Project Title	Prototype of HPC/Data Infrastructure for On-demand Services
Project Acronym	PHIDIAS
Grant Agreement No.	INEA/CEF/ICT/A2018/1810854
Start Date of Project	01.09.2019
Duration of Project	36 Months
Project Website	www.phidias-hpc.eu

D4.3.1 - Sentinel 5 Precursor service V0

Work Package	**WP 4, Use Case 1 - Satellite data**
Lead Author (Org)	**Hervé THEVENON (SPASCIA)**
Contributing Author(s) (Org)	**Pascal PRUNET (SPASCIA)**
Due Date	**01.03.2021**
Date	**28.05.2021**
Version	**V0.4**

Dissemination Level

X	PU: Public
	PP: Restricted to other programme participants (including the Commission)
	RE: Restricted to a group specified by the consortium (including the Commission)
	CO: Confidential, only for members of the consortium (including the Commission)

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Versioning and contribution history

Version	Date	Author	Notes
0.1	21.03.2021	Hervé THEVENON, Pascal PRUNET (SPASCIA)	TOC and V0.1
0.2	21.04.2021	Nicolas PASCAL (Université de Lille)	Review
0.3	05.05.2021	Hervé THEVENON (SPASCIA)	Revised edition
0.4	28.05.2021	Florian PIFFET (CINES)	Final edition

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Table of Contents

List of Figures

List of Tables

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Executive Summary

This deliverable showcases the work done on the automatic detection and geolocation of NO_2 emission sources using solely S5P Level 2 NO_2 data.

Several sites that emit NO_2 at different geographic scales were successfully detected and geolocated, with a prediction rate above 90%.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

1 Automatic detection and geolocation

1.1 Datasets

1.1.1 Primary dataset

Our primary dataset consists of 2 years of level 2 NO_2 concentrations in netcdf format, from October 2018 till November 2020. The data was downloaded from AERIS/ICARE that mirrors ESA's repository.

The processor versions used to produce the data is 1.3.0 at best, with no backward compatibility issue. The next processor version – 1.4.0 – was enforced early December 2020. It is not backward compatible with 1.3.0 and earlier, therefore we will be required to use the new RPRO series for the purpose of using data from January 2021.

As S5P orbits Earth 14 times a day, 2 years of data represents approximately 10,000 orbits, as many netcdf files for the OFFL series we elected to use, and these 10,000 files contain approximately 15 billion observations.

1.1.2 Database storage – Spatial index

In order to improve the usability of the dataset, the data was imported in a Postgresql database. By usability we mean that:

- we deal with a single data source instead of 10,000 netcdf files,
- the data source can filter, aggregate and sort data, which operations can not be performed in netcdf files, - the results can be made available as CSV files for further processing, or directly visualised in QGIS.

The import of the observations held in the netcdf files was performed without any filtering (neither on vertical column density, or cloud coverage, or quality assurance flag).

Each observation added to the database is assigned a bespoke spatial index that provides better retrieval performances than native PostGIS indexing. This spatial index is built through a bespoke SQL function that takes a tuple consisting of a longitude `lon` and a latitude `lat` in decimal degrees as arguments, and returns a positive integer `ndx = 180 * [ ROUND(lon+360) % 360 ] + [ ROUND(lat+180) % 180 ]`.

For the moment, all observations are added in a single table. Partitioning will be in order to keep the retrievals speedy as the database grows.

The production environment will keep ingesting and indexing new data as new granules become available.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

1.1.3 Dataset used to develop and validate the processing method

The main dataset used for this work is centred on the Kusile power station, located 90 km East-North-East of
Johannesburg, South Africa. The area covered consists of the 1º x 1º tiles whose defining corner lies within 300 kilometres from the Kusile power station. This area is interesting for it counts numerous coal power stations, at different stage of operation, as well as open coal mining and disposal areas (where ashes being spread to stop smouldering).

We used a (lon, lat) grid based on a quarter of the native resolution of the data produced by TROPOMI under such latitudes. This grid was selected to provide usefulness to the geolocation. Useful geolocation *pinpoints* the emission sources.

