

# **Air4EU**

**Air Quality Assessment for Europe: from local to continental scale**



6th Framework Programme- Policy oriented Research  
Priority 8.1 Topic 1.5 Task 2

## **Individual case study report: Paris Comparisons of data assimilation methods at urban scale**

Deliverable:	D7.1 Part xx
Dissemination level:	PU
Editor:	N-N.
Version:	x.x
Date:	Date
Contract:	503596

## LIST OF AIR4EU PARTNERS

<b>Partic. no.</b>	<b>Participant name</b>	<b>Participant short name</b>	<b>Country</b>
1	Netherlands Research Organisation	TNO	NL
2	Norsk Institut for Luftforskning	NILU	NO
3	Aristotle University Thessaloniki	AUT	GR
4	University of Stuttgart	IER	DE
5	University of Hertfordshire	UH	UK
6	Universidade de Aveiro	UAVR	PT
7	AIRPARIF	AIRPARIF	FR
8	Agenzia per i Trasporti Autoferrotramviari e la Mobilità del Comune di Roma S.p.A.	ATAC	IT
9	Environment Agency	EA	UK
10	City Development Authority of Prague	URM	CZ
11	Enveco	ENVECO	GR
12	Gemeentewerken Rotterdam	GW	NL
13	Milieudienst Rijnmond	DCMR	NL
14	City of Oslo, Public Health Authority	OPHA	NO

## DOCUMENT HISTORY

Date:	Editor:	Version:	Text:
20/10/2006	AIRPARIF	1	
27/10/2006	AIRPARIF	1	Updates from Sam-Erik Walker comments

## LIST OF CONTRIBUTORS

Charlotte Songeur		
Olivier Sanchez		
Nadège Blond		

# Table of Content

<b>1</b>	<b>ABSTRACT .....</b>	<b>2</b>
<b>2</b>	<b>CASE STUDY DESCRIPTION .....</b>	<b>2</b>
2.1	BACKGROUND .....	2
2.1.1	<i>Monitoring Network</i> .....	2
2.1.2	<i>Integrated modelling systems</i> .....	4
2.2	AIM AND DESCRIPTION .....	7
2.3	RELEVANCE TO RECOMMENDATIONS IN AIR4EU .....	7
<b>3</b>	<b>METHODOLOGY .....</b>	<b>8</b>
3.1	PRESENTATION OF THE ASSIMILATION DATA TECHNIQUES IMPLEMENTED IN ESMERALDA IN THE FRAMEWORK OF AIR4EU .....	8
3.1.1	<i>Statistical Interpolation</i> .....	9
3.1.2	<i>Kriging</i> .....	11
3.2	EVALUATION METHODOLOGY .....	14
3.2.1	<i>Jackknife</i> .....	14
3.2.2	<i>Leave-one-out</i> .....	14
<b>4</b>	<b>RESULTS .....</b>	<b>15</b>
4.1	PRESENTATION OF THE STUDY RESULTS .....	15
4.1.1	<i>Comparison with the model output</i> .....	15
4.1.2	<i>Comparisons between the data assimilation methods</i> .....	18
4.1.3	<i>Example of the 20<sup>th</sup> June 2005</i> .....	19
4.1.4	<i>Method limits</i> .....	19
4.2	UNCERTAINTY ASSESSMENT .....	19
<b>5</b>	<b>CONCLUSION AND DISCUSSION .....</b>	<b>20</b>
5.1	ASSESSMENT OF THE CASE STUDY .....	20
5.2	IMPROVEMENTS IN ASSESSMENT DERIVED FROM THE CASE STUDY .....	20
5.3	RECOMMENDATIONS RESULTING FROM THE CASE STUDY .....	20
<b>6</b>	<b>BIBLIOGRAPHY: .....</b>	<b>21</b>
<b>7</b>	<b>ANNEX I .....</b>	<b>22</b>
7.1	OZONE CONCENTRATION MAPS FOR THE 20 <sup>TH</sup> JUNE 2005: .....	22
7.2	OZONE CONCENTRATION MAPS FOR THE 14 <sup>TH</sup> JUNE 2006: .....	23
7.3	STANDARD DEVIATION MAPS, LINKED WITH ANALYSIS UNCERTAINTIES: .....	23
7.4	OZONE CONCENTRATION CORRECTIONS BY DATA ASSIMILATION: .....	24

# 1 Abstract

Air quality assessment can be performed using:

- Monitoring network covering the region of concern.
- An integrated mesoscale modelling system.
- Assimilation techniques, which couple observations and simulation results.

Within the frame of the AIR4EU project, AIRPARIF chose to focus on data assimilation methods which allow improving air pollutant concentration estimation and reducing uncertainties from dispersion models.

Two data assimilation methods are currently used in AIRPARIF, that is anisotropic statistical interpolation and innovation intrinsically Kriging. These methods are applied to the POLLUX system outputs for daily air quality index mapping. In the Paris case study, we aimed for two objectives: first, to extend the assimilation processes to the new ESMERALDA forecast system and then to get an evaluation of the 2 techniques for ozone.

After a short description of these methods and their adaptation to the ESMERALDA system, we present their evaluation against measurements collected from the monitoring network. The total error made on ozone concentrations is then estimated and compared for the two methods.

Standard deviation maps have been also plotted. Those maps allowed us to visualize the spatial distribution of ozone concentration uncertainties for the two data assimilation methods.

## 2 Case study description

### 2.1 Background

Air quality assessment in Ile-de-France region is performed using different tools:

- A monitoring network covering all the region.
- An inter-regional emission inventory elaborated jointly with the Esmeralda project partners (Haute-Normandie, Nord Pas-de-Calais, Picardie, Centre and Champagne-Ardenne regions).
- 2 integrated mesoscale modelling systems POLLUX and ESMERALDA.

#### 2.1.1 Monitoring Network

The AIRPARIF monitoring network includes 46 automatic air quality stations in Ile-de-France. They are classified as:

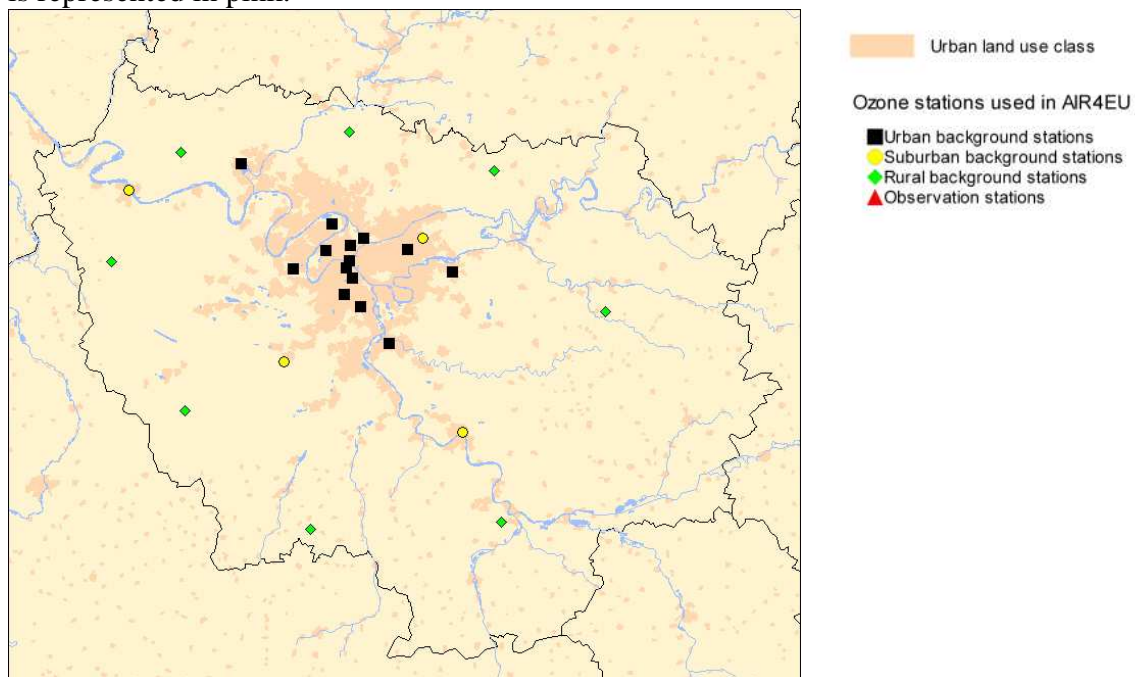
- “Urban background” stations (25).
- “Suburban background” stations (6) located in the suburbs of Paris agglomeration.
- “Regional background” (8) stations located outside the urban area, almost in the countryside.
- “Traffic stations” (6) located directly on streets (few meters from traffic) with heavy traffic conditions.
- 1 observation station situated on Eiffel Tower at 300m.

Those stations monitor pollutants such as CO, SO<sub>2</sub>, NO<sub>x</sub>, PM<sub>10</sub>, PM<sub>2.5</sub>, BTEX and O<sub>3</sub> depending on their classification and location.

