# Optimal Analysis of In Situ Data in the Western Mediterranean Using Statistics and Cross-Validation

JEAN-MICHEL BRANKART AND PIERRE BRASSEUR

*GeoHydrodynamics and Environment Research (GHER), Liège, Belgium*

(Manuscript received 20 September 1994, in final form 3 July 1995)

ABSTRACT

To study the Mediterranean general circulation, there is a constant need for reliable interpretations of available hydrological observations. Optimal data analyses (in the probabilistic point of view of objective analysis) are fulfilled using an original finite-element technique to minimize the variational principle of the spline procedure. Anyway, a prior statistical knowledge of the problem is required to adapt the optimization criterion to the purpose of this study and to the particular features of the system. The main goal of this paper is to show how the cross-validation methodology can be used to deduce statistical estimators of this information only from the dataset. The authors also give theoretical and/or numerical evidence that modified estimators—using generalized cross-validation or sampling algorithms—are interesting in the analysis optimization process. Finally, results obtained by the application of these methods to a Mediterranean historical database and their comparison with those provided by other techniques show the usefulness and the reliability of the method.

## 1. Introduction

A direct observation of the sea is an essential source of information and knowledge of the marine systems. The difficulty in interpreting this type of information lies in its two most important features. The measures are very sparse in space and time due to the cost of oceanographic cruises. Furthermore, not only are they representative of the subject of the study, they are also affected by all processes taking place in the ocean. How is it possible to extract the relevant information from a dataset? The solution is closely linked to the possibility of estimating the quality of information, that is, error evaluation. Statistical methods have basically been developed to provide an error estimation. Objective analysis (Gandin 1963; Bretherton et al. 1976) has become very popular in oceanography because error maps and analysis may be derived simultaneously. On the contrary, a priori, the spline analysis method has no relation with the minimization of an error. It is defined by a variational principle weighing two terms: a measure of the "difference between the solution and the data" and a measure of the "smoothness of the solution" (Wahba and Wendelberger 1980; McIntosh 1990).

However, these two methods are very similar and a great benefit can be derived from an equivalence theorem (McIntosh 1990; Brasseur 1994). As a consequence, the spline method actually receives a reliable statistical interpretation. The benefit mainly lies in the numerical efficiency of the spline analysis method—especially if we use an efficient finite-element technique to solve the variational problem (Brasseur et al. 1991). Nevertheless, whether the spline analysis or the objective analysis is used, the selection criterion (i.e., minimization of the global error) must be based on some prior statistical information about both the field to be reconstructed and the noise affecting the data. Yet, the database is the only available information source. The main purpose of this paper is precisely *to build an efficient procedure in order to extract this statistical information from the database.*

Up to now, oceanographers have faced great difficulties in selecting the optimization criterion of the objective analysis method. Indeed, the computation of the statistical properties (correlation function and signal-to-noise ratio) necessary to *give sense* to the error minimization has been rendered very difficult for lack of observations sets dense enough to generate reliable estimations by classic averaging procedures (such as in meteorology). Consequently, this problem has often been circumvented by making quite arbitrary assumptions about the statistical features. Yet, it has been demonstrated that the correlation length and especially the signal-to-noise ratio govern the main features of the analysis (by selecting the optimal degree of representation of the structures contained in the data). We will give some examples of the critical importance of these parameters. Moreover, the noise variance estimation directly influences the computed error amplitude and, accordingly, must be very cautiously com-

*Corresponding author address:* Dr. Jean-Michel Brankart, GeoHydrodynamics and Environment Research, Université de Liège, Sart Tilman B5, B-4000 Liège, Belgium.
E-mail: brankart@modb.oce.ulg.ac.be

puted to give a meaning to the error maps produced by any analysis scheme.

Today, with an increase of experimental work, the quality of this prior statistical information can be improved, especially in the context of climatological analyses of a basin as well covered as the Mediterranean. The purpose of the Mediterranean Oceanic Data Base project (MODB, part of the MAST Mediterranean Targeted Project supported by the European Commission) is precisely to gather most of the hydrological (in situ) observations collected in the Mediterranean since the beginning of the century, to bank and check the dataset, and finally, to propose analysis methods and results at the climatological timescale. The different techniques described in this paper will be applied in this particular context.

## 2. Statistical basis of data analysis

### *a. Data analysis target*

To perform data analysis of some features of the ocean, the first step is to define the purpose of the study, which lies in the relation between the real evolutive field and the target of the analysis. Whatever the target field—noted $\phi_t$—(synoptic, climatological at the monthly or seasonal timescale, climatic, etc.), it can be considered as a known function (including averaging over a time period, filtering of small spatial scales, etc.) of reality. As will be seen later, the characteristics of data analysis are closely related to the definition of the target field, as, for example, the properties of the dataset. A dataset from only one cruise could be convenient for a synoptic target field, but climatological analyses need a large historical database.

Moreover, we will call the "reconstructed field" (noted $\phi$) the result of the analysis (for example, objective or spline analysis that we will briefly describe later) of an appropriate dataset. Considering the fundamental underdetermination of this problem, each method needs an optimization criterion to select the solution. Thus, the analysis method (i.e., the criterion) must be adapted to both the purpose of the study and the particular features of the system.

The challenge of data analysis is to minimize the difference between these two fields (target and reconstructed fields: $e = \phi - \phi_t$), *given the information contained in the dataset,* and to provide an estimation of this difference. Of course that objective may only be achieved if some prior information on the target field has been extracted from the dataset and if the analysis method is flexible enough to take this information into account. An interesting way of investigating is to consider this problem from a probabilistic point of view. The use of probability to quantify the extent of our knowledge and the application of simplifying hypothesis to compute them in practice are the key features of this approach.

### *b. Statistics and hydrological fields*

First, we will call the *background field* (noted $\phi_b$) the best prior estimate of the target field: the first prior information essential to the solution of the problem. For lack of a better key, it may, for example, be computed as the mean or the linear regression of the data. For data analysis, differences with respect to this background alone will be considered because hypothesis (H2 and H3) on the statistical structure of the field (or probability distributions) are made much more realistic.

We suppose that the prior knowledge on the target field (difference with respect to the background) that we are looking for is expressed in terms of probability. Let $N_{\mathbf{x}}(\phi)$ be the probability distribution characterizing our knowledge of the target field at the point $\mathbf{x}$, and $F_{\mathbf{x},\mathbf{y}}(\phi_x, \phi_y)$ be the conditional probability distribution for the points $\mathbf{x}$ and $\mathbf{y}$ (the probability that the field at $\mathbf{y}$ lies between $\phi_y$ and $\phi_y + \delta\phi_y$, given that the field at $\mathbf{x}$ lies between $\phi_x$ and $\phi_x + \delta\phi_x$). All the statistical properties of the target fields could be derived from this information. For example, its covariance function is

$$c(\mathbf{x}, \mathbf{y}) = \iint \phi_x\phi_y F_{\mathbf{x},\mathbf{y}}(\phi_x, \phi_y) N_{\mathbf{x}}(\phi_x) d\phi_x d\phi_y. \tag{2.1}$$

This technique should be considered as a reliable and powerful way of expressing our prior knowledge of the target field. Nevertheless, considering we have to gain access to this knowledge only through the dataset, some simplifications or hypothesis are obviously needed to make the method applicable. The four most classic ones used in the literature are the following (Gandin 1963; Bretherton et al. 1976):

H1—All the probability distributions are supposed to be Gaussian. (It should be noted that this hypothesis is not absolutely required for most of the analysis methods but makes their explicit probabilistic interpretation much easier.)

H2—The mean of $N_{\mathbf{x}}(\phi)$ is zero. The background field is supposed to have been chosen to get this important property.

The distribution $N_{\mathbf{x}}(\phi)$ is then entirely determined by only one parameter, the first-order moment, the background error variance $\epsilon^2$, and may be written as

$$N_x(\phi) = \frac{1}{(2\pi\epsilon^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{\phi^2}{\epsilon^2}\right). \tag{2.2}$$

H3—Homogeneity and/or isotropy of the characteristics of the field.

The background error variance is then independent from the position, and the conditional probability distribution only depends on the distance $r = |\mathbf{x} - \mathbf{y}|$

between the two points. Given H1, H2, and H3, (if we introduce the correlation function $\gamma$) the conditional probability distribution becomes

$$F_r(\phi_x, \phi_y) = \frac{1}{[2\pi\alpha^2(r)]^{1/2}} \times \exp\left\{-\frac{1}{2}\frac{[\gamma(r)\phi_x - \phi_y]^2}{\alpha^2(r)}\right\}. \quad (2.3)$$

These expressions lead to the covariance function

$$c(r) = \epsilon^2\gamma(r). \quad (2.4)$$

The application of Bayes theorem gives the expression of the resulting variance

$$\alpha^2(r) = \epsilon^2[1 - \gamma^2(r)]. \quad (2.5)$$

H4—When the dataset is very small, it is often necessary to use supplementary hypotheses for the shape of the correlation function, that is, to impose the shape of the function $\gamma(r)$.