Native 5.5 km x 3.5 km (since August 2019) plus or minus half the longitude and latitude is not proper pinpointing material. Number of power stations are located within 6 km from each other. They would remain undistinguishable under the native resolution.
Disaggregation – the process of losing granularity on one dimension of a dataset to the benefit of another dimension – allows to gain insight on the geographical dimensions at the expense of the time dimension. We disaggregated 2 years of data.
The quarter of the native resolution (quarter in longitude, and quarter in latitude) was decided empirically by comparing the Shannon's entropy of the original data set with the entropy of the disaggregated dataset within an area of interest.
In information theory, the entropy provides the amount of information (or surprise information) in a dataset. We reasoned that disaggregation must not "create" information, therefore the entropy of the disaggregated dataset must be lesser than the entropy of the original dataset. The quarter of the native resolution proved to be a safe limit.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

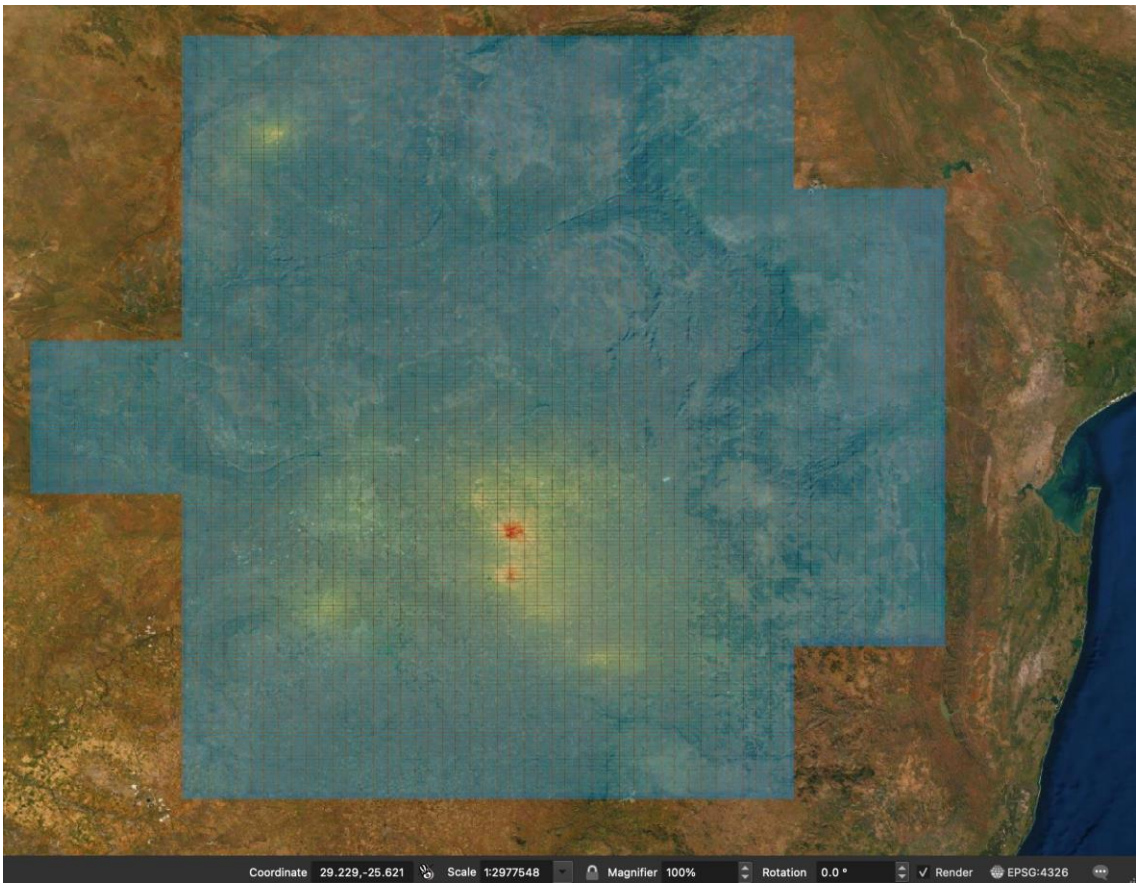

Figure 1 – Average NO_2 concentrations resulting from the processing described in section 1.1.3. Both peaks are Matla (North) and Secunda (South) power stations.