In the frame of the AIR4EU project, we focused on Ozone and used 26 monitoring stations classified as follow:

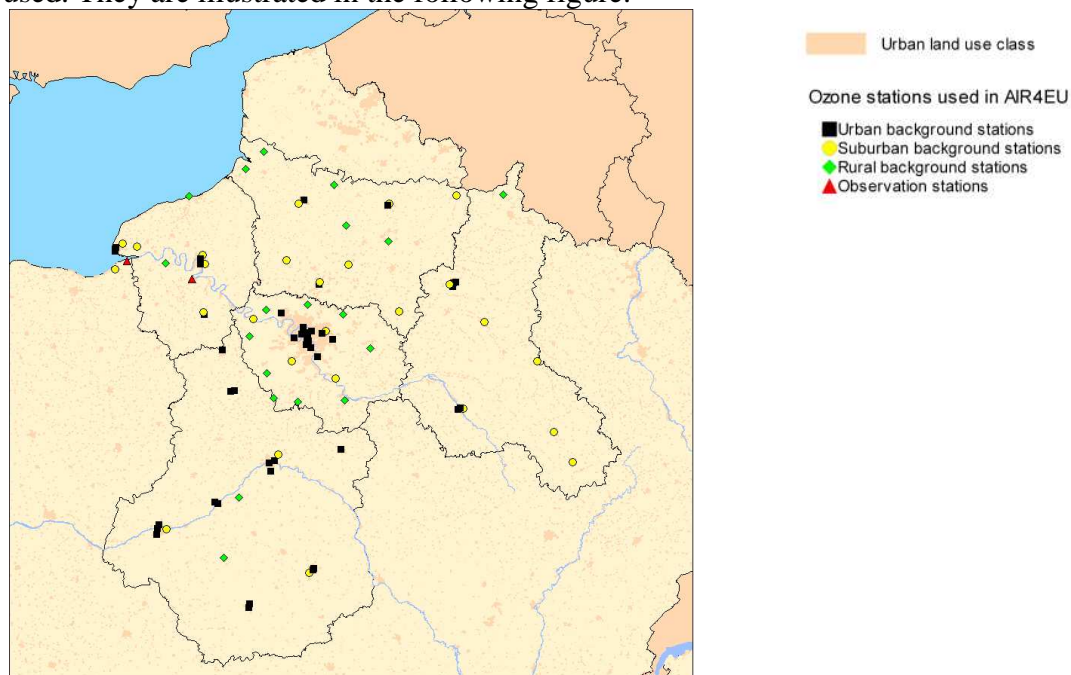
- 14 “urban background” stations in the agglomeration of Paris.
- 4 “suburban background” stations
- 8 rural background stations

The map in Figure 1 shows the location of the air quality monitoring network in Ile-de-France/Paris used in the framework of AIR4EU. The agglomeration of Paris spatial coverage is represented in pink.



**Figure 1: Ozone monitoring stations in Ile-de-France used in the AIR4EU project**

In the frame of the Esmeralda project, monitoring stations from region partners have been also used. They are illustrated in the following figure:



**Figure 2: Ozone monitoring stations in Ile-de-France used in the AIR4EU project**

## 2.1.2 Integrated modelling systems

### 2.1.2.1 Emission Inventory

Within the frame of the Esmeralda project, the region partners shared also their experience and knowledge of their home region in order to elaborate a common emission inventory.

On the following maps, the gridded inter-regional NMVOC and NO<sub>x</sub> annual emissions are presented.

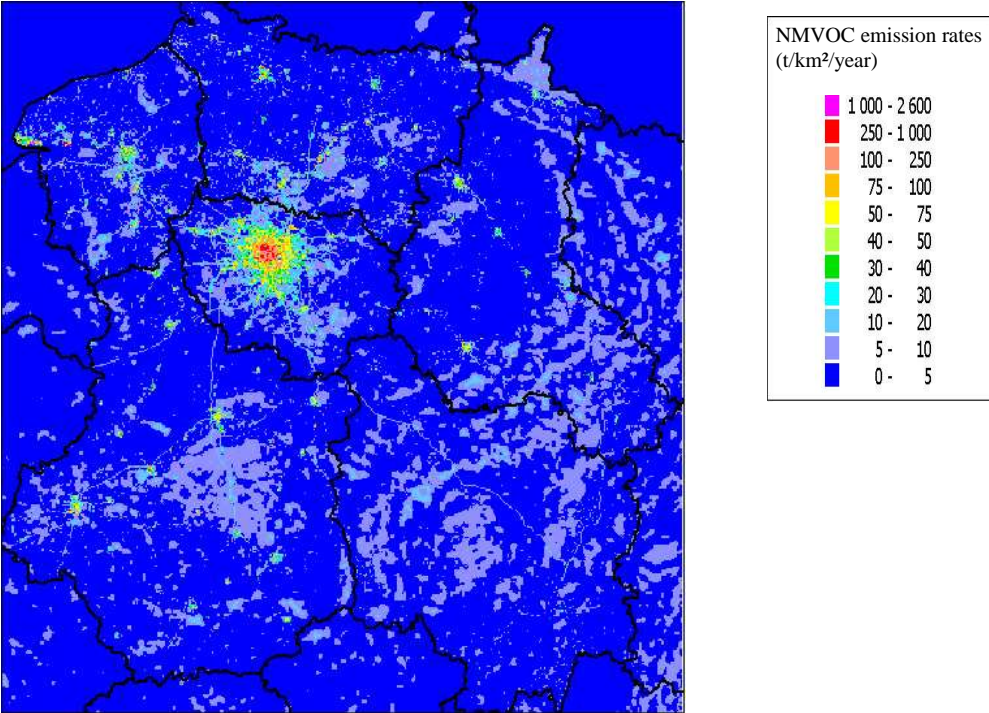


Figure 3: NMVOCs Emission rates : t/km2/year

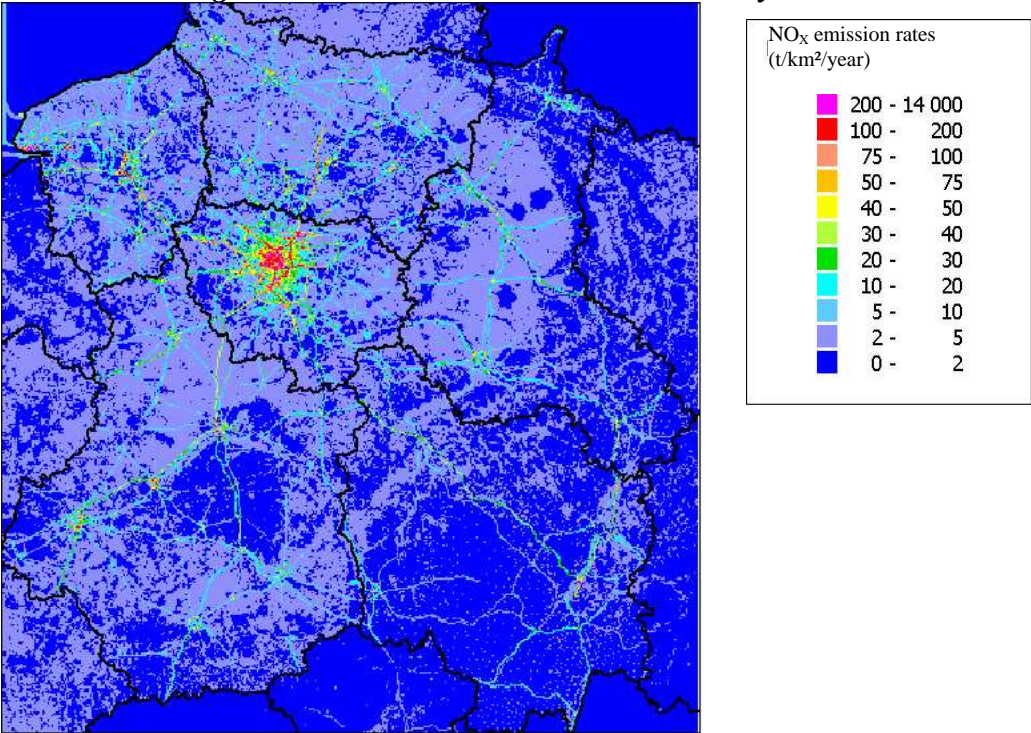


Figure 4: NO<sub>x</sub> Emission rates: t/km2/year

Ile-de-France is of special interest among these 6 regions because of its particular situation. Actually, Ile-de-France region spatial coverage is 12,000 km<sup>2</sup> (2% of the whole country) with 12 millions inhabitants (20% of the French population). Within this region, Paris covers about 105 km<sup>2</sup> and has 2.144 millions inhabitants. As a consequence, air pollution in Paris is mainly related to road traffic emissions. In order to better estimate those emissions, a real-time evaluation system has been developed in the frame of the HEAVEN European project. “Real time” traffic data are collected by the system and used in a traffic model in order to provide information on the whole traffic network, that is about 35000 portion of roads. This information is updated every hour. “Real time” traffic emissions are then calculated using the COPERT III methodology and a detailed description of the fleet composition, specific of the Ile-de-France area.

### **2.1.2.2 Air quality forecast systems**

The POLLUX and Esmeralda forecast systems use the chemical transport model CHIMERE developed by LMD/IPSL (Laboratoire de Météorologie Dynamique/Institut Pierre Simon Laplace). Emissions are from the inter-regional inventory and from the outputs of the HEAVEN system.

The POLLUX system is operational since 2001 and uses as meteorological forcing, outputs from ARPEGE (run by Météo-France). The system is run daily, on 2 domains:

- A 6km resolution domain covering the Ile-de-France region.
- A 3km resolution domain covering the Paris agglomeration.

Boundary conditions are calculated using back-trajectories from Météo-France following the back-plume algorithm from Vautard and al.

The Esmeralda system has its own meteorological interface. The mesoscale meteorological model MM5 is used as meteorological processor for CHIMERE. It is run daily and provides meteorological forecasts from Day-1 to Day+2. MM5 is forced with GFS (Global Forecast System) products from NCEP (National Centers for Environmental Prediction). CHIMERE is run over 7 domains:

- A 9km resolution inter-regional domain covering the 6 Esmeralda partner regions
- 6-3km resolution domains over the 6 regions.

The mother domain uses as boundary conditions outputs from PREV’AIR, the national air quality forecast system: this system is run by INERIS over Europe with a 0.5° spatial resolution.

