The Complete Data Fusion for a Full Exploitation of Copernicus Atmospheric Sentinel Level 2 Products

Nicola Zoppetti[1], Simone Ceccherini[1], Bruno Carli[1], Samuele Del Bianco[1], Marco Gai[1], Cecilia Tirelli[1], Flavio Barbara[1], Rossana Dragani[2], Antti Arola[3], Jukka Kujanpää[4], Jacob C.A. van Peet[5,6], Ronald van der A[5] and Ugo Cortesi[1]

[1] Istituto di Fisica Applicata "Nello Carrara" del Consiglio Nazionale delle Ricerche, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Italy
[2] European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK
[3] Finnish Meteorological Institute, Atmospheric Research Centre of Eastern Finland, P.O.Box 1627, 70211 Kuopio, Finland
[4] Finnish Meteorological Institute, Space and Earth Observation Centre, P.O. Box 503, FI-00101 Helsinki, Finland
[5] Royal Netherlands Meteorological Institute, Utrechtseweg 297, 3731 GA De Bilt, The Netherlands
[6] Vrije Universiteit Amsterdam, Department of Earth Sciences, Amsterdam, The Netherlands

Supplementary material

Fusion of 1000 pixels in coincidence

Here, the CDF is applied to 1000 coincident L2 measurements that refer to the same true profile, the same AK matrix and the same CM but have different (noise) errors $\boldsymbol{\delta}_i$ randomly generated according to Eq. (3). The 1000 products have been simulated according to the specification of S4 platform and thermal infrared (TIR) band. It is noted that the particular type of product is of secondary importance in this example, which aims to evaluate the behaviour of the fusion of many coincident measurements of the same type that only differ by the random error. In the left panel of **Fig.S1** the profile obtained fusing 1000 coincident L2 products is compared with their arithmetic average, with the true profile and with the a priori profile. Since in this case the 1000 pixels are coincident in space and time, no coincidence error $\boldsymbol{\delta}_{coinc,i}$ was added in the CDF formulas of Eqs.(6).

In the right panel of **Fig S1**, the deviations of the fused profile (hereafter indicated with FUS), of the average value of the L2 measurements (indicated as <L2>) and of the a priori profile from the true profile are shown. In the same panel, the estimate of the total error standard deviation $\boldsymbol{\sigma}_{\text{total}}$ that characterize each of the 1000 L2 profiles (calculated as the root square of $\mathbf{S}_{total}$, Eq. (5)), the estimate of the total error standard deviation of FUS profile $\boldsymbol{\sigma}_{f\,total}$ (calculated as the root square of $S_{f\,\text{total}}$, Eqs. (6)) and the estimate of the total error standard deviation of the average of the L2 measurements (calculated by dividing $\boldsymbol{\sigma}_{total}$ by $\sqrt{1000}$, as if no bias is present) are also represented. It is worth noticing that $\boldsymbol{\sigma}_{total}/\sqrt{1000}$ is much smaller than the observed (<L2> minus true) differences, suggesting the presence of a bias. A clear similarity of these differences with the shape of the (a-priori minus true) profile can be observed indicating a link between this bias and the a priori information. The fused profile provides instead a better representation of the true profile with residuals that are consistent with the estimated errors, although these are much larger than $\boldsymbol{\sigma}_{total}/\sqrt{1000}$.

