

## Optimal reduction of the ozone monitoring network over France

Lin Wu<sup>a,b,\*</sup>, Marc Bocquet<sup>a,b</sup>, Matthieu Chevallier<sup>a</sup>

<sup>a</sup> Université Paris-Est, CEREAs, Joint Laboratory École des Ponts ParisTech-EDF R&D, Marne la Vallée, France

<sup>b</sup> INRIA, Paris-Rocquencourt Research Center, France

### ARTICLE INFO

#### Article history:

Received 19 January 2010

Received in revised form

6 April 2010

Accepted 7 April 2010

#### Keywords:

Air quality

Ozone monitoring

Network design

Geostatistics

### ABSTRACT

Ozone is a harmful air pollutant at ground level, and its concentrations are measured with routine monitoring networks. Due to the heterogeneous nature of ozone fields, the spatial distribution of the ozone concentration measurements is very important. Therefore, the evaluation of distributed monitoring networks is of both theoretical and practical interests. In this study, we assess the efficiency of the ozone monitoring network over France (BDQA) by investigating a network reduction problem. We examine how well a subset of the BDQA network can represent the full network. The performance of a subnetwork is taken to be the root mean square error (RMSE) of the hourly ozone mean concentration estimations over the whole network given the observations from that subnetwork. Spatial interpolations are conducted for the ozone estimation taking into account the spatial correlations. Several interpolation methods, namely ordinary kriging, simple kriging, kriging about the means, and consistent kriging about the means, are compared for a reliable estimation. Exponential models are employed for the spatial correlations. It is found that the statistical information about the means improves significantly the kriging results, and that it is necessary to consider the correlation model to be hourly-varying and daily stationary. The network reduction problem is solved using a simulated annealing algorithm. Significant improvements can be obtained through these optimizations. For instance, removing optimally half the stations leads to an estimation error of the order of the standard observational error ( $10 \mu\text{g m}^{-3}$ ). The resulting optimal subnetworks are dense in urban agglomerations around Paris (*Île-de-France*) and Nice (*Côte d'Azur*), where high ozone concentrations and strong heterogeneity are observed. The optimal subnetworks are probably dense near frontiers because beyond these frontiers there is no observation to reduce the uncertainty of the ozone field. For large rural regions, the stations are uniformly distributed. The fractions between urban, suburban and rural stations are rather constant for optimal subnetworks of larger size (beyond 100 stations). By contrast, for smaller subnetworks, the urban stations dominate.

© 2010 Elsevier Ltd. All rights reserved.

### 1. Introduction

Ozone is a harmful atmospheric pollutant that, when presents in the boundary layer, damages human health, crop production and many materials (Pleijel et al., 2007; Mauzerall et al., 2005). It is therefore crucial to know its past, present and future concentrations. To this end, diverse sources of information have to be combined for a better ozone assessment, such as observations, model simulations, and statistical inferences. Due to the heterogeneity of the available information and the intrinsic heterogeneity of ozone fields, the spatial distribution of the ozone concentration measurements is very important. Determining the optimal positioning of ozone monitoring sites, is referred to as the ozone network design problem.

Network design has been extensively considered in environmental science (Fedorov and Hackl, 1994; Müller, 2007). These investigations differ in the design criterion, the assumption on the underlying statistics, and the algorithm for the selection of monitoring sites.

One important scientific concern for ozone network design is that the design criterion should adequately account for the purpose of the monitoring network. Consequently the network design is inevitably problem-specific. For example, a sparse network is often sought to be augmented (Nychka and Saltzman, 1998; Rayner, 2004). Conversely for dense networks, when measuring instruments and their maintenance are expensive, there is a need to evaluate the efficiency of the network, so that scarce resources can be optimally allocated (Nychka and Saltzman, 1998; Fuentes et al., 2007). Examples of criteria are the maximal uncertainty reduction (computed by the entropy) or the minimal root mean square error (RMSE) of ozone estimations for the network adjustments.

Almost all the design criteria are related to the estimation of ozone concentrations. The estimation methods can be roughly catalogued

\* Corresponding author at: Université Paris-Est, CEREAs, Joint Laboratory École des Ponts ParisTech-EDF R&D, Marne la Vallée, France.

E-mail address: [Lin.Wu@cerea.enpc.fr](mailto:Lin.Wu@cerea.enpc.fr) (L. Wu).

into three groups: the observation-free, simulation-free, and advanced ones. The observation-free methods simulate ozone status using mathematical models that describe the chemistry-transport phenomena (Reynolds et al., 1973; Russell and Dennis, 2000). By contrast, in the simulation-free methods, the ozone concentrations are estimated mainly based on in situ ozone observations. In general, these estimations are made with statistical methods, e.g. regression estimations (Di Carlo et al., 2007), time series predictions, and kriging methods in which spatial correlations among observations are used for a Best Linear Unbiased Estimation (BLUE) (Kriging, 1951). The advanced methods combine the information from both the model simulations and the statistics of observations, for instance, the data assimilation algorithms (Wu et al., 2008) and the ensemble forecast techniques (Mallet et al., 2009).

The optimal deployment of a network is a combinatorial optimization problem. For dense networks, the search space is quite large. Usually, frequent evaluations of the design criterion are needed. This hampers the application of time-consuming estimation methods, i.e. the observation-free and advanced methods. Among simpler statistical methods, since spatial data are concerned, the geostatistical methods, such as kriging, are the most appropriate in this context.

In classical kriging methods, the underlying random field is often assumed to be spatiotemporally stationary, and there is no historical data from which the kriging can benefit. However, ozone fields have complex nonstationary structures, and statistics based on historical data is appropriate. Hence there are two key issues: the treatment of non-stationarity and the usage of climatological statistics. Efforts adapting the classical kriging methods for ozone estimations can be found in Blond et al. (2003); Wackernagel et al. (2004); Fuentes et al. (2007).

The objective of this paper is twofold. First, we will examine the kriging methods, with an emphasis on the non-stationarity hypothesis and on the usage of statistics, applied to the regional dense BDQA (*Base de Données sur la Qualité de l'Air*) network. Then, given the winning kriging method, the efficiency of the BDQA network will be evaluated by investigating a network reduction problem, in which the subnetwork best representing the whole network is to be determined. For network reduction problems, it is not mandatory to estimate ozone concentrations at locations different from those of the network stations. Therefore, the root mean square error for the estimations of hourly ozone means over the network can be seen as a simple and objective criterion to evaluate the subnetworks. Hopefully, the solutions of the network reduction problem will be helpful to construct practical design rules, such as the removal of redundant stations and the maintenance of important stations.

Caution is advised when analyzing the reduction results. They should not be over interpreted, because the reduction experiments covering longer period should be conducted. Statistics on the performance of the optimally reduced networks across years should be gathered and analyzed. Nevertheless, our objective is to illustrate the methodology of network reduction and to assess, at least a posteriori, the efficiency of the full network and the best subnetworks.

The paper is organized as follows. The network reduction methodology is detailed in Section 2; it covers the ozone field, the correlation models, the kriging methods, the design criterion and the reduction algorithm. Section 3 presents the setup of the network reduction experiment. We report the kriging comparisons and the reduction results in Section 4. Conclusions are provided in Section 5.

## 2. Methodology

### 2.1. Ozone field

From the probabilistic point of view, the ozone concentration is seen as a spatiotemporal random field

$$\{Z^k(\mathbf{s}), \mathbf{s} \in \mathcal{D}, k \in \mathbb{N}\}, \quad (1)$$

where  $k$  denotes the discrete time index and  $\mathcal{D} \subset \mathbb{R}^d$  (for two-dimensional field  $d = 2$ ). The points in the field domain  $\mathcal{D} = \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$  can either be cells of a regular grid or the stations from an observational network.

It is impossible to compute the probability density function (PDF) of the underlying random vector, because of the high dimension of  $\mathcal{D}$  and the nonlinearity of the chemical evolution. Two approaches, leading to approximations, can be used: the mathematical modeling and the spatial statistics. For the former, chemistry and transport is discretized into an evolution equation

$$\mathbf{z}^{k+1} = \mathcal{M}_k(\mathbf{z}^k) + \varepsilon^{k+1} \quad (2)$$

where  $\mathbf{z} = \{z_i(\mathbf{s}_j), i \in \mathcal{I}_s, j \in \mathcal{I}_d\}$  is the vector of species concentrations for certain species indices set  $\mathcal{I}_s$  and for spatial indices set  $\mathcal{I}_d = \{1, \dots, n\}$  (often a regular grid),  $\mathcal{M}_k$  is an operator of advection-diffusion-reaction type, and  $\varepsilon$  is the error vector usually assumed to be Gaussian with zero mean and a covariance structure. For a review of the chemistry-transport models (CTMs), please see Sportisse (2007).

Spatial statistics or geostatistics (Matheron, 1962; Cressie, 1993; Wackernagel, 2003), originally developed for mining operations, aims at spatial predictions, from observed field values, taking into account the spatial correlation of the random field. If the underlying field is temporally stationary, we can drop the time index  $k$  in formulation Eq. (1). The random variable  $Z(\mathbf{s})$  is then assumed to be of the form

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}, \quad (3)$$

where

$$\mu(\mathbf{s}) = \mathbb{E}[Z(\mathbf{s})] \quad (4)$$

is the mean function that describes large-scale spatial variation, and  $\delta(\cdot)$  the error field (also called residue) with zero mean that accounts for the spatial correlation structure. The field  $\delta$  is an aggregate error field that includes errors from smaller scales and different sources. This error field is usually assumed to be stationary so that statistical inferences can be more easily conducted.

