https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

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

Correspondence to: Nicola Zoppetti (N.Zoppetti@ifac.cnr.it)

Abstract.

The new platforms for Earth observation from space are characterized by measurements made with great spatial and temporal resolution. While this abundance of information makes it possible to detect and study localized phenomena, on the other hand it may be difficult to manage this large amount of data in the study of global and large scale phenomena.

A particularly significant example is the use by assimilation systems of level 2 products that represent gas profiles in the atmosphere. The models on which assimilation systems are based are discretized on spatial grids with horizontal dimensions of the order of tens of kilometres in which tens or hundreds of measurements may fall.

A simple procedure to overcome this problem is to extract a subset of the original measurements. However, this procedure involves a loss of information and is therefore justifiable only as a temporary solution. A more refined solution is to resort to the so-called fusion algorithms, capable of compressing the size of the dataset limiting the information loss. A novel data fusion method, the Complete Data Fusion, was recently developed to merge a-posteriori a set of retrieved products in a single product. In the present paper, the Complete Data Fusion method is applied to ozone profile measurements simulated in the thermal infrared and ultraviolet bands, in a realistic scenario, according to the specifications of the Sentinel 4 and 5 missions of the Copernicus programme. Then the fused products are compared with the input profiles; comparisons show that the output products of data fusion have in general smaller errors and higher information contents. The most significant improvement is an increased vertical resolution together with a reduction of the errors. The comparisons of the fused with the fusing products are presented both at single fusion grid-box scale and with a statistical analysis. The grid box size impact was also evaluated, showing that the Complete Data Fusion method can be used with a wide range of grid-box size, the quality of the products improving with larger grid boxes.

Introduction

In the context of the Copernicus programme (https://sentinel.esa.int/web/sentinel/missions) of the European Union, the European Space Agency is responsible for the Space Component consisting of a novel set of Earth Observation (EO) satellite missions for environmental monitoring applications: the Sentinels. Each mission focuses on a specific aspect of EO. In particular, the geostationary mission Sentinel-4 (S4) and the two Low Earth Orbit missions (Sentinel-5p and Sentinel 5 (S5)), referred to as the atmospheric Sentinels, are dedicated to monitoring air quality, stratospheric ozone, ultraviolet surface radiation and climate.

The atmospheric Sentinels will provide an enormous amount of data with unprecedented accuracy and spatio-temporal resolution. In this scenario, a central challenge is to enable a generic data user (for example, an assimilation system) to exploit

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

such a large amount of data. Data fusion algorithms, such as the Complete Data Fusion (CDF) (Ceccherini et al., 2015), can be particularly well suited to reduce the data volume that users need to access and handle while retaining the information content of the whole level 2 (L2) products. In other words, whenever the user does not need the full spatial and temporal resolution, but wants to exploit without loss of information the global coverage of the observations, an algorithm such as CDF is particularly useful since it is able to reduce the data volume of the input products to that corresponding to the required space and time resolution, while using all the available observations.

The CDF input is any number of L2 profiles retrieved with the optimal estimation technique and characterized by their a-priori information, covariance matrix (CM) and averaging kernel (AK) matrix. The output of the CDF is a single product (also characterized by an a-priori, a CM and AK matrices) which collects all the available information content.

This work is based on the simulated data produced in the context of the Advanced Ultraviolet Radiation and Ozone Retrieval for Applications project (AURORA, Cortesi et al.,2018), funded by the European Commission in the framework of the Horizon 2020 programme. The project regards the sequential application of fusion and assimilation algorithms to ozone profiles simulated according to the specifications of the atmospheric Sentinels.

The use of synthetic data allows evaluating the performances of the algorithm also in terms of differences between the products of interest and a reference truth, represented by the atmospheric scenario used in the procedure to simulate the L2 products. On the other hand, the absence of systematic errors in the simulated measurements limits the study to ideal measurement conditions. However, the CDF algorithm intrinsically provides a mechanism to include different kinds of errors into the analysis. For instance, Ceccherini et al. (2018) discussed how interpolation and coincidence errors can be accounted for and Ceccherini et.al. (2019) explicitly introduces the treatment of systematic errors.

This work is divided in two parts. In the first part, we describe the datasets and methodologies (the L2 simulation procedure and the CDF) used in the present paper. In the second part, the quality of the fused products obtained from L2 profiles that are not perfectly co-located in space and in time with each other is analysed. To account for the geo-temporal differences in the L2 profiles, a coincidence error is added to the fused product error budget. The fused and standard L2 products are compared and assessed in terms of their information content, highlighting the better data quality provided by the fusion. Finally, we also show that the CDF can be applied with different coincidence grid-box sizes, allowing for different compression factors of the Level 2 input data volume.

The application of CDF to L2 products simulated with the characteristics expected from the atmospheric Sentinel 4 and 5 allows to establish the possible benefits in case of real Sentinel data.

Material and methods

Atmospheric scenario and ozone climatology

Two basic external sources have been used to generate the database of the standard L2 ozone products used in this work: the ozone climatology and the atmospheric scenario.

The ozone climatology was used as a priori information for both the simulation of L2 products and the calculation of the CDF. The atmospheric scenario represents the true state of the atmosphere and is used for both the simulation of L2 products and the quality assessment of the fused ones.