### *c. Statistics and hydrological data*

The database is a set of measures of the real field ($d_i$ at the points $\mathbf{x}_i$, $i = 1, \cdots, N$). The measures are affected by instrumental errors as are all measures. Moreover, the noise is defined as the difference between the data and the objective of the analysis (i.e., the target field), which is different from the real field:

$$\delta_i = d_i - \phi_t(\mathbf{x}_i). \quad (2.6)$$

Thus, the resulting noise affecting the data depends on the relation existing between these two fields. It may include the small-scale variability variance if this relation contains filtering or interannual variability for climatological study of historical dataset.

To optimize the analysis method, the characteristics of this noise should be known. Here we do not have any choice; using a statistical methodology is the only way of representing the characteristics of the noise. (We use statistics basically because it is necessary to take the characteristics of the noise into account.) Let $M_{\mathbf{x}_i}(\delta)$ be the probability distribution of the noise at the point $\mathbf{x}_i$ (i.e., the probability distribution for the observation given the target field). Once more, hypotheses are necessary to make this information computable:

H5—The probability distributions are supposed to be Gaussian.

H6—The mean of $M_{\mathbf{x}_i}(\delta)$ is supposed to be zero.

The distribution $M_{\mathbf{x}_i}(\delta)$ is then entirely determined by only one parameter, the first-order moment, the noise variance $\sigma^2$, and may be written as

$$M_{\mathbf{x}_i}(\delta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2}\frac{\delta^2}{\sigma^2}\right). \quad (2.7)$$

H7—Homogeneity of the characteristics of the noise.

H8—The probability distributions are supposed to be uncorrelated from one point to another.

Nevertheless the data are the only source of information available. Despite the noise, a method must be found to extract the characteristics of the noise and all the statistical information required on the target field from the data. This is the challenge addressed in this paper, because after the computation of this statistical (or probabilistic) knowledge of the field, it becomes possible to adapt the optimization criterion so that it gives the best statistical approximation of the ideal criterion [i.e., to bring the reconstructed field as close as possible to the (unknown) target field].

### *d. Data analysis methods*

The two previous sections have described the prior probability distributions characterizing our knowledge of the target field with respect to the background and the data with respect to the target field. Using Bayes's theorem, we can deduce the posterior probability distribution characterizing our knowledge of the target field everywhere, knowing that we obtained data somewhere. In this view, analysis methods may be interpreted as an (even indirect) application of probabilistic criteria to solve the inversion problem.

The approximation of the ideal criterion by any analysis method is closely related to the knowledge of the probability features of both the field and the noise. But before attempting to compute it, let us examine how the analysis methods operate to take into account this prior information and what the relations between this and the hypothesis on the statistical structure of the field are. On this aspect, we will also briefly compare the two analysis methods in order to eventually be able to choose the most appropriate to our study.

#### 1) OBJECTIVE ANALYSIS

This method is probably the oldest and the most widespread data analysis procedure in oceanography. Originally, the foundation of this method was the minimization of an error estimation, assuming some prior statistical knowledge of the unknown (target) field and the noise. In fact, this knowledge may be directly deduced from the probabilistic distributions, and it has been proven that the method is consistent with the probabilistic point of view presented above (Lorenc 1986). But the procedure only requires the knowledge of the correlation function of the target field and the noise variance. The reduction of the prerequisite information in these two elements is the consequence of the use of the hypotheses: H2, H6, H7, and H8. Thus, the method theoretically enables one to take very precise and particular features of the target field into account. Nevertheless, in practice, the lack of a priori information generally leads one also to admit the hypotheses H3 and H4.

The covariance matrix of the data **A** is defined as

$$A_{ij} = \epsilon^2\gamma(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2\delta_{ij}. \qquad (2.8)$$

The objectively reconstructed field may be written as (Gandin 1963; Bretherton et al. 1976)

$$\phi(\mathbf{x}) = \sum_{j=1}^{N}\sum_{i=1}^{N} A_{ij}^{-1} c(\mathbf{x}_i, \mathbf{x}) d_j. \qquad (2.9)$$

This solution can be interpreted as the mean of the posterior probability distribution of the target field that has (indirectly) been deduced from its supposed prior probability distribution (background information) combined with the information contained in the dataset (with regard to the noise variance). The standard deviation of this probability distribution is also computed by objective analysis in order to produce error maps:

$$e(x) = \epsilon^2 - \sum_{j=1}^{N}\sum_{i=1}^{N} A_{ij}^{-1} c(\mathbf{x}_i, \mathbf{x}) c(\mathbf{x}_j, \mathbf{x}). \qquad (2.10)$$

### 2) Spline analysis

The point of view developed in the spline analysis method is quite different. The criterion is defined as the minimum of a variational principle (Wahba and Wendelberger 1980; McIntosh 1990):

$$J[\phi(\mathbf{x})] = \int_{\mathcal{D}} S[\phi(\mathbf{x})] d\mathbf{x} + \mu \sum_{k=1}^{N} [\phi(\mathbf{x}_k) - d_k]^2, \qquad (2.11)$$

where $S[\;\;]$ is a positive definite smoothing operator. In this case, the most general expression of $S[\;\;]$ is

$$S[\phi] = \sum_{i=0}^{m} \alpha_i L_i[\phi],$$

where

$$L_i[\phi] = \sum_{\omega_1+\omega_2+\cdots+\omega_n=i} \beta(i, \omega_j)\left(\frac{\partial^i \phi}{\partial x_1^{\omega_1}\partial x_2^{\omega_2}\cdots\partial x_n^{\omega_n}}\right)^2 . \qquad (2.12)$$

Moreover, $\mathcal{D}$ is the domain where the analysis is performed, and $\mu$ the parameter that controls the weighing between the smoothing of the solution and its compatibility with the data.

In spite of the completely different features of the procedure, it has been demonstrated that, under certain conditions, this method (norm spline: $\alpha_0 \neq 0$) is exactly equivalent to an objective analysis reconstruction. Indeed, if the domain $\mathcal{D}$ used in the spline method is supposed to be infinite, the solution of the variational principle minimization may be written similarly to that of objective analysis (Wahba and Wendelberger 1980; McIntosh 1990; Brasseur 1994), if we choose the correlation function as

$$c(\mathbf{x}, \mathbf{y}) = \mu\sigma^2 K(\mathbf{x}, \mathbf{y}), \qquad (2.13)$$

where $K(\mathbf{x}, \mathbf{y})$ is the reproducing kernel of the Hilbert space, whose norm is defined by

$$\int S[\phi(\mathbf{x})] d\mathbf{x}. \qquad (2.14)$$

As this kernel is the solution of the following Green equation (McIntosh 1990)

$$\sum_{i=0}^{m} (-1)^i \alpha_i M_i[K(\mathbf{x}, \mathbf{y})] = \delta(\mathbf{x} - \mathbf{y}), \qquad (2.15)$$

where

$$M_i[\;\;] = \sum_{\omega_1+\omega_2+\cdots+\omega_n=i} \beta(i, \omega_j) \times \left(\frac{\partial^{2i}}{\partial x_1^{2\omega_1}\partial x_2^{2\omega_2}\cdots\partial x_n^{2\omega_n}}\right), \qquad (2.16)$$

it is possible to find the Fourier transform of the solution easily when the operator is isotropic:

$$\beta(i, \omega_j) = \frac{i!}{\omega_1!\omega_2!\cdots\omega_n!}. \qquad (2.17)$$

The solution is

$$\tilde{K}(k) = \frac{1}{\sum_{i=0}^{m} \alpha_i k^{2i}}, \qquad (2.18)$$

where $k$ is the norm of the wavenumber and $\tilde{K}$ the Fourier transform of $K(r)$, $r = |\mathbf{x} - \mathbf{y}|$.

Analytically, it is possible to find several shapes of correlation functions consistent with this equation. In other words, in this case, even if this method is completely equivalent to the former, the choice of the correlation function is not completely free any more, it is parametric. Consequently, the application of this method (such as it is presented here) absolutely requires the hypotheses H3 and H4. (Note that although these hypotheses are not necessary in objective analysis they are, nevertheless, generally used.)

