Maximum-Likelihood Estimation of Forecast and Observation Error Covariance Parameters. Part I: Methodology

DICK P. DEE

General Sciences Corporation, Laurel, Maryland, and Data Assimilation Office, NASA/Goddard Space Flight Center, Greenbelt, Maryland

ARLINDO M. DA SILVA

NASA/Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 19 November 1997, in final form 11 August 1998)

ABSTRACT

The maximum-likelihood method for estimating observation and forecast error covariance parameters is described. The method is presented in general terms but with particular emphasis on practical aspects of implementation. Issues such as bias estimation and correction, parameter identifiability, estimation accuracy, and robustness of the method, are discussed in detail. The relationship between the maximum-likelihood method and generalized cross-validation is briefly addressed.

The method can be regarded as a generalization of the traditional procedure for estimating covariance parameters from station data. It does not involve any restrictions on the covariance models and can be used with data from moving observers, provided the parameters to be estimated are identifiable. Any available a priori information about the observation and forecast error distributions can be incorporated into the estimation procedure. Estimates of parameter accuracy due to sampling error are obtained as a by-product.

1. Introduction

This paper concerns the estimation of unknown forecast and observation error covariance parameters for an atmospheric data assimilation system from observational residuals. The method we present is based on maximum-likelihood covariance parameter estimation as described by Dee (1995). In a companion paper (Dee et al. 1999, hereafter referred to as Part II) we describe three different applications, involving univariate as well as multivariate covariance models, and data from both stationary and moving observing systems.

The simplest example of a covariance parameter, which is required for atmospheric data assimilation, is the error standard deviation of rawinsonde height measurements at a fixed pressure level. The traditional method (Gandin 1963; Rutherford 1972) for estimating this parameter computes the least squares fit of an isotropic correlation model to the sample spatial correlations of a collection of *observed-minus-forecast residuals.* These are the differences between the heights reported by the instrument and the corresponding (interpolated) predictions obtained from a short-term forecast. The data typically span a period of 1–3 months, and involve quality-controlled reports from a fixed set of stations. The sample correlations among all pairs of stations can be plotted as a function of separation distance, together with a curve representing a fitted correlation model (see Fig. 1). By extrapolating this curve to the origin one can determine the ratio between the observation and forecast error standard deviations.

This procedure can be generalized in several ways. For example, wind error standard deviations can be estimated by separating observed-minus-forecast wind residuals into their longitudinal and transverse components (Daley 1985). Hollingsworth and Lönnberg (1986) and Lönnberg and Hollingsworth (1986) used this technique in their extensive and groundbreaking study of the statistical structure of multivariate forecast errors. They also showed how the same procedure can be used to estimate the vertical correlations of forecast and observation errors from multilevel observed-minus-forecast residuals. Additional applications involving anisotropic, multivariate, and nonseparable covariance models have been described by Thiébaux et al. (1986; 1990), Bartello and Mitchell (1992), and Devenyi and Schlatter (1994).

A different method must be used to study the error characteristics of observations originating from moving

Corresponding author address: Dr. Dick P. Dee, NASA/GSFC Data Assimilation Office, Mail Code 910.3, Greenbelt, MD 20771.
E-mail: ddee@dao.gsfc.nasa.gov