Following from the usefulness expectation set above, an area of interest was defined as containing two or more sources (e.g. two power stations) less than 10 kilometres apart. The area containing the power stations of Matimba and Medupi was identified as suitable. The size of the area of interest was set to approximately 30 km x 30 km. The precise area was enclosed between (27.412E, 23.816S) and (27.722E, 23.573S).

The re-gridding process assigns the vertical column density, eastward and northward wind components of each observation to the node of the target grid that is the nearest to the centre of the observation. Only observations where vertical column density is positive, cloud coverage is below 40%, and quality assurance flag is greater than 60% were used. The three components of each node were subsequently averaged to produce the dataset used in the following.
As wind components were added with processor 1.3.0 (March 2019), the mean concentration is built on more observations that the mean wind components.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

1.2 Processing

1.2.1 Overview

Core algorithm

The processing of the dataset resulting of the disaggregation described in section 1.1.3 is performed in three steps, each with a very specific perspective:

- Step 1: the dataset is considered as an image that needs enhancement, specifically noise reduction,
- Step 2: smooth concentration gradients obtained in step 1 are further processed in order to infer what concentrations would look like without the combined effect of the wind and photochemical decay of NO_2,
- Step 3: emission peaks are selected by iteratively (a) selecting the highest level of emission found in step 2 and (b) using a recursive gradient descent algorithm in order to infer and tag out the locations whose emissions are likely due to the peak.

Multi-scale approach

Experience shows that this process is advantageously applied from wide to narrower field of view, in order to identify the weakest sources that are overshadowed by the strongest ones.

1.2.2 Core algorithm

Step 1 is a classic gaussian noise cancelling technique in image processing (https://arxiv.org/pdf/1505.03489.pdf), applied to all the components of the dataset: vertical column density, eastward wind, and northward wind.

a) The *smoothing mask* is calculated by applying a normalised 5x5 gaussian filter to each node of the dataset surrounded by at least two pixels in the four directions (E, N, W, S).
b) The *difference* between the original dataset and the smoothing mask is calculated. Effectively this is the basic arithmetic difference between each component of each node.
c) The *extrema* of the difference built in step 1b are calculated.
d) The vertical column density of the difference is *thresholded*. Values of vertical column density lower than 90% of the maximum vertical column density are replaced by the minimum vertical column density. The choice of 90% is arbitrary.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

e) Step 1e is the last step: the *smoothed* dataset is obtained by adding together the thresholded dataset (1d) with the smoothing mask (1a). All the components are included.

Step 2 is a naive retro-advection process.

a) We create a copy T (for Target) of the smoothed dataset S (for Source or Smooth) obtained in step 1e.
b) Let's consider nodes s of S and t of T such as $(lon_t, lat_t) = (lon_s, lat_s)$.
c) The wind components (u_s, v_s) are used to calculate the location (lon_h, lat_h) of the puff located at s, one hour in the future.
d) (lon_h, lat_h) is approximated to (lon_f, lat_f), where f is a node of grid S.
e) The reasoning is that the de-noised concentration c_f of f is the remainder of the decayed concentration c ($c \neq c_s$) that was located in s before the observation was captured. We assign to c_t the quotient of c_f and an arbitrary hourly decay rate factor in]0,1[.

Step 3 is a bespoke classification process.

a) From the two-dimensional map of concentrations that is the product of step 2, we create a list of nodes, sorted by decreasing concentrations.
b) Each node is associated with class *to_be_classified* (we use 9999 - any arbitrarily large number will do).
c) The first node N with the largest concentration and class *to_be_classified* is assigned class 1.
d) The concentration value associated with each of the 8 nodes around N is tested against N's concentration e_N. All the nodes of class *to_be_classified* with a smaller concentration value than e_N are assigned to class *uninteresting* (typically 0). The process is repeated recursively until no more nodes are found.
e) The process is repeated from step (c) where the assigned class is being incremented – the second highest largest concentration gets class 2, the third gets class 3, and so on. The process is repeated an arbitrary number of times. In practice, we don't find it useful to go over 50 sources within a dataset. The multiscale approach gives more interesting results (as discussed in section 1.2.3).