In short, from this theoretical comparison, we may say that the equivalence has been established for some parametric forms of the statistical structure in objective analysis and, without taking the boundaries into account, in the spline method. But the advantages of the spline method mainly lie in the use of an efficient finite-element numerical technique to solve the variational problem (see Brasseur et al. 1996). This original implementation of the method is entirely different from the numerical technique used by Wahba and Wendelberger (1980), but it has only been built for two-di-

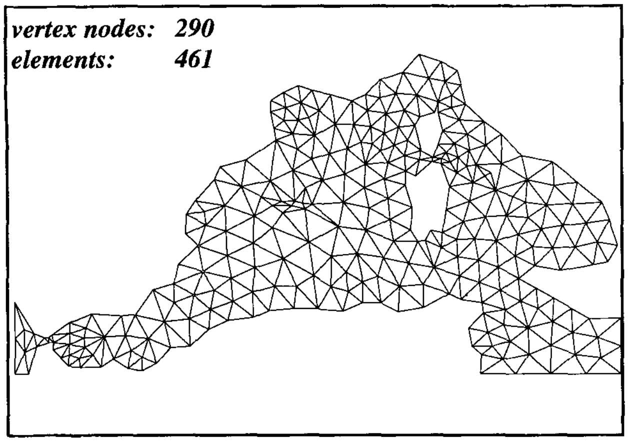

FIG. 1. Finite-element mesh computed to solve the variational problem (spline analysis) on the western Mediterranean Sea.

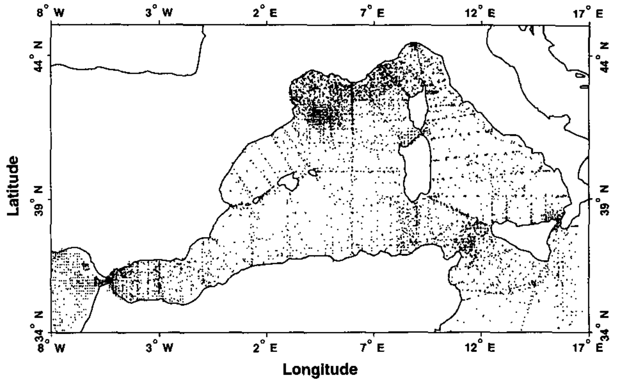

FIG. 3. Horizontal distribution of the casts of the MED2 database for the western Mediterranean Sea.

mensional problems and maximum second-order derivatives that are allowed in the variational principle ($m = 2$).

For example, to cover the western Mediterranean, the finite-element mesh of Fig. 1 has been generated. The grid is composed of 461 triangular finite elements and 1044 (vertex and interface) nodes for 1624 degrees of freedom. On each triangle, the unknown function $\phi^e$ is decomposed as

$$\phi^e(x, y) = \sum_{m=1}^{12} q^e_m W_m(x, y) , \qquad (2.19)$$

where $x$, $y$ is a local coordinate system; $w_m$ are the shape functions (third-order polynomials) used to approximate the solution on the triangle; and $q^e_m$ are the connectors. These connectors are the new unknowns of the minimization problem and they entirely determine the solution. By identifying the connectors between adjacent elements, a predetermined level of continuity (first order) is guaranteed over the whole domain; they constitute the degrees of freedom of the problem.

By introducing the shape (2.19) of the solution in (2.14), using the knowledge of the shape functions, we obtain a quadratic expression for the variational principle in terms of the connectors (where the connectivity of the mesh has been taken into account):

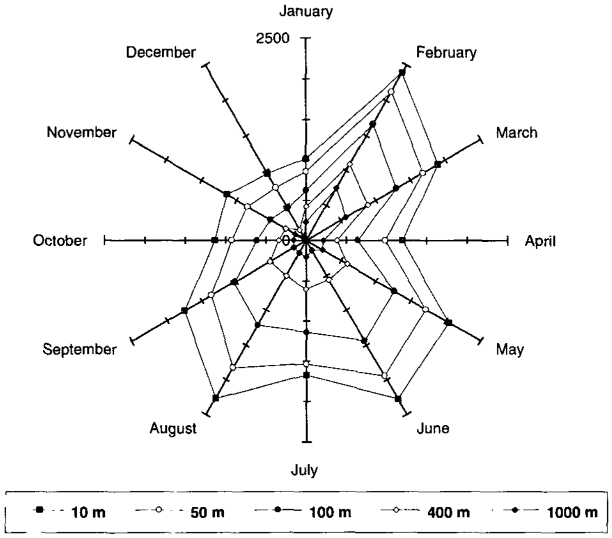

FIG. 2. Size of the database: number of observations available on the western part of the Mediterranean Sea with respect to the depth and the time period (monthly).

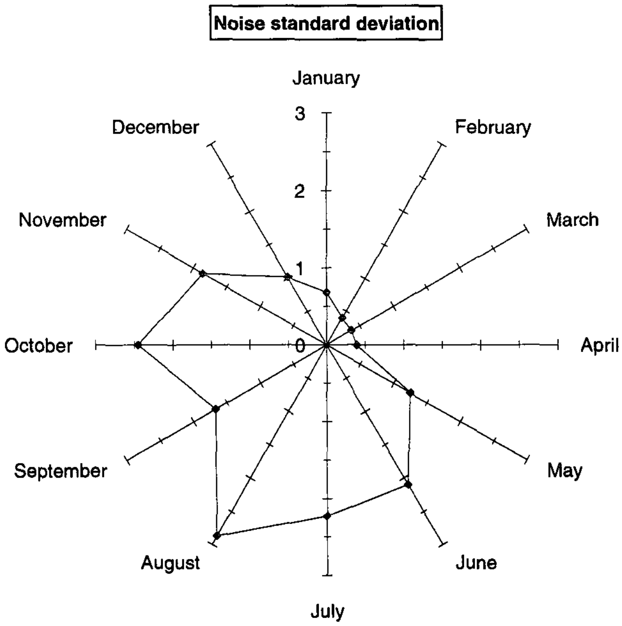

FIG. 4. Month-to-month variability of the noise standard deviation (°C) of the temperature field (for monthly climatological analysis) in the western Mediterranean Sea computed by the classic statistical method.

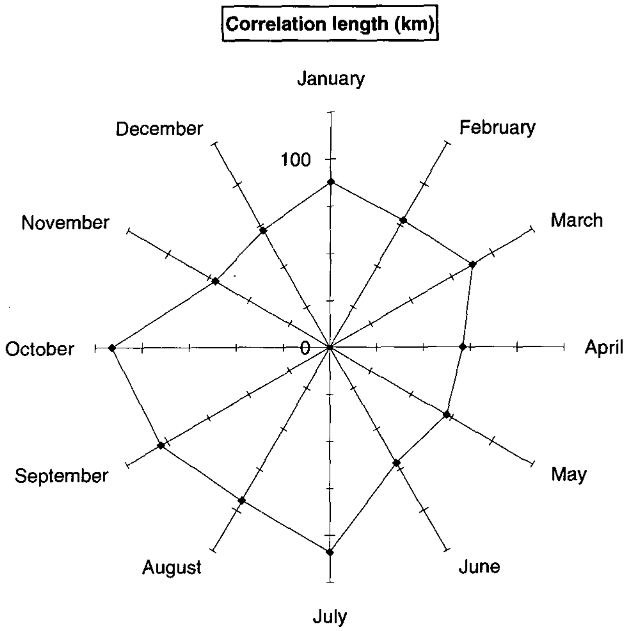

FIG. 5. Month-to-month variability of the correlation length (km) of the temperature field (for monthly climatological analysis) in the western Mediterranean Sea computed by the classic statistical method.

$$J(\mathbf{q}) = \mathbf{q}^{\mathrm{T}}\mathbf{K}\mathbf{q} - 2\mathbf{q}^{\mathrm{T}}\mathbf{g} + \mu \sum_{k=1}^{N} d_k^2, \qquad (2.20)$$

which leads to a linear algebraic system

$$\mathbf{K}\mathbf{q} = \mathbf{g}, \qquad (2.21)$$

where **g** depends on the data ($d_k$). If the elements are correctly sorted, the resulting matrix **K** is very sparse and leads to a computation time (for the resolution of the linear system) roughly proportional to the power $5/2$ of the number of degrees of freedom. Consequently, the main part of computation time does not depend on the size of the dataset. As for objective analysis, it is proportional to the cube of the number of data, which makes the finite-element method especially efficient for large datasets.

Moreover, the result of the finite-element technique is a de facto continuous field. Consequently, there is no need to introduce an additional interpolation operator to deduce the value of the reconstructed field on any location. Finally, the mesh is limited by the boundaries of the real marine domain. In the coastal zones (where the equivalence with objective analysis is no more guaranteed), the method is not rigorously isotropic and homogeneous any more; however, this feature is beneficial to prevent data information from crossing over land barriers (islands, peninsulas, etc.) and makes the solution more realistic. Consequently, this method will be used for the next steps of our work, by taking advantage from its statistical interpretation.