One might consider that if the mean function  $\mu$  takes the expectation (deterministic part) of the CTM simulation Eq. (2), and if the error vector  $\varepsilon$  takes the same spatial structure as the error field  $\delta$  in the spatial estimation Eq. (3), the two approaches would be identical. Nevertheless the length of the model grid interval might be tens of kilometers for regional applications, and the subgrid phenomena can only be approximated by physical parameterizations. By contrast, the site observations may have much smaller scales accounting only for local fluctuations of the species around the sites. The use of model simulations as the expected means might be far from satisfactory, although Blond et al. (2003) reported successful results for a refined horizontal model resolution of  $6 \text{ km} \times 6 \text{ km}$ . In geostatistics, parametric models are proposed for the mean function  $\mu$  (Wackernagel et al., 2004).

Temporal ozone predictions are seldom conducted using geostatistical methods. The linear combinations of the observations, or other terms derived from observations, are likely to be insufficient for temporal prediction (Coman, 2008, ch. 3). This may due to i) the temporal evolution of the ozone field is highly nonlinear; ii) the temporal correlation is poorly known; and iii) the future ozone field is clearly related to more factors than ozone concentrations alone, e.g. the meteorological conditions, the precursor emissions, and the deposition parameterization.

The geostatistical methods are popular in producing analyzed maps of geophysical fields from the observations that represent the local information. These methods benefit from the spatial correlation models for successful spatial predictions. In reality, such spatial correlations are time-varying. Note that there are no repeatable experiments in geophysics. The statistical inference has to be based on only one realization (time series) of the underlying stochastic process. Hence the ergodic hypothesis is needed, so that the statistics make sense. Unfortunately the ergodicity implies the temporal stationarity. One artifact is that the timescale for the statistics may not be long enough to avoid the ergodicity breaking. Furthermore, the value of the random field at one site is not only related to other spatial observations, it can also be connected to other spatial factors, e.g. the meteorological conditions and the site typology and altitude. The geostatistical framework can deal with these additional factors, but the determination of the factor set is a non-trivial issue.

The core of mathematical models is a numerical scheme that solves Eq. (2) without the error term. Important factors, e.g. the gas-phase chemistry, the deposition parameterization, the meteorological data, and the emissions, are provided to mathematical models to generate the diurnal cycle of ozone for short-range forecasts. The ability to predict is valuable. Unfortunately the huge computational load hampers the numerical optimization for the network reduction. The information from the CTM could be helpful, but the CTM cannot be resorted too often, such as inside the optimization iterations.

### 2.2. Spatial correlation model

The spatial correlation model is essential in spatial prediction, because it helps to determine how the information from observations is dispatched in the multi-sites estimator. Considering the temporally stationary random field Eq. (3), the covariance between two arbitrary sites  $\mathbf{s}_i, \mathbf{s}_j$  is

$$\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\} = \mathbb{E}[(Z(\mathbf{s}_i) - \mu(\mathbf{s}_i))(Z(\mathbf{s}_j) - \mu(\mathbf{s}_j))] = \mathbb{E}[\delta(\mathbf{s}_i)\delta(\mathbf{s}_j)]. \quad (5)$$

If the covariance has the form

$$\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\} = C(\mathbf{s}_i - \mathbf{s}_j), \quad (6)$$

the function  $C(\cdot)$  is, by definition, the covariance function model (referred to as covariance model for short hereafter). Let  $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$  be the separation vector, and  $h = \|\mathbf{h}\|$  the distance. The covariance model  $C(\cdot)$  is isotropic, if it depends only on the distance but not the direction of  $\mathbf{h}$ . When  $\mathbf{s}_i$  and  $\mathbf{s}_j$  coincide where the field admits variance, one has

$$\text{var}\{Z(\mathbf{s})\} = \text{cov}\{Z(\mathbf{s}), Z(\mathbf{s})\} = C(\mathbf{0}). \quad (7)$$

We use isotropic covariance model in this study.

The covariance model  $C(\cdot)$  needs to be a positive-definite function (see p. 84 in Cressie (1993) for definition). Let  $\mathbf{T}$  denote the transpose of vectors or matrices, and let  $\boldsymbol{\eta}$  be the column vector  $[\dots\delta(\mathbf{s}_j)\dots]^T, j \in \mathcal{I}_d$ , the covariance matrix  $\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^T]$  has to be positive semi-definite. We refer to Balgovind et al. (1983); Gaspari and Cohn (1999) for the construction of positive-definite functions. One popular choice is the exponential model

$$C(h) = \sigma^2 e^{-\frac{h}{L}}, \quad (8)$$

where  $\sigma^2$  is the a priori variance of the error field, and  $L$  is the correlation length. For two points with distance much longer than  $L$ , their correlation approaches to zero. However, in our case, the daily ozone concentrations are correlated at long distance, e.g. between the Paris region and the urban areas along Mediterranean

coast. The daily insolation seems to produce correlations for ozone concentrations over all the French territory. We thus use a nested model

$$C(h) = c_0 + \sigma^2 e^{-\frac{h}{L}}, \quad (9)$$

where  $c_0$  is the background correlation for long distances. The variance for the field is thus  $c_0 + \sigma^2$ . This nested covariance function is positive definite if  $c_0$  is positive.

The evolution of ozone concentrations shows a diurnal cycle. We assume that the ozone field is daily stationary, but time-varying during the day. The 24 h of a day are partitioned into several bins. For convenience, we suppose each bin has the same time length. Let  $\mathcal{I}_b = \{1, \dots, B\}$  be the bin indices set, where  $B$  is the total number of bins. That is, for hourly bins,  $B$  equals to 24. The covariance model is then

$$C(h) = c_{0,b} + \sigma_b^2 e^{-\frac{h}{L_b}}, \quad b \in \mathcal{I}_b \quad (10)$$

One simple estimator of the covariance Eq. (5) for a given bin is

$$\tilde{C}(\mathbf{s}_i, \mathbf{s}_j) = \frac{1}{N_t} \sum_{t=1}^{N_t} (Z^t(\mathbf{s}_i) - \bar{Z}(\mathbf{s}_i))(Z^t(\mathbf{s}_j) - \bar{Z}(\mathbf{s}_j)), \quad (11)$$

with

$$\bar{Z}(\mathbf{s}_i) = \frac{1}{N_t} \sum_{t=1}^{N_t} Z^t(\mathbf{s}_i), \quad i \in \mathcal{I}_d, \quad (12)$$

where  $N_t$  is the number of observations along time for site  $\mathbf{s}_i$ ,  $N_t$  is the number of mutually available observation pairs along time for site pair  $(\mathbf{s}_i, \mathbf{s}_j)$ , and  $Z^t(\mathbf{s}_i)$  denotes the random variable for the ozone field at time index  $t$  and site  $\mathbf{s}_i$ . The estimator for the covariance model  $C(\cdot)$  is then

$$\hat{C}(h) = \frac{1}{|N(h)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(h)} \tilde{C}(\mathbf{s}_i, \mathbf{s}_j), \quad (13)$$

where  $N(h)$  denotes the set of the site pairs with distance equal to  $h$ :

$$N(h) = \{(\mathbf{s}_i, \mathbf{s}_j) : h = \|\mathbf{s}_i - \mathbf{s}_j\|; i, j \in \mathcal{I}_d\}, \quad (14)$$

and  $|N(h)|$  is the set cardinality. In practice, the pairs set can be adapted to:

$$N_{\mathcal{T}}(h) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in \mathcal{T}(h)\}, \quad (15)$$

where  $\mathcal{T}(h)$  is some tolerance region around  $h$ . The tolerance region is chosen so that  $|N_{\mathcal{T}}(h)|$  is sufficiently large for statistical reasons, e.g. bigger than 30 pairs.

We remark that the choice of  $N_t$  is a balance between the statistical consideration, for which long time windows are preferable, and the daily stationarity assumption which would be less relevant in the long term. We consider that one month in the summer is a proper choice. The hourly covariance model can also be found in Blond et al. (2003). In this study, we will present more detailed investigations.

The variogram (for the variance of the field increments  $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$ ) is more popular than the covariance model in geostatistics. The estimation of the variogram usually requires intrinsic stationarity Cressie (1993, p. 40), that is, the first two moments (mean and variance) of the field increments are stationary. Hence the variogram works for the random field of which the mean and the variance may not exist. By contrast, the estimation of the covariance model in geostatistics often imposes second-order stationarity Cressie (1993, p. 53), that is, the first two moments (mean and covariance) of the random field are stationary. Note that for ozone, it is reasonable to assume that the mean and covariance

are finite. Furthermore, the mean function  $\mu(\mathbf{s})$  could be estimated using statistics on historical ozone data. Therefore the ozone mean likely bears no spatial stationarity. For these reasons, we use covariance model instead of variogram. It would be better if the covariance model is interpreted as an approximation of the second-order moment (covariance) for the ozone field.

### 2.3. Spatial interpolation

Following geostatistical methodology, the spatial interpolation consists in estimating the random field  $Z$  at ungauged sites  $\mathcal{U} = \{\mathbf{s}_{p+1}, \mathbf{s}_{p+2}, \dots, \mathbf{s}_n\}$  given observations at gauged sites  $\mathcal{G} = \{\mathbf{s}_1, \dots, \mathbf{s}_p\}$ . Let us define the vectors of concentrations

$$\mathbf{y}^T = [Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_p)], \tag{16}$$

$$\mathbf{x}^T = [Z(\mathbf{s}_{p+1}), Z(\mathbf{s}_{p+2}), \dots, Z(\mathbf{s}_n)]. \tag{17}$$

Suppose that the statistics on the field are known, e.g. the following covariance matrices

$$\Sigma_{xy} = \mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^T], \tag{18}$$

$$\Sigma_{yy} = \mathbb{E}[(\mathbf{y} - \mathbb{E}(\mathbf{y}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^T]. \tag{19}$$

From these statistics, and the observations of  $Z$  on  $\mathcal{G}$ , one wishes to infer the concentrations  $\mathbf{x}$ , through an estimator  $\hat{\mathbf{x}}$ .