The ozone climatology was derived from McPeters and Labow (McPeters and Labow, 2012) and directly provides the a priori profile. The CM of the a priori profile is obtained setting the diagonal terms equal to the square of standard deviation of the McPeters and Labow climatology where this standard deviation is larger than 20% of the a priori profile and to the square of 20% of the a priori profile otherwise. The off-diagonal elements are calculated using a correlation length of 6 km. The correlation length is used to reduce oscillations in the simulated profiles and the value of 6 km is typically used for nadir ozone profile retrieval (Liu et al., 2010, Kroon et al., 2011, Miles et al., 2015).

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

The atmospheric scenario is taken from the Modern Era-Retrospective analysis for Research and Applications version 2 (MERRA2) reanalysis (Gelaro et al., 2017). The MERRA2 data are provided by the Global Modelling and Assimilation Office (GMAO) at NASA Goddard Space Flight Center. This reanalysis covers the recent time of remotely sensed data, from 1979 through the present.

L2 Product Simulation

The simulation algorithm has been originally formalized in the context of the AURORA project, aiming at an efficient computational process. The L2 retrieved state is simulated on a fixed vertical grid with a 3 km step, by the linear approximation given in Eq. (1):

$$\hat{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x}_t + (\mathbf{I} - \mathbf{A})\boldsymbol{x}_a + \boldsymbol{\delta} \quad (1)$$

where $\boldsymbol{x}_t$ is the true state of the atmosphere represented by the atmospheric scenarios, $\boldsymbol{x}_a$ is the a priori estimate of the state vector provided by the ozone climatology, $\boldsymbol{\delta}$ is the uncertainty in retrieved value due to measurement noise, and $\mathbf{A} = \partial\hat{\boldsymbol{x}}/\partial\boldsymbol{x}_t$ is the AK matrix (Rodgers, 2000, Ceccherini et al., 2003; Ceccherini and Ridolfi, 2010) calculated according to Eq. (2):

$$\mathbf{A} = \left(\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} \quad (2)$$

In Eq. (2), $\mathbf{K}$ is the Jacobian matrix of the forward model, the superscript T represents the transpose operator, $\mathbf{S}_y$ is the CM of the observations and $\mathbf{S}_a$ is the CM of the a priori profile. The retrieval error $\boldsymbol{\delta}$ is calculated applying the gain matrix $\mathbf{G}$ (Rodgers, 2000) to an error $\boldsymbol{\varepsilon}$ on the observations randomly taken from a Gaussian distribution with average equal to zero and CM given by $\mathbf{S}_y$:

$$\boldsymbol{\delta} = \mathbf{G}\boldsymbol{\varepsilon} = \left(\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{K}^T\mathbf{S}_y^{-1}\boldsymbol{\varepsilon} \quad (3)$$

The $\mathbf{S}$ and $\mathbf{S}_{total}$ CMs associated to the retrieval error $\boldsymbol{\delta}$ (introduced in Eq. (3)) and to the total error $\boldsymbol{\delta}_{total}$ (i.e. the difference between the simulated and the true profiles that is equal to the random $\boldsymbol{\delta}$ plus the so called smoothing error caused by the limited vertical resolution, see also Eq. (7)) are given by Eq. (4) (Rodgers, 2000, Ceccherini and Ridolfi, 2010), and Eq. (5) (Rodgers, 2000), respectively:

$$\mathbf{S} = <\boldsymbol{\delta}\boldsymbol{\delta}^T> = \left(\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1}\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K}\left(\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1} \quad (4)$$

$$\mathbf{S}_{total} = <\boldsymbol{\delta}_{total}\boldsymbol{\delta}_{total}^T> = \left(\mathbf{K}^T\mathbf{S}_y^{-1}\mathbf{K} + \mathbf{S}_a^{-1}\right)^{-1} \quad (5)$$

It should be noted that through the term $\boldsymbol{\delta}$ it is possible to simulate additional error components with respect to the random one considered in this study and this fact adds flexibility to the simulation method.

The above formulation was used to simulate ozone profiles in the two spectral bands (UV and TIR) for both S4 and S5, after considering the instrument specifications and accounting for the differences in the two spectral bands. In particular, if a fixed geo-location is considered, starting from the same true profile and the same a priori information, the L2 products of the different instruments are obtained by the choice of the suitable Jacobian matrix $\mathbf{K}$ and of the CM $\mathbf{S}_y$ that have been synthetized using the technical requirements of the considered platforms and their foreseen performances. A detailed description of the specific characteristics of different L2 products goes beyond the scope of this article and can be found in (Cortesi et al. 2018).

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

This simulation method offers indisputable advantages in terms of speed of calculation compared to a complete retrieval.

The CDF method

In this Section, we briefly recall the formulas of the CDF method (Ceccherini et al., 2015). We assume to have *N* independent simultaneous measurements of the vertical profile of an atmospheric species that can be referred to the same geo-location. Performing the retrieval of the *N* measurements, we obtain *N* vectors $\hat{\boldsymbol{x}}_i$ (*i*=1, 2, …, N) that provide independent estimates of the profile, here assumed to be represented on a common vertical grid. Using as inputs these *N* measurements, the CDF produces as output a single product characterized by a profile $\mathbf{x}_f$, an AK matrix $\mathbf{A}_f$ and a CM matrix $\mathbf{S}_f$ with the procedure summarized by Eq.s (6). These three quantities are function of the input products ($\boldsymbol{x}_i$, $\mathbf{A}_i$, $\mathbf{S}_i$), hereafter referred to as fusing products, and depend on the a priori information ($\boldsymbol{x}_a$, $\mathbf{S}_a$) used as a constraint for the fused product.