## 3. Cross-validation as optimization criterion: A tool for computing statistical properties from the data

This section aims at providing a reliable method for computing the prior statistical properties of both the field and the noise to optimize the analysis criterion with respect to the particular features of both the system and the objectives of the study. But before dealing with this problem, let us consider a completely different way of approaching the ideal analysis criterion, a procedure (introduced by Wahba 1980) based only on the dataset—cross-validation. The first subsection will describe the interest of this method and its usefulness in the scope of our study. It will be fully justified only in the scope of a statistical interpretation of the procedure in the second subsection.

### *a. Foundations of the method*

Originally, this method has been developed to optimize one (or a few) important parameter(s) characterizing a given data analysis scheme. The procedure consists in successively eliminating one measure from the full database and performing reconstructions with the incomplete datasets thus constituted. The variance

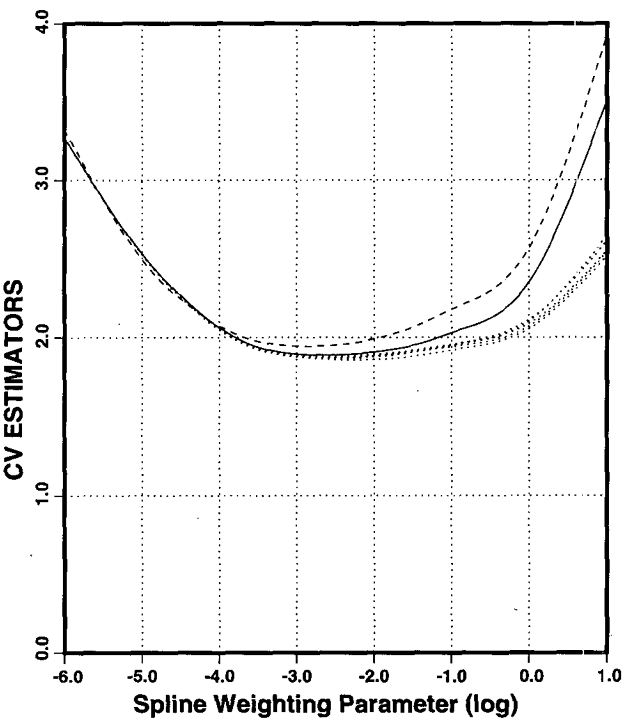

FIG. 6. Various estimations of $\theta$ (cross-validation estimator, in degrees Celsius) with respect to $\mu$ (spline analysis parameter). The dotted curves show the generalized cross-validation estimate $\theta^G$ for five different random vector **z**. Both other curves represent $\theta^S$. The dashed one for samples of 100 observations, and the continuous one for samples of 50 observations. (Summer temperature field at 10-m depth.)

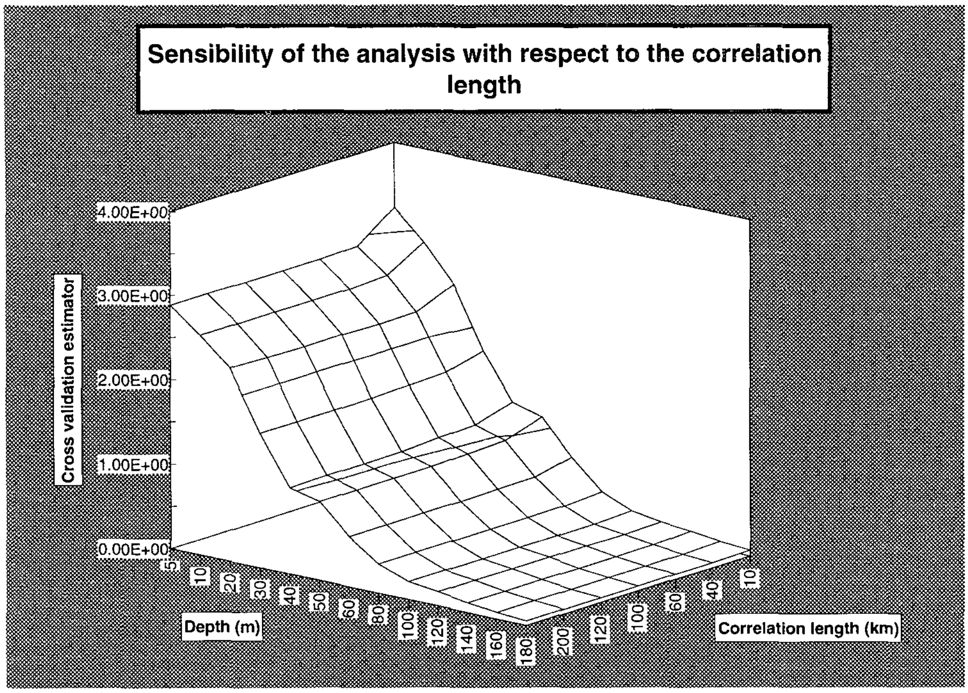

FIG. 7. This figure relates the variation of the optimum presented on Fig. 6 (and for the same test experiment) to the correlation length for some depths.

of the misfits between these reconstructed fields and the corresponding eliminated data can be considered as a statistical indicator of the quality of the analysis. The optimization can then be iteratively achieved by minimizing this indicator with respect to the selected parameter(s) (Wahba 1980). (For example, in the fourth section, this procedure will be used to determine the best background field.)

Moreover, we will show that this optimization technique enables one to compute estimators of the two main statistical features: the background error variance $\epsilon^2$ and the noise variance $\sigma^2$. It will be fairly true if the optimized parameter(s) is (are) much more important than the others with respect to the quality of the reconstruction, so that the quite arbitrary assumptions (H3 and H4) are not of critical importance in these calculations.

At this point, two questions immediately come to mine. 1) Is it not possible to use the classic direct method to compute these statistical properties? 2) Is it still relevant to compute these statistical features if we can directly optimize the analysis method by cross-validation?

The first answer mainly comes from a practical point of view. Indeed, the classic direct method has an important limit. It is only possible to obtain reliable results with a huge dataset that is supposed to be characterized by the same statistical properties. [See, e.g., for meteorological observations, Julian and Thiebaux (1975).] In practice with our database (see section 4), these conditions are only approximately realized near the sea surface. We will nevertheless perform this method to describe its limitations better and to compare the results with those provided by cross-validation estimators.

Second, we claim, as an answer to the second question, that the knowledge of the statistical features of the field and the noise remains interesting in itself. It is then possible to clearly distinguish the statistical study [including quality control, which will be the subject of further studies in the frame of the procedure developed by Lorenc and Hammon (1988)] and the analysis. The reason for this separation is the need for a possibility to extend the use of the statistical parameter thus computed and the database thus controlled to situations for which these tasks are impossible to perform (because of the cost of the procedure or the lack of data). Moreover once cross-validation has been performed, it requires only a very weak supplementary cost. Working out a reliable confirmation of a result always improves it.

### *b. Statistical interpretation of the method*

Let $\phi^{(k)}$ be the field reconstructed when the measure $d_k$ has been withdrawn from the dataset. For each value of the parameter and for each measure $k$, we compute the differences