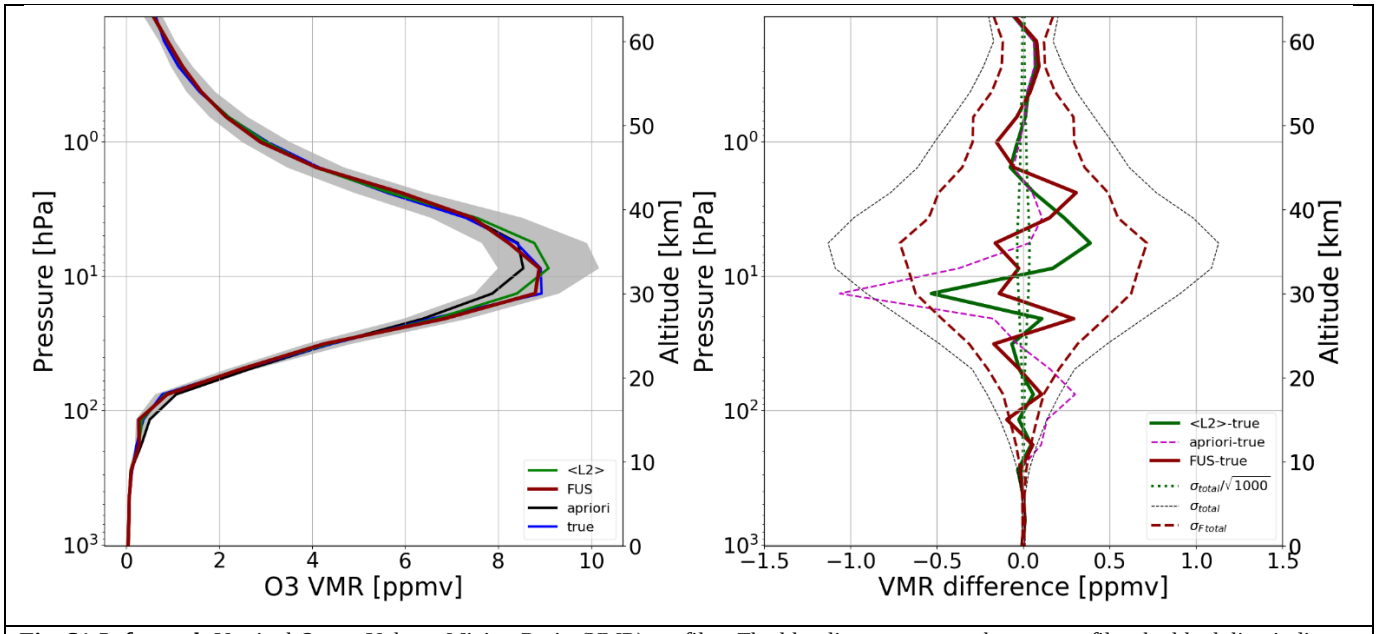

Fig. S1 ***Left panel:*** *Vertical Ozone Volume Mixing Ratio (VMR) profiles. The blue line represents the true profile; the black line indicates the a priori used in the L2 simulations (the same for all the 1000 L2 products) that is also the one used to constrain the fused profile, represented in dark red. The green line represents the average of the 1000 simulated profiles and the grey shaded area centred on the green line represents the standard deviation of the L2 total error (the same for all the 1000 L2 products).* ***Right panel:*** *the green and the dark-red lines represent respectively the difference between the <L2> average and the FUS profile from the true profile. The magenta dashed profile represents the difference between the a priori and the true profiles. The dotted black line represents the standard deviation of the total error estimate of each of the 1000 L2 measurements. The dashed dark-red line represents the standard deviation of the total error estimate of the FUS product. The dotted green line is the standard deviation of L2 total error estimate divided by √1000*

Recalling Eq. (8), it is the term $(\mathbf{I}-\mathbf{A})(\boldsymbol{x}_a-\boldsymbol{x}_t)$ that causes the bias observed in the right panel of **Fig. S1**. **Fig. S2** compares the amplitude of the bias term $(\mathbf{I}-\mathbf{A})(\boldsymbol{x}_a-\boldsymbol{x}_t)$ with the mean total error. For illustration, the total errors computed when only considering either 5 or 10 individual measurements are also plotted. As it can be noticed, the mean total error tends to the bias term as the number of profiles increases. When a large number of profiles are considered (order of 1000) the mean total error substantially coincides with the bias itself.

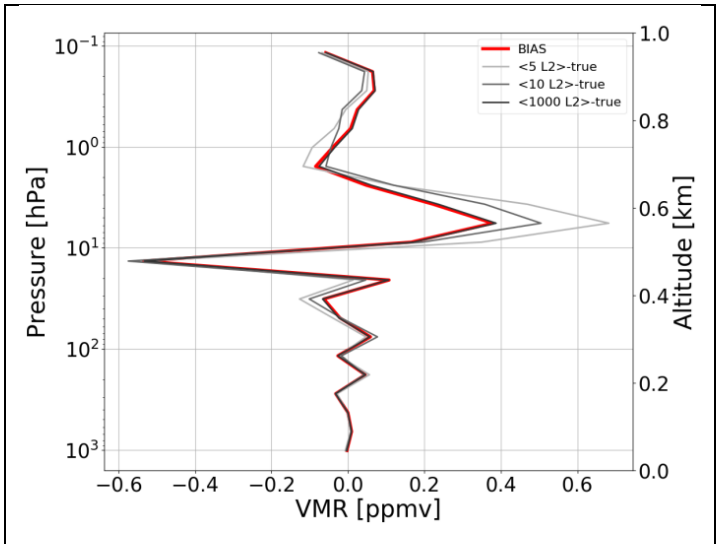

Fig.S2*: comparison of the bias term in Eq.) and the difference between <L2> average and true profile (i.e. mean total error) with different numbers of averaged L2 profiles.*

Statistical analysis for a large domain (1°x1°)

Fig. S3 shows the *SF DOF* obtained in the case of a 1°x1° resolution (**Tab. 3**). A test of the flexibility of the data fusion procedure is the objective of this analysis and, for simplicity, the same coincidence error used for the higher resolution grid

(0.5°x0.625°) was adopted. The adaptive choice of the amount of coincidence error to be used in the fusion is currently an open issue in the CDF development and is discussed in Ceccherini et al. (2019). Like in **Fig.5**, the *SF DOF* increases linearly with the logarithm of the number of L2 fusing profiles, and with a similar rate of growth so that **Fig.S3** looks like an extrapolation of **Fig.5**, for greater values of *N*. This is because the same types of L2 measurements as in the previous case are being fused.