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

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

$$\mathbf{S}_{f\,total} = \left(\sum_{i=1}^{N} \mathbf{A}_i^T \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i + \boldsymbol{S}_a^{-1}\right)^{-1} \quad (6)$$

$$\mathbf{A}_f = \left(\sum_{i=1}^{N} \mathbf{A}_i^T \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i + \mathbf{S}_a^{-1}\right)^{-1} \sum_{i=1}^{N} \mathbf{A}_i^T \tilde{\mathbf{S}}_i^{-1} \mathbf{A}_i$$

$$\boldsymbol{\alpha}_i = \hat{\boldsymbol{x}}_i - (\mathbf{I} - \mathbf{A}_i)\boldsymbol{x}_{ai} = \mathbf{A}_i \boldsymbol{x}_t + \boldsymbol{\delta}_i + \mathbf{A}_i \boldsymbol{\delta}_{coinc,i}$$

$$\tilde{\boldsymbol{S}}_i = \mathbf{S}_i + \mathbf{A}_i \mathbf{S}_{coinc,i} \mathbf{A}_i^T$$

A coincidence error characterized by a CM $\mathbf{S}_{coinc}$ is added if the input products are not coincident in time and space. When CDF is applied to not perfectly coincident products, the diagonal elements of $\mathbf{S}_{coinc}$ are calculated as the 5% of profile estimates in the ozone climatology. The off-diagonal elements of $\mathbf{S}_{coinc}$ are obtained applying an exponential decay with a correlation length of 6 km (Ceccherini et al. 2018).

The formulas of Eqs.(6) refer to the case of measurements made on the same vertical grid. In general, also an interpolation error may be needed (Ceccherini et al. 2018), but since it does not apply to the present study (the L2 products have been simulated on the same vertical grid) it is not considered in the following discussion.

Arithmetical average and biases

Before proceeding, it is necessary to clarify why the arithmetic average of the profiles cannot be considered as a good option to represent a set of products retrieved with optimal estimation techniques.

To do this, we consider *N* coincident L2 measurements (*i*=1,..*N*) 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 total error expression for the *i-th* measurement is given in Eq. (7) that can be easily derived from Eq. (1).

$$\boldsymbol{\delta}_{i,\text{total}} = \hat{\boldsymbol{x}}_i - \boldsymbol{x}_t = (\mathbf{I} - \mathbf{A}_i)(\boldsymbol{x}_a - \boldsymbol{x}_t) + \boldsymbol{\delta}_i \quad (7)$$

Considering that the individual measurements are co-located in space and time, thus they refer to the same truth, the same a priori profile and the same AK matrix **A**, the mean total error is equal to:

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

$$< \boldsymbol{\delta}_{i,\text{total}} >= \langle \hat{\boldsymbol{x}}_i \rangle - \boldsymbol{x}_t = \ldots = (\mathbf{I} - \mathbf{A})(\boldsymbol{x}_a - \boldsymbol{x}_t) + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\delta}_i \quad (8)$$

It follows that the averaging process reduces the random component of the total error, but does not reduce the bias, due to the a priori information and equal to the term $(\mathbf{I} - \mathbf{A})(\boldsymbol{x}_a - \boldsymbol{x}_t)$ of Eq.(8), which therefore becomes a dominant component. The existence of this bias is one of the reasons why the arithmetic mean cannot be considered as a reference algorithm to collect the information of several products into one. Further reasons concern the choice of a suitable AK matrix to be assigned to the average and the management of possible coincidence and interpolation errors. An explicative comparison of the application of CDF and standard averages in the case of 1000 coincident L2 products is reported in the supplementary material section.

Results and discussion

Fusion in realistic spatial and temporal resolution conditions: the L2 Datasets

To analyse the behaviour of CDF in realistic spatial and temporal resolution conditions four sets of measurements were considered. These measurements correspond to the cloud free observations that were possible between 9:00am and 10:00am on the 1st April 2012. **Tab.1** lists the L2 product types, namely S4-TIR, S4-UV1, S5-TIR and S5-UV1, used in the remaining of the article.

L2 Type	Platform	Band	Number of simulated measurements	Minimal distance between measurements across x along track [km]
S4-TIR	S4	TIR	35594	5.7 x 7.4
S4-UV1	S4	UV1	35594	
S5-TIR	S5	TIR	8023	12.2 x 12.3
S5-UV1	S5	UV1	570	46.2 x 46.7
TOTAL			79781	

Tab.1: *Characteristics of the simulated measurements. For S4 platform across-track is South-North direction and along-track is East-West direction.*

Single grid-box analysis (0.5ºx0.625º)

We consider first the case of a single grid-box (**Fig.1**). In the selected grid-box, 118 measurements were available (55 of S4-TIR, 55 of S4-UV1, 8 of S5 TIR, no S5 UV1). The cell has the size of 0.5 degrees in latitude and 0.625 degrees in longitude, centred on the Egina Island in the Aegean Sea. The cell size has been chosen to be comparable with the assimilation grid used in the AURORA project. We assign the geo-location of the fused product to be the barycentre of the horizontal coordinates of the L2 measurements in the grid-box. In this particular case, since the horizontal distribution of the 118 L2 profiles is quite homogeneous, the barycentre is practically placed at the centre of the grid-cell.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