$$\theta_k = d_k - \phi^{(k)}(\mathbf{x}_k). \qquad (3.1)$$

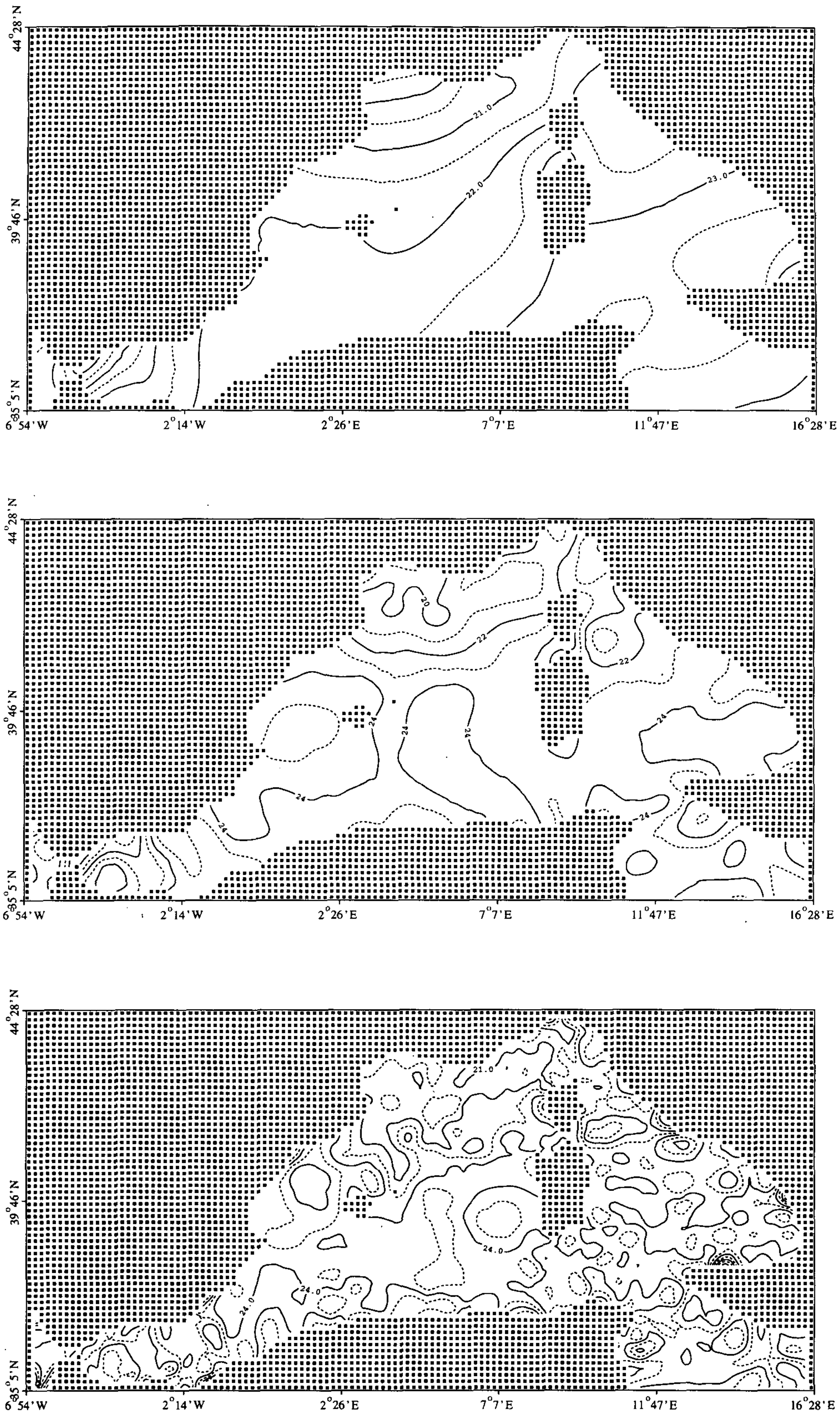

FIG. 8. Summer temperature field (°C) at 10-m depth reconstructed by the spline analysis method. The only difference between these three analyses is the value of the signal-to-noise ratio: (a) 0.1, (b) 1.0, (c) 10.

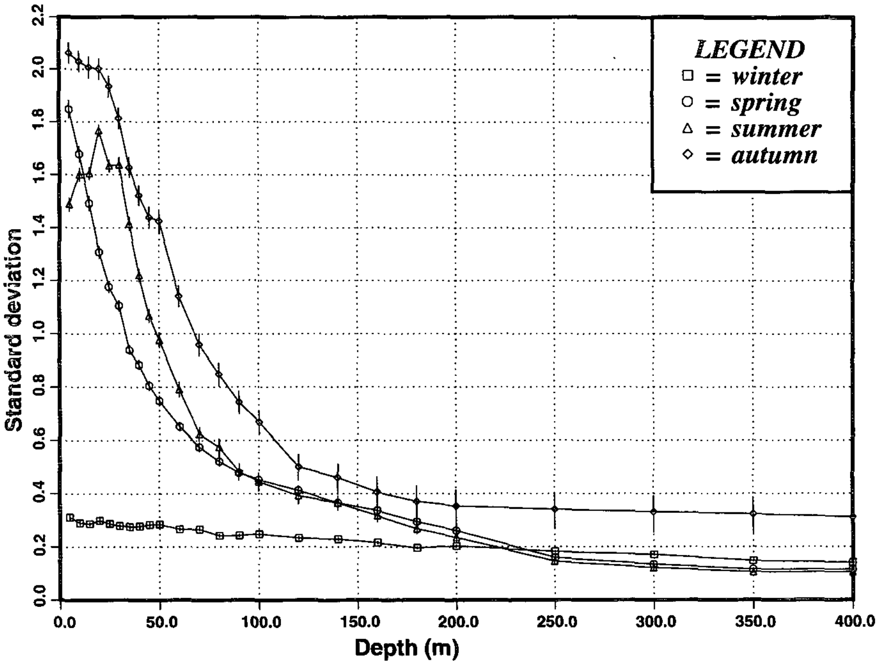

FIG. 9. Noise standard deviation (°C) of the temperature field for seasonal climatological analysis in the western Mediterranean Sea computed by the generalized cross-validation procedure on the basis of the MED2 database.

Craven and Wahba (1979) have shown that the value of the parameter minimizing

$$\theta^2 = \langle \theta_k^2 \rangle = \frac{1}{N} \sum_{k=1}^{N} \theta_k^2 \qquad (3.2)$$

is "an amazingly good estimate" of the optimal parameter (minimizing the true error $e$). The cross-validation optimization criterion (minimization of $\theta$) is thus a good approximation of the ideal criterion with respect to the selected parameter.

As soon as the optimal parameter is computed, we can provide some other estimators useful to compute the error standard deviation. First, let us define

$$\hat{\theta}_k = d_k - \phi^*(\mathrm{x}_k), \qquad (3.3)$$

where $\phi^*$ denotes the optimal reconstruction (using the whole dataset). To interpret this new estimator, which can be easily computed, the objective analysis theory (Gandin 1963; Bretherton 1976) gives an interesting expression of the difference between the data and the optimal solution (where **A** is still the covariance matrix of the data):

$$\hat{\theta}_k^2 = \sigma^4 A_{kk}^{-1}. \qquad (3.4)$$

It should be noted that this is exact whether the hypotheses on the statistical structure of the field are correct or not. It is interesting to use it to link this estimator to the statistical properties that we try to evaluate:

$$\hat{\theta}^2 = \langle \tilde{\theta}_k^2 \rangle = \frac{1}{N} \sum_{k=1}^{N} \tilde{\theta}_k^2 \sim \sigma^4 \frac{\mathrm{Tr}\mathbf{A}^{-1}}{N}. \qquad (3.5)$$

From these two estimators and their interpretation a range for the value of the noise variance can already be given:

$$\sigma^2 \left( \frac{\sigma^2}{\sigma^2 + \epsilon^2} \right) \leqslant \hat{\theta}^2 \leqslant \sigma^2 \leqslant \theta^2 \leqslant \sigma^2 + \epsilon^2 = \omega^2, \qquad (3.6)$$

where $\omega^2$ is the whole data variance with respect to the background. This variance can always be computed easily. (The problem comes from the difficulty of splitting this whole data variance into the noise variance and the background error variance.)

Moreover, it is possible to find an exact relation between these two estimators whether the hypotheses needed to perform the reconstruction are correct or not (Craven and Wahba 1979):

$$\theta_k = \frac{\hat{\theta}_k}{1 - R_{kk}}, \qquad (3.7)$$

where **R** is the influenced matrix of the optimal analysis:

$$\phi^*(\mathrm{x}_k) = \sum_{j=1}^{n} R_{k_j} d_j. \qquad (3.8)$$

Then rewriting (2.9) at the data points $\mathrm{x}_k$ (following Bretherton et al. 1976):

$$\phi^*(\mathrm{x}_k) = \sum_{j=1}^{N} \sum_{i=1}^{N} (A_{ik} - \sigma^2\delta_{ik})A^{-1}_{ij}d_j. \qquad (3.9)$$

By identifying (3.8) and (3.9), we obtain the influence matrix of objective analysis:

$$R_{kj} = \delta_{kj} - \sigma^2 A^{-1}_{kj}. \qquad (3.10)$$

Equation (3.7) consequently becomes

$$\theta_k = \frac{\hat{\theta}_k}{\sigma^2 A^{-1}_{kk}}. \qquad (3.11)$$

By eliminating $A^{-1}_{kk}$ between (3.4) and (3.11), we directly deduce

$$\sigma^2 = \theta_k\hat{\theta}_k. \qquad (3.12)$$

If the optimum is known (by minimization of $\theta^2$, if the basic hypotheses are quite correct), we may suppose that this always gives the true value of the noise standard deviation [always coherent with (3.6)]. However in practice we observe some variations with respect to $k$. Thus, it is preferable to average this estimator over the dataset and write

$$\sigma^2 \sim \zeta = \langle\theta_k\hat{\theta}_k\rangle. \qquad (3.13)$$

Each of these estimators is computed as the mean $\bar{\chi}$ of a given number $n$ of realizations of a random variable $\chi_k$. In this case the variance of the error on the result ($\eta_\chi$) may be estimated from the variance of this random variable

$$\eta_\chi^2 = \frac{\langle(\chi_k - \bar{\chi})^2\rangle}{n}. \qquad (3.14)$$

Once more, computing $\epsilon^2$ and $\sigma^2$ from these estimators is an inverse problem. It is possible to show that the resulting probability distribution for the solution is (Tarantola 1987)

$$p(\epsilon^2, \sigma^2) = \exp[-J(\epsilon^2, \sigma^2)], \qquad (3.15)$$

where

$$J(\epsilon^2, \sigma^2) = \frac{1}{\eta_\zeta^2}(\sigma^2 - \zeta)^2 + \frac{1}{\eta_\omega^2}(\epsilon^2 + \sigma^2 - \omega^2)^2 + \begin{cases} \frac{1}{\eta_{\hat{\theta}}^2}(\sigma^2 - \hat{\theta}^2)^2, & \sigma^2 < \hat{\theta}^2 \\ 0, & \hat{\theta}^2 \leqslant \sigma^2 \leqslant \theta^2 \\ \frac{1}{\eta_\theta^2}(\sigma^2 - \theta^2)^2, & \theta^2 > \sigma^2 \end{cases} \qquad (3.16)$$

from which the solution (mean and standard deviation) may be easily computed.

