} + \boldsymbol{\sigma}_i, \qquad (2)$$

where $\mathbf{I}$ represents the identity matrix. In Eq. (2) we notice that the vector

$$\boldsymbol{\alpha}_i = \hat{\mathbf{x}}_i - \left(\mathbf{I} - \mathbf{A}_i\right)\mathbf{x}_{ai}, \qquad (3)$$

which is obtained from known quantities, is an estimate of the vector $\mathbf{A}_i\mathbf{x}_{true}$, made of the components of $\mathbf{x}_{true}$ along the averaging kernels and, as such, corresponds to a new indirect measurement of the true profile made in the vector space generated by the averaging kernels. From Eqs. (2) and (3) we see that the new measurement provided by the vector $\boldsymbol{\alpha}_i$ has the same errors $\boldsymbol{\sigma}_i$ as the retrieved profile (consequently its CM is also given by $\mathbf{S}_i$), but does not depend on the a priori profile $\mathbf{x}_{ai}$ (which instead contributes as a bias to $\hat{\mathbf{x}}_i$).

Since the vectors $\boldsymbol{\alpha}_i$ are indirect measurements in the form of $\mathbf{A}_i\mathbf{x}_{true}$, we can perform a simultaneous fit of these measurements minimizing the following cost function:

$$c(\mathbf{x}) = \sum_{i=1}^{N}\left(\boldsymbol{\alpha}_i - \mathbf{A}_i\mathbf{x}\right)^T \mathbf{S}_i^{-1}\left(\boldsymbol{\alpha}_i - \mathbf{A}_i\mathbf{x}\right) + \left(\mathbf{x} - \mathbf{x}_a\right)^T \mathbf{S}_a^{-1}\left(\mathbf{x} - \mathbf{x}_a\right), \qquad (4)$$

where $\mathbf{x}_a$ and $\mathbf{S}_a$ are an a priori profile and its CM that we may want to use as a constraint of the solution. This constraint depends on the ill conditioning of the simultaneous fit and is in general different from the constraints used in the individual retrievals.

The minimum of $c(\mathbf{x})$ is obtained for $\mathbf{x}$ equal to:

$$\mathbf{x}_f = \left(\sum_{i=1}^{N}\mathbf{A}_i^T\mathbf{S}_i^{-1}\mathbf{A}_i + \mathbf{S}_a^{-1}\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{A}_i^T\mathbf{S}_i^{-1}\boldsymbol{\alpha}_i + \mathbf{S}_a^{-1}\mathbf{x}_a\right). \qquad (5)$$

This relationship provides a new estimate of the profile determined with the data fusion of N different profiles.

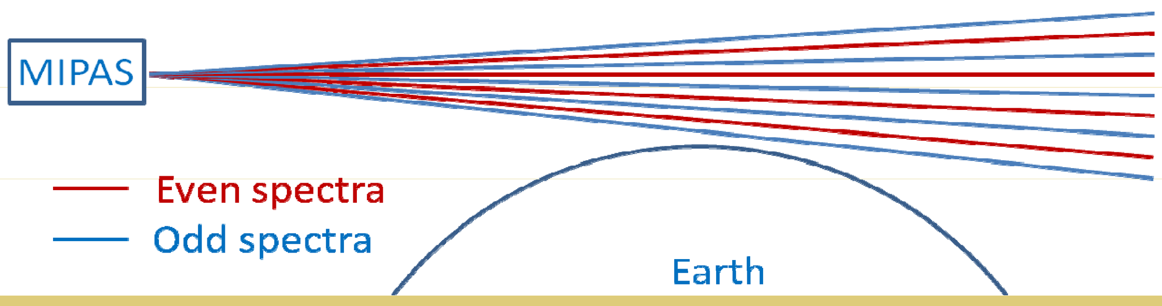
If the AKMs $\mathbf{A}_i$ are equal to the identity matrix from Eq. (3) it follows that $\boldsymbol{\alpha}_i = \hat{\mathbf{x}}_i$ and Eq. (5) coincides with the weighted mean of the N measurements (which in turn reduces itself to the arithmetic mean when all the measurements are considered to have the same $\mathbf{S}_i$). Therefore, Eq. (5) is a generalization of weighted mean in the case of AKMs different from the identity matrix and, for its capability of considering all the features of the measurements that are being combined, we shall refer to it as *complete fusion* formula.