The 6 internal domains are forced by the outputs from the mother domain.

### 2.1.2.3 Data assimilation

In addition to these systems, AIRPARIF performs data assimilation in order to inform the Ile-de-France population about the current air quality situation. This has been set-up under the frame of the POLLUX system:

- Air quality index maps are produced every day for the previous day: it deals with ozone, nitrogen dioxide, sulphur dioxide and particulate matter (PM10). Data assimilation is performed using outputs from the forecast system for the day before.
- In an operational way, outputs from the forecast system are corrected with the pollutant measurements collected from the monitoring network every day at 16:00 (LT) to produce air quality index maps.
- Ozone and nitrogen dioxide maximum concentration patterns are also produced daily with a different data assimilation system.

Finally, the results of the data assimilation processes published on the AIRPARIF web site are air quality index maps or concentration maps based on pollutant maximum concentrations for ozone, nitrogen dioxide and sulphur dioxide and on pollutant mean concentrations for particulate matter (PM10).

The data assimilation techniques used in AIRPARIF are sequential and various:

- To produce the concentration patterns, statistical interpolation is performed on ozone concentrations and Innovation Kriging on nitrogen dioxide concentrations.
- To produce the air quality index patterns, the co-kriging technique is performed.

For the moment, the assimilated concentration fields are not used as initial conditions for the forecast systems, as the concentration evolution in the vertical direction can't be easily estimated.

The aim of data assimilation implementation in AIRPARIF was to improve air quality assessment and information to public, as assimilation allows reducing uncertainties inherent to measurements and modelling (Deliverable 4.2).

#### Data assimilation for air quality index maps

The air quality index maps are produced using the ISATIS© geostatistical software. Co-kriging is performed on pollutant concentrations with model outputs as co-factor.

Like all geostatistical techniques, co-kriging uses spatial correlations to interpolate data. This assimilation technique is used with fixed variograms (based on annual mean concentrations), that is to say they are not updated with the current measurement data.

#### Data assimilation for ozone and nitrogen dioxide concentrations:

The programs used for ozone and nitrogen dioxide maps have been developed within the frame of the PHD of Nadège Blond (from LMD) [1] “*Assimilation de données photochimiques et prévision de la pollution troposphérique*”. The author has applied and compared several sequential data assimilation techniques for air pollution cartography over the Ile-de-France region: In particular, statistical interpolation using innovative method to model covariance matrices and innovation kriging have been applied to Ile-de-France. Within this work, representativeness errors of measurement data are evaluated and taken into account in the calculations.

Further more, variogram modelling is performed at every time step and uncertainties are calculated.

## **2.2 Aim and description**

Even if several assimilation techniques are applied daily in AIRPARIF to produce concentration or air quality index mapping, for annual and inter-annual calculations, only measurements are used to estimate compliance with standards and national or European thresholds. Moreover, those techniques have not been adapted to the Esmeralda system which represents a strong improvement of the previous forecast system.

As a result, AIRPARIF has set-up a data assimilation project which objectives are the following:

- First, to harmonize and extend the assimilation processes in the daily air quality map production. In the harmonisation process, the best assimilation technique should be selected and adapted to the pollutant nature.
- Secondly, to adapt those techniques to the Esmeralda system and finally to produce air pollution maps and associated uncertainties over the 6 different regions of the project.
- Thirdly, in the yearly air quality assessment, to elaborate the calculation process in order to produce the annual maps combining monitoring and modelling.
- And finally, to produce uncertainty mapping of the air pollutant patterns, based on the selected assimilation process.

As a consequence, the Paris case study is part of that project as AIRPARIF could benefit the scientist feed-back on the evaluation of the assimilation techniques.

As a result, the Paris case study should lead to a better understanding of the assimilation techniques:

- Evaluate assimilation parameters in the pollution mapping process.
- Evaluate assimilation techniques according to the pollutant type and dispersion behaviour.
- Analyse uncertainty mapping.

Within the frame of the AIR4EU Paris case study, ozone assimilation has been tested and evaluated for Statistical Interpolation and Innovation Kriging. The selected period of interest is summer 2005 for which the outputs of the Esmeralda system and the monitoring data from the 90 stations presented in figure 2 have been extracted.

## **2.3 Relevance to recommendations in AIR4EU**

The main objective of data assimilation processes is to reduce uncertainties in modelled concentrations. As a consequence, the Paris case study investigates several recommendations relatives with model uncertainties (Ref D4.1-§8). The two AIRPARIF forecast systems have been evaluated against monitoring data for a long time [2]. Those evaluations are based on statistical indexes (Ref D4.1-§8.2).

The results of those validation processes have shown some good agreement of predicted ozone peaks with measurements but uncertainties arose when looking at hourly values. Nevertheless, those uncertainties are usual to all dispersion or chemical transport models

- Ozone and nitrogen dioxide peaks occur when wind speeds are weak. This is often the case in the Paris area for ozone and nitrogen dioxide. In particular, the latter pollutant peaks occur during stable atmospheric conditions with low boundary layer height in the early morning or in the evening during or after the rush hours. Stable and low wind speed condition modelling is still of great challenge to forecast and especially in urban areas like Paris where urban meteorology is difficult to model.
- Uncertainties in emission data are inherent to the manipulated kind of data: actually, emission methodologies use strong assumptions on emission processes (Ref D4.1-§4.6) as they are not always well understood and temporal profiles are usually roughly evaluated. (Ref D4.1-§4.5).
- Speciation profiles used for aggregated species, re-aggregation and the associated chemical schemes are also rough estimates of the atmospheric chemical processes in dispersion models. Turbulent vertical transport is also factors that are quite uncertain in dispersion modelling.

This is important for modellers to keep in mind all that kind of uncertainty sources. This is also those constraints that point-out the data assimilation usefulness for air quality assessment at any time resolution.

As a consequence, the Paris case study investigated several recommendations relevant with the “Integration and data assimilation for assessing air pollution at urban/agglomerate scales” paragraph. These include:

- Model uncertainties that has been evaluated through model bias estimation and error covariance matrix “modelling” (Ref D4.1-§9.1.1).
- Representativeness errors that have been evaluated through an observation bias (Ref D4.1-§9.1.2).
- Assimilation assumptions relevant with the selected data assimilation technique (Ref D4.1-§9.2)

### 3 Methodology

#### 3.1 Presentation of the assimilation data techniques implemented in ESMERALDA in the framework of AIR4EU

In order to use model outputs with observations, we use two methods to analyse ozone and nitrogen dioxide concentrations in POLLUX system (for Ile-de-France):

- Statistical Interpolation (for O3)
- Kriging (for NO2)

Some definitions:

$Z^t$  average real state

$Z^b$  model output and  $\epsilon^b = Z^b - Z^t$  the model error

$Y^\circ$  observations

$H$  is the observation operator representing the interpolation for grid cells to the measure points

$\epsilon^\circ = Y^\circ - HZ^t$  is the observation error vector (the sum of an instrument error and an representativeness error).

$Z^a$  analysis vector. This is the resulting vector after the model output correction.

### 3.1.1 Statistical Interpolation

General formulation :

$$Z^a = Z^b + K[Y^o - HZ^b] \quad (1)$$

$K$  is the weight matrix of the linear interpolation. Weights are calculated in order to get the best possible analyses, that is by minimisation of error variance:

$$K = BH^T [HBH^T + R]^{-1} \quad (2)$$

To calculate  $K$ , the model error covariance matrix ( $B$ ) and the observation error covariance matrix ( $R$ ) have to be calculated.

The general expression of the analysis can be finally written as follows:

$$Z^a(s_a) = Z^b(s_a) + \sum_{k=1}^p w_k(s_a)(Y^o(s_k) - Z^b(s_k)) \quad (3)$$

If the observations and the simulations (model output) are not biased, the weights are solution of the following system:

$$\forall k = 1, \dots, p = \text{number of stations} \quad \sum_{l=1}^p w_l(s_a)[B_{k,l} + R_{k,k}\delta_{k,l}] = B_{a,k} \quad (4)$$

$s_a$  grid node,  $s_k$  and  $s_l$  measure points.

We also need to estimate the model and observation biases to subtract them from  $Z^b$  and  $HZ^b$  in the equation (1). The problem is that we don't know these elements (error covariance matrices and bias). They have to be estimated.

Estimation of the global bias (addition of model bias and observation bias) :

The main difficulty in estimating the model bias and the observation bias is to evaluate what is due to the model and what is due to the observation.