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

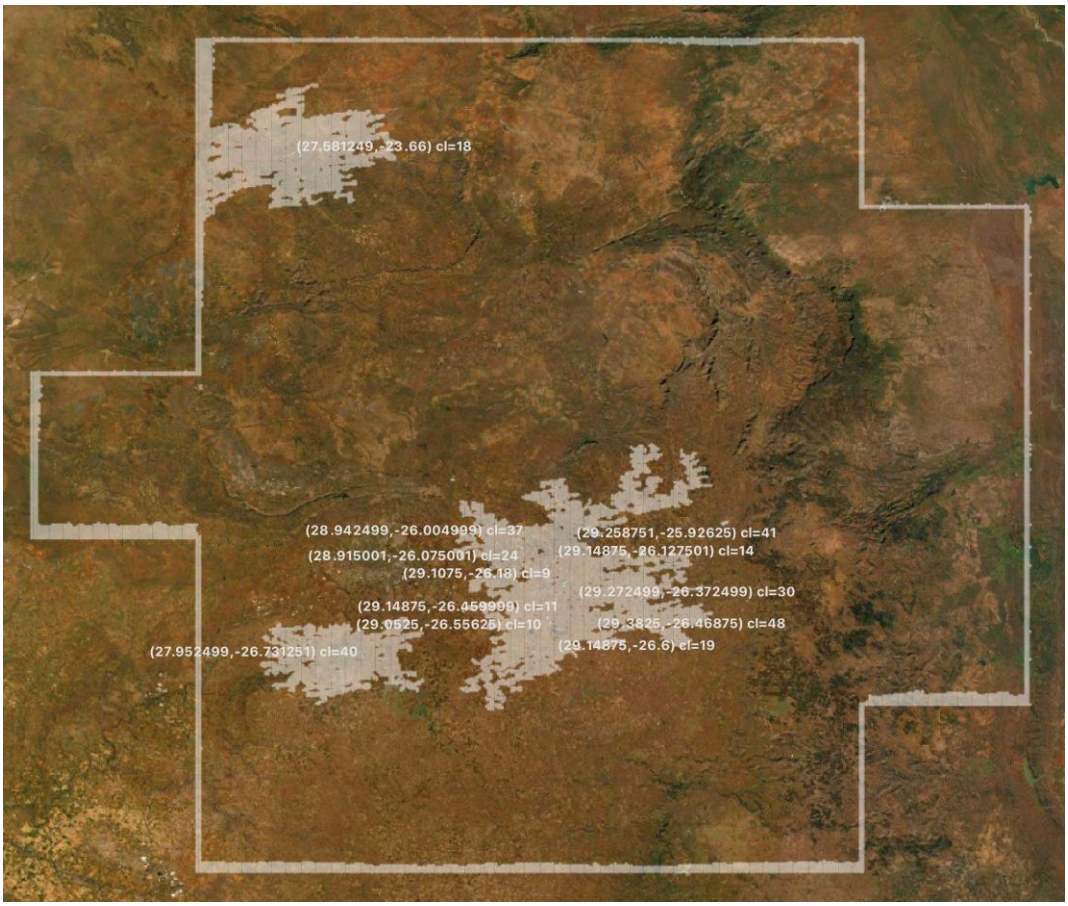

Figure 2 – Result of the full processing on the entire area defined in 1.1.3.

1.2.3 Multi-scale approach

The processing of a large area will tend to highlight the strongest areas of emission or accumulation of NO_2 while hiding weaker areas. This phenomena is demonstrated through Figure 3, with a specific focus on the area covering the northern part of Johannesburg and Pretoria (North of Johannesburg).

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Figure 3 – Top left: the area Pretoria ~ Johannesburg (blue frame) is super-imposed on the original area.

Bottom left: close up showing NO_2 average concentrations related to (a) in the upper left, northern Johannesburg invisible on the entire map (Figure 1), and (b) NO_2 concentrations related to Matla and Secunda power stations as they were seen on Figure 1. Right: The full processing described in 1.2.2 specifically applied to the Pretoria ~ Johannesburg area, showing 11 potential sources projected on the Google Terrain Hybrid basemap in QGIS.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

2 Evaluation of the results

2.1 Qualitative interest of the retro-advection (step 2)

It is reasonable to question the added-value of the coarse retro-advection process presented in step 2 of section 1.2.2. We believe it can be helpful despite its gross shortcomings.