### c. *Derived estimates: Sampling and generalized cross-validation*

Performing cross-validation using (3.1) and (3.2) poses an important difficulty—the numerical cost. In-

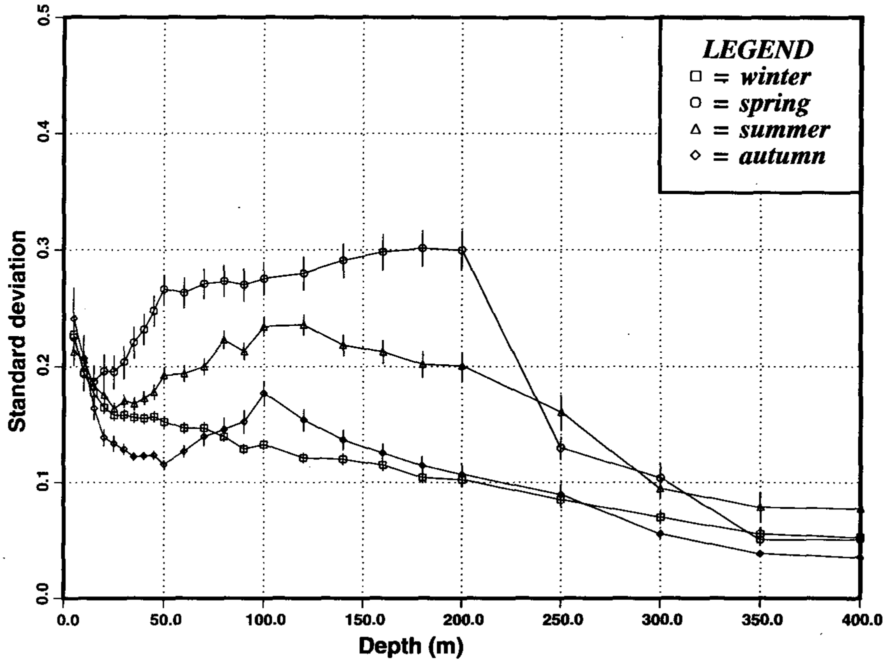

FIG. 10. Noise standard deviation (psu) of the salinity field for seasonal climatological analysis in the western Mediterranean Sea computed by the generalized cross-validation procedure on the basis of the MED2 database.

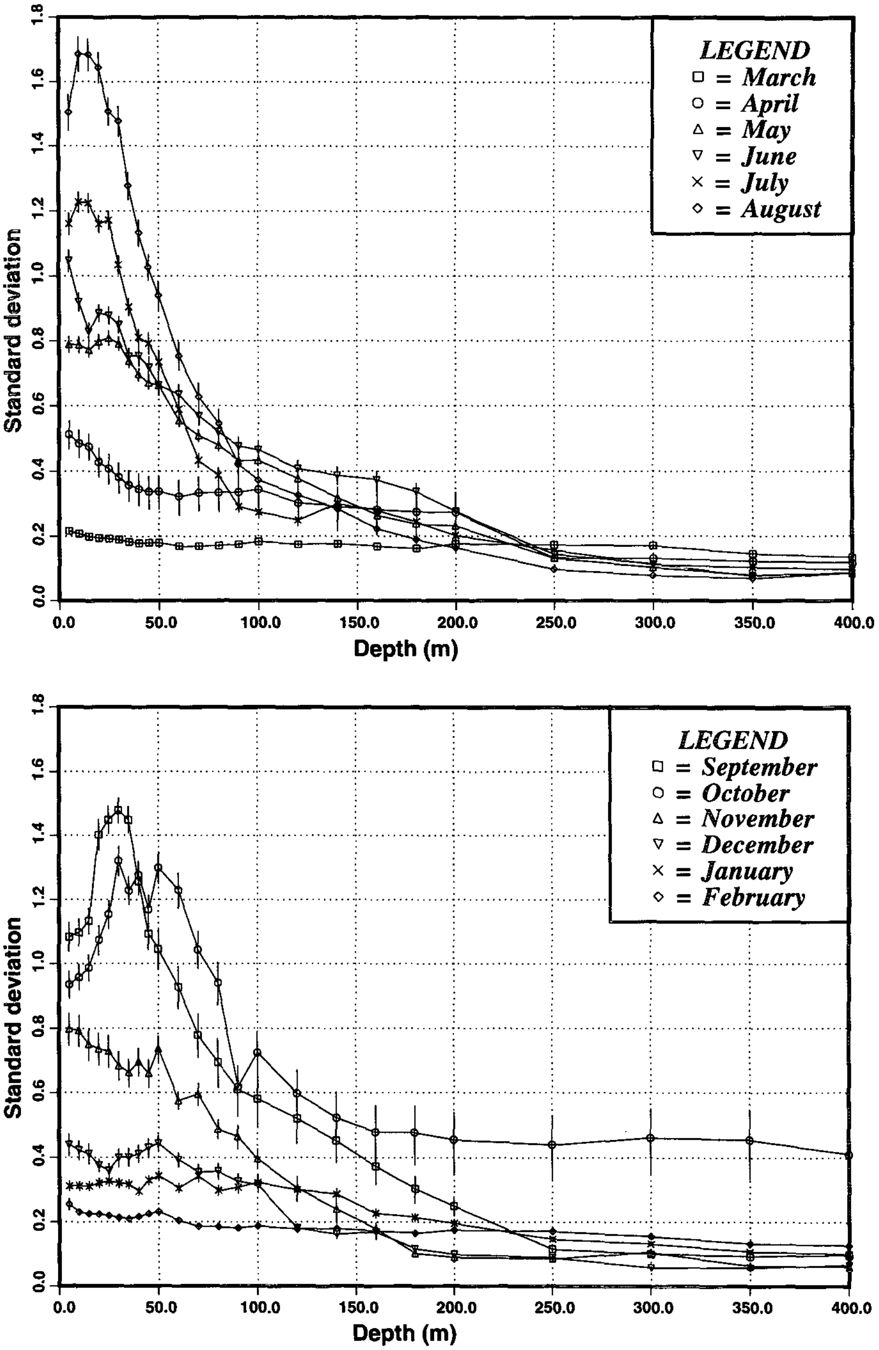

FIG. 11. Noise standard deviation (°C) of the temperature field for monthly climatological analysis in the western Mediterranean Sea computed by the generalized cross-validation procedure on the basis of the MED2 database. (a) Increase from March to August. (b) Decrease from September to February.

deed this method requires as many reconstructions as there are measures in the dataset, for each parameter. So, it is inapplicable on a large scale in this original shape.

This problem can be solved in two ways. First, by using the generalized cross-validation estimate (Craven and Wahba 1979; Golub et al. 1979). It has been shown that

$$\theta^G = \frac{\hat{\theta}}{N^{-1}\,\mathrm{Tr}(\mathbf{I} - \mathbf{R})} \qquad (3.17)$$

may be considered as a worthwhile alternative to $\theta$.

