living planet symposium | VIENNA 23–27 June 2025

Mutual Gap-Filling of Sentinel-5p Datasets

Giorgia De Moliner[1], Alessandro D'Ausilio[2], Camillo Silibello[2], Giovanni Lonati[1]

[1] Department of Civil and Environmental Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, 20133, Italy
[2] ARIANET-Suez srl, Via Benigno Crespi 52, Milan, 20159, Italy

THE RISK OF MISSING OUT

The EU's Copernicus **Sentinel-5p** (S5p) satellite has been operating in a sun-synchronous orbit since 2018, proving **high-resolution** (3.5x5km) **daily global observations of the atmospheric composition** and delivering critical EOs datasets to inform decision-making across various sectors, from public health to climate mitigation.

Despite significant advancements in remote sensing, these EO datasets still suffer from **complex, large-scale missing values** that vary across chemical species. These gaps stem from **sensor limitations**, **thick clouds**, **snow/ice**, or low-quality data caused by the **challenges of retrieving column totals from spectral information**. As an example, Figure 1 shows the percentage of missing data of NO_2, SO_2, CO and HCHO on winter and summer.

Missing data **hinder trend analyses**, **introduce biases** into observational records, and **obscure physical dependencies** among trace gases. Moreover, when EO data are used as predictors in Machine Learning (ML) applications, gaps can **invalidate other predictors**, **significantly reducing the size of training datasets**. Furthermore, training ML models on low-quality gap-filled data can result in inaccurate or biased outcomes, as well as reduced interpretability.

High-quality gap-filling is therefore essential to **unlock the full potential** of air quality EO datasets, ensuring they deliver **valuable**, **reliable**, and **usable** information products.

SENTINEL-5P MISSING RATES

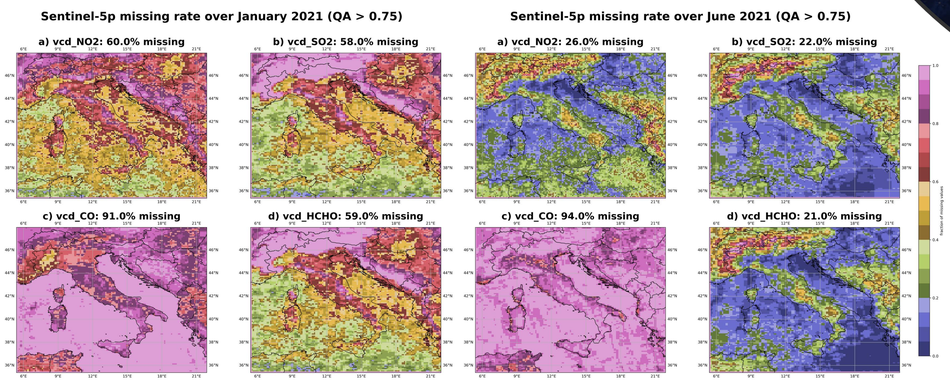


Figure 1. Percentage of invalid data of Sentinel-5p observations in January 2021 (left) and in June 2021 (right). Quality flag is set to > 0.75, filtering low-quality retrievals.