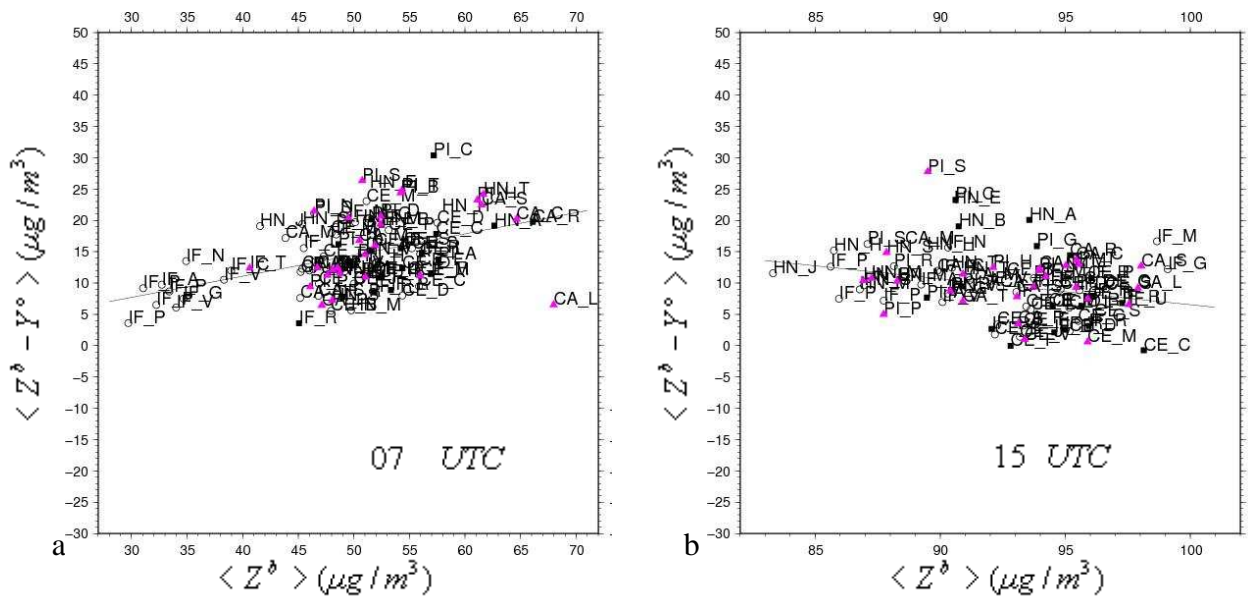
The *average concentration simulated by the model* is used as a criterion and the model bias is supposed to depend on it linearly (figure 5). An estimation of *the model bias* at the measure point can therefore be performed and based on this relationship, *the model bias* on grid nodes can be also evaluated: it is actually evaluated by the linear relation.

The observation bias (or more precisely representativeness error) is deduced from the following:

$$\begin{aligned} E[\mathcal{E}^o(s_k)] - E[\mathcal{E}^b(s_k)] &= E[Y^o(s_k) - Z^t(s_k)] - E[Z^b(s_k) - Z^t(s_k)] \\ &= E[Y^o(s_k) - Z^b(s_k)] \end{aligned} \quad (5.a)$$

$$\Rightarrow E[\mathcal{E}^o(s_k)] = E[Y^o(s_k) - Z^b(s_k)] + E[\mathcal{E}^b(s_k)] \quad (5.b)$$

The observations  $Y^o$ , the simulated concentrations  $Z^b$  and the observation operator  $H$  have been evaluated.  $E[\varepsilon^b(s_k)]$  has just been calculated, and *the observation bias* can now be estimated using equation 5.b.

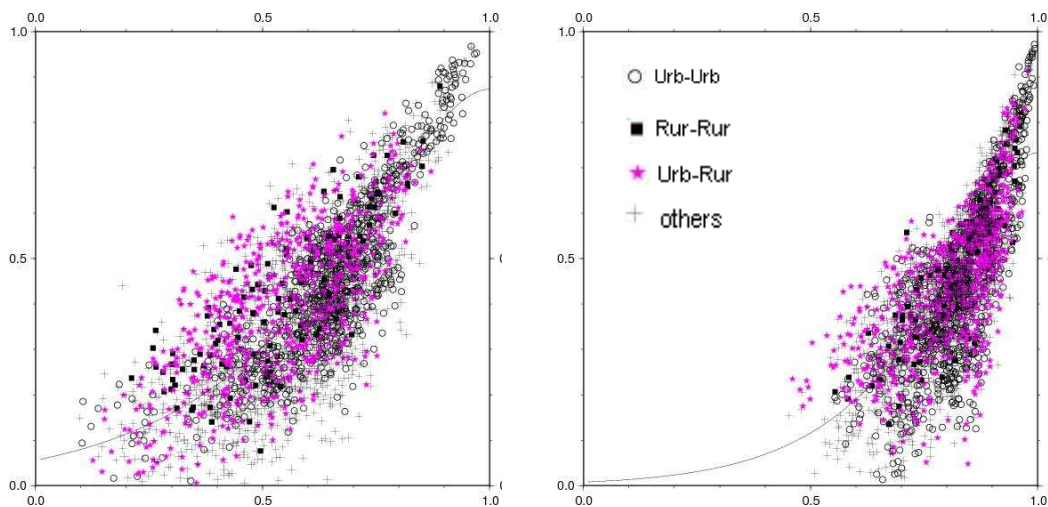


**Figure 5: Bias at 07h and 15h UTC calculated over the summer 2005 for Esmeralda area. The empty rounds represent the urban stations, the black squares the rural stations and the mauve triangles all the others stations.**

Currently, the observations in the programs are supposed not to be biased.

Estimation of error covariance matrices:

A heterogeneous and anisotropic correlation model is considered. Innovation correlations are plotted against the observed ozone concentration correlations. As a consequence, the correlogram is heterogeneous and anisotropic (figure 2). *The observed ozone correlations* are used as criteria to estimate the model error covariances.



**Figure 6: Innovation correlations against observed concentration correlations (the category “others” corresponds to the correlations with all the others stations). Urb means urban station, Rur, rural station**

The  $B_{a,k}$  terms are estimated by using *the simulated ozone concentration correlations* because we don't have observations at all grid nodes. The  $B_{k,l}$  are also estimated by the modelled concentration correlations in order to avoid discontinuity problems.

A correlation model is adjusted with the following function:

$$f_c(c) = \left(1 + \frac{|1-c|}{L}\right) \exp\left(-\frac{|1-c|}{L}\right), \text{ with } -1 \leq c \leq 1 \text{ and } L \geq 0 \quad (6)$$

$B$ , is now obtained by multiplying the correlation model,  $f_c$ , by the innovation standard deviations to obtain covariances.

The variance of the model error is estimated with the innovation variance and the value of the correlation model when the model concentration correlation is equal to 1.

The weights of the interpolation and the analysis vector can now be estimated. The analysis error variance is obtained by calculating:

$$\sigma_a^2(s_a) = B_{a,a} - \sum_{k=1}^p w_k(s_a) B_{a,k} \quad (7)$$

which is equivalent to the data assimilation uncertainties.

### 3.1.2 Kriging

The general formulation for Kriging is the following:

$$Z^a(s_a) = \sum_{k=1}^p w_k(s_a) Y^o(s_k) \quad (8)$$

In this method, the observations are used to estimate the values at any node of the grid. Instead of observations, in the method implemented in AIRPARIF, Innovations are used in order to use model outputs and observations: innovations are differences between observed and modelled concentrations.

Applied to innovations, the general formulation becomes:

$$X^a(s_a) = \sum_{k=1}^p w_k(s_a) X^o(s_k) \quad (9)$$

$$Z^a(s_a) = Z^b(s_k) + \sum_{k=1}^p w_k(s_a) X^o(s_k) \quad (10)$$

In this method, the variogram is used to model spatial variations. Errors of predicted values, estimated from their spatial distributions are then minimised.

Looking at equations, Kriging and Statistical Interpolation formulations are very close. The only difference is in the estimation of the interpolation weights. In the Kriging method, they are calculated with statistics on *spatial* data instead of *spatial and temporal* data in the Statistic Interpolation method.

In the Paris test case, an Intrinsically Kriging is used, which supposes that  $X^t(s) = Z^t(s) - Z^b(s)$  is intrinsically stationary instead of only stationary in ordinary Kriging. This is the preferred method as it avoids problems of infinite variances:

$$\forall (s_1, s_2), E[X^t(s_1) - X^t(s_2)] = 0 \quad (11.a)$$

$$\forall (s_1, s_2), \frac{1}{2} \text{Var}[X^t(s_1) - X^t(s_2)] = \gamma(s_1, s_2) \quad (11.b)$$

$\gamma(s_1, s_2)$  is the *semi-variance* of differences between two  $X^t$  values. Those differences are called increments. The method consists in supposing that those increments are stationary of order 2 (instead of the values for the ordinary Kriging).

The weights of the linear interpolation are solutions of the following linear system:

$$\forall k = 1, \dots, K, \sum_{l=1}^p w_l(s_a) [\gamma_{k,l} + R_{k,k} \delta_{k,l}] - \mu = \gamma_{k,a} \quad (12.a)$$

$$\sum_l w_l = 1 \quad (12.b)$$

$s_a$  grid node,  $s_k$  and  $s_l$  measure points and  $\mu$  Lagrange parameter.

Measure points separated by the same vector  $|s_k - s_l|$  are used to calculate the semi-variances. Classes of distance are defined and the semi-variances are calculated with every station  $k$  and  $l$  separated by the distance  $|s_k - s_l| = h_j$ . As observations are not numerous enough, a tolerance is considered:

$$N_j = \left\{ (s_k, s_l); \left| |s_k - s_l| - (2j-1) \frac{L}{2} \right| < \frac{L}{2} + \tau \right\} \quad (13.a)$$

$\tau$ , tolerance and  $L$  step of class.

$$h_j = \frac{1}{N_j} \sum_{N_j} |s_k - s_l| \quad (13.b)$$

Instead of plotting a correlogram, we plot a variogram considering the classes of spatial differences. The variances are calculated for the  $J$  classes:

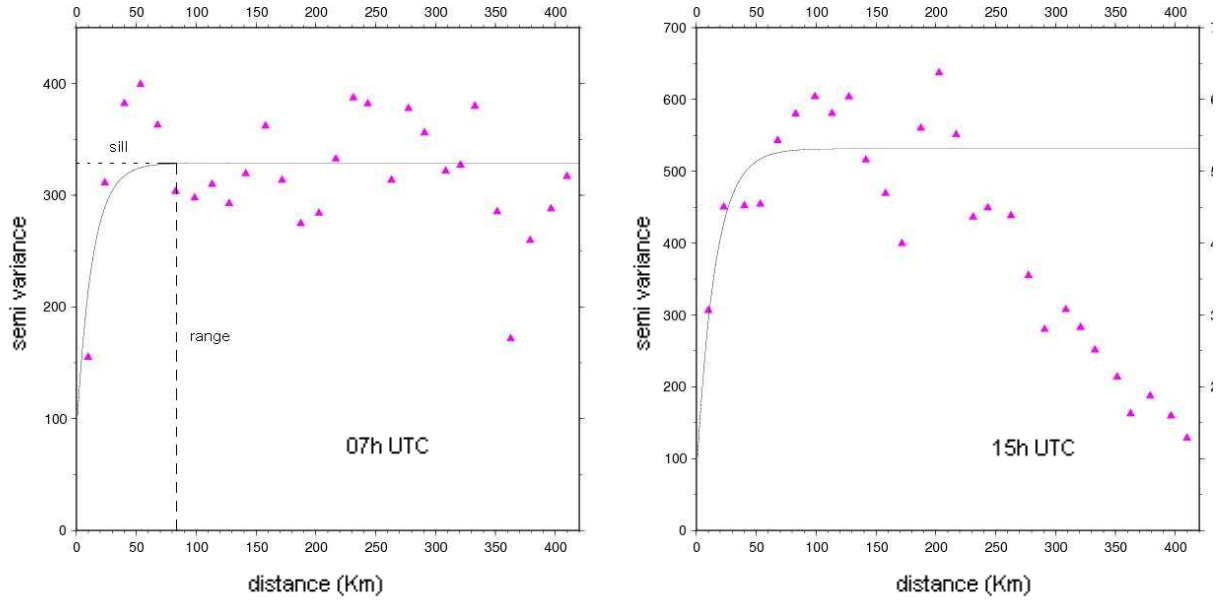
$$2\gamma^o(h_j) = \frac{1}{|N_j|} \sum_{N_j} (X^o(s_k) - X^o(s_l))^2 \quad j = 1, \dots, J \quad (14)$$

Whatever the class, these variances always contain the observation error variance:

$$2\gamma_{k,l}^o = E[(X^o(s_k) - X^o(s_l))^2] = 2\gamma_{k,l} + \frac{1}{2} (R_{k,k} + R_{l,l} - 2R_{k,l} \delta_{k,l}) \quad (15)$$

The following exponential function has been used:

$$\gamma(h) = c \left( 1 - \exp\left(-\frac{h}{a}\right) \right) + \sigma_0^2 \quad (16)$$



**Figure 7: Exponential variograms at 7h UTC and 15h UTC**

Some different functions can be used to fit variograms:

- Gaussian  $\gamma(h) = c \left( 1 - \exp\left(-\frac{h^2}{a^2}\right) \right) + \sigma_0^2$
- Cubic  $\gamma(h) = c_0 + c \left[ 7\left(\frac{h}{a}\right)^2 - \frac{35}{4}\left(\frac{h}{a}\right)^3 + \frac{7}{2}\left(\frac{h}{a}\right)^5 - \frac{3}{4}\left(\frac{h}{a}\right)^7 \right]$  if  $h < a$   
 $\gamma(h) = c_0 + c$  if  $h \geq a$

To estimate  $c$ ,  $a$  and  $\sigma_0^2$  coefficients, we use *Levenberg-Marquardt Method* (also called *Marquardt Method*) for non linear model.

The exponential function is able to model large range of variograms.

Finally, the weights of the interpolation and the analysis vector can be estimated. The analysis error variances can also be determined. The latter can be seen as kriging uncertainties:

$$\sigma_a^2(s_a) = \sum_{k=1}^p w_k(s_a) * \gamma_{a,k} - \gamma_{a,a} - \mu \quad (17)$$

## **3.2 Evaluation methodology**

To evaluate the analysis method efficiency, analysis error statistics are calculated for each hour: the statistical parameters used are standard deviation and bias. In order to do it objectively, Jackknife method and leave-one-out method are used. More over, maps of uncertainties can be plotted.

The methods have been tested for June 2005. The covariance error matrices, for the Statistical Interpolation, have been estimated using July, August and September 2005 data.

### **3.2.1 Jackknife**

This is a resampling method (comparable with Bootstrap method) used to estimate, as well as possible, statistical parameters when the sample is rather weak. Jackknife is used to minimise the bias and the variance of the classical standard deviation estimator  $\sigma$ . In our case, this method have been chosen because of its easy implementation.

An error sample for each measure point at any hour is obtained thanks to analysis results and observations. The Jackknife method is then applied to those samples in order to estimate the standard deviations.

An error sample for the maximum values of concentration is also calculated.

### **3.2.2 Leave-one-out**

To create the error samples, a leave-one-out method can also be used. Sequentially, observations collected at one measure point are omitted in the analysis process. These omitted observations are afterward statistically compared with the resulting analyzed field.

Leave-one-out allows checking the robustness of analysis method.

## 4 Results

### 4.1 Presentation of the study results

#### 4.1.1 Comparison with the model output

As for the analysis error samples, simulation error samples for ozone are calculated for each measure point at any hour during June 2005. The biases and standard deviations of each sample are then compared.

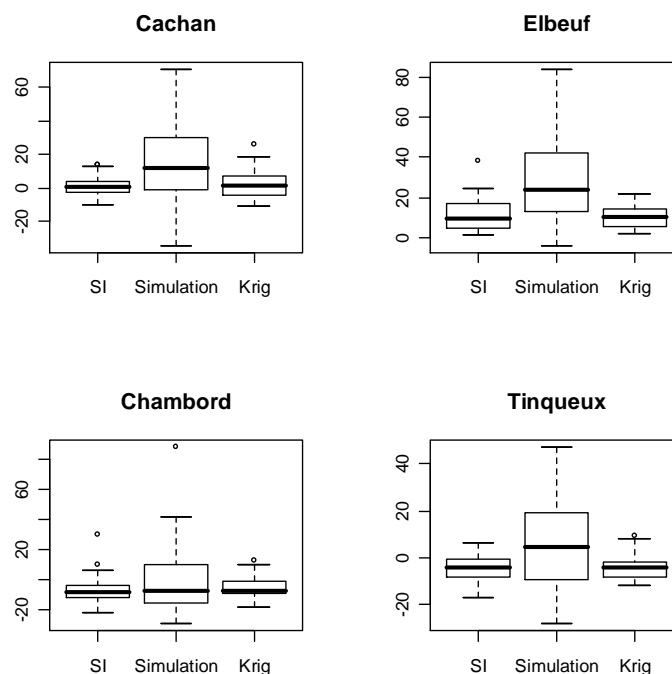
In the next table, biases and standard deviations for each monitoring station are compared for June 2005:

Station	Biais simulation	Biais IO	Biais IO (leave)	Biais Krigeage	SD simulation	SD IO	SD IO (leave)	SD Krigeage
Ile de France								
Aubervilliers	14.93	-0.85	-1.06	0.50	26.00	7.27	8.87	7.86
Cachan	17.68	1.06	1.21	2.63	26.31	6.36	7.58	8.80
Cergy_Pontoise	9.14	0.20	0.25	-1.10	23.59	9.24	12.14	7.65
Garches	18.64	3.00	4.21	3.48	30.55	10.22	14.38	10.23
Gennevilliers	10.93	-2.72	-3.12	-3.45	23.71	7.88	8.99	7.77
Lognes	10.50	2.78	3.43	-1.16	24.13	11.21	13.75	9.99
Mantes	13.98	1.49	1.91	1.86	30.11	12.65	16.09	7.92
Melun	10.38	-0.57	-0.81	0.12	21.80	9.55	12.88	7.87
Montgeron	20.07	8.05	9.78	6.79	24.33	9.86	11.99	10.07
Neully	28.13	13.91	15.78	14.10	31.39	13.91	15.73	11.86
Pans_01	13.96	-4.06	-4.66	-1.41	25.81	6.66	7.59	7.39
Pans_06	11.64	-6.12	-7.03	-3.94	24.96	7.78	8.91	10.19
Pans_13	19.79	2.49	2.76	4.82	22.58	5.42	6.15	6.34
Pans_18	18.82	2.18	2.47	3.76	24.29	6.96	8.03	8.96
Tremblay	17.14	5.49	7.69	2.67	25.10	18.62	26.09	16.15
Les Ulis	5.32	-5.88	-7.49	-4.49	20.85	7.94	10.07	8.31
Villennoble	13.75	2.37	3.00	0.36	20.07	5.29	6.57	5.82
Vitry	12.39	-3.96	-4.36	-1.55	21.06	6.92	7.61	5.14
Fremainville	8.46	-0.93	-1.19	-1.80	26.72	9.97	12.66	5.81
Montge	1.71	-6.13	-7.96	-5.03	16.14	8.80	11.36	7.60
Saints	2.38	-3.78	-4.85	-4.18	16.14	9.70	12.44	7.32
St Martin du Tertre	4.15	0.02	0.01	-4.45	24.95	10.70	14.19	7.26
Bois Herpin	0.96	-4.88	-6.44	-5.74	25.04	8.70	11.47	7.69
Fontainebleau	6.61	-3.00	-3.44	-2.11	23.90	8.61	9.85	5.21
Prunay	4.60	-2.42	-3.32	-2.90	28.80	9.88	13.64	9.19
Rambouillet	6.04	-1.40	-1.71	-1.48	26.25	9.58	11.85	8.23
Champagne-Ardennes								
Cernay	12.39	3.53	4.10	3.75	17.41	5.85	6.81	4.67
Chaumont	7.24	0.54	0.66	-1.73	18.24	6.07	7.50	3.52
La Tour	8.86	0.29	0.36	0.75	17.85	7.02	8.29	5.13
Langres	11.90	4.24	5.88	1.56	20.19	9.22	12.76	5.34
Maine	15.64	6.31	7.37	7.04	18.71	6.00	7.02	5.27
Munty	5.25	-3.50	-4.04	-2.70	16.90	5.40	6.23	4.61
Revin	6.88	3.05	4.17	1.12	14.81	9.76	13.35	6.92
Saint-Dizier	10.38	2.99	3.49	0.45	13.77	9.25	11.13	5.96
Saint-Memmie	6.19	-0.93	-1.07	-0.86	18.49	8.63	9.95	4.72
Saint-Parres	7.50	-0.39	-0.44	-0.70	20.62	8.55	10.05	7.45
Sainte-Savine	6.79	-1.45	-1.67	-1.37	16.34	5.42	6.31	3.87
Tingieux	4.14	-4.13	-4.72	-3.97	17.99	5.99	6.86	5.52
Centre								
Blais centre	-4.10	-5.61	-6.33	-2.97	13.22	5.46	6.18	3.91
Blais nord	-2.36	-9.03	-10.21	-5.45	18.55	8.77	9.95	6.27
Bouges sud	9.57	2.26	2.65	0.90	20.33	8.27	9.65	6.04
Chambard	-0.46	-5.93	-6.41	-4.58	25.53	10.65	11.53	6.65
Chateauroux_Sud	2.68	-3.67	-4.39	-2.19	20.76	6.12	7.30	5.11
Deols	2.68	-3.74	-4.39	-2.12	19.10	7.25	8.51	4.04
Dreux_Nord	12.96	2.26	2.60	1.27	19.08	10.77	12.34	8.52
Fulbert	8.63	-0.74	-0.83	-2.92	25.19	8.61	9.71	7.41
Gibonces	11.79	4.93	5.75	3.02	20.91	8.42	9.82	3.95
Jardin_Botanique	7.96	0.21	0.24	0.45	23.69	8.67	10.02	7.22
Joye les Tours	2.86	-4.45	-5.14	-3.54	19.88	8.85	7.91	4.22
La Bruyère	17.21	10.28	11.81	9.43	22.32	6.70	7.72	6.26
La Source	8.31	1.59	1.71	-0.28	20.97	7.53	8.12	3.90
Leblanc	7.79	0.87	1.03	-1.09	19.96	7.20	8.42	3.22
Lucé	19.71	10.21	11.43	7.84	19.10	7.07	7.91	5.41
Oysonville	9.21	1.94	2.31	0.85	23.01	8.70	10.31	6.83
Prefecture	19.32	12.65	13.31	9.81	20.20	8.76	9.28	4.38
St Jean	4.78	-1.12	-1.19	-3.18	22.62	8.13	8.64	2.94
Ville aux Dames	3.96	-2.57	-2.90	-2.56	22.56	9.20	10.40	7.25
Manigay-les-usages	6.23	-7.30	-7.76	-6.47	29.40	9.75	10.38	4.00
Haute Normandie								
Bois_Geosmonautes	9.00	-4.87	-5.50	-3.51	24.03	10.54	11.94	6.64
Elbeuf	27.15	11.71	12.96	10.04	20.87	9.01	9.97	5.97
Forest de Brotonne	17.46	3.22	3.57	3.22	21.61	15.49	17.11	8.21
Hemot	14.86	3.21	4.51	2.88	23.56	7.22	10.28	6.86
Honfleur	16.89	3.34	4.66	3.94	22.22	9.40	13.19	6.10
Mare Rouge	9.36	-3.20	-4.24	-2.99	21.15	5.27	6.81	6.33
Mesnil_Estard	8.04	-3.97	-4.42	-3.96	20.92	7.64	8.49	5.10
Montvilliers	12.39	-1.63	-2.05	-0.70	19.82	4.57	5.68	5.40
Palais de justice	15.00	0.27	0.31	2.36	22.33	8.77	10.21	6.16
Phare Ailly	22.75	9.03	12.36	6.57	20.71	11.09	15.17	8.38
Setteville	10.61	-2.46	-2.75	-1.96	21.42	9.69	10.84	6.29
St_Romain	14.68	-0.78	-0.93	0.64	20.04	12.54	14.76	7.34
Touques	5.82	-6.02	-7.96	-4.77	24.20	17.61	23.18	13.04
Tyssandier	13.07	1.25	1.43	0.87	23.91	10.40	11.89	5.45
Val_Itton	9.89	-2.02	-2.29	-2.32	21.55	8.67	9.96	4.61
Picardie								
Albert	0.15	-5.70	-6.75	-5.66	17.55	7.75	9.17	6.66
Arras	4.82	-6.20	-8.32	-3.85	20.74	8.24	11.07	6.17
Beaumont	7.00	0.93	1.16	-1.61	19.38	7.87	9.96	5.72
Château Thierry	8.46	2.25	2.85	0.97	20.29	9.33	11.85	6.52
Crécy	11.29	1.99	2.70	2.50	23.00	12.18	16.50	9.01
Desbordes	-0.29	-6.13	-6.99	-3.95	18.81	10.50	11.95	5.16
Faiencerie_Creil	8.19	1.72	2.15	0.77	17.64	6.98	9.00	6.01
Hirson	3.67	-2.23	-2.77	-1.15	13.72	6.71	8.21	4.10
Paul Bart	-1.18	-6.80	-7.71	-4.53	15.38	6.90	7.82	4.88
Philippe Roth	5.26	-0.67	-0.76	1.05	16.54	7.19	8.09	4.52
Roye	4.18	-2.24	-2.43	-0.44	18.83	10.20	11.04	5.84
Saint_Leu	5.93	-3.45	-3.96	-3.22	19.58	12.64	14.53	8.14
Salouel	20.04	11.36	12.55	9.71	17.09	10.12	11.18	6.29
St Gobain	8.79	3.81	4.46	2.48	21.05	12.35	14.43	7.41
Moyennes	9.73	0.06	0.09	-0.01	21.36	8.81	10.64	6.60
Moyennes (urbain)	11.23	0.83	0.97	0.93	21.61	7.84	9.20	6.38

Table 1: biases and standard deviation (SD), at 15h UTC, on Esmeralda ozone stations, for Statistical Interpolation (SI), Statistical Interpolation with leave-one-out (SI l-o-o) and Kriging.

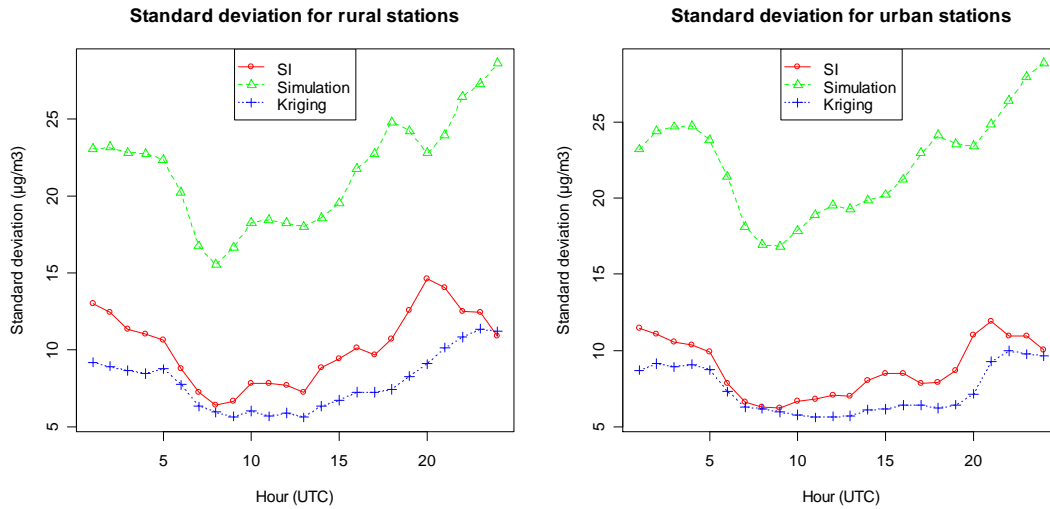
First of all, the observed differences between Statistical Interpolation sample biases and Statistical Interpolation sample biases using a leave-one-out method are rather weak (station averages at 15h UTC:  $0.06 \mu\text{g}/\text{m}^3$  and  $0.09 \mu\text{g}/\text{m}^3$  with leave-one-out). The standard deviations are a little bit more important, it means that some errors are bigger using leave-one-out method (station averages at 15h UTC:  $8.81 \mu\text{g}/\text{m}^3$  and  $10.64 \mu\text{g}/\text{m}^3$  with leave-one-out). But those differences are weak enough to consider the assimilation data methods as robust.

Secondly, biases and standard deviations at the stations are on average much better for analyses (Statistical Interpolation and Kriging) than for Simulations. Nevertheless, this is not true for all stations; actually, for stations with low initial bias, this statistical parameter could be slightly increased. This is the case for the Tinquex and Chambord stations. On the contrary, important biases are significantly reduced. This is illustrated in the next boxplots for the Cachan, Elbeuf Chambord and Tinquex stations.

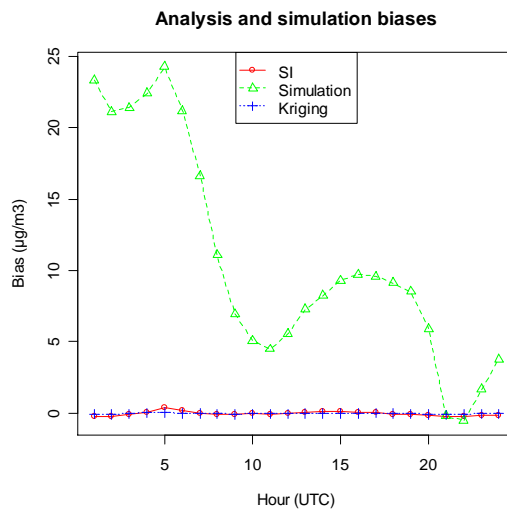


**Figure 8: Error boxplot, for analyses and simulation the Cachan, Elbeuf, Chambord and Tinquex stations at 15h UTC.  
SI on the left, simulation in the middle and Kriging on the right.**

In the figure 8, the average time evolutions of the standard deviations for SI, Kriging and simulations are presented for urban and rural stations. In figure 9, the mean hourly time evolutions of biases are plotted.