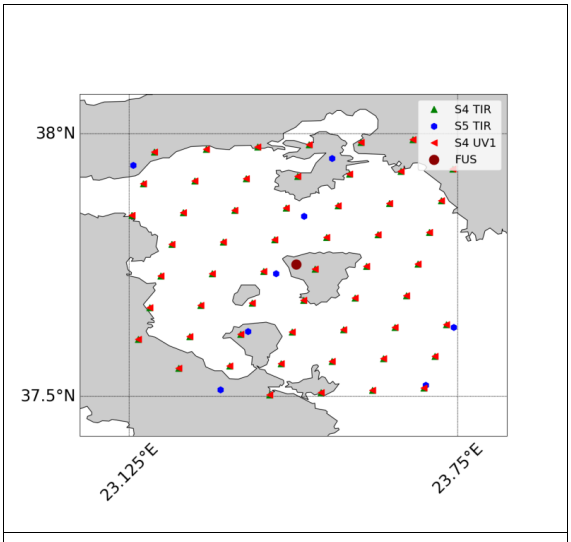

Fig.1: *geographical distribution of the simulated L2 measurements and geo-location of the fused product. The black dashed lines represent the borders of the 0.5°x0.625° grid cells.*

Fig.2 shows with green lines the differences between each L2 profile and the corresponding true profile, with a red line the difference between the fused profile and the mean truth (computed as average of the 118 true profiles), with a black solid-dot line the average of the estimated standard deviation of total error of the individual L2 measurements $\boldsymbol{\sigma}_{total}$, and with a red solid-dot line the estimated standard deviation of the total error of the fused profile $\boldsymbol{\sigma}_{f\,total}$. The last two quantities have been calculated as the square root of the diagonals of the $\mathbf{S}_{total}$ and $\mathbf{S}_{f\ total}$ CMs given by Eqs. (5) and (6) respectively. **Fig. 2** shows that the fused product is in better agreement with its truth than the individual profiles with their own, and presents a smaller estimated total error than the individual L2 products.

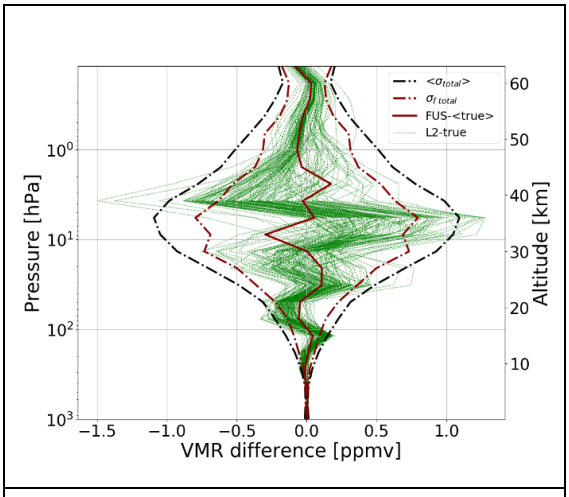

Fig.2:*differences between L2 profiles and their true profiles (green lines), difference between the fused profile and the average of the true profiles (dark red continuous line), the average of* $\boldsymbol{\sigma}_{total}$ *of L2 simulations (black dashed lines),* $\boldsymbol{\sigma}_{f\,total}$ *(dark red dashed lines).*

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

The representation of a retrieved profile is always a compromise between the amplitude of the errors and the vertical resolution. The latter can be quantified by the AKs, which ideally would be equal to the identity matrix in the case of a profile that has a vertical resolution equal to that defined by the sampling grid. Diagonal elements with values smaller than 1 correspond to a loss of vertical resolution. **Fig.3** compares the diagonal elements of the AKs of the L2 products and those of the fused product. **Fig.3** shows that the fused product has better vertical resolution than all the other L2 products at all pressure levels. We have also computed the number of Degrees Of Freedom (DOFs), given by the sum of the diagonal elements of the AK matrix (Rodgers, 2000), for both L2 and fused products, and reported the values in the text box of **Fig.3**. The number of DOFs of the fused product is about twice the number of DOFs of the best L2 one.

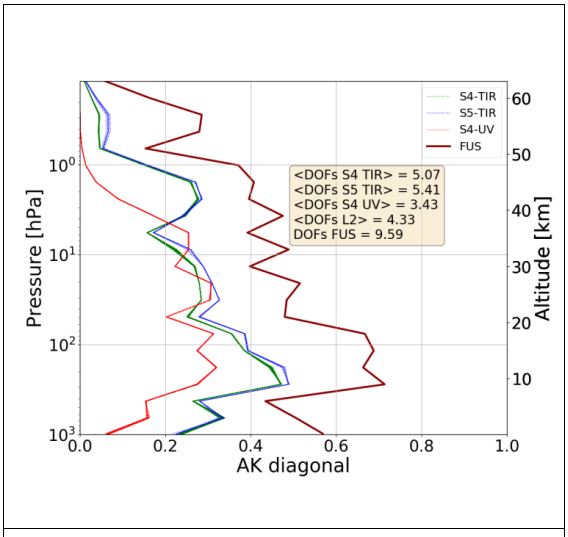

Fig.3: *AKs diagonal of: S4-TIR products (red lines), S5-TIR products (blue lines) S4-UV products (red lines) and FUS product (dark red line). In the text box the average number of DOFs for each type of L2 product, the average number of DOFs for all L2 products and the number of DOFs of the FUS product are reported.*