The best linear unbiased estimator (BLUE) is based on the linear Ansatz

$$\hat{\mathbf{x}} = \mathbf{L}\mathbf{y}, \tag{20}$$

where  $\mathbf{L}$  is a matrix of  $\mathbb{R}^{(n-p) \times p}$ . The true concentration vector is of the form  $\mathbf{x} = \hat{\mathbf{x}} + \epsilon$ , with  $\epsilon$  the vector of estimation errors (residues). The assumption of absence of bias imposes that:  $\mathbb{E}[\epsilon] = 0$ , that is,  $\mathbb{E}[\mathbf{x}] = \mathbf{L}\mathbb{E}[\mathbf{y}]$ . The errors can thus be written as

$$\epsilon = \mathbf{x} - \mathbb{E}[\mathbf{x}] - \mathbf{L}(\mathbf{y} - \mathbb{E}[\mathbf{y}]). \tag{21}$$

The estimator is optimal in the sense that the total variance of the estimation errors  $\text{Tr}(\epsilon\epsilon^T)$  is minimal, which leads to

$$\hat{\mathbf{L}} = \Sigma_{xy}\Sigma_{yy}^{-1}. \tag{22}$$

The BLUE estimator (also called simple kriging) and the corresponding minimal error is thus

$$\hat{\mathbf{x}} = \Sigma_{xy}\Sigma_{yy}^{-1}\mathbf{y}. \tag{23}$$

$$\epsilon = \min_{\mathbf{L}} \mathbb{E}[\text{Tr}(\epsilon\epsilon^T)] = \text{Tr}(\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}). \tag{24}$$

#### 2.3.1. Ordinary kriging

For one simple case, the field expectations on all the sites are taken to be uniformly null. Consequently, the field values are considered as large fluctuations. The geostatistical techniques can be performed directly on the field values. There is no bias and the BLUE estimator can be applied, so that the formulae above are valid.

Let us define  $\mathbf{I}_p$  and  $\mathbf{I}_{n-p}$  as the unit vectors in  $\mathbb{R}^p$  and  $\mathbb{R}^{n-p}$  respectively:

$$\mathbf{I}_q = (1, 1, \dots, 1)^T, \tag{25}$$

with  $q = p$  or  $q = n - p$ . If the average field value is uniform but not necessarily null, that is,  $\mathbb{E}[\mathbf{x}] = c\mathbf{I}_{n-p}$  and  $\mathbb{E}[\mathbf{y}] = c\mathbf{I}_p$  where  $c$  is

a positive constant, then the removal of the bias  $\mathbb{E}[\epsilon] = c\mathbf{I}_{n-p} - c\mathbf{L}\mathbf{I}_p$ , implies that

$$\mathbf{L}\mathbf{I}_p = \mathbf{I}_{n-p}. \tag{26}$$

This can be enforced as a constraint in the minimization of  $\text{Tr}(\epsilon\epsilon^T)$ . Then the best estimator is

$$\hat{\mathbf{L}} = \left( \Sigma_{xy} - \frac{(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\mathbf{I}_p - \mathbf{I}_{n-p})\mathbf{I}_p^T}{\mathbf{I}_p^T\Sigma_{yy}^{-1}\mathbf{I}_p} \right) \Sigma_{yy}^{-1}, \tag{27}$$

$$\begin{aligned} \epsilon &= \min_{\mathbf{L}} \mathbb{E}[\text{Tr}(\epsilon\epsilon^T)] \\ &= \text{Tr} \left( \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \right. \\ &\quad \left. + \frac{(\Sigma_{xy}\Sigma_{yy}^{-1}\mathbf{I}_p - \mathbf{I}_{n-p})(\mathbf{I}_p^T\Sigma_{yy}^{-1}\Sigma_{yx} - \mathbf{I}_q^T)}{\mathbf{I}_p^T\Sigma_{yy}^{-1}\mathbf{I}_p} \right). \end{aligned} \tag{28}$$

#### 2.3.2. Kriging about the hourly means

In the previous subsections, we present the classical geostatistics results. In this and following subsections, we will adapt the kriging to the air quality context.

Geostatistics was first developed for the spatial estimation of geological ore. The true representation to be estimated corresponds to a single sampling of a random field. In particular, often, no prior statistics are available on the distribution of the ore. However, in the context of air quality, a large database of concentration records may be available. For instance for ozone, hourly measurements are recorded, so that statistics of hourly means of ozone on each measurement site can be extracted.

Assume that the average hourly concentrations are accessible on  $\mathcal{U}$  and  $\mathcal{G}$ :

$$\mathbb{E}[\mathbf{x}] = \mu_x \quad \text{and} \quad \mathbb{E}[\mathbf{y}] = \mu_y. \tag{29}$$

Then the previous BLUE analysis can be performed on the fluctuations around these mean values. Only these fluctuations are considered as random: each concentration is the sum of a fluctuation and a deterministic part (the mean). Therefore

$$\hat{\mathbf{x}} = \mu_x + \mathbf{L}(\mathbf{y} - \mu_y), \tag{30}$$

with  $\mathbf{L}$  given either by Eq. (22) (Simple Kriging about the Means) or by Eq. (27) (Ordinary Kriging about the Means).

A usual type of kriging is an estimator with the form being the linear combination of the observations at the gauged sites. This estimator can also be assumed linearly correlated to not only the observations, but also other factors (termed as external drift) at the ungauged sites. In the cases where the statistics at all the sites are available, one can simply choose the mean statistics as the external drift, instead of parameterization methods e.g. the universal kriging in Wackernagel et al. (2004).

#### 2.3.3. Consistent kriging about the hourly means

The BLUE analysis around the means is independent from the actual values of  $\mu_x$  and  $\mu_y$ . If one assumes that the fluctuations are of the same physical nature as these means, can the means be made consistent with the BLUE analysis? That would imply that the estimator applied to  $\mu_y$  would yield  $\mu_x$ . If this is so, one can perform the BLUE analysis on the fluctuations, imposing that

$$\mu_x = \mathbf{L}\mu_y. \tag{31}$$

Then the best estimator is

$$\hat{\mathbf{x}} = \left( \Sigma_{xy} - \frac{(\Sigma_{xy}\Sigma_{yy}^{-1}\mu_y - \mu_x)\mu_y^T}{\mu_y^T\Sigma_{yy}^{-1}\mu_y} \right) \Sigma_{yy}^{-1}\mathbf{y}, \quad (32)$$

$$\begin{aligned} \varepsilon &= \min_{\mathbf{L}} \mathbb{E}[\text{Tr}(\epsilon\epsilon^T)] \\ &= \text{Tr} \left( \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \right. \\ &\quad \left. + \frac{(\Sigma_{xy}\Sigma_{yy}^{-1}\mu_y - \mu_x)(\mu_y^T\Sigma_{yy}^{-1}\Sigma_{yx} - \mu_x^T)}{\mu_y^T\Sigma_{yy}^{-1}\mu_y} \right). \end{aligned} \quad (33)$$

It is also possible to constrain additionally  $\mathbf{L}$  to satisfy the ordinary kriging condition  $\mathbf{L}\mathbf{p} = \mathbf{I}_{n-p}$ . This leads to new BLUE estimators (more intricate though). However, we have checked that it does not significantly improve the spatial interpolation, at least in the context of ozone estimation. That is why the results are omitted here.

#### 2.4. Reduction criterion

The network design problem can be described by:

$$\xi^* = \underset{\xi}{\text{argmin}} \Psi(\xi), \quad (34)$$

where  $\xi$  is a potential network configuration, and  $\Psi$  is a certain scalar criterion. The choice of the criterion is influential (Müller, 2007; Abida et al., 2008). With different criteria, the resulting optimal networks might be quite different.

Let  $\mathbb{B}$  be the binary set  $\{0, 1\}$ ,  $\xi \in \mathbb{B}^n$  is then the vector that describes the network configuration. The  $i$ -th component of the configuration vector  $\xi_i$  is 1 if site  $\mathbf{s}_i$  is selected in the subnetwork, otherwise  $\xi_i = 0$ .

Define the complete network  $\mathcal{A} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$ , let the gauged set  $\mathcal{G} \equiv \{\mathbf{s}_i : \xi_i(\mathbf{s}_i) = 1\} \subset \mathcal{A}$ , and the ungauged set  $\mathcal{U} \equiv \{\mathbf{s}_i : \xi_i(\mathbf{s}_i) = 0\} \subset \mathcal{A}$ . The network reduction problem is then

$$\xi^* = \underset{\xi \in \mathbb{B}^n}{\text{argmin}} \left( \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} (\hat{z}^o - z^o)^2 \right)^{\frac{1}{2}}, \quad (35)$$

with the constraint

$$\sum_{i=1}^n \xi_i = p, \quad (36)$$

where  $\mathcal{O}$  is the indices set for the observations available at ungauged sites  $\mathcal{U}$  during a given period,  $|\mathcal{O}|$  is the set cardinality,  $p$  is the size of the gauged subnetwork,  $\hat{z}^o$  is the kriging result for an ungauged site at a given time step using observations at gauged sites  $\mathcal{G}$ , and  $z^o$  is the observation which is observed for the same site and at the same time step as that of the corresponding  $\hat{z}^o$ .

This criterion is the  $\text{RMSE}$  for the spatial predictions based on the selected subnetwork. It is expected that the optimal subnetwork under this criterion best represents the whole network  $\mathcal{A}$  among all the subnetworks of size  $p$ . Although the criterion Eq. (35) has a quadratic form, it bears complex nonconvex structures on  $\xi$ . The search space, with its size being  $\binom{n}{p}$ , is colossal even for moderate  $n$  and  $p$ .