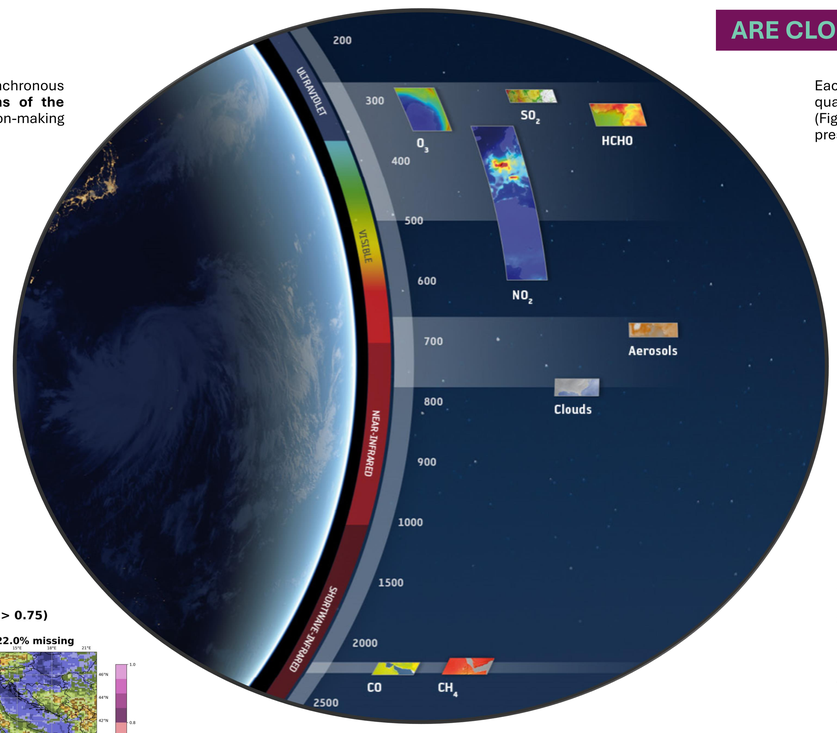


Figure 2. TROPOMI's spectral range (source: ESA, 2017).

ARE CLOUDS THE MAIN CAUSE OF MISSING VALUES?

Each observation is accompanied by a quality index, the **qa value**, that grades the quality of the retrieval result. When plotted against an indicator of the cloud shielding (Figure 3), it can help in assessing how much observations are influenced by the presence of clouds on the scene.

In the case of **SO_2**, the graph is straight-forward. No data under cloudy-sky conditions (cf > 0.5) are paired with a high-quality flag (qa > 0.75). **Ghost columns below the clouds can't be retrived and quality drops**. There is a **linear correlation** between the two: the higher the cloud fraction, the lower the reliability of the data. On the other hand, the vertical column densities (VCD) seems to not be a polarizing factor as clouds are. The same conclusions can be drawn for **HCHO** as well.

On the contrary, in the case of **CO**, observations are clearly **independent of the cloud fraction**. This is due to the SICOR algorithm, which uses the instrument sensitivity to CO above the cloud to infer the entire column by suitably scaling the a-priori profile. Moreover, **over the ocean, meaningful retrievals rely purely on cloudy observations**, as the ocean surface is typically very dark in the shortwave infrared spectral range and so clear-sky captures generally have too few light for the sensor to be effective. This is also visible in Figure 1, in January in particular. Quality and VCD seems to be linked. Under 0.01 mol/m2, qa was always below the recommended value of 0.5, hinting on the **difficulty of the sensor to correctly catch low CO concentrations**.

For **NO_2**, three regimes can be identified. In the first, the filtering acts effectively as a cloud mask: it is the instance of high-quality retrievals (qa > 0.75), for which the lower the cf, the higher the qa. **Optically thick clouds as a matter of fact limit satellite sensitivity** to NO_2 below the cloud by blocking its view. In the second regime, corresponding to medium-to-low quality retrievals (0.3 < qa < 0.75), the pattern appears reversed – only cloudy-sky observations are present. **Interestingly, in this quality range, data seem largely unrelated to cloud fraction.** This may be due to multiple scattering within the cloud and strong backscatter from the cloud top, which can enhance sensitivity to NO_2 near or above the cloud. As a result, **observed signals may be either attenuated or amplified in the presence of clouds, leading to qa values that are, to some extent, independent of cloud cover.**

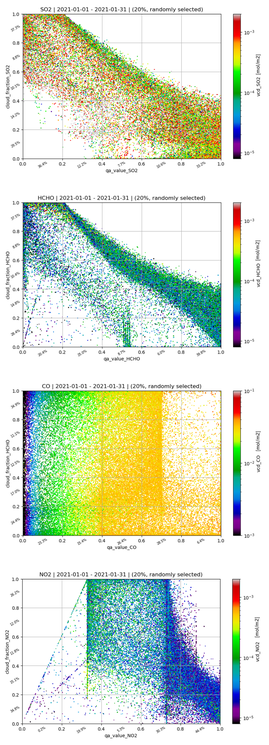


Figure 3. Scatter plots showing the relationship between the quality filter and the cloud fraction. Each point represents a S5p retrieval. From top to bottom, NO2, SO2, CO and HCHO.