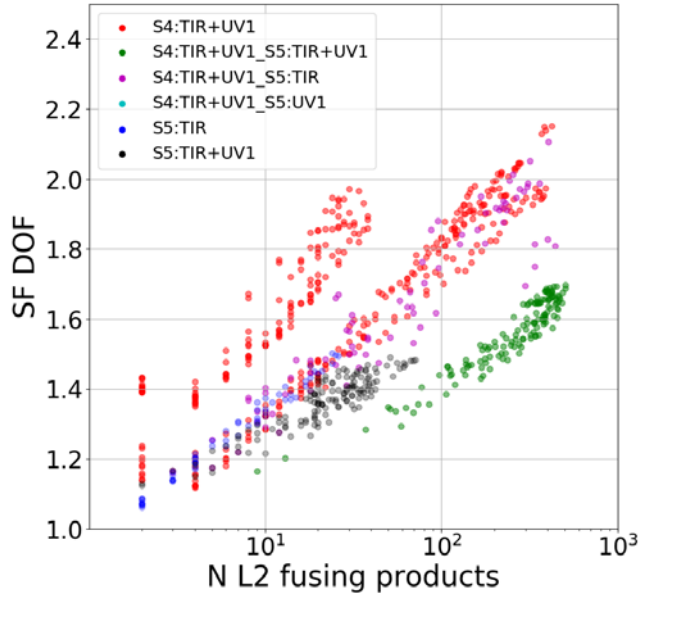

Fig S3*: scatter plot of SF DOF as a function of the number of L2 products fused in each coincidence grid cell; different colours represent different FUS types.*

Fig. S4 shows the *SF AK* and *SF ERR* now computed for the coarse resolution grid and for the 775 FUS products considered in **Tab.3**. The greater number of fusing observations with respect to **Fig.6** produces a general improvement for both the vertical resolution and the total error, although in the figures it is difficult to detect the first improvement because of the logarithmic scale. The CDF method can be used with a wide range of grid-box size and data compression and the quality of the products generally improves with larger cells. An upper limit to the grid-box size is caused by the requirement of a coincidence error amount, which degrades the quality of the fused product.

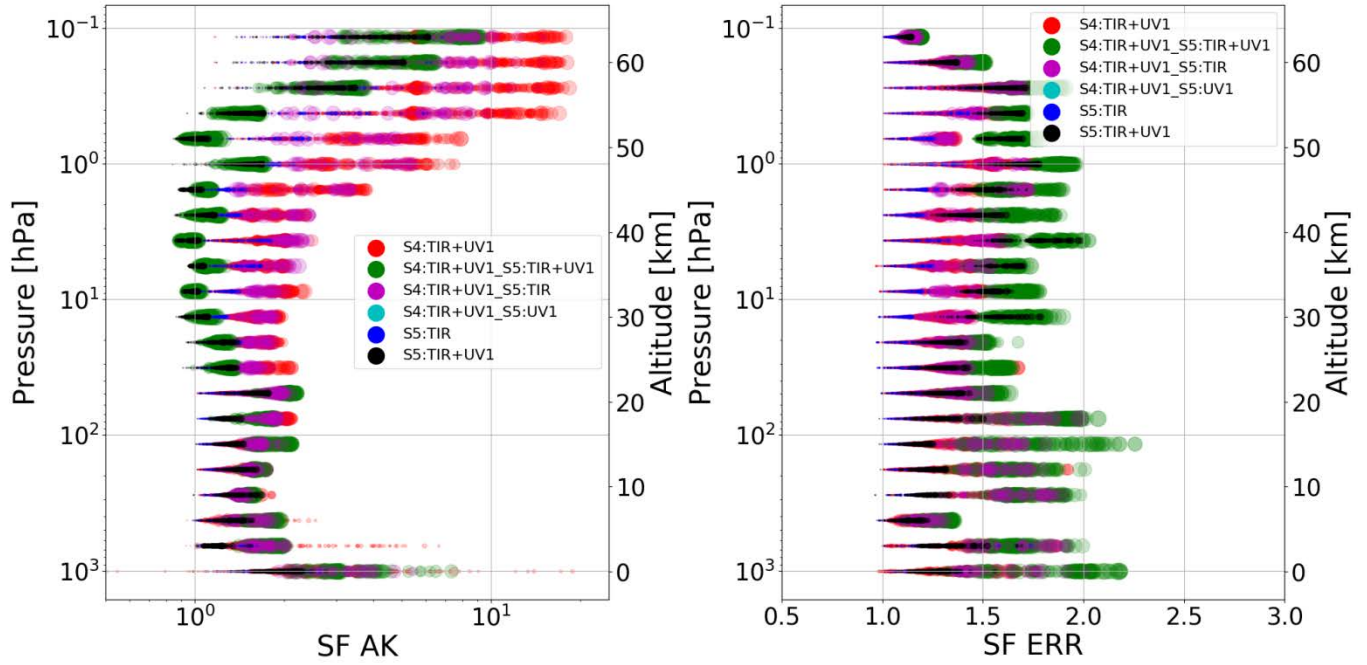

Fig. S4. ***Left panel:*** *SF AK versus vertical level.* ***Right panel:*** *SF ERR versus vertical level. In both panels, different colours of the symbols represent the FUS type, different sizes of the symbols represent the number of products that have been fused. The maximum symbol size shown in the legend corresponds to N=504.*