#### 2.5. Optimization using simulated annealing

The reduction criterion Eq. (35) has multiple minima. One popular solver for such optimization problem is the simulated annealing algorithm. This stochastic method (Metropolis et al., 1953;

Kirkpatrick et al., 1983) has the ability to escape from local minimum  $\xi^{(i)}$  to a new configuration  $\xi^{(i+1)}$  with an acceptance probability. For the classical Boltzmann annealing, this probability equals to

$$P(\xi^{(i)}, \xi^{(i+1)}, \tau) = \min \left( 1, e^{-\frac{\Psi(\xi^{(i+1)}) - \Psi(\xi^{(i)})}{\tau}} \right) \quad (37)$$

where  $\tau$  is a global parameter which is an analog of temperature. Often  $\tau$  is initially high, and the iterative process probes large-scale variation of  $\Psi$ . When  $\tau$  decreases according to certain cooling schedule, the iterations search for finer variations. By carefully choosing the cooling schedule, the global minimum can be approached to some precision which can be arbitrarily small. For the geometric cooling schedule,

$$\tau^{(K+1)} = \alpha\tau^{(k)}, \quad (38)$$

where  $\alpha \in [0, 1]$  is a decreasing factor, and usually  $k$  coincides with  $i$ . The procedure that generates new candidate  $\xi^{(i+1)}$  should favor the configurations similar to the current configuration  $\xi^{(i)}$ , so that much inferior candidates are excluded for a more efficient search. The new candidates are therefore chosen from the neighborhood of  $\xi^{(i)}$ .

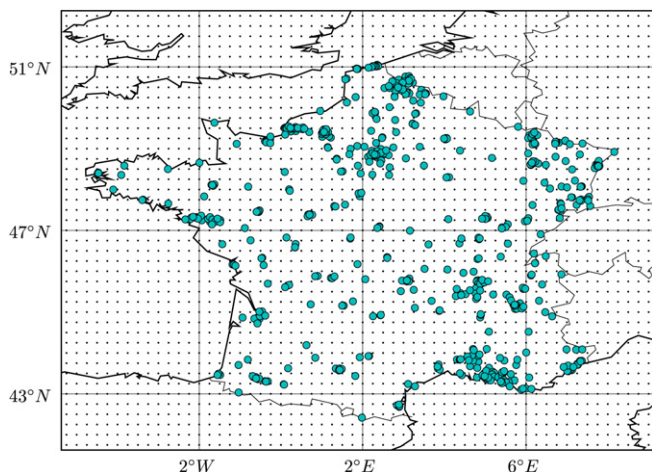
The simulated annealing algorithm has been applied in several network design problems (Banjevic, 2004; Fuentes et al., 2007; Abida et al., 2008; Saunier et al., 2009). All these applications adopt Boltzmann annealing and geometric cooling schedule, but differ in the neighborhood assignment and in the definition of  $\Psi$  and  $\xi$ . In our study, the Boltzmann annealing and the geometric cooling are also employed. We generate the new configurations using a swap procedure: keep an array of the indices of  $n$  sites; divide the index array into two parts, that is,  $p$  indices for selected sites and  $n-p$  indices for the estimation sites; randomly choose one index from each part, and switch the two indices. By this way, we randomly choose one estimation site, and randomly replace one previously selected site with this newly chosen estimation site. Note that the constraint Eq. (36) is automatically satisfied. In practice, the tuning of the parameters values, especially for the initial and final temperatures, is necessary to obtain a satisfactory solution.

### 3. Experiment setup

Ozone concentrations peak during summer, which is the most risky scenario for human health and crop production. The time window of the experiment is thus set to be from 1st July 2001 at 0100 UTC to 1st August 2001 at 0000 UTC. Note that the time window does not cover the whole summer season, so that the daily stationarity assumption likely holds.

The sites  $\mathcal{A} = \mathcal{G} \cup \mathcal{U}$  are taken to be stations from the BDQA network (information available at <http://www.atmonet.org>). There are 678 BDQA stations within the domain that covers French territory ( $[41.75^\circ \text{ N}, 5.25^\circ \text{ W}] \times [52.5^\circ \text{ N}, 9.25^\circ \text{ E}]$ ). These stations (see Fig. 1) are located in typologically different areas, such as urban districts and regional areas of cities, industry sites, and traffic lines. The isotropic hypothesis might be regarded as too idealized for this case. However, there is a practical concern to evaluate the network as a whole. Furthermore, Blond et al. (2003) show that anisotropic considerations bring a positive impact, but the improvements are not significant (see Fig. 7 in that paper). The improvements are especially less significant for daily ozone peaks.

Hourly (averaged) ozone observations are collected during July 2001. There are only 351 stations having observations within this



**Fig. 1.** Map of the BDQA Network. The circles indicate the locations of the stations, and the points show the POLAIR3D model grid-cell centers.

period. These stations may not always have observations throughout the experiment. No specific techniques are considered for the treatment of missing values. We simply do not conduct the spatial interpolation for a certain station at a given date, when there is no observation for this station at that date. The observational error is approximately  $10 \mu\text{g m}^{-3}$  (Flemming et al., 2003).

We use the POLYPHEMUS/POLAIR3D air quality model, which has been validated in Sartelet et al. (2007). The simulations are performed with an integration time step of 5 min. The hourly ozone simulation results are used when necessary. The RMSE for the hourly ozone simulations over all the BDQA stations for July 2001 is  $27.7 \mu\text{g m}^{-3}$ . Note that this error includes the error arising from the interpolations of ozone concentrations between the model grid and BDQA sites. The domain has a  $58 \times 43$  grid, and the model mesh has a  $0.25^\circ$  horizontal resolution (see Fig. 1). Blond et al. (2003) investigate a much finer resolution of 6 km, which is a proper scale compared with that of the physical phenomena represented by the hourly averaged observations from the stations. For regional applications, such refined resolution is not practical, since, in general, the meteorological data and the emissions have much coarser resolutions.

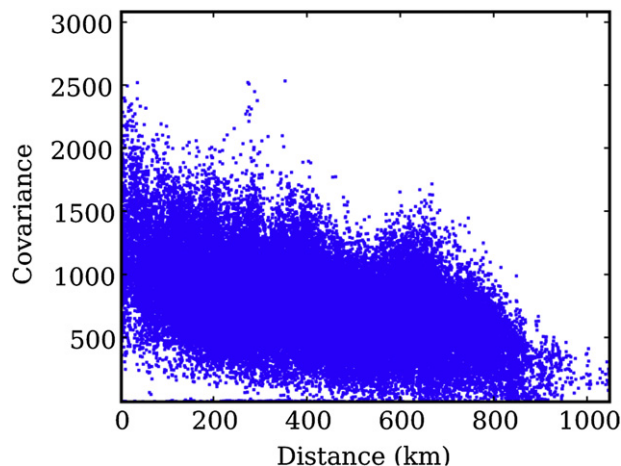
The POLYPHEMUS/POLAIR3D simulations are used to test the sensibility of kriging performance with respect to covariance models (Section 4.2.1) and mean statistics (Section 4.2.2).

## 4. Results and discussions

### 4.1. Covariance model

We investigate the diurnal covariance model by dividing a day into several contiguous bins. Each bin has the same time length. In order to evaluate the impact of the bin length, we set the length value to 24 h, 12 h, 6 h, 4 h, 3 h, 2 h, and 1 h respectively. Hence accordingly, we have 1, 2, 4, 6, 8, 12, and 24 bins for the corresponding bin lengths. For each bin, one can calculate the covariance value  $\hat{C}(h)$  according to Eq. (13). With  $h$  taking all the possible distances of the station pairs in the site pair set  $N(h)$  (defined by Eq. (14)), a cloud of covariance values can be obtained. We plot such a covariance cloud at 1500 UTC in Fig. 2. Note that the ozone observations used in the calculation of these cloud points are those during July 2001.

Now we examine the covariance within the tolerance regions. The length of the tolerance region (hereafter denoted  $L_T$ ) is set to 30 km,



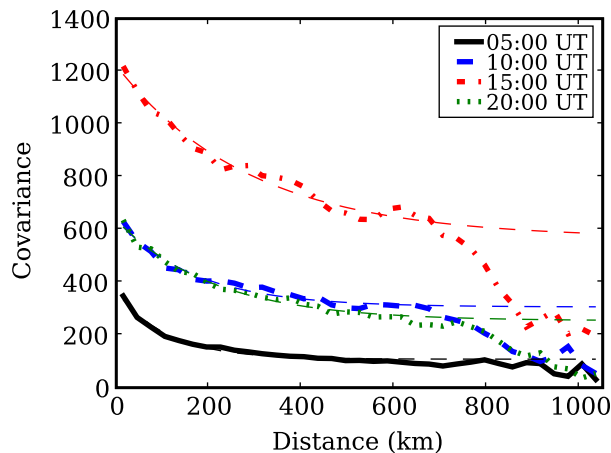
**Fig. 2.** The covariance cloud at 1500 UTC. The unit for covariance is in  $(\mu\text{g m}^{-3})^2$ .

so that there are enough site pairs for each tolerance region. The covariance clouds are averaged within contiguous tolerance regions  $\mathcal{T}(h_i) = [h_i - L_T/2, h_i + L_T/2]$ , for  $h_i = i \times L_T - L_T/2, i \in \mathbb{N}$ . By this way, the curves of the regionalized covariances can be obtained (see Fig. 3). The covariances beyond 600 km (about half the domain size) are considered as spurious correlations. They are discarded in accordance with usual geostatistical recommendations.