TOWARDS A MULTIVARIATE ANALYSIS

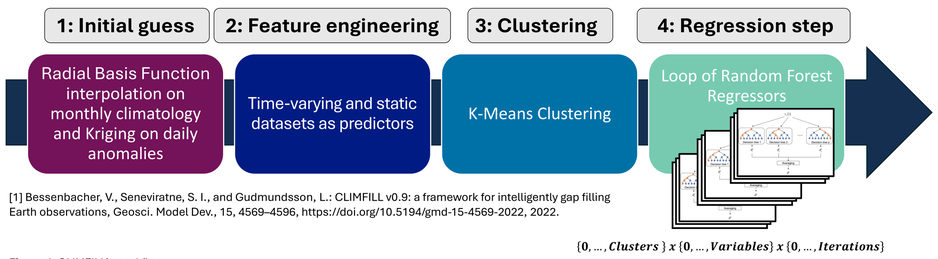
Each trace gas leaves a unique absorption signature in the backscattered radiances, caught by the optical instrument. The spectral signature is used to infer the presence of a trace gas. **Different trace gases are retrived from different wavelength ranges**, namely on bands where their spectral signature stands out the most. The **TROPOMI** multispectral instrument, aboard S5p, covers the spectrum **from UV to SWIR**. Figure 2 shows which wavelengths are used to infer each chemical species.

As a result, retrived polluant maps may be characterized by patterns of missing data that differ from one another. In other words, when the data of a trace gas is captured, not necessarily all the others are retrived as well. The **non-overlapping missingness patterns** of trace gases can be exploited to carry out a **multivariate analysis** – an **innovative approach** in the field of air pollution modeling.

METHODS

An advanced ML-based gap-filling methodology designed to address gaps in EOs of Land–Climate Interaction, **CLIMFILL** (V. Bessenbacher, 2022 [1]), was selected and its performance in reconstructing S5p TROPOMI Level 2 datasets was evaluated with respect to **naïve** and a **state-of-the-art spatio-temporal interpolation technique**. CLIMFILL employs **multiple Random Forest** (RF) **Regressors** and leverages **temporal autocorrelation**, **spatial neighborhood information**, and **cross-variable dependencies** in a truly mutual gap-filling framework. It is computationally feasible for large datasets.

CLIMFILL's workflow is presented in Figure 4. It consists of four main steps. In the first one missing values are filled using spatial interpolation to produce a **first guess**. The second step is a **feature engineering** where dynamic and static predictors are introduced to build the training dataset. In the third step, **clustering**, data points are grouped based on their multivariate similarity, without considering their location in time or space. The final step refines the gap-filling of the first step, leveraging ML. Within each cluster, first guesses are improved by repeatedly learning and predicting with **a loop of RFs** the missing values using information from predictors and other variables until the results stabilize or a maximum number of iterations is reached. In this way, real-world observations are directly used to fill gaps in other trace gas records, without the need for any donor maps.



[1] Bessenbacher, V., Seneviratne, S. I., and Gudmundsson, L.: CLIMFILL v0.9: a framework for intelligently gap filling Earth observations, Geosci. Model Dev., 15, 4569–4596, https://doi.org/10.5194/gmd-15-4569-2022, 2022.

Figure 4. CLIMFILL's workflow.

MATERIALS

This approach was tested on, tropospheric **NO_2**, **SO_2**, **CO** and tropospheric **HCHO** retrievals as observed by TROPOMI sensors from **January 1st, 2021, to June 30th, 2021**, over the Italian domain. Missing rates are shown in Figure 1. Data were filtered with a quality flag of **> 0.75**. Lowering the threshold was not leading to any significant improvement for CLIMFILL and results are not presented here. Both observations and predictors were regridded to match the resolution of **0.1°x0.1°.** 10% of the data of each variable have been removed and fictitious but realistic missing values have been introduced to allow results assessments. Inevitably, the **evaluation** was performed on mostly clear-sky observations, where data were available for comparisons.

For the current case study, 11 variables of the **ERA5 reanalysis** are used (as displayed in the correlation heatmap below), together with 8 **Land Use** classifications, a **DEM**, a population density distribution and the spatial distances from the domain corners and the center. An optimal number of **50 clusters** was picked with the elbow method for the current case study.

RESULTS

CLIMFILL is evaluated with respect to two others gap-filling methods, the naïve technique of infilling the period mean and the spatial interpolation of step one, in reconstructing the whole picture. **Monthly means differ from one gap-filling to another, as well as from the monthly mean computed from the original dataset** as streamed by the satellite. The latter is influenced by its partiality as seen above, and therefore the data aggregation risks to spark biased and untrustworthy results. The Mean Absolute Error (MAE) metric was chosen for the testing as it is relatively little influenced by outliers.

Results strongly differ from variable to variable (Figure 6)**.** An improvement was seen for **NO_2 in January**, when the MAE was reduced by CLIMFILL, where a 14.5% reduction was registered. Interestingly, **the retrieved map has a lower peak value in the Po valley compared to both the spatial interpolated maps and the gappy one**. It is not the case of June, where MAE for NO_2 grows back. In warmer time periods, **CO and HCHO gap-fillings were proved to find CLIMFILL beneficial**, refining first guess interpolation by 30 to 70%, significantly departing from the naïve infilling.