**Figure 9: Hourly Standard deviation plots - Analyses and Simulation plots on average on urban stations and rural stations. SI in red, simulation in green and Kriging in blue.**

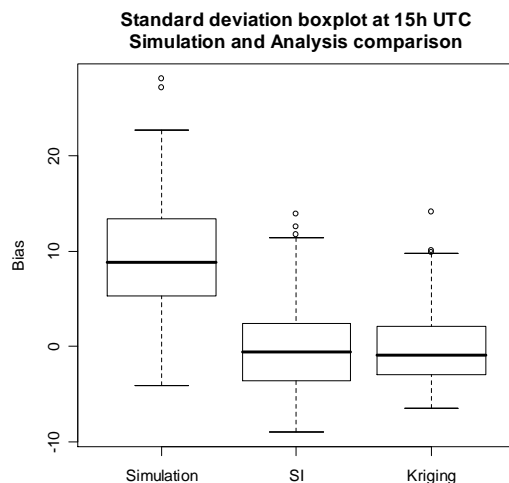


**Figure 10: Hourly Bias plots - Analyses and simulation plots on average on Esmeralda ozone stations. SI in red, simulation in green and Kriging in blue.**

On the previous graphs, we can notice a significant improvement of the standard deviations between the two data assimilation methods compared with the simulations. They are decreased by a factor above 2. As a result, data assimilation allows decreasing the error variability and also the biases as it can be seen in figure 9.

Further more, the percentage of stations which biases are above  $10 \mu\text{g}/\text{m}^3$  decreases from 44% for simulations to 4-3% for the analyses. For rural stations, the percentage is reduced to 7%.

The bias variability on all stations can be plotted using boxplot:

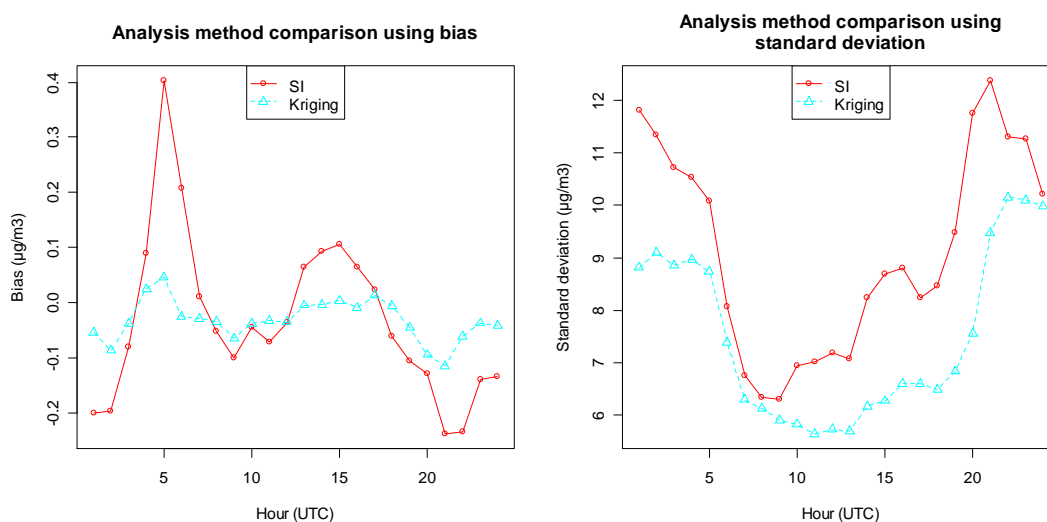


**Figure 11: Bias boxplot, for analyses and simulation for the Esmeralda ozone stations at 15h UTC. SI in the middle, simulation on the left and Kriging on the right.**

All these results show that applying data assimilation methods allow decreasing significantly model biases and uncertainties.

#### 4.1.2 Comparisons between the data assimilation methods

It's interesting to compare the two methods, using error biases and standard deviations.



**Figure 12: Hourly Biases and standard deviations - SI and Kriging on average over Esmeralda ozone stations. SI is in red and Kriging is in blue.**

Biases and standard deviations are both better for Kriging at any hour of the day. These results has been established using only analysis results at measure points: this is not surprising as the variograms used in the kriging method are adjusted every hour whereas in the Statistical Interpolation method, the covariance matrices are calculated using a climatology. In order to compare the two methods on the whole grid, the error variance have been plotted for each method. The results are presented in the *Uncertainty assessment* paragraph.

### 4.1.3 Example of the 20<sup>th</sup> June 2005

Analyses have been made on the 20<sup>th</sup> June 2005.

This day, the simulation forecasted a plume on *Centre* region (South of the domain). Observations made in the plume, were not so important. Analyses allowed to decrease the concentrations predicted in this area while preserving the simulation shape. The Maps are presented in Annex I.

### 4.1.4 Method limits

When the predicted plume is situated between measure points, analyses can't correct it. This could be a problem, especially when the model results are wrong.

The 14<sup>th</sup> June 2006, is a good example: the analysis map (by Kriging) is presented in Annex I.

## 4.2 Uncertainty assessment

Error maps can be plotted for SI and kriging, as in both methods, an error variance estimation can be calculated (equations 7 and 17). However, before assessing and comparing the error maps, we must keep in mind that the assumptions made in the two methods are completely different:

- In the SI method, standard deviations depend on the climatology of the selected hour: from one day to another, the error maps only differ with the number of valid measurement data.
- On the contrary, in the kriging method, the variograms which represent the spatial consistency of data with respect to the distance are evaluated at every time step: as a result, the error maps could be very different from one day to another depending on the concentration patterns.

Error Maps for the 20<sup>th</sup> June 2005 are presented in Annex I.

At the stations, Kriging gave better concentration values than Statistical Interpolation, but the differences were very weak. However, over the whole grid, the SI error is clearly smaller than the Kriging error. In the Kriging error maps, we can notice a lower uncertainty close to the stations than further. The uncertainty increase as we go away from the stations is linked with the variogram range. Between stations, errors strongly depend on the variogram sill. Further work should therefore be performed in order to evaluate whether these results are specific to the selected day. A study over June 2005 would be probably more representative. Besides, average standard deviations over July, August and September for the kriging method should be performed in order to be directly comparable to the standard deviations calculated for Statistical interpolation.

Nevertheless, it could be noticed that the spatial distribution of monitoring stations is very heterogeneous over the Esmeralda domain and could explain the shape of the kriging error maps: in Ile de France, the monitoring network is dense enough while in less populated areas like in the North, East and South of Ile de France, it is not the case.

On the SI error maps, the standard deviation patterns are smoother. The reason is that we consider, in the implemented method, correlations between stations and grid nodes instead of distances.

## 5 Conclusion and discussion

### 5.1 Assessment of the case study

Within the framework of the Paris case study, two analysis methods have been used to estimate, as well as possible, the real pollutant concentrations in the atmosphere. These methods use model outputs and observations from the monitoring network.

The first method is the Anisotropic and heterogeneous Statistical Interpolation. This method is an adaptation of the classical Statistical Interpolation used in meteorology: it no more considers covariances as function of distances between stations but correlations.

The second method is Intrinsically Kriging, which is used in mining sciences. This is a Geostatistic method. Kriging is used to avoid stationnarity assumption, previously supposed (in SI case).

A comparison between these methods and the model outputs concludes with a significantly better estimation of pollutant levels using data assimilation methods.

Concerning the differences between methods, we can say that Kriging seems to be a better method on measure points while Statistical Interpolation is better further. Nevertheless, some additional work should be undertaken to compare precisely the two methods. This will be probably done in the next few weeks.

### 5.2 Improvements in assessment derived from the case study

This case study allowed us to get a better understanding of the SI and Kriging data assimilation methods. Moreover, assimilation seems to be a very promising tool for AIRPARIF needs:

- **Annual statistics.** Actually, for annual air quality assessment, assimilated concentration patterns are significantly more reliable than the model outputs. Further more, using those annual concentrations with the associated uncertainties will allow us to define risk assessment for limit value exceedances.
- **Network optimization.** Assimilation can also be used as a powerful tool to establish the relevance of some stations or to determine areas where monitoring stations should be needed.

### 5.3 Recommendations resulting from the case study

The recommendations from the AIR4EU deliverables allowed us to get general information about data assimilation: some key ideas about uncertainties and the methods were also pointed-out. Nevertheless, the case study showed us that it is absolutely necessary to have sound mathematical and scientific knowledge to make correct use of these methods. For such algorithm implementation, statistical knowledge is essential, especially the variogram and covariance matrix concepts.

## 6 Bibliography:

[1] Nadège Blond. Assimilation de données photochimiques et prévision de la pollution atmosphérique.