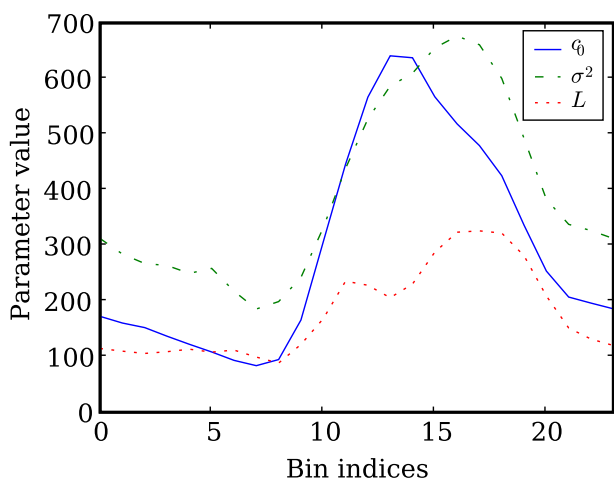
Let  $\Theta = [c_{0,b}, \sigma_b^2, L_b]$  be the vector of unknown parameters for the nested covariance function  $C(\cdot)$  in Eq. (10). The parameter  $\Theta$  is determined by solving the ordinary least-square fitting problem

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{N_h} \hat{C} \left( (h_i) - C(h_i) \right)^2 \quad (39)$$

where  $N_h$  is the total number of the tolerance regions, and  $h_i$  is the center of the  $i$ -th region on which the regionalized covariance is  $\hat{C}(h_i)$ . For each bin, the parameter vector  $\Theta$  is calibrated using the Levenberg–Marquardt algorithm. These regionalized covariances are fitted from observations over all the available BDQA stations for July 2001. The fitting results for 1 h bins are shown in Figs. 3 and 4. The value  $C_{0,b} + \sigma_b^2$  is the spatially stationary variance for the error field  $\delta(\cdot)$ . The correlation length  $L_b$  is smaller at night than at



**Fig. 3.** The regionalized covariance and fitted 1 h covariance model for four representative hours. The fitted covariance models are in thinner dash line. The unit for covariance is in  $(\mu\text{g m}^{-3})^2$ . Note that the correlations beyond 600 km are not taken into account for calibration.



**Fig. 4.** The calibrated parameter values with respect to bin indices for 1 h bins. The unit for  $c_0$ ,  $\sigma^2$  is in  $(\mu\text{g m}^{-3})^2$ , and the unit for  $L$  is in km.

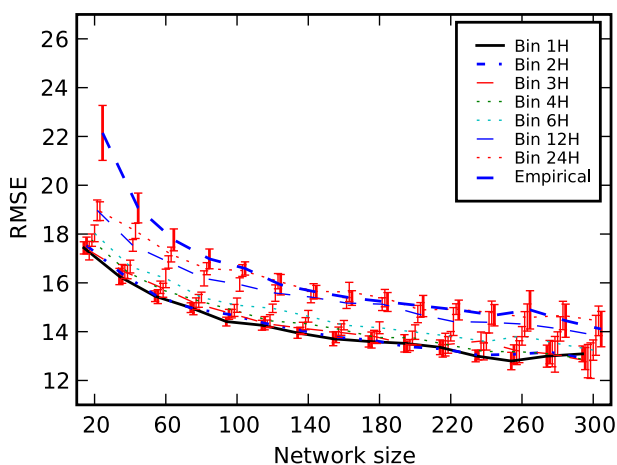
daytime, and peaks at noon and in the afternoon. The covariances are lower at night than during the day. This is probably due to a much more developed photochemistry during daytime (insolation, heavy traffic emissions at rush hours, etc.). Moreover, the ozone field during night is much more spatially heterogeneous.

#### 4.2. Kriging validation

The kriging performance is of great importance for the network reduction problem Eq. (35). In this section, we investigate the sensitivity of the kriging performance to the covariance model, the ozone mean statistics, and the kriging type. The kriging performance is evaluated by the  $\text{RMSE}$  in  $\mu\text{g m}^{-3}$  between the observed and kriged ozone hourly means at ungauged stations. The size of the subnetwork is thus the number of gauged stations. In the sequel, the unit of the  $\text{RMSE}$  is omitted for convenience.

##### 4.2.1. Sensitivity to the covariance model

The impact of the diurnal covariance model on the performance of simple kriging about the means is shown in Fig. 5. Both the covariance model calibration and the kriging validation are performed during July



**Fig. 5.** The performance of simple kriging about the means during July 2001 with respect to different diurnal covariance models. The error bar indicates the standard deviation of  $\text{RMSE}$  for 10 randomly chosen networks with the corresponding network size.

2001, except for an empirical covariance model based on the following Balgovind parameterization (Wu et al., 2008):  $c_0 = 0$ ,  $\sigma = 20 \mu\text{g m}^{-3}$  and  $L = 100 \text{ km}$ , the parameters of the covariance models are calibrated as in Section 4.1 (over all the available BDQA stations for July 2001). Each of these covariance models has been tested on each of the following subnetwork sizes: 20, 60, 100, 140, 180, 220, 260 and 300 stations. Moreover, for each of these experiments, the result is the average on a set of ten randomly sampled subnetworks, so as to guarantee the robustness of the results.

The kriging performance increases when augmenting the network size, since there is more data from gauged sites. The slight oscillation of the kriging performance for large subnetworks is due to the fact that the marginal information gain through increasing the subnetwork size is less important than the accumulation of kriging errors. Kriging with the calibrated 24 h covariance model shows evident improvement against that with the empirical covariance model for small subnetworks. This is normal, since, when less observations are available for kriging, precise correlation will contribute to a more accurate estimation. When more observations are available, the kriging error and the gain using a precisely calibrated covariance model are rather comparable. The consideration of diurnal cycle of the covariance models significantly improves the kriging performance. The 1 h covariance model is the best, probably because it best characterizes the diurnal cycle as detailed in Fig. 4.

The worst kriging  $\text{RMSE}$  for calibrated isotropic covariance model in Fig. 5 is about  $19 \mu\text{g m}^{-3}$  for subnetworks of 20 stations. This is better than the hourly  $\text{RMSE}$  of the POLYPHEMUS/POLAIR3D simulations ( $27.7 \mu\text{g m}^{-3}$ ). The kriging performance justifies that the isotropic correlation is a reasonable assumption for ozone estimation in our application.

In principle, the observations for covariance model calibrations and for kriging should be independent from each other. Now we conduct 1 h covariance model calibrations based on the observations from two years, 2000 and 2001. A smaller network (referred to as the synchronous network hereafter) with 311 stations available for both years is used instead of the complete network with 351 available stations. Consequently slightly less observations are available for the kriging validation, e.g. about 30–40 less observations at given time dates.

We conduct 1 h diurnal covariance model calibrations with different observation sets. The observations for calibrations are taken from the synchronous network for four different periods of the years 2000 and 2001. The time periods for these observation sets differ in their durations and in their starting and ending dates, but all cover the month of July. We also consider the POLAIR3D simulations at the regular model grid during July 2001 for calibration purpose. Note that the parameters of the covariance model fitted with the synchronous network are slightly different from those with the complete network (1.5% relative difference on average for the observation set during July 2001).

The simple kriging about the means is performed for July 2001 with the 1 h covariance models calibrated above. The validation results are shown in Fig. 6. The kriging performance is quite stable with different calibrated covariance models. This is true even for the covariance model fitted to the POLAIR3D simulations. With this desirable result, we conclude that the assumption of daily stationarity for covariance models is favorable in the context of spatial ozone estimation. The correlation for the summer scenario, e.g. the high insolation levels and the heavy emissions, can be determined through statistics in a very robust manner.

##### 4.2.2. Sensitivity to the mean statistics

For the kriging about the means introduced in Section 2.3.2, the mean statistics is considered as a factor to which the ozone estimations are linearly correlated. In this section, we examine the

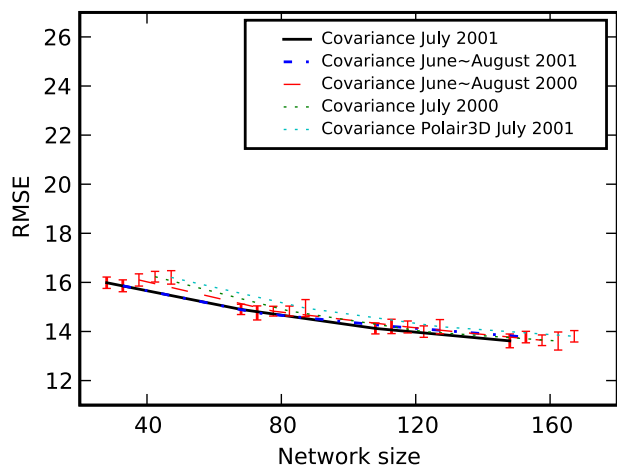


Fig. 6. The performance of simple kriging about the means during July 2001 with respect to 1 h diurnal covariance models calibrated based on different observations sets. The error bar indicates the standard deviation of RMSE for 10 randomly chosen networks.

sensitivity of the kriging performance to different sets of mean statistics. The observations are taken from four different periods of the years 2000 and 2001, which are exactly the same as those in Section 4.2.1. The POLAIR3D means are computed based on the bilinear interpolations of the POLAIR3D simulations from model grid points to site locations.

Ordinary kriging about the means is used, since it is popular in geostatistical applications. For the ordinary kriging presented in Section 2.3.1, the ozone field is assumed to be uniformly distributed, but the exact value of the constant a priori mean is not needed. The kriging validation is conducted with the synchronous network. The 1 h diurnal covariance model is adopted with its parameter values calibrated based on the observations during July 2001.

The performance of ordinary kriging about the means with respect to different mean statistics is shown in Fig. 7. The POLAIR3D means bear no positive impact against the case of constant mean. This is probably due to the over smoothed mean surface generated by the POLAIR3D simulations which do not capture the smaller spatial scales near the observation sites. By contrast, significant

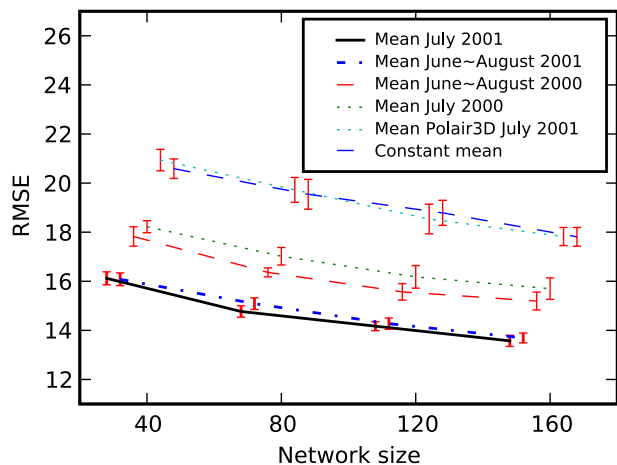


Fig. 7. The performance of ordinary kriging about the means during July 2001 with respect to different settings for the mean statistics. The error bar indicates the standard deviation of RMSE.

improvements over the case of constant mean are obtained for the kriging with the mean statistics calculated from different sets of observations. This is even true with the mean statistics during summertime in 2000. The classical geostatistical method (constant mean case) has poor performance, because it makes no use of the a priori information implicitly inferred through mean statistics. The mean statistics can be considered as an aggregate factor, which synthesizes the influence of the site-related information say the site altitude, and of the seasonal patterns, e.g. the meteorological data and emissions.