The rationale is demonstrated in Figure 4, where the impact of step 2 is demonstrated on two urban areas separated by 15 kilometres. Step 2 splits the mass of the northern urban area in two roughly equivalent masses, that may be attributed to each urban area respectively North and South of the river that flows from West to South-East. The paler block right of the pixel labeled "class 31" can also be associated with another area of activity consisting of 3 towns located 3 km from one another.

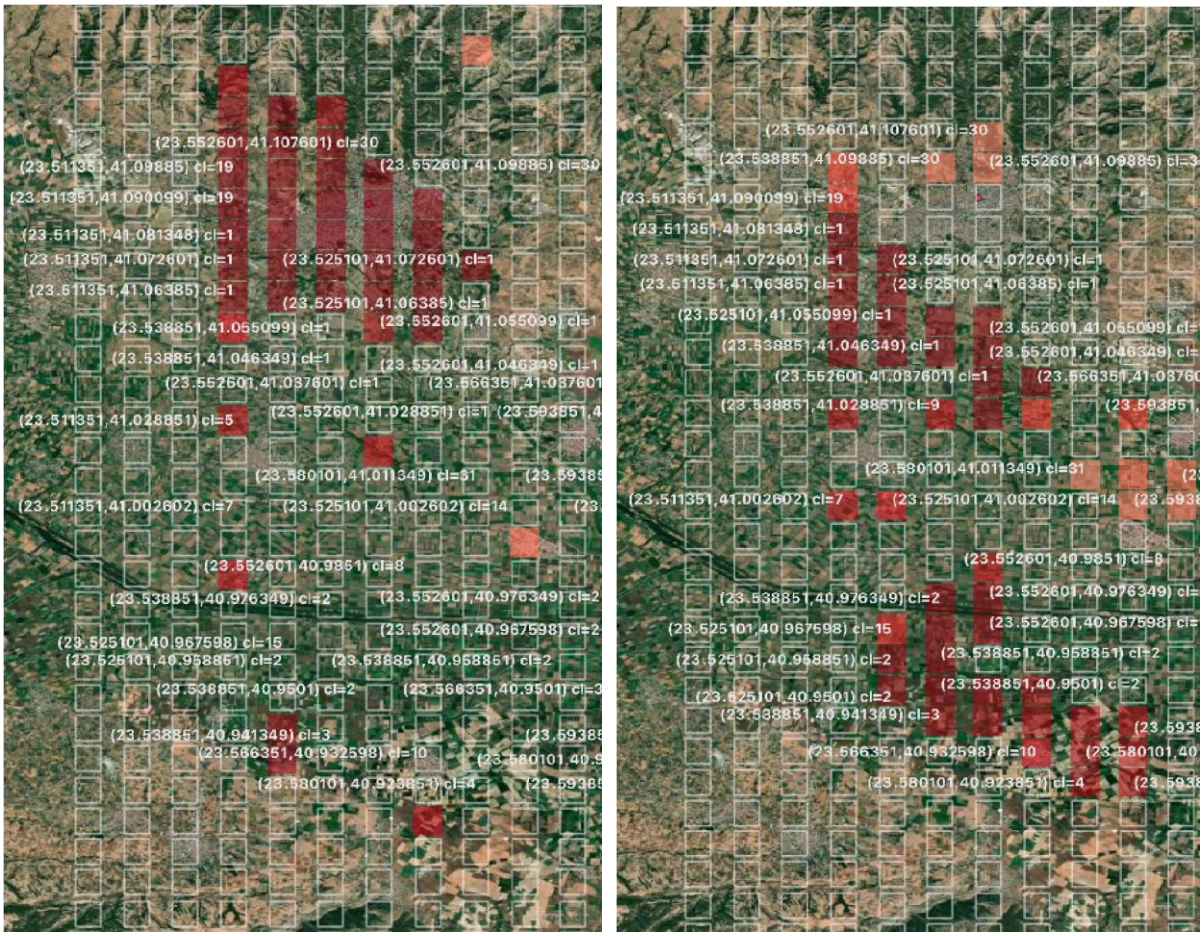

Figure 4 – Left screenshot is without retro-advection. The northern agglomeration gets the spotlight.