Statistical analysis for a large domain (0.5ºx0.625º)

While the analysis of the previous paragraph focuses on a particular cell, here an overview of the CDF behaviour is presented, referring to all the 1939 fusion grid-boxes with size of 0.5 degrees in latitude and 0.625 degrees in longitude in which more than one of the 79781 L2 products considered in **Tab.1** is placed. The fused products can be classified depending on the types of L2 measurements falling inside the coincidence grid cell. Since S4-TIR and S4-UV1 products are in perfect coincidence and S5-UV1 products have a horizontal spacing larger than the cell size, only six fused product types (FUS type), listed in **Tab.2**, effectively occur. In this table, the FUS type and its description are reported together with the following complementary data:

- Ncells: the number of grid-boxes characterized by the considered FUS type.
- <NL2>: the mean number of individual L2 fusing profiles per grid-box.
- Max NL2: the maximum number of individual L2 fusing products per grid-box.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

FUS Type	Description	*Ncells*	*<NL2>*	*max NL2*
S4:TIR+UV1	Two or more S4 pixels, no S5 pixels.	908	29.3	160
S4:TIR+UV1_S5:TIR+UV1	Two or more S4 pixels, one or more S5_TIR pixel, one or more S5_UV1 pixel.	245	114.7	163
S4:TIR+UV1_S5:TIR	Two or more S4 pixels, one or more S5_TIR pixel, no S5_UV1 pixels.	299	69.4	165
S4:TIR+UV1_S5:UV1	Two or more S4 pixels, one or more S5_UV1 pixel, no S5_TIR pixels.	2	20	37
S5:TIR+UV1	No S4 pixels, one or more S5_TIR pixels, one or more S5_UV1 pixels.	247	11.1	24
S5:TIR	No S4 pixels, two or more S5_TIR pixels, no S5_UV1 pixels.	238	6.2	14
	TOTAL	1939	41.1	165

Tab.2: *types and characteristics of fused product when a coincidence grid cell size of 0.5°x0.625° is used.*

The left-hand side panel of **Fig.4** shows the geographical distribution of the FUS products. Different colours have been used to classify the fused data according to their provenance type. The irregular geographical coverage is due to the realistic distribution of the cloud free measurements. The histogram in the right-hand side panel of **Fig.4** shows number of cells that contain a given number of measurements, divided in different colours depending on the FUS type. The FUS cells in which only S5 platform L2 products fall are characterized by a small number of L2 measurements, while when S4 products are present, many L2 measurements can be present.

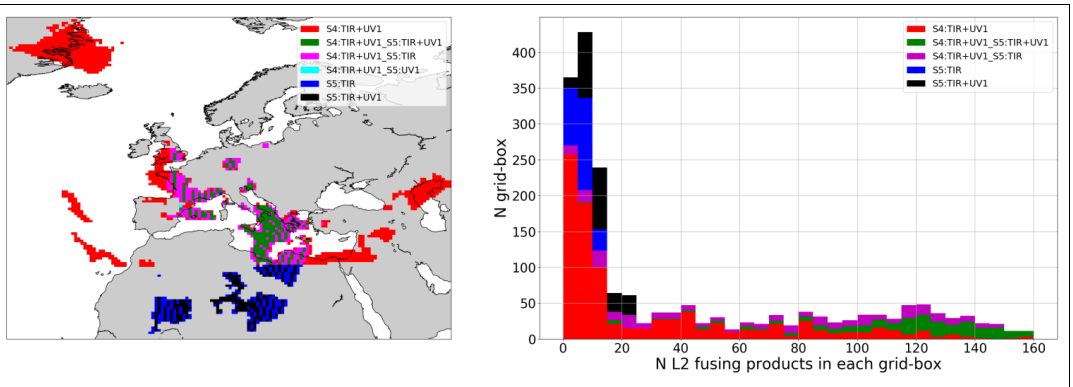

Fig.4. Left panel*: geographical distribution of FUS products differentiated by FUS type where the effect of the lower resolution of S5-UV1 respect to the other L2 products is the cause of the periodic FUS type transitions in the Mediterranean area.* ***Right panel:*** *histogram of the number of cells with a given number of L2 measurements differentiated by FUS type.*

With the selected grid-box size and the multitude of different products that are present in each cell, the question is which product can be used in alternative to the fusion process in those operations in which a single product is requested in each grid-box. Since the averaging process is affected by a large bias error, a viable alternative is the use of the best fusing product present in the cell and we want to compare the CDF result with this product. This comparison is the so called Synergic Factor (SF), introduced by Aires et al. (2012).