#### 4.2.3. Sensitivity to the kriging type

Apart from the optimality condition of minimum variance, additional conditions can be taken into account in the kriging, e.g. the de-biasing constraints detailed in Section 2. In this section, we investigate the sensitivity of the kriging performance to different kriging types, e.g. the simple kriging about the means (SKM), the ordinary kriging about the means (OKM) with the constraint Eq. (26), and the consistent kriging about the means (CKM) with the constraint Eq. (31).

The comparison results are shown in Fig. 8. The site means are computed from the observations of July 2001. The results of the classical ordinary kriging are plotted for comparison purpose. The constraint of constant mean introduced in OKM brings almost no impact in RMSE against that of SKM. In the kriging about the means, the fluctuations around the means are corrected instead of the ozone concentrations. These fluctuations are expected to have zero means, therefore the constraint of constant mean is automatically satisfied and almost has no impact. By contrast, the de-biasing constraint imposing on the site means in CKM slightly improves the kriging performance. Such constraint might be more important for the cases in which factors other than site means are introduced as external drifts.

#### 4.3. Network reduction

##### 4.3.1. Reduction results with a reference algorithmic setting

In Fig. 5, one can observe that different subnetworks of the same size may have different performances. It is desirable to compare the optimal solution of the network reduction problem Eq. (35) with a set of randomly distributed subnetworks. If the improvement through optimal network reduction is small against the best random trials, it would imply that the ozone field might be

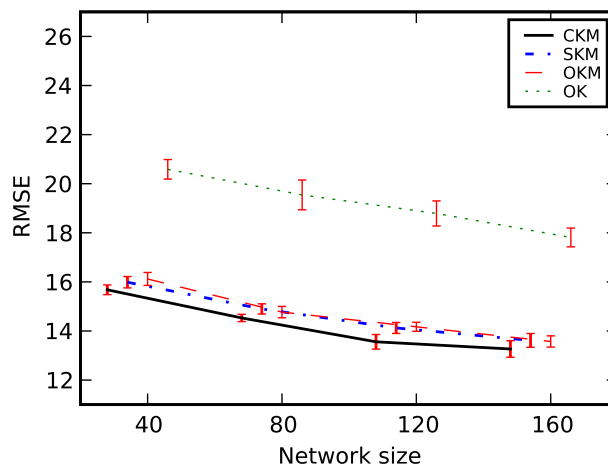


Fig. 8. The kriging performance during July 2001 with respect to kriging types. The error bar indicates the standard deviation of RMSE.



homogeneous and stationary, and that there is no need for mathematical optimization since a few random tests fulfill the objective.

In this section, we perform the network reduction with the same algorithmic setting as that in Fig. 5. In summary, simple kriging about the means and the 1 h nested covariance model are employed; the ozone observations during July 2001 are used for both means calculations and covariance model calibrations; the complete BDQA network is covered; and the geometric cooling schedule is adopted with the decreasing factor  $\alpha$  set to 0.99.

The kriging performance of optimal subnetworks of different size is shown in Fig. 9. The relative improvements in kriging  $RMSE$  for optimal subnetworks are from 12% to 58% with respect to the random subnetworks. It is striking that an optimal subnetwork of half the BDQA network, guarantees an average error of about  $10 \mu\text{g m}^{-3}$  (5 ppb) for the hourly ozone means on the ungauged sites. Since this is much lower than the average model simulation error and of the same order as the observational error, this result validates the interest of the optimization technique. The improvement increases with respect to the network size. The main reason is probably that, with more stations included in the subnetworks, the heterogeneous ozone field can be better represented based on heterogeneous station observations.

The size of the largest subnetworks is 350. In this case, only one station is excluded from the full network, and the  $RMSE$  between kriged and observed concentrations is computed for this excluded station (also called leave-one-out validation). In Fig. 9, the last (biggest) error bar shows the standard deviation of  $RMSE$  for all the 351 possible subnetworks in the leave-one-out validation. The optimal subnetwork has the minimal  $RMSE$ . The histogram of the leave-one-out  $RMSE$  is shown in Fig. 10. The distribution is roughly log-normal, which demonstrates the heterogeneous nature of the network reduction problem.

One may argue that the optimal reduction results might be misleading, since heterogeneity arises whereas our ozone estimation is based on the hypothesis of isotropic correlation. In this regard, we would like to recall that our isotropic kriging method has been justified by the cross validation results ( $19 \mu\text{g m}^{-3}$  for the worst case). The anisotropic considerations slightly improve the kriging results Blond et al. (2003). In our sensitivity studies, it has been shown that the kriging results are much more sensitive to the historical mean statistics (Fig. 7) than to the correlation model (Fig. 6). The network reduction methodology assembles a set of tools, e.g. the spatial non-stationarity in historical mean statistics

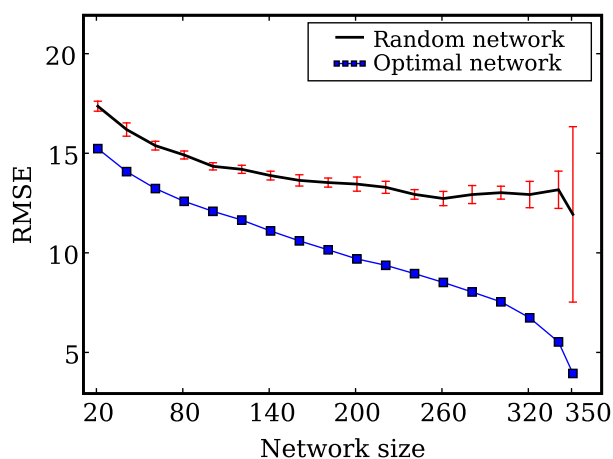


Fig. 9. The kriging performance of subnetworks with different sizes. The optimal subnetworks are obtained with the reference algorithmic setting for network reduction, and the random subnetworks are the same as those for 1 h covariance model cases in Fig. 5.

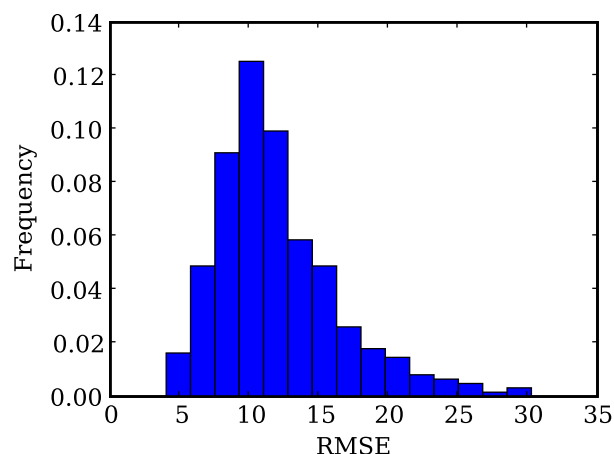


Fig. 10. Histogram of the leave-one-out validation for all the 351 subnetworks.

and station observations, and the station selection procedure. The spatial interpolation is only part of the methodology. The ozone heterogeneity produced by the entire reduction methodology is a reasonable result. Yet it is noted that better anisotropic correlation models could improve the kriging performance and deserve further investigations.

The maps of the optimal subnetworks with different sizes are shown in Fig. 11. The stations of the optimal subnetworks are heterogeneously distributed. Redundant stations in the urban agglomerations around Paris (*Île-de-France*) and Nice (*Côte d'Azur*) are removed. However, the network remains dense within these regions as compared to the rest of France. More stations are selected in the frontier regions. For the large rural regions, the stations are uniformly distributed. More stations are selected in the large urban agglomerations where the ozone field is more heterogeneous. The network remains dense near the frontiers because beyond these frontiers, no observations are available and the ozone field is uncertain. The ozone concentrations could be heterogeneous near Belgium and over Alsace regions. The optimization results are consistent with the intuition that the stations in the regions where the ozone field is uncertain should be selected in the optimal subnetworks. This is also proposed in Fuentes et al. (2007).

The fractions of several types of stations present in the optimal subnetworks are shown in Fig. 12. There is no saturation in any of the station types, as the subnetwork's size increases. Most of the stations in the smallest optimal subnetworks are urban stations. This may be due to the high concentrations and strong heterogeneity in urban areas. With a limited number of stations, the subnetworks have to be representative of these urban areas. When the size of the subnetworks increases, the fraction of suburban and rural stations increases, whereas the fraction of urban stations decreases. Indeed, once a few stations are placed in the main high concentrations urban areas, the estimation errors in rural areas will play a more important role, and the need for background stations increases. For subnetworks with size bigger than 100, the fractions between the urban, suburban and rural stations remain stable. About half of the stations in optimal subnetworks are urban stations. There are always more suburban stations than rural stations, and their percentages vary from about 10% to 30%.

The criterion is evaluated a posteriori in this algorithmic setting, since both the statistics and the network optimization are based on the observations during July 2001. Therefore, the optimal subnetworks best represent the whole network for July 2001, but the findings might not be optimal for the subsequent years.