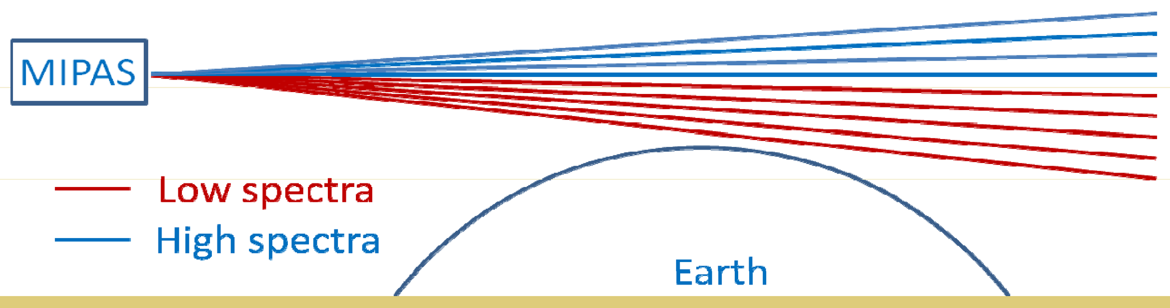
It is possible to analytically demonstrate that in a linear approximation the solution obtained with complete fusion coincides with the solution obtained with simultaneous retrieval.

APPLICATION TO REAL DATA

We experimentally evaluate the performance of the new data fusion approach applying it to an ozone vertical profile retrieved from MIPAS observations.

We divided the observations consisting of 27 spectra in two different ways. Firstly, we obtained two data sets taking spectra at alternate tangent altitudes (referred to as *even* and *odd spectra*). In this way, we obtain two data sets that cover approximately the same vertical range. In the second place, we obtained two data sets including the 13 highest spectra in the first one and the remaining 14 spectra in the second one. In this way, we obtain data sets that cover two complementary altitude ranges: the first one approximately from 29 km to 71 km (referred to as *high spectra*) and the second one from 7 to 29 km (referred to as *low spectra*).





FUSION OF EVEN AND ODD SPECTRA

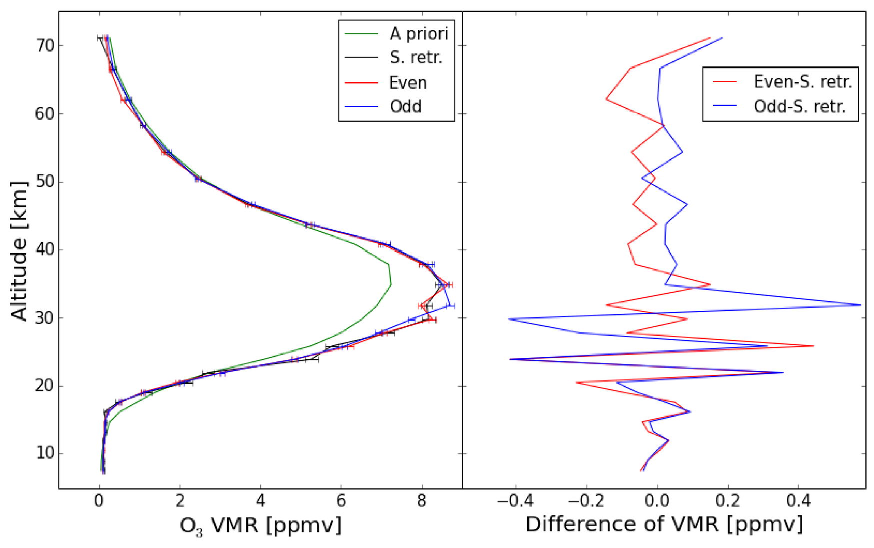


Fig. 1. Left panel: a priori profile (green line) and profiles retrieved using all spectra (black line), even spectra only (red line) and odd spectra only (blue line). Right panel: differences between the profiles obtained with reduced data sets and simultaneous retrieval in the case of even spectra (red line) and in the case of odd spectra (blue line).