The *SF DOF*, defined by Eq. (9), is a pure number that can be calculated for every FUS pixel by the ratio of the number of DOFs of the FUS product and the maximum number of DOFs of the L2 measurements that have been fused. In this equation the index *l* enumerates the vertical levels and the index *i* enumerates the L2 products fused in each grid-box.

$$SF\ DOF = \frac{\sum_l \mathbf{A}_{f,ll}}{\max_{i \in L2} \sum_l \mathbf{A}_{f,ll}} \quad (9)$$

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

When *SF DOF* is larger than one, the FUS product carries more information than the individual L2 measurements. **Fig.5** shows that the *SF DOF* computed for all the fused products (and plotted as a function of the number of L2 profiles in each grid-box) is always larger than one. This means that the information content of the fused product is always larger than that of the standard L2 retrievals. It is also worth noticing that *SF DOF* increases approximately linearly with the logarithm of the number of fusing products, although the proportionality depends on the FUS type. The two different clusters of red symbols (S4:TIR+UV1) are caused by the different latitude bands in which these products are distributed (see also left panel of **Fig.4**). It is important to underline that the improvement in vertical resolution, which cannot be obtained with the arithmetic averaging, is the most demanding requirement (in terms of observation time and instrument sensitivity) in remote sensing observations and, considering the significant gain obtained relative to the single product selection, is the most important feature of fused products.

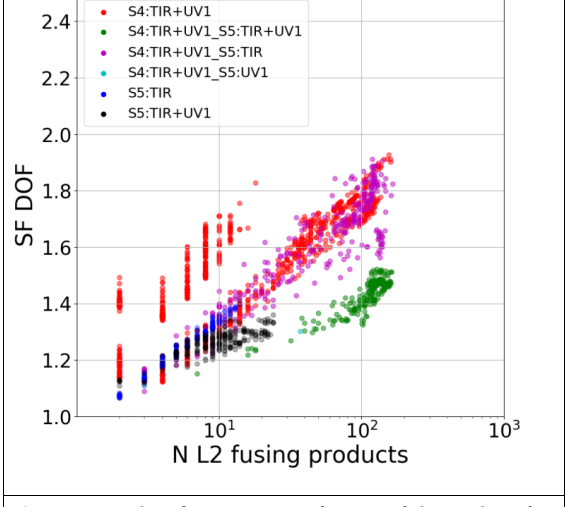

Fig.5: *scatter plot of SF DOF as a function of the number of L2 measurements fused in each coincidence grid cell; different colours represent different FUS types.*

While *SF DOF* is a scalar quantity, both *SF AK* and *SF ERR*, defined by Eqs. (10) and (11), are vertical profiles of pure numbers. *SF AK* represents an expansion on the vertical dimension of *SF DOF* and, in particular, is calculated, level by level, as the ratio between the diagonal elements of the AK matrix of the FUS product and the maximum of the corresponding elements of the AK matrices of the fusing L2 measurements.

A value of *SF AK* larger than 1.0 at a specific vertical level (indicated by the index *l*) means that, at that level, the vertical resolution of the FUS product is better than that of all the individual products.

$$SF\ AK_l = \frac{\mathbf{A}_{f,ll}}{\max_{i \in L2} \mathbf{A}_{i,ll}} \quad (10)$$

The *SF ERR* (Eq. (11)) at a given level is calculated as the ratio between the minimum total error of the L2 measurements that have been fused and the total error of the FUS product. A value of *SF ERR* larger than 1.0 means that at a specific level the error of the FUS product is smaller than that of all the individual products.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

$SF\ ERR_l = \frac{\min_{i \in L2} \boldsymbol{\sigma}_{total,i,l}}{\boldsymbol{\sigma}_{f\ total,l}}$	(11)

The SFs defined by Eqs. (10) and (11) provide a conservative comparison because the fused product is compared with the L2 product that at that level has the largest diagonal value in its AK matrix and with the one that has the smallest total error at the same level (generally these are two distinct L2 products).

Fig.6 shows the *SF AK* (left panel) and *SF ERR* (right panel) profiles for the 1939 FUS products considered in **Tab.2**. We have used different colours to denote the provenance of the L2 data contributing to the fused products and different symbol size to infer the number of L2 fusing measurements (the larger the symbol size, the larger the number of L2 fusing profiles). As mentioned above, the merit of the fused product in terms of SF is higher than that of the L2 retrievals if the *SF AK* and *SF ERR* are greater than 1. The significant improvement obtained with the fused products is confirmed by **Fig. 6**. It can also be noticed that considering symbols of the same colour the symbol size (*N*) tends to increase moving horizontally in the graph (same vertical level) from left to right (SF increasing) denoting that, for each FUS type, SF increases with *N*. This is not in contradiction with the fact that symbols with different colours (FUS types) and different sizes (*N*) can share the same position (SF, vertical level) on the graph.

A few *SF AK* values are slightly smaller than one in the middle to upper stratosphere. This is likely caused by the coincidence errors that have to be added in the fusion process.