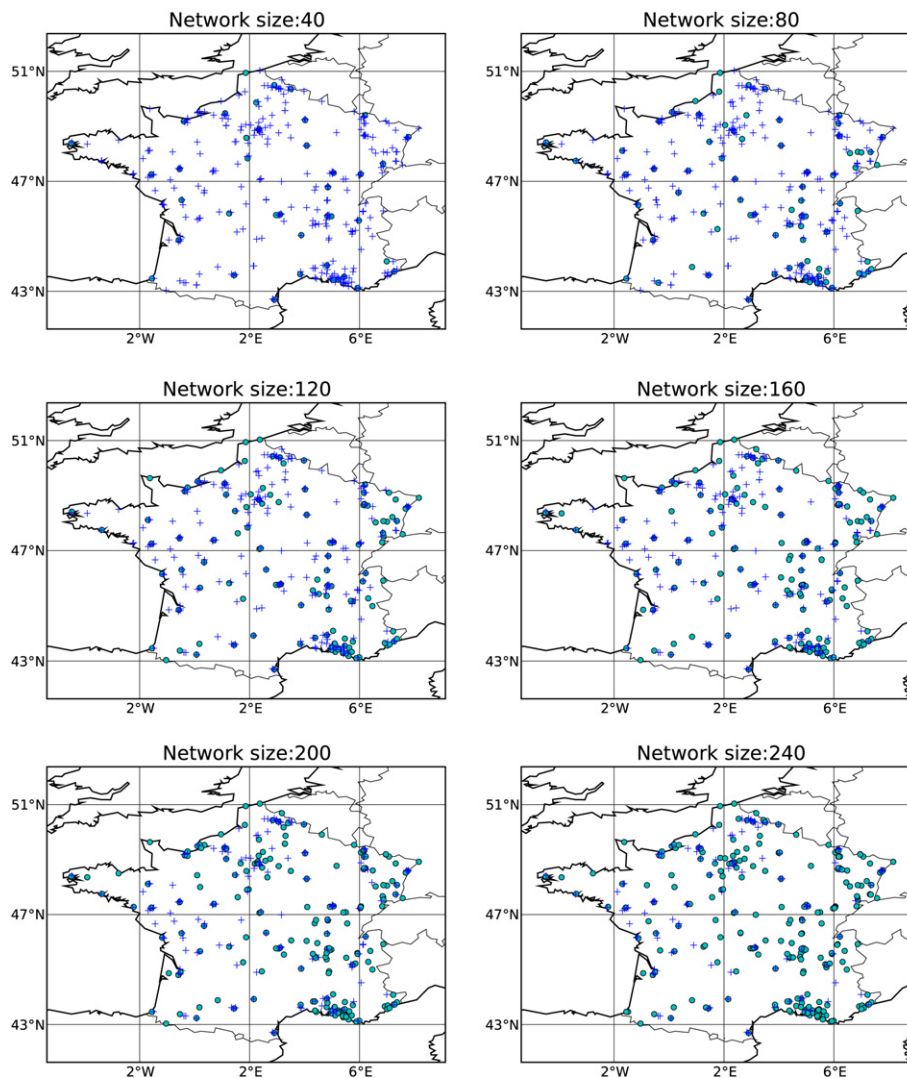


Fig. 11. The maps of the optimal subnetworks with different sizes. The circles are stations included in the optimal subnetworks, and the plus signs show the excluded stations.

Nevertheless, we hope that the constructive rules derived from the optimization results remain helpful to conduct the network reduction. This is certainly true if the evolution of the ozone field shows similar pattern and magnitude in summer time through the years. Beyond several years, climatological and mitigation policy effects will undoubtedly impact the photochemistry analysis, so that the optimality of the chosen network will have to be questioned again.

Our reduction criterion is defined as the kriging performance over the BDQA network. Other criteria could also be defined on a uniform coverage (e.g. a regular grid). For example, the kriging performance on a  $0.25^\circ \times 0.25^\circ$  grid over France (Fig. 1), might be expected to be another proper reduction criterion. Unfortunately, this criterion is not practical due to the following facts: i) in general, there are no observations at grid points to evaluate the kriging performance; and ii) the uncertainty of the kriging results over France is an entangled issue. Nevertheless, a simplified criterion of this kind to evaluate a subnetwork could be the  $RMSE$  between the kriging results over France with this subnetwork and those with the complete BDQA network.

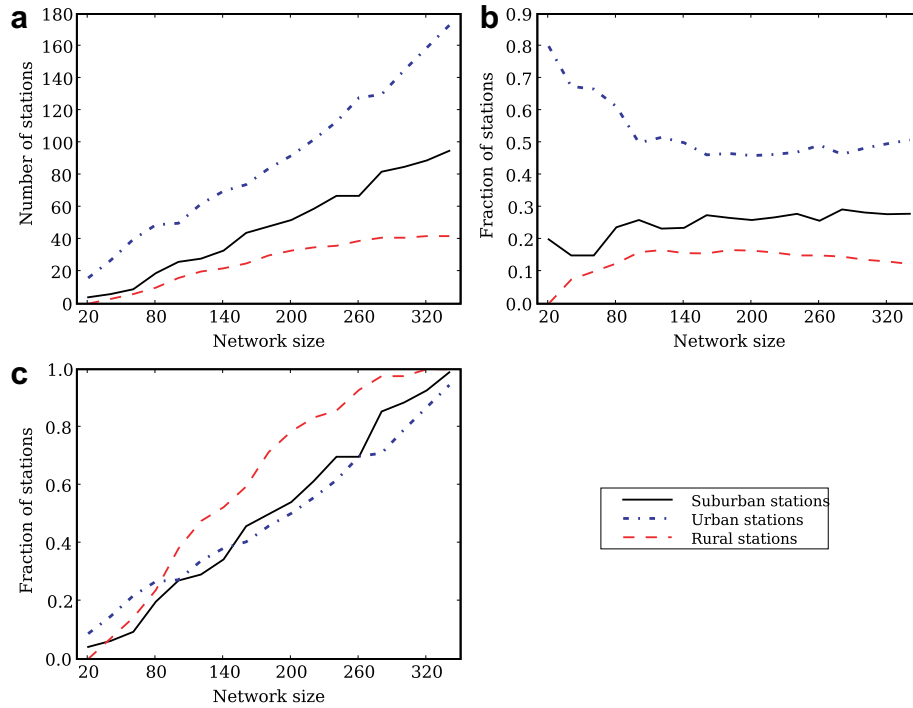
In Fig. 13, we evaluate this new criterion with respect to the optimal subnetworks (obtained under the original reduction

criterion) and the random subnetworks in Fig. 9. It seems that the optimal subnetworks remain quite optimal under this new criterion. This is reasonable since the BDQA network is dense for regional applications, and since the stations in optimal subnetworks are located more likely in the area where the ozone concentrations are high and uncertain.

#### 4.3.2. Sensitivity of reduction results to algorithmic settings

The rationale of the network reduction problem is that the optimal subnetworks should represent well the heterogeneous ozone field. Since this heterogeneity is formulated and estimated implicitly by the statistics based on observations during summer time, it would be crucial to evaluate the sensitivity of the reduction results to the statistical inference methodology. In addition, the setting for the optimization algorithm may also have impacts on the reduction results.

In this section, we perform network reductions under different algorithmic settings shown in Table 1. These different settings make alternative choices, as compared to the reference algorithmic setting in Section 4.3.1, in one or several following configurations: the mean statistics, the covariance model, the kriging type, and the annealing schedule.



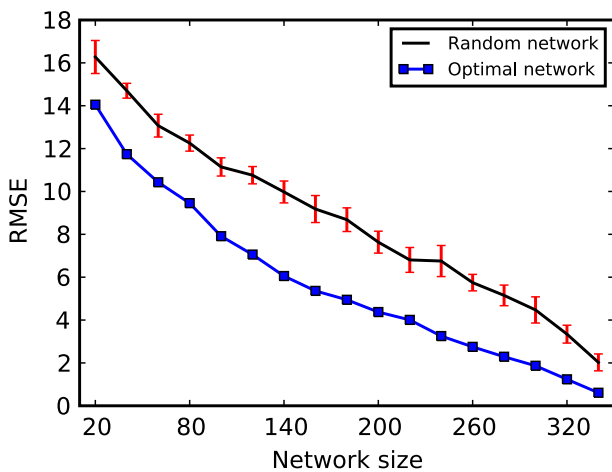
**Fig. 12.** Statistics on the reduction results: (a) the number of stations of given types in the optimal subnetworks; (b) the fraction of stations of given types with respect to stations of all the six types in the optimal subnetworks; and (c) the fraction of stations of a given type in the optimal subnetworks with respect to all the stations of that type in the full network. Other types of stations, e.g. traffic and industrial ones, have much smaller portions in the optimal subnetwork. Their statistics are therefore not shown in this figure.

The alternative configurations are the 1 h covariance model calibrated with observations from June to August in 2000, the mean statistics based on observations from June to August in 2000, the consistent kriging about the means, and a fast annealing schedule (decreasing factor  $\alpha$  set to 0.7).

The synchronous network in Section 4.2.1 is employed to examine the sensitivity of the reduction results. For simplicity, the size of the subnetworks is set to 120. The kriging performance of the optimal subnetworks with respect to different algorithmic settings is shown in the last column in Table 1. The mean statistics is the most influential factor. Other factors, e.g. the kriging type and

the annealing schedule, have little impact on the  $RMSE$  for the resulting optimal subnetworks. The consistent kriging about the best mean statistics (setting  $c$  in Table 1) has the best kriging performance. These findings are consistent with the results in Section 4.2. Note that the kriging is performed with optimal subnetworks in this section, whereas in Section 4.2, random subnetworks are employed.

The statistics on the selected stations are shown in Table 2. It is found that the optimal subnetworks are robust to algorithmic changes. The distribution of the optimal subnetworks is very similar, and many stations are even identical in different optimal subnetworks (only about 20–30% difference). The fraction between urban, suburban and rural stations remains almost unchanged. The robustness of the reduction results justifies the applicability of these results for practical contexts. The optimal reduction is mainly determined by the problem itself, e.g. the summer scenario say the emissions and meteorological conditions.