However, these improvements are not always driven by other S5p variables. For CO and HCHO, they are influenced by co-observations, whereas this is not the case for NO_2, where additional predictors play a more significant role. This is most likely due to the **difficulty in establishing correlations between pollutants that lack strong physical linkages**. Even though they are retrieved from the same observation, **ML struggles to identify meaningful dependencies**. The correlation heatmap shown in Figure 5 illustrates the linear correlations between covariates, which are generally weak or absent for most variable pairs.

Moreover, **some CLIMFILL estimates do not differ significantly from other gap-filled maps**, which are also less computationally demanding. This outcome can be attributed to the **missingness patterns** of the variables. The proportion of valid observations available for only a single variable at a time accounts for less than 5% of all data collected during the first six months of 2021. This is illustrated in Table 1, which shows the percentages of having data for one variable while others are missing. **The fact that pollutants are observed at different wavelengths is beneficial, as is the independence of some retrievals from physical obstacles affecting other variables** (i.e., CO and clouds). However, such favorable observations appear to be too limited to consistently support the training of a machine learning algorithm. **Expanding the temporal span of the training dataset might help increase the number of favorable cases.**

To assess the potential role of models in helping reconstructing observations, **CAMS regional air quality reanalysis** were added as predictors. Total vertical columns were computed without accounting for TROPOMI's Averaging Kernels, as they were not available over the whole domain**. Only in the case of NO_2 CLIMFILL estimates were improved, by 20%. For all other pollutants, introducing model simulations did not benefit** RFs in their training phase. Note that S5p observations are as of now only assimilated in the global reanalysis, which are then used as boundary conditions to run higher-resolution regional models. Regional reanalysis were considered independent from S5p observations.

QA > 0.75	Variables	% of valid data	NaNs in *all others* when current valid	NaNs in *two others* when current valid	NaNs in *one other* when current valid	% of *all* variables being valid
	CO	9,57%	0,48%	0,97%	0,82%	7,30%
	HCHO	60,90%	0,39%	8,15%	45,06%	7,30%
	NO2	57,91%	2,00%	3,88%	44,73%	7,30%
	SO2	59,96%	1,36%	6,63%	44,34%	7,30%

Table 1. Proportion of valid observations across all pixels retrieved during the January–June 2021 period, followed by the percentages of having data for one variable while others are missing, by trace gas.

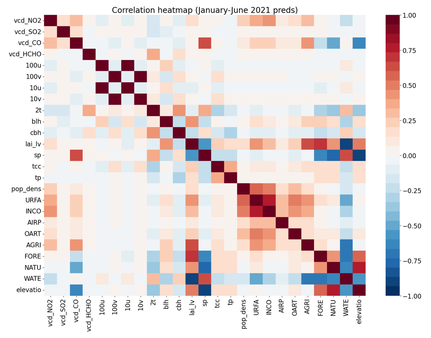


Figure 5. The correlation heatmap of the covariates fed into the RFs loop. It shows linear relations between each pair.