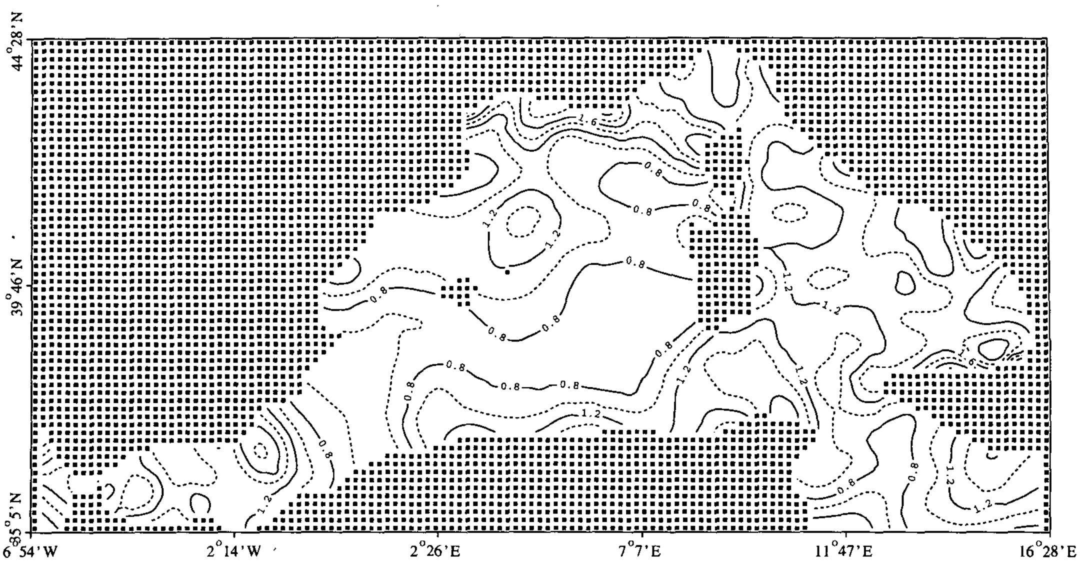

FIG. 12. Map of the noise standard deviation (°C) of summer temperature at 10-m depth.

Even if the influence matrix $R$ is accessible only with high numerical cost, the trace of the matrix may be efficiently computed by the randomized trace estimation method (Wahba 1990):

$$\theta^G \sim \frac{\hat{\theta}}{\frac{1}{N}(\mathbf{z}^\mathrm{T}\mathbf{z} - \mathbf{z}^\mathrm{T}\mathbf{R}\mathbf{z})}, \qquad (3.18)$$

where $\mathbf{z}$ is a Gaussian random vector (whose mean is zero and standard deviation is one) that may easily be supplied by an automatic random number generator. Equation (3.18) requires only one supplementary reconstruction to compute $\theta^G$ from $\hat{\theta}$. Indeed, we use the analysis procedure to compute $\mathbf{Rz}$ from $\mathbf{z}$. In practice, we will observe that this efficient procedure definitely provides a very reliable estimator that can advantageously replace $\theta$.

Nevertheless, we will examine a second way of reducing the cost of (3.1) and (3.2). Instead of only eliminating one measure from the database for each reconstruction, we may eliminate a *sample* of data. We define $\phi^{(S_p)}$ as the reconstructed field when the measures $d_k$ included in the sample $S_p$ ($p = 1, \cdots, N_S$; $N_S \ll N$) have been withdrawn from the dataset:

$$\theta_k^S = d_k - \phi^{(S_p)}(\mathbf{x}_k), \quad d_k \in S_p \qquad (3.19)$$

and following

$$\theta^{S^2} = \langle \theta_k^{S^2} \rangle. \qquad (3.20)$$

If the data in each sample are scattered and few enough within the whole dataset, we may neglect the difference between $\phi^{(k)}(\mathbf{x}_k)$ and $\phi^{(S_p)}(\mathbf{x}_k)$. Consequently, it will also be valid to replace $\theta_k$ and $\theta$ by their estimation using sampling $\theta_k^S$ and $\theta^S$.

The quality of this approximation also depends of the parameter of the analysis method. Indeed the shorter the characteristic length of the reconstructed field becomes (when $\mu$ is increasing, for example), the less the sampling approximation fits the true value of $\theta$. Consequently, the objective of the sampling guidelines is to ensure the quality of the approximation at least for the parameters inferior or equal to the optimal value.

On the other hand, the cost of the procedure is directly related to the number of subsets. (The number of reconstructions required is equal to the number of samples used.) The problem of determining the smallest number of samples ensuring the quality of the approximation may be solved in the following way. The whole dataset is divided into subsets by progressively splitting the whole physical domain into rectangular boxes. The subsets are then generated by selecting one observation per box. The procedure enables one to take several types of condition into account. For example, we may fix the number and the size of the samples or introduce a minimum length separating two measures included in the same sample and a maximal size ratio for any sample with respect to the dataset.

In the latter case, all the data cannot always be included in the subsets without getting too many little samples. (In some situations, a great amount of data is concentrated in a small area.) But this is not absolutely necessary to perform the cross-validation. The objective is to provide a statistical set *large* and *representative* enough to give sense to the estimators. For that matter, in order to test the reliability of these estimators, the method enables us to compute statistical confidences for each of them. In short, a few samples of correctly scattered data could form a statistical set sufficient to perform a reliable and much cheaper cross-validation.

At this point, there is still no reason to prefer the sampling procedure over the generalized cross-validation. The latter provides an equivalent, if not better, estimate at lower cost. The only drawback of the generalized cross-validation procedure is that it provides only a *global* estimate. It does not matter as far as we do not intend to give up the hypothesis of homogeneity of the statistical structure. But, in the context of the analysis of oceanic hydrological fields, it may become interesting to relax this hypothesis, replace the means in (3.2), (3.5), or (3.13) by a smoothing reconstruction using the spline method, and detect by the way some inhomogeneous characteristics of the statistical structure. Once more, an objective way of evaluating the correct degree of smoothness to give to the solution is cross-validation. It will prevent us from exaggerating the quantity of information that can be extracted by this procedure and allow us to infer all significative patterns.

## 4. Application to a historical database of the western Mediterranean Sea

In this section, the general ideas above developed are applied to solve a practical problem. In the frame of the study of the general circulation in the Mediterranean Sea (Brasseur and Haus 1991), reliable climatological analyses of the hydrological data are constantly needed. (Whether for model initialization and validation or for data assimilation purpose, such fields have to be based on reliably computed information.) The target fields of our study have been defined to contain the general circulation pattern of the sea only. (In fact, only horizontal sections will be reconstructed and then superposed to obtain the 3D field.) It is thus supposed that all small-scale phenomena have to be filtered off. Moreover, they are supposed to characterize the climatology of a given period of the year (month or season). Thus, its relation with the real evolutive field includes averaging over this period from year to year. As a consequence, we must expect that the noise affecting the data will include, in addition to the observational error variance, small-scale phenomena and interannual variability. (The latter term may be considered as the largest part of the noise variance.)

### *a. Description of the database*

To perform such climatological analysis, a large historical database is needed for the western Mediterranean Sea. But, up to now, no exhaustive historical database has been compiled with all the in situ data collected. Nevertheless, two major files containing a substantial part of the experimental work are available for the purpose of this study: the French Bureau National des Données Océaniques (Brest, France) file that exclusively contains hydrographic casts for the Mediterranean region, and the National Oceanographic Data Center (Washington, D.C.) file, which is a subset of a larger World Ocean data bank. The MED2 database (used in this work) has been obtained by merging of these two original datasets (Brasseur and Haus 1991). However, in the future, the quality of the data bank (and consequently the results of the analyses) will greatly benefit from the MODB project (briefly described in the introduction).

Whatever the data analysis procedure is, the quality of the reconstruction and the reliability of the statistical study strongly depend on the data density. Figure 2 shows the total number of data available on the western part of the sea with respect to depth and time period. On the other hand, Fig. 3 gives a typical example of the horizontal distribution of the data.

### *b. Choosing the background field*

As we have said in the second section, the background field is the best prior estimate of the target field. The whole analysis procedure will be performed on the data anomalies with respect to the background field. The choice of this field is thus preliminary to any analysis treatment of the data.

For lack of better keys, it could be chosen as the mean or the linear regression of the data, but in a domain as large as the Mediterranean, where the (target) field may show important variations from one region to another, these simple solutions would lead to systematic errors in the solution.

An interesting alternative could be to compute the background field using the seminorm spline reconstruction procedure [i.e., using (2.14) and (2.15) without an underived term in the smoothing operator: $\alpha_0 = 0$]. There is an important restriction in the use of this method: the background field must be smoothed enough so as not to be influenced by the noise affecting the data.

The cross-validation procedure enables us to test the relative quality of different analysis procedures. So it will also be possible to study the influence of the background field on the best analysis that can be computed from it. Such an experience shows that the quality of the final result can be significantly improved if the background field has been carefully chosen, and the first tests tend to prove that the very best analysis is obtained by choosing the optimal norm-spline reconstruction (in the sense of the cross-validation criterion) as a background field. Thus, as far as the background field is smooth enough, it will not significantly influence the computation of the noise standard deviation.