Right screenshot is with retro-advection. Both the northern and southern agglomerations are identified.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

2.2 Qualitative interest of the retro-advection (step 2)

Evaluation of predictions requires a reference to compare the predictions with. We built such a reference by visual inspection of an arbitrarily selected area within the boundaries of the dataset described in 1.1.3. For the purpose of comparison, this reference was built on the same grid as the dataset (Figure 5).

The reference area spans 24 x 22 pixels, that is approximately 25 km in longitude and 20 km in latitude. The visual inspection was performed with QGIS and the ESRI terrain basemap (http://www.arcgis.com/home/ item.html?id=10df2279f9684e4a9f6a7f08febac2a9. World Imagery. Last update: Feb 25, 2021). Pixels whose more than half of the surface contains coal power stations, open mining areas, or disposal areas where tagged as "sources", and the rest as "not sources". Out of the 528 pixels, 71 were tagged as sources and 457 as not sources. The resulting prevalence of sources is 13.4%.

The comparison aimed at creating a confusion matrix for the purpose of evaluating the performance of the predictor in correctly detecting sources and clean areas. Several confusion matrices were created, each allowing to evaluate the predictor with a given tolerance in order to account for the uncertainties of the current processing.

As a result we identified the true and false positives as per the following algorithm, and calculated the false negative and true negative as the complements. Given tolerance T in number of pixels, (i,j) a pixel, predictor(i,j) the map of predicted sources, and reference the map obtained through visual inspection, the algorithm is:

```
IF predictor(i,j) == true
    IF there is at least 1 pixel tagged as source in area [(i-T, j-T), (i+T, j+T)] of reference
        result    =    ”True
Positive”              ELSE
result = “False Positive”
```

The results are provided in table 1, for each level of tolerance tested. On a per pixel basis, the predictor is not helpful for the probability of actually having a source under the same pixel is only 22%. Performance is much better from tolerance +/- 3 pixels, for the same probability is above 80%, and the probability of not having a source when no source is detected in above 90%.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Table 1 – Performance of the predictor (including the naive retro-advection) according to different levels of tolerance

	Tolerance (+/-)	0		3		4		5		6	
Predictor says:	**'source'**	TP=6	FP=21	22	5	24	3	25	2	26	1
	'no source'	FN=65	TN=436	49	452	47	454	46	455	45	456
	Accuracy	83.7%		89.8%		90.5%		90.9%		91.3%	
	p(TP), p(FP)	1.1%	4.0%	4.2%	0.9%	4.5%	0.6%	4.7%	0.4%	4.9%	0.2%
	p(FN), p(TN)	12.3%	82.6%	9.3%	85.6%	8.9%	86.0%	8.7%	86.2%	8.5%	86.4%
Probability of having a source around given positive prediction		**22.2%**		**81.5%**		**88.9%**		**92.6%**		**96.3%**	
Probability of NOT having a source around given negative prediction		**87.0%**		**90.2%**		**90.6%**		**90.8%**		**91.0%**	

Table 2 – Performance of the predictor (without retro-advection) according to different levels of tolerance.

	Tolerance (+/-)	0		3		4		5		6	
Predictor says:	**'source'**	TP=3	FP=13	12	4	15	1	15	1	15	1
	'no source'	FN=68	TN=444	59	453	56	456	56	456	56	456
	Accuracy	84.7%		88.1%		89.2%		89.2%		89.2%	
	p(TP), p(FP)	0.6%	2.5%	2.3%	0.8%	2.8%	0.2%	2.8%	0.2%	2.8%	0.2%
	p(FN), p(TN)	12.9%	84.1%	11.2%	85.8%	10.6%	86.4%	10.6%	86.4%	10.6%	86.4%
Probability of having a source around given positive prediction		**18.8%**		**75.0%**		**93.8%**		**93.8%**		**93.8%**	
Probability of NOT having a source around given negative prediction		**86.7%**		**88.5%**		**89.1%**		**89.1%**		**89.1%**	

Larger reference maps in different parts of the world would be needed to ascertain the performance of the processing chain.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Figure 5 – Result of the visual inspection of the reference area. 1 stand for sources, 0 for not sources.

The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.