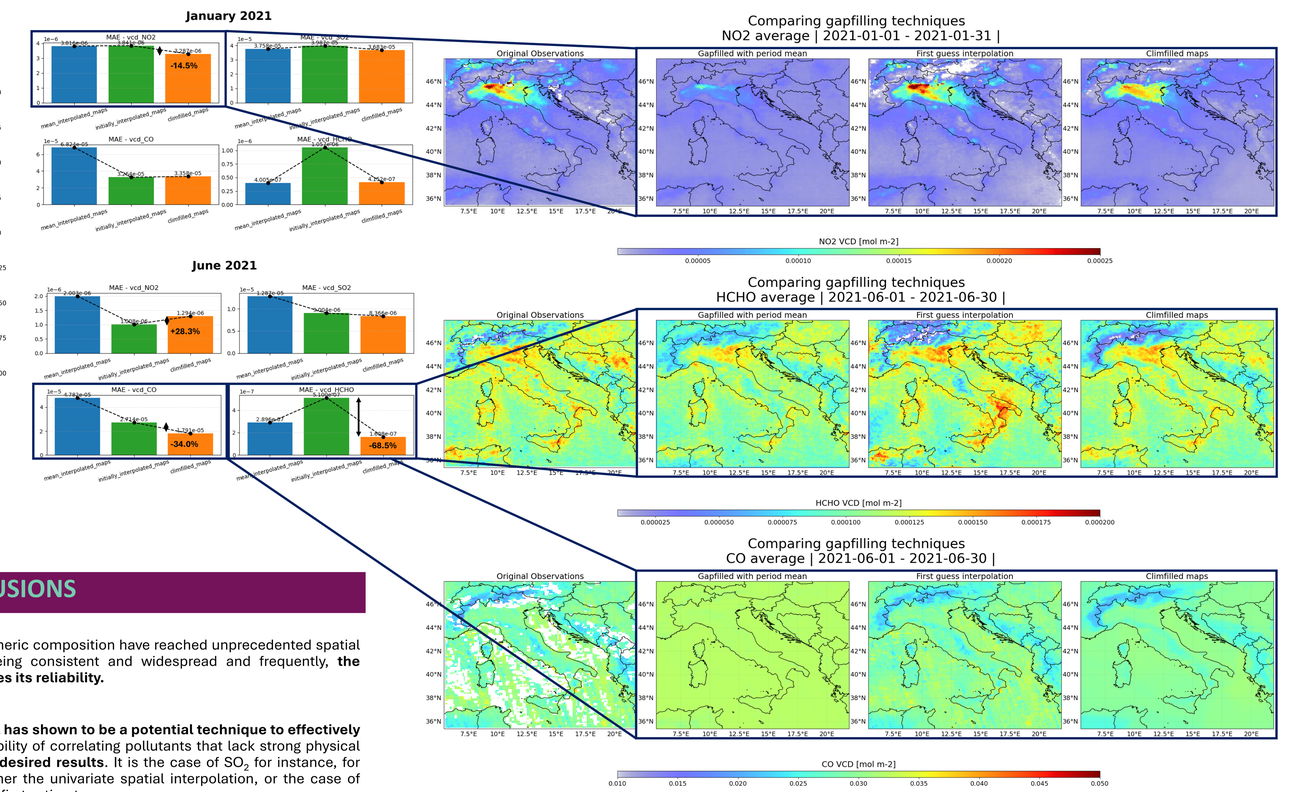


Figure 6. On the left, the Mean Absolute Errors (MAEs) of the three gap-filling techniques – highlighted by different colors – are shown. They are computed with respect to original observations. Results are displayed for January 2021 (top) and June 2021 (bottom), broken down by trace gas. On the right, monthly averages of NO2 (top), HCHO (center) and CO (bottom) are presented. Each subplot shows a different gap-filling technique. The first to the left is the monthly mean of the original, not gap-filled dataset. For NO2, January 2021 is pictured, whereas for HCHO and CO, June 2021.

CONCLUSIONS

Remote sensing techniques able to monitor the atmospheric composition have reached unprecedented spatial and temporal resolutions. Yet, missing data keep being consistent and widespread and frequently, **the impartiality of the data hinders its usability and reduces its reliability.**

By capturing complex non-linear relationships, **CLIMFILL has shown to be a potential technique to effectively infill data gaps in air quality EOs**, hinting on the feasibility of correlating pollutants that lack strong physical linkages. **However, it does not consistently yield the desired results**. It is the case of SO_2 for instance, for which the multivariate analysis does not help any further the univariate spatial interpolation, or the case of June's gap-filling of NO_2, in which RFs even worsened the first estimate.

Further analysis and testing are necessary to address the limitations, which primarily stem from applying the technique to datasets that differ significantly from those used in **previous applications**. Unlike the current dataset, earlier applications involved variables that were strongly interrelated, aggregated into monthly time series, and characterized by observations that were more likely to occur when others were absent. Widening the **temporal span of the training dataset** might help improve performance.

The study also examined whether incorporating simulated concentration maps as predictors enhances the gap-filling process. It was proven that, apart from NO_2, using **CAMS regional reanalyses** as predictors hardly contribute. It is most likely due to the well-documented bias between model simulations and trace gases satellite observations. Additionally, the study analyzed the occurrence of missing data in S5p products, identifying patterns and investigating the role of clouds. **SO_2 and HCHO were shown to be more constrained by clouds, whereas CO and NO_2 were, to different extents, less dependent on them.**

POLITECNICO MILANO 1863

ARIANET

REFERENCE CONTACT:
Giorgia De Moliner
PhD Candidate, Politecnico di Milano
giorgia.demoliner@polimi.it