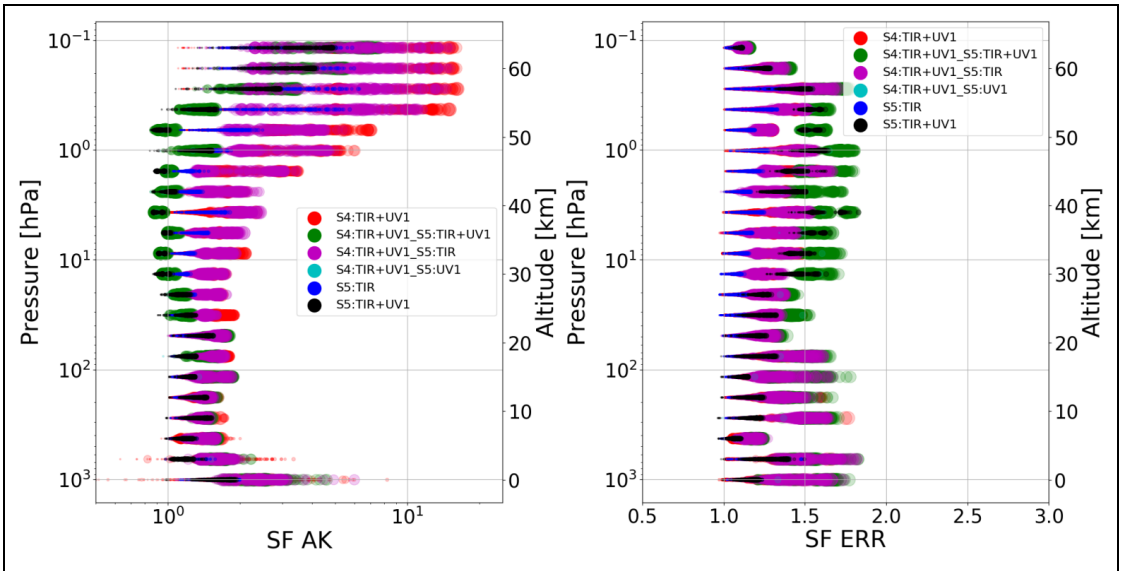

Fig.6. Left panel: *SF AK versus vertical level.* ***Right panel:*** *SF ERR versus vertical level. In both panels, different colours of the symbols represent the FUS type, different sizes of the symbols represent the number of measurements that have been fused. The maximum symbol size shown in the legend corresponds to N=160.*

Statistical analysis on a coarse horizontal resolution (1°x1°)

We have seen that, starting from 79781 L2 measurements (**Tab.1**), when a coincidence grid-box with size 0.5°x0.625° is used, the number of fused profiles is 1939 (**Tab.2**), with a reduction of the data volume of more than a factor 40.

Tab.3 provides a summary of the number of fused profiles and the provenance of the L2 profiles that contribute to them for a fusion grid resolution of 1°x1°. In this case, the total number of FUS products is 775 with a reduction of data volume of more than a factor 100.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

FUS Type	Description	N_{cells}	$<NL_2>$	*max NL2*
S4:TIR+UV1	Two or more S4 pixels, no S5 pixels.	354	73.1	420
S4:TIR+UV1_S5:TIR+UV1	Two or more S4 pixels, one or more S5_TIR pixel, one or more S5_UV1 pixel.	140	289.4	504
S4:TIR+UV1_S5:TIR	Two or more S4 pixels, one or more S5_TIR pixel, no S5_UV1 pixels.	79	115.4	442
S4:TIR+UV1_S5:UV1	Two or more S4 pixels, one or more S5_UV1pixel, no S5_TIRpixels.	0	0	0
S5:TIR+UV1	No S4 pixels, one or more S5_TIR pixels, one or more S5_UV1 pixels.	142	26.2	71
S5:TIR	No S4 pixels, two or more S5_TIR pixels, no S5_UV1 pixels.	60	8.9	26
TOTAL		775	102.9	504

Tab.3: *Like in Tab.2 but with a grid-box size of 1°x1°.*

The Synergic factors *SF DOF*, *SF AK* and *SF ERR* have been considered also in this case and the figures (similar to **Fig.5** and **Fig.6**) are reported in the supplementary material. In summary, the greater number of fusing observations in each fusion cell produces a further improvement for both the vertical resolution and the total error proving that 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 coincidence error amplitude, which increases with the geographical variability degrading the quality of the fused product.

Conclusions

This paper presents a sensitivity study of the Complete Data Fusion technique, applied to L2 measurements simulated with the characteristics expected for the atmospheric Sentinels 4 and 5. This analysis allows to evaluate the performances of the CDF algorithm in ideal conditions (i.e., with no systematic errors added) and to quantify the possible benefits of the application of CDF to real Sentinel data.

In particular, we show the application of CDF to a single cell with size of 0.5 degrees in latitude and 0.625 degrees in longitude in which more than 100 L2 products are fused. Results show that the fused product is characterized by higher information content, smaller errors and smaller residuals (i.e. anomaly from the true profile) compared to individual L2 products. The information content being, with its improvement of the vertical resolution, the most important achievement.

This analysis is then extended to a larger domain consisting in 79781 L2 products subdivided in 1939 grid boxes with 0.5°x0.625° size. In this case the comparison of L2 products and CDF output are carried on in terms of synergic factors. This analysis shows that the CDF can be applied to a wide range of situations and that the benefits of the fusion strongly depend on the number of the measurements that are fused together and from their characteristics. It is also shown that CDF can be run customizing grid resolutions, e.g. to match the resolution requirements of the process that will ingest the products, with full exploitation of all the available measurements.

As the fused products are traced back to a regular, fixed horizontal grid and, as shown here, are not affected by the bias introduced by the a priori information, they can be considered as a new type of level 3 products with improved quality (reduced bias) and the same characteristics (AK included) with respect to L2 products.