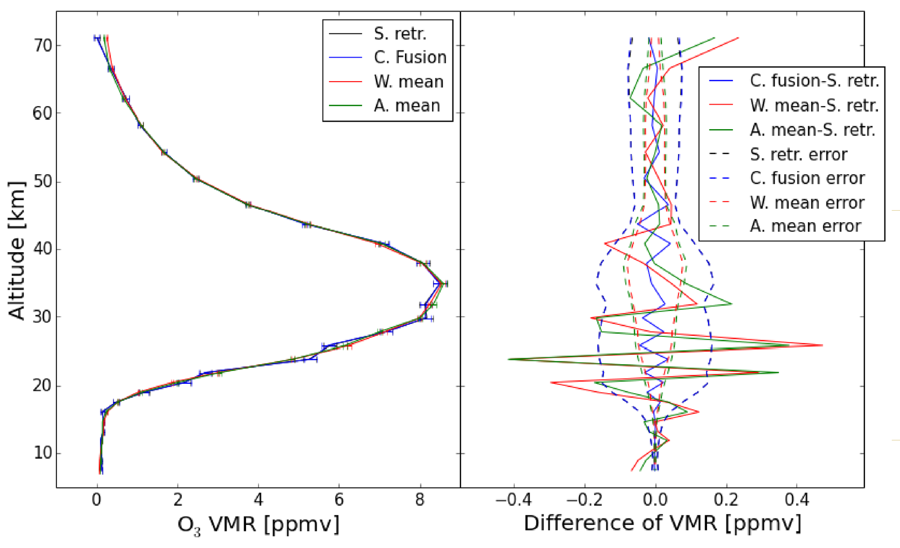


Fig. 2. Left panel: profile obtained using the simultaneous retrieval of all spectra (black line) and profiles obtained with complete fusion (blue line), weighted mean (red line) and arithmetic mean (green line) of the two profiles retrieved from even and odd spectra. The right panel shows the differences, with respect to the profile of the simultaneous retrieval, of profiles obtained with complete fusion (blue line), weighted mean (red line) and arithmetic mean (green line). The dashed lines show the error amplitudes.

Table 1. Number of degrees of freedom (NDOF) of the profiles reported in Figs. 1 and 2.

Simultaneous retrieval	Even spectra	Odd spectra	Complete Fusion	Weighted mean	Arithmetic mean
23.6	11.6	11.7	23.6	9.7	11.7

FUSION OF HIGH AND LOW SPECTRA

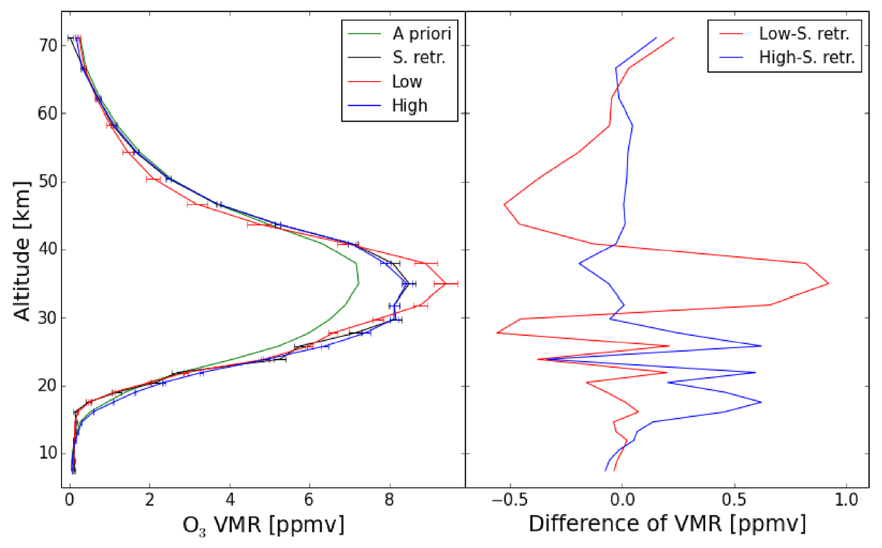


Fig. 3. Left panel: a priori profile (green line) and profiles retrieved using all spectra (black line), low spectra only (red line) and high spectra only (blue line). Right panel: differences between the profiles obtained with reduced data sets and simultaneous retrieval in the case of low spectra (red line) and in the case of high spectra (blue line).