**Fig. 13.** The performance of the simplified criterion ( $RMSE$  between the kriging results on the regular grid over France using the complete BDQA network and the subnetworks) defined on a uniform coverage with respect to the optimal and random subnetworks in Fig. 9.

**Table 1**

Different algorithmic settings under which the network reductions are performed. Each row defines an algorithmic setting different from the reference one in Section 4.3.1. The plus marks in each row indicates that the marked configurations are alternated for the algorithmic setting defined by that row. The unmarked configurations in that row are the same as those corresponding configurations for the reference algorithmic setting (ref.). The kriging performance of the optimal subnetworks of size 120 with respect to different algorithmic settings is listed in the last column.

	Mean	Covariance model	Kriging	Annealing	$RMSE$
(ref.)					11.333
(a)	+				12.424
(b)		+			11.375
(c)			+		11.065
(d)				+	11.338
(e)	+	+	+		12.235

**Table 2**

The statistics on the stations included in the optimal subnetworks which are obtained under the alternative algorithmic settings defined in Table 1. These subnetworks are of size 120. The row labeled as “difference” shows the number of stations which are selected in those alternative optimal subnetworks but are excluded in the optimal subnetwork obtained under the reference algorithmic setting. The last three rows show the difference between the total number of stations of the corresponding station type (i.e. urban, suburban and rural) included in these alternative optimal subnetworks and that counted with the reference optimal subnetwork.

Algorithmic setting	(a)	(b)	(c)	(d)	(e)
*Difference	32 (27%)	20 (17%)	33 (28%)	21 (18%)	36 (30%)
Urban	2	−3	1	0	1
Suburban	−1	4	0	1	1
Rural	−3	−1	−3	−4	−3

## 5. Conclusion

As a first step towards a consistent and useful definition of the network reduction problem, we have examined and developed a geostatistical methodology, in which the ozone field is assumed to be daily stationary and isotropic. This methodology also benefits from the statistics on the ozone historical data. The kriging methods have been tested for ozone estimation based on the observations of the BDQA network over France. It has been shown that the mean statistics and the diurnal covariance model have significant impact on the geostatistical estimation performance. Considering additional de-biasing constraints slightly improves the kriging performance.

The ozone heterogeneity can be accounted for by our network reduction methodology (ozone kriging and station selection procedure using simulated annealing). We have evaluated the efficiency of the BDQA network by investigating a network reduction problem, in which the optimal subnetwork of a given size with the best estimation performance has been determined. It has been found that the gain of the optimal subnetwork against non-optimized subnetworks is significant. Optimally keeping half of the BDQA network leads to an average error of about  $10 \mu\text{g m}^{-3}$  (5 ppb) in hourly ozone concentrations, which is commensurate with model simulation error and standard observational error.

The optimal subnetworks are dense in the large urban areas and near the country borders. This finding is also consistent with the intuition that stations in areas where observed ozone concentrations are high and present strong spatial heterogeneity (because of high precursor emissions), should be included in the optimal subnetworks.

More efficient networks can thus be constructed by reducing the network size with minimal decrease in the estimation performance. Simple rules can be derived from the reduction results for the guidance of the practical network construction. For example, potential redundant stations should be removed, and the network should remain dense in those regions with high and heterogeneous ozone values.

The typology of stations has been studied as the optimal subnetwork size is increased. Beyond 100 stations, the fractions between urban, suburban and rural stations are rather constant. For smaller subnetworks, the urban stations play a more significant role.

It has been found that the network reduction results are very robust to the algorithmic setting for the kriging and the combinatorial optimization. This finding favors the application of the reduction results in practical context. Nevertheless, it is important to notice that, for practically applicable rules, one has to validate the reduction results for the observations over several years. The effect of missing observations has to be assessed. In addition, the kriging and reduction results have their proper spatiotemporal

context. These results might not be the same when the spatio-temporal scales change, e.g. from regional to local spatial scale, or from hourly to daily or weekly temporal scale. Therefore, further investigations are needed for these reduction results of immediate and practical use.

Other issues are also worth investigating, for instance, the extension of this work to the case of multiple pollutants, e.g. the network design for both ozone and nitrogen dioxide. The optimal redistribution of the stations of the reduced network is a natural follow-up of the reduction problem. It is different in that the ozone field is to be estimated at locations where no observations are available. Such locations could be a regular grid over France. Simulations of chemistry-transport models could provide information at these locations.

## Acknowledgements

This research was supported by the French Research Network on Sustainable Development (*Réseau R2DS Île-de-France 2007-21*). The authors acknowledge an early but useful discussion with Philippe Lameloise and later discussions with Anne Kauffmann's team from AIRPARIF, constructive discussions in INERIS with Laurence Rouil's team, helpful discussions with Anda Ionescu and Adriana Coman, the help from Victor Winiarek on preparing the POLAIR3D simulations, and a careful reading of the manuscript by Christian Seigneur.

## References

- Abida, R., Bocquet, M., Vercauteren, N., Isnard, O., 2008. Design of a monitoring network over France in case of a radiological accidental release. *Atmos. Env.* 42, 5205–5219.
- Balgovind, R., Dalcher, A., Ghil, M., Kalnay, E., 1983. A stochastic-dynamic model for the spatial structure of forecast error statistics. *Mon. Wea. Rev.* 111, 701–722.
- Banjevic, M., 2004. Optimal network designs in spatial statistics. Ph.D. thesis, Stanford University.
- Blond, N., Bel, L., Vautard, R., 2003. Three-dimensional ozone data analysis with an air quality model over the Paris area. *J. Geophys. Res.* 108, D23.
- Coman, A., 2008. Modélisation spatio-temporelle de la pollution atmosphérique urbaine à partir des mesures d'un réseau de surveillance de la qualité de l'air. Ph.D. thesis, Université Paris-Est.
- Cressie, N.A.C., 1993. *Statistics for Spatial Data*, Revised Edition. Wiley, New York.
- Di Carlo, P., Pitari, G., Mancini, E., Gentile, S., Pichelli, E., Visconti, G., 2007. Evolution of surface ozone in central Italy based on observations and statistical mode. *J. Geophys. Res.* 112, D10316.
- Fedorov, V.V., Hackl, P., 1994. Optimal experimental design: spatial sampling. *Calcutta Stat. Assoc.* 44, 57–81.
- Flemming, J., van Loon, M., Stern, R., 2003. Data assimilation for CTM based on optimum interpolation and Kalman filter. In: Paper presented at 26th NATO/CCMS International Technical Meeting on Air Pollution Modeling and Its Application, NATO Comm. on the Challenges of the Mod. Soc., Istanbul.
- Fuentes, M., Chaudhuri, A., Holland, D.M., 2007. Bayesian entropy for spatial sampling design of environmental data. *Environ. Ecol. Stat.* 14, 323–340.
- Gaspari, G., Cohn, S.E., 1999. Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.* 125, 723–757.
- Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. *Science* 220 (4598), 671–680.
- Krige, D.G., 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. Chem. Metall. Min. Soc. S. Africa* 52, 119–139.
- Mallet, V., Stoltz, G., Mauricette, B., 2009. Ozone ensemble forecast with machine learning algorithms. *J. Geophys. Res.* 114, D05307.
- Matheron, G., 1962. *Traité de géostatistique appliquée*. Tome I. Editions. Technip, Paris.
- Mauzerall, D.L., Sultan, B., Kim, N., Bradford, D.F., 2005. NOx emissions from large point sources: variability in ozone production, resulting health damages and economic costs. *Atmos. Env.* 39, 2851–2866.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21 (6), 1087–1092.
- Müller, W.G., 2007. *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, third ed. Springer-Verlag.
- Nychka, D., Saltzman, N., 1998. Design of air quality monitoring networks. Lecture Notes in Statistics 120. In: Nychka, D., Cox, L., Piegorsch, W. (Eds.), *Case Studies in Environmental Statistics*. Springer-Verlag, pp. 51–76.
- Pleijel, H., Danielsson, H., Emberson, L., Ashmore, M., Mills, G., 2007. Ozone risk assessment for agricultural crops in Europe: further development of stomatal

- flux and flux response relationships for European wheat and potato. *Atmos. Env.* 41 (14), 3022–3040.
- Rayner, R.J., 2004. Optimizing CO<sub>2</sub> observing networks in the presence of model error: results from transcom 3. *Atmos. Chem. Phys.* 4, 413–421.
- Reynolds, S.D., Roth, P.M., Seinfeld, J.H., 1973. Mathematical modeling of photochemical air pollution-I: formulation of the model. *Atmos. Env.* 7, 1033–1061.
- Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modeling. *Atmos. Env.* 34 2283–2234.
- Sartelet, K.N., Debry, E., Fahey, K., Roustan, Y., Tombette, M., Sportisse, B., 2007. Simulation of aerosols and gas-phase species over Europe with the polyphemus system: part I—model-to-data comparison for 2001. *Atmos. Env.* 41, 6116–6131.
- Saunier, O., Bocquet, M., Mathieu, A., Isnard, O., 2009. Model reduction via principal component truncation for the optimal design of atmospheric monitoring networks. *Atmos. Env.* 43, 4940–4950.
- Sportisse, B., 2007. A review of current issues in air pollution modeling and simulation. *J. Comp. Geosci.* 11 (2), 159–181.
- Wackernagel, H., 2003. *Multivariate Geostatistics*, third ed. Springer-Verlag, Berlin.
- Wackernagel, H., Lajaunie, C., Blond, N., Roth, C., Vautard, R., 2004. Geostatistical risk mapping with chemical transport model output and ozone station data. *Ecol. Modell.* 179, 177–185.
- Wu, L., Mallet, V., Bocquet, M., Sportisse, B., 2008. A comparison study of data assimilation algorithms for ozone forecasts. *J. Geophys. Res.* 113, D20310.