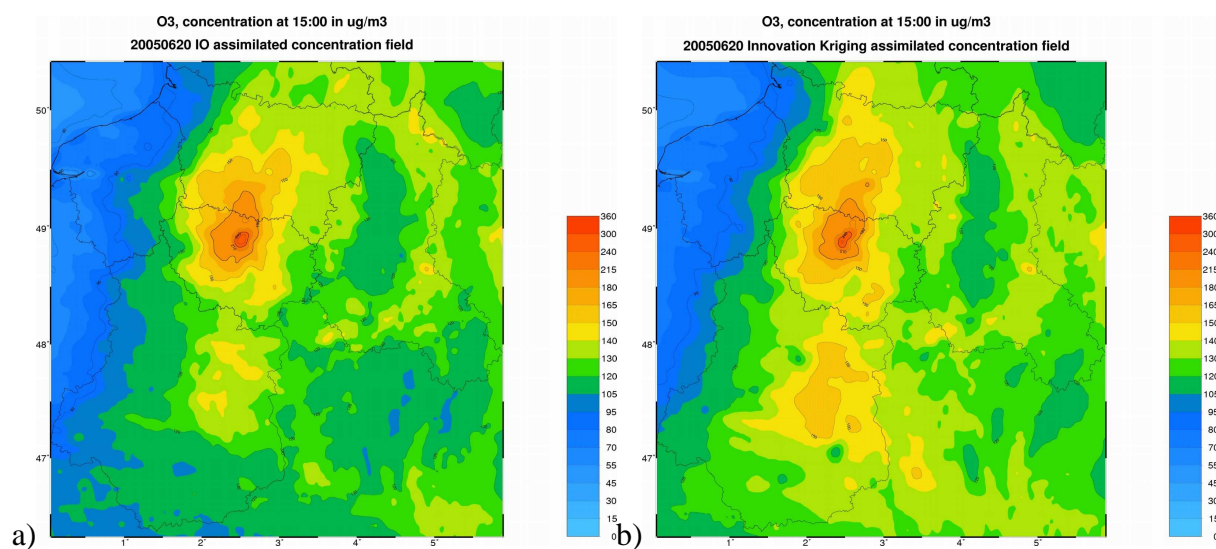
[2] Robert Vautard, Matthias Beekman, Jean Roux, Dominique Gombert. Validation of a hybrid forecasting system for the ozone concentrations over the Paris area.

### Elements

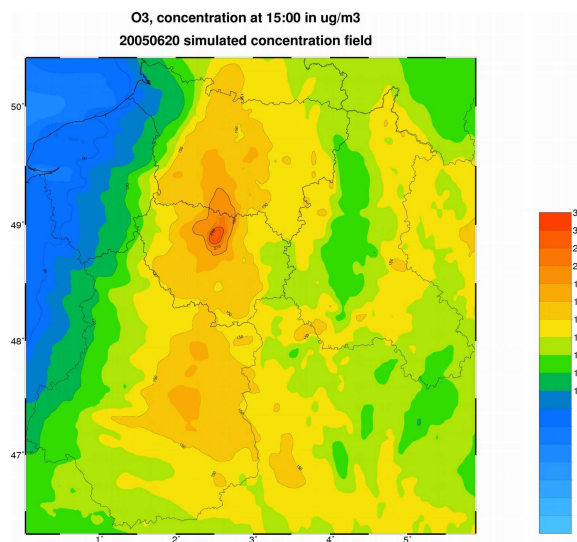
- What has been performed
  - Adapt the OI and innovation kriging to Esmeralda system.
  - Evaluation of the two methods.
- The future developments
  - Kriging, variogram estimation...

## 7 Annex I

### 7.1 Ozone concentration maps for the 20<sup>th</sup> June 2005:



*Figure 13: Analysed ozone concentration ( $\mu\text{g}/\text{m}^3$ ), june,20,2005 at 15h UTC. a) SI, b) Kriging*



*Figure 14: Simulated ozone concentration ( $\mu\text{g}/\text{m}^3$ ), june 20 2005 at 15h UTC.*

## 7.2 Ozone concentration maps for the 14<sup>th</sup> June 2006:

Assimilated ozone concentrations for the 14th June 2006 at 15:00  
( $\mu\text{g}/\text{m}^3$ )

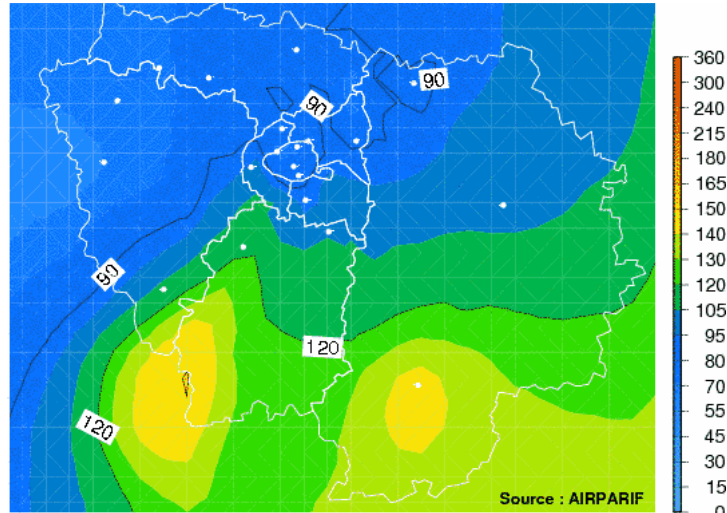


Figure 15: Analysed ozone concentration ( $\mu\text{g}/\text{m}^3$ ), 14<sup>th</sup> June 2006 at 15h UTC. Outputs from the POLLUX systems are used

## 7.3 Standard deviation maps, linked with analysis uncertainties:

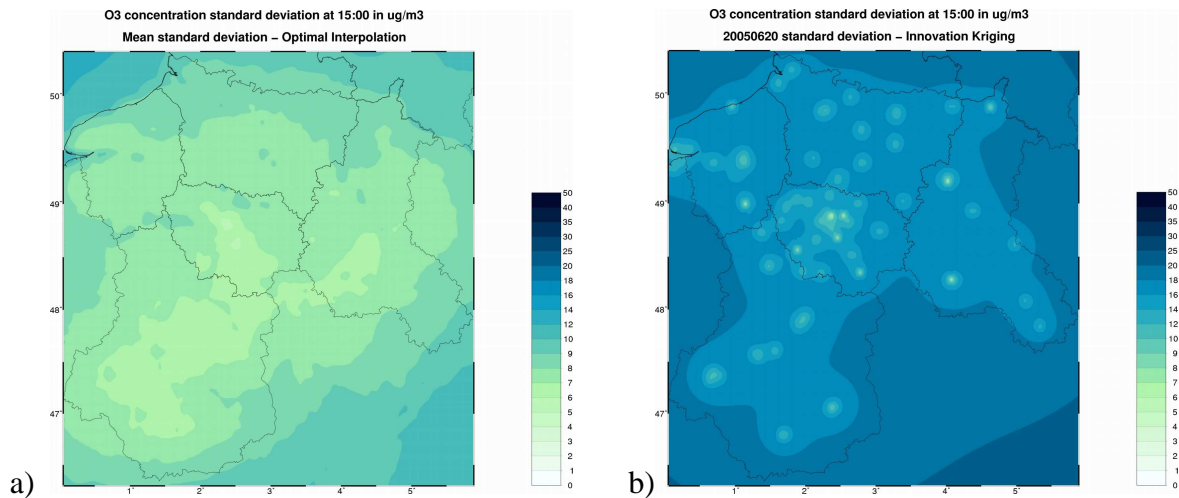
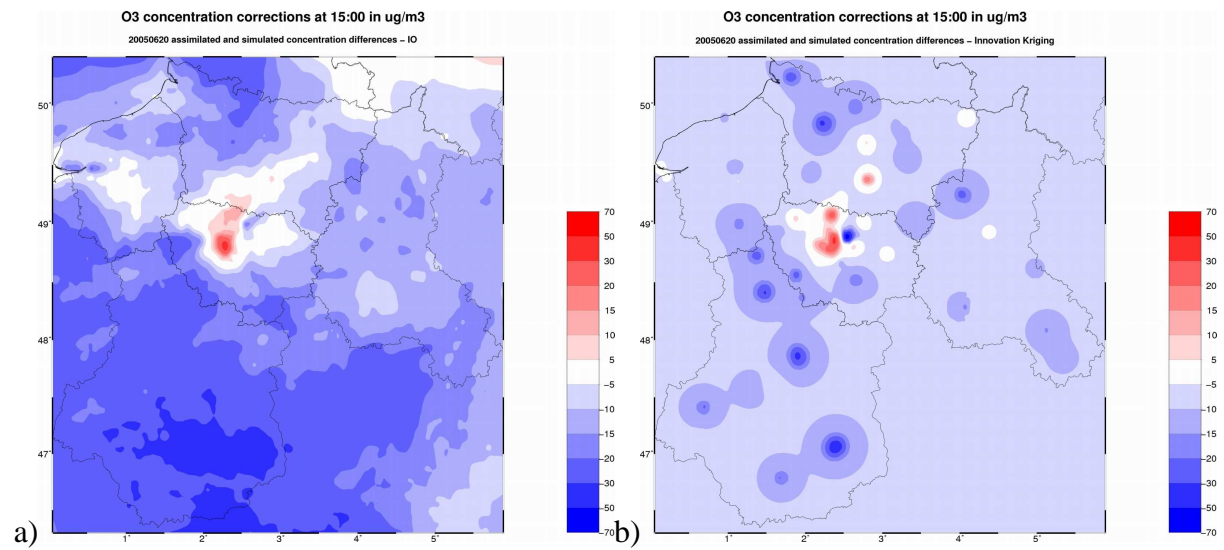


Figure 16: Standard deviation average for ozone concentration estimation by Statistical Interpolation (a) and Kriging (b).

## 7.4 Ozone concentration corrections by data assimilation:



**Figure 17: Corrections ( $\mu\text{g}/\text{m}^3$ ) applied to simulated ozone concentrations by Statistical Interpolation (a) and Kriging (b).**