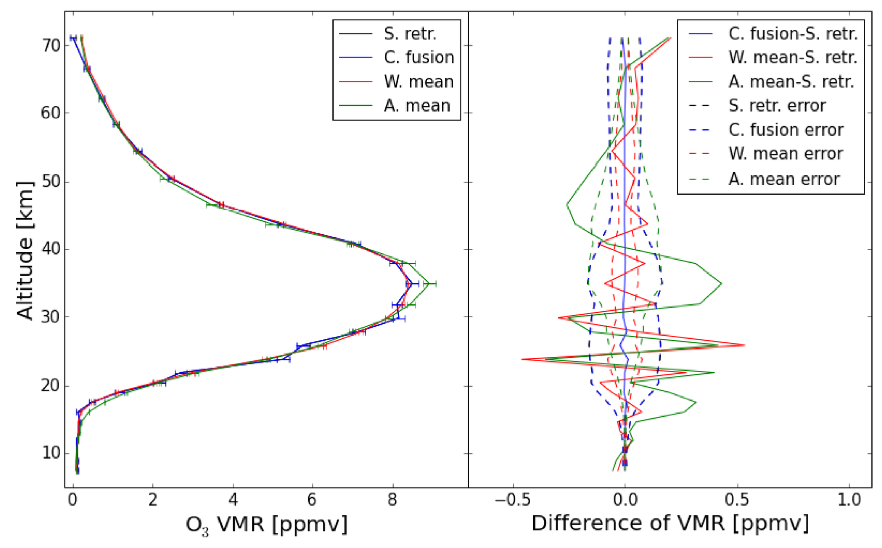


Fig. 4. Left panel: profile obtained using the simultaneous retrieval of all spectra (black line) and profiles obtained with complete fusion (blue line), weighted mean (red line) and arithmetic mean (green line) of the two profiles retrieved from low and high spectra. The right panel shows the differences, with respect to the profile of the simultaneous retrieval, of profiles obtained with complete fusion (blue line), weighted mean (red line) and arithmetic mean (green line). The dashed lines show the error amplitudes.

Table 2. Number of degrees of freedom (NDOF) of the profiles reported in Figs. 3 and 4.

Simultaneous retrieval	Low Spectra	High Spectra	Complete Fusion	Weighted mean	Arithmetic mean
23.6	9.0	10.6	23.6	11.4	9.8

CONCLUSIONS

Comparing the results of the two tests, we see that they are very similar independently of the type of complementarity of the measurements: interlaced observations in the first test and complementary altitude ranges in the second test. The complete fusion reproduces very well the results of the simultaneous retrieval both from the point of view of the values, which differ by quantities much smaller than the errors, and from the point of view of the error estimates (in Figs. 2 and 4 the black dashed lines are exactly under the blue dashed lines). The arithmetic and weighted means differ from the simultaneous retrieval by much larger quantities and are characterized by errors that are much smaller than the observed differences. This apparent contradiction is explained by the analysis of NDOF. The profiles retrieved from reduced data sets have, as expected, a NDOF that is about half the NDOF of the profile obtained from the simultaneous retrieval. The profile obtained with the complete fusion has the same NDOF as simultaneous retrieval while weighted and arithmetic means have a NDOF of the same order as profiles retrieved from reduced data sets. The weighted mean and the arithmetic mean are the result of an averaging process that reduces the errors, but cannot provide a better NDOF. Accordingly, the small NDOF prevents an adequate representation of the shape of the profile so that the differences with respect to the more realistic representation provided by the simultaneous retrieval are significant and, in particular, larger than the reduced errors. The simultaneous retrieval and the complete fusion best exploit the available NDOF at the cost of a slightly larger retrieval error.

These results show that data fusion when performed using either weighted or arithmetic mean provides worse results than those obtained with the simultaneous retrieval. However, when the more rigorous method of the complete fusion is used, the data fusion reproduces very well the results of the simultaneous retrieval (from the point of view of values, of error estimates and of NDOF) and, therefore, can be considered equivalent to the simultaneous retrieval.

S. Ceccherini, B. Carli and P. Raspollini, *Equivalence of data fusion and simultaneous retrieval*, Optics Express, Vol. **23** No. 7, 8476-8488 (2015).