### *c. Overview of the classic method—Results*

Classically, estimations of the statistical features of the target field and the noise are directly computed by averages over the whole dataset (or over area where these properties may be considered as homogeneous). For this purpose, we consider all pairs of data that can be formed in the dataset. Each of these pairs enables the computing of an element of the statistical set whose mean would be the value of the correlation function for the distance separating the data of the pair. With the pairs for which the correlation may be a priori considered to be maximal (because of the short spatial distance separating the data), it is possible to give an estimation of the noise variance.

Despite the various refinements in the computation of the statistical structure from these multiple interdependent statistical sets, this method will only produce reliable results if these sets are very large, especially for very densely distributed data. In the frame of our dataset in the scope of the western Mediterranean Sea study, this condition is only sufficiently guaranteed at low depth (<50 m). Moreover, even in this case, the procedure cannot be automatic; the results have to be continuously checked.

For further comparison with cross-validation, on Figs. 4 and 5, we have given results obtained at 10-m depth for monthly historical datasets using the classic procedure. Figure 4 shows the month-to-month variability of the noise variance. Concerning the correlation function, we compute only a main feature: the correlation length. [Here, it has been defined as the length $l$ for which the function $\exp(-r^2/l^2)$ best suits the computed shape for short distances $r$.]

### d. *Cross-validation of an analysis method*

As we have observed in the previous section, the cross-validation procedure mainly consists of finding the method parameters leading to the minimization of the $\theta$ estimator. Fortunately, one parameter is much more important than the others: the parameter $\mu$ of the variational problem, which is directly related to the noise-to-signal ratio of the objective analysis procedure [Eq. (2.16)]. Figure 6 shows some estimations of the $\theta$ estimator with respect to the parameter $\mu$ for a test experiment (summer temperature field at 10-m depth); a minimum can be easily computed. The dotted curves show the generalized cross-validation estimate $\theta^G$ for several (five) different random vectors $\mathbf{z}$. The two other curves are two cross-validation estimates using the sampling approximation $\theta^S$. The continuous one with samples of 50 observations is the most precise; the dashed one has been computed with samples of 100 observations. As explained in section 3, there is a discrepancy between the different curves for large values of $\mu$: the larger the spline weighting parameter, the shorter the characteristic length of the reconstructed field, and the poorer the quality of the sampling approximation. Anyway, it does not matter, as far as the optimum is well represented.

To examine the influence of other parameters, Fig. 7 shows the evolution of this optimum with respect to the correlation length for some depths. The weak dependence (except for very short correlation lengths) with respect to this parameter has led us to fix it to a constant and quite arbitrary value (80 km: according to the direct analysis) for the following applications. This intriguing weak dependence with respect to the correlation length mainly comes from the high density of the dataset. As a consequence, the climatological main structures are well represented by the observations themselves and a horizontal characteristic length of the analysis does not influence the solution any more.

On the contrary, the crucial importance of the signal-to-noise ratio should not be denied (see also Provost 1987). To illustrate this feature on the shape of the solution, Fig. 8 displays three analyses of the same data subset (corresponding to summer situation) for different parameters $\mu$ of the spline method (which is directly related to the signal-to-noise ratio). The medium value corresponds to the optimum deduced from cross-validation: Fig. 6. The variation in the shape of the solution can be linked to the spectral interpretation of the method: the higher the signal-to-noise ratio is supposed to be, the less the solution has to be smoothed.

### e. *Cross-validation to compute statistics*

Finally, to achieve the goal of this paper, the cross-validation procedure should be used to compute the statistical structure in the particular situation presented above. By a rigorous application of the method, we obtain the following results.

First, let us consider the case of the reconstruction of seasonal climatological fields. Figures 9 and 10 show the variation with depth of the *noise standard deviation* for temperature and salinity, respectively. This has been computed using the generalized cross-validation estimator from the dataset covering the whole basin. Then, for monthly climatological analysis Figs. 11a and 11b show the noise standard deviation as a function of depth.

Some general features of the noise variability may be deduced from these figures. First, in depth (below 200-m depth), the noise standard deviation is quite constant throughout the year. (The results obtained for October should be considered as doubtful in depth because of the lack of data.) Near the surface it is more important during summer than during winter. This feature is linked either to more important small-scale phenomena during summer or to a larger interannual variability. Figure 11a shows the increase in the standard deviation from March to August, whereas Fig. 11b depicts its decrease from August to March (above 200 m).

It should also be noted that the confidence range has been computed from the probability distribution of the solution with a confidence of 95%. However, this estimation does not take the bias introduced by the lack of representativeness of the dataset itself into account. (This would have been impossible.) So these confidence ranges are only exact if the database is an unbiased sample of reality, which is never the case. So, they have to be considered with care.

Finally, Fig. 12 gives an example of the use of the sampling procedure (followed by a spline reconstruction in place of only averaging) to explore the inhomogeneities in the noise standard deviation across the Mediterranean Sea. Once more, we give the example of the summer temperature field at 10-m depth. The reconstruction of this field from local information has been optimized using the generalized cross-validation procedure.

## 5. Conclusions

As a conclusion, the application of the cross-validation procedure on the MED2 database to compute the main statistical features of both the target field and the noise clearly shows that the method is efficient to select the criterion defining any analysis method. In fact, there is a second "cross-validation": a direct application of the cross-validation method allowed to determine the optimum that may now be validated and extended to other situations by using its statistical interpretation. (It should be remembered that this procedure is greatly simplified by the predominant importance of only one parameter.)

From the numerical experiments we can also repertoriate some specific features of the method. The first one is the reliability of the results compared to those of the direct method. Afterward, at first sight, the numerical cost of cross-validation is very high. But we have showed that some developments of the procedure enable a substantial reduction of the cost and give this method a great flexibility. The trouble caused by the numerical cost of cross-validation is also strongly reduced by using the numerically efficient spline method (which consists of minimizing a variational principle with a finite-element technique) to fulfill the analysis.

This benefit would have been lost without a statistical interpretation of the spline method emerging from an equivalence theorem with objective analysis. Indeed, we have to compute some statistical properties of the particular situation that can be linked to the parameters of the analysis method. For that purpose it has been demonstrated that the probabilistic point of view was convenient and coherent with the objective analysis requirements and, as a consequence, with the spline analysis method. A complete link between the way of expressing these properties (probabilities) and their reliable computation (cross-validation) has consequently been set up thanks to the mathematical structure (and numerical features) of the two most widespread analysis methods: objective analysis and spline analysis.

*Acknowledgments.* The authors are greatly indebted to Prof. J. Nihoul for his permanent support and advice. This work was partially funded by the Commission of European Communities (Grants MAS2 CT93.0055 and MAS2 CT93.0075.BE) and the National Fund for Scientific Research of Belgium.

### REFERENCES

Brasseur, P., 1991: A variational inverse method for the reconstruction of general circulation fields in the Northern Bering Sea. *J. Geophys. Res.,* **96,** 4891–4907.

——, 1994: Reconstitution de champs d'observations océanographiques par le Modèle Variationnel Inverse: Méthodologies et Applications. Ph.D. dissertation, University of Liège, 262 pp.

——, and J. Haus, 1991: Aplication of a 3-D variational inverse model to the analysis of ecohydrodynamic data in the Northern Bering and Southern Chuckchi Seas. *J. Mar. Syst.,* **1,** 383–401.

——, J. M. Beckers, J. M. Brankart, and R. Schoenauen, 1996: Seasonal temperature and salinity fields in the Mediterranean Sea: Climatological analyses of an historical data set. *Deep-Sea Res.,* in press.

Bretherton, F., R. Davis, and C. Fandry, 1976: A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Res.,* **23,** 559–582.

Craven, P., and G. Wahba, 1979: Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Num. Math.,* **31,** 377–403.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields.* Israel Program for Scientific Translations, 242 pp.

Golub, G. H., M. Heath, and G. Wahba, 1979: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21,** 215–223.

Julian, P., and H. J. Thiebaux, 1975: On some properties of correlation functions used in optimum interpolation schemes. *Mon. Wea. Rev.,* **103,** 603–616.

Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.,* **112,** 1177–1194.

——, and O. Hammon, 1988: Objective quality control of observations using Bayesian methods—Theory and practical implementation. *Quart. J. Roy. Meteor. Soc.,* **114,** 515–543.

McIntosh, P., 1990: Oceanographic data interpolation: Objective analysis and splines. *J. Geophys. Res.,* **95,** 13 529–13 541.

Provost, C., 1987: The variational inverse method revisited. *Ann. Geophys.,* **5B**(3), 213–220.

Tarantola, A., 1987: *Inverse Problem Theory.* Elsevier, 612 pp.

Wahba, G., 1990: *Spline Models for Observational Data.* SIAM, 169 pp.

——, and J. Wendelberger, 1980: Some new mathematical methods for variational objective analysis using splines and cross-validation. *Mon. Wea. Rev.,* **108,** 1122–1143.