Data availability

The data of the simulations presented in the paper are available from the authors upon request.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Atmospheric Measurement Techniques Discussions

Author contributions (according to CRediT https://casrai.org/credit/)

N. Zoppetti: Conceptualization, Methodology, Software, Writing – Original Draft, Writing – Review & Editing, Investigation, Data curation, Visualization S. Ceccherini: Conceptualization, Methodology, Investigation, Writing – Review & Editing B. Carli: Conceptualization, Methodology Writing – Review & Editing, Supervision S. Del Bianco: Investigation, Data curation, Project Administration M. Gai: Investigation, Data curation C. Tirelli: Investigation, Data curation, Project Administration F. Barbara: Resources R. Dragani: Investigation, Data curation, Writing – Review & Editing A. Arola: Investigation, Data curation J. Kujanpää: Investigation, Data curation R. Van Der A: Investigation, Data curation U. Cortesi: Funding Acquisition, Project Administration

Competing interests.

The authors declare that they have no conflict of interest.

Acknowledgments

The results presented in this paper arise from research activities conducted in the framework of the AURORA project (http://www.aurora-copernicus.eu/) supported by the Horizon 2020 research and innovation programme of the European Union (Call: H2020-EO-2015; Topic: EO-2-2015) under Grant Agreement N. 687428.

Financial support.

This research has been supported by the European Commission, H2020 (AURORA, grant no. 687428).

References

Aires,F., Aznay,O., Prigent,C., Paul,M. and Bernardo,F.: Synergistic multi-wavelength remote sensing versus a posterior combination of retrieved products: Application for the retrieval of atmospheric profiles using MetOp-A. J GEOPHYS RES, Vol. 117, D18304, https://doi.org/10.1029/2011JD017188, 2012.

Ceccherini, S., Carli, B., Pascale, E., Prosperi, M., Raspollini, P. and Dinelli, B.M.: Comparison of measurements made with two different instruments of the same atmospheric vertical profile, Appl. Opt.,42,6465–6473, https://doi.org/10.1364/AO.42.006465, 2003.

Ceccherini, S., and Ridolfi, M.: Technical Note. Variance-covariance matrix and averaging kernels for the Levenberg-Marquardt solution of the retrieval of atmospheric vertical profiles, Atmos. Chem. Phys., 10**,** 3131-3139, https://doi.org/10.5194/acp-10-3131-2010, 2010.

Ceccherini, S., Carli, B., and Raspollini, P.: The average of atmospheric vertical profiles, Opt. Express, 22,24808-24816, https://doi.org/10.1364/OE.22.024808, 2014.

Ceccherini, S., Carli, B., and Raspollini, P.: Equivalence of data fusion and simultaneous retrieval, 2015. Opt. Express, 23, 8476-8488, https://doi.org/10.1364/OE.23.008476, 2015.

https://doi.org/10.5194/amt-2019-436
Preprint. Discussion started: 18 December 2019
© Author(s) 2019. CC BY 4.0 License.

Ceccherini, S., Carli, B., Tirelli, C., Zoppetti, N., Del Bianco, S., Cortesi, U., , Kujanpää, J., and Dragani, R.: Importance of interpolation and coincidence errors in data fusion. Atmos. Meas. Tech., 11, 1009–1017, https://doi.org/10.5194/amt-11-1009-2018, 2018.

Ceccherini S., Zoppetti N., Carli B., Cortesi U., Del Bianco S., and Tirelli C.: The cost function of the data fusion process and its application. Atmos. Meas. Tech., 12, 2967–2977, https://doi.org/10.5194/amt-12-2967-2019, 2019.

Cortesi, U., Ceccherini, S., Del Bianco, S., Gai, M., Tirelli, C., Zoppetti, N., Barbara, F., Bonazountas, M., Argyridis, A., Bós, A.,Loenen, E., Arola, A., Kujanpää, J., Lipponen, A., Nyamsi, W.W., van der A, R., van Peet, J., Tuinder, O., Farruggia, V., Masini, A., Simeone, E., Dragani, R., Keppens, A., Lambert, J.-C., van Roozendael, M., Lerot, C., Yu, H., and Verberne, K.: Advanced Ultraviolet Radiation and Ozone Retrieval for Applications (AURORA): A Project Overview, Atmosphere, 9, 454, https://doi.org/10.3390/atmos9110454, 2018.

Gelaro, R., McCarty, W., Max J. Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J. Climate, 30, 5419-5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Kroon, M., de Haan, J. F., Veefkind, J. P., Froidevaux, L., Wang, R., Kivi, R. and Hakkarainen, J. J.: Validation of operational ozone profiles from the Ozone Monitoring Instrument, J. Geophys. Res., 116, D18305, https://doi.org/10.1029/2010JD015100, 2011.

Liu, X., Bhartia, P. K., Chance, K., Spurr, R. J. D., and Kurosu, T. P.: Ozone profile retrievals from the Ozone Monitoring Instrument, Atmos. Chem. Phys., 10, 2521-2537, https://doi.org/10.5194/acp-10-2521-2010, 2010.

McPeters, R.D., and Labow, G.J.: Climatology 2011: An MLS and sonde derived ozone climatology for satellite retrieval algorithms, J. Geophys. Res., 117, D10303, https://doi.org/10.1029/2011JD017006, 2012.

Miles, G. M., Siddans, R., Kerridge, B. J., Latter, B. G., and Richards, N. A. D.: Tropospheric ozone and ozone profiles retrieved from GOME-2 and their validation, Atmos. Meas. Tech., 8, 385-398, https://doi.org/10.5194/amt-8-385-2, 2015.

Rodgers, C.D.: Inverse Methods for Atmospheric Sounding: Theory and Practice. Vol. 2 of Series on Atmospheric, Oceanic and Planetary Physics. World Scientific: Singapore, 2000.