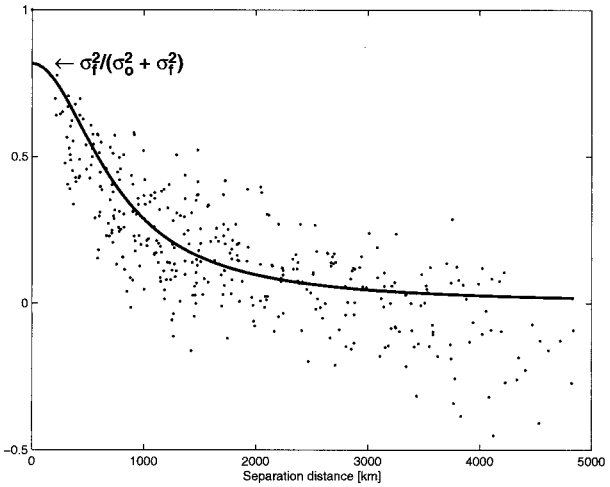

FIG. 1. Illustration of the traditional method for determining observation and forecast error standard deviations from observed-minus-forecast residuals, by extrapolating an isotropic correlation model fit to zero separation distance. Each dot represents the sample correlation between 500-hPa residuals at a pair of North American rawinsonde stations producing at least 50 simultaneous nighttime reports during the month of February 1995. See Part II for details.

platforms such as ships, aircraft, and satellites. These data compose the vast majority of atmospheric observations and potentially contain a wealth of information about forecast errors as well. The observations represent the *only* real source of information about forecast errors: all other benchmarks such as verifying analyses are derived from them. For this reason we feel it is important to develop additional techniques for extracting maximum information from the observations themselves.

The purpose of this paper is to present a flexible and mathematically rigorous covariance parameter estimation method that does not involve a priori restrictions on the nature of either the observing system or the covariance model. We estimate the covariance parameters by maximizing the likelihood function of the parameters given the data. Advantages of the maximum-likelihood method over other estimation methods are that 1) it can incorporate information about the error distributions to the extent that such information is available and 2) it is consistent with current operational atmospheric data assimilation systems, all of which may be regarded as particular implementations of the maximum-likelihood method to the problem of estimating the state of the atmosphere from observations and model information (Lorenc 1986; Cohn 1997). Furthermore, the maximum-likelihood method generates useful information about the accuracy of the parameter estimates along with the estimates themselves.

In this study as well as in Part II we especially try to address the limitations of the method, some of which are intrinsic to the problem at hand. For example, the traditional method described earlier relies on the basic assumption that the spatially uncorrelated component of the residual arises from observation errors, while the spatially correlated component represents forecast errors. This illustrates two fundamental aspects of the problem of estimating the error statistics for one particular source of information by comparing it with another. The first is that the error distributions associated with each source must be sufficiently different in order that their respective parameters can be reliably estimated from the residuals. This is the issue of *parameter identifiability* that we discuss in section 4c: *no* method can produce meaningful estimates of poorly identifiable parameters.

The second fundamental aspect of the problem is that it is necessary to describe each of the error distributions in sufficient detail, in order that the only remaining unknowns are the parameters one wishes to estimate. The simple example described earlier actually involves a myriad of simplifying assumptions about forecast and observation errors, such as their statistical independence, their local homogeneity in space and time, and the isotropy of forecast error correlations. This raises the question of *robustness* of the parameter estimates; that is, the sensitivity of the estimates to the various modeling assumptions involved in the formulation of the problem. For example, it can be seen from Fig. 1 that the estimate of the observation and forecast error standard deviations obtained by extrapolation must depend, to some extent, on the choice of the isotropic model used to represent the forecast error correlations. We discuss this issue in section 4e and devote considerable attention to it in Part II.

The organization of this paper is as follows. In section 2 we define forecast and observation errors and discuss covariance modeling in general terms. Appendix A describes a few different isotropic correlation models that we use here and in Part II. In section 3 we describe the relationship between the covariance models and the data, represented by the observed-minus-forecast residuals. We also consider the possibility of using residuals obtained from two separate observing instruments. The heart of this paper is in section 4, where we discuss in detail the application of the maximum-likelihood method to the problem of estimating any unknown covariance parameters from the data. We briefly refer to the generalized cross-validation method (Wahba and Wendelberger 1980), which is reviewed in appendix B. Section 5 contains concluding remarks.

2. Covariance models

Suppose that the n-vector $\mathbf{w}_k^f$ is a model forecast valid at time t_k, and $\mathbf{w}_k^t$ is the unknown true state of the atmosphere at that time. It is convenient to define both quantities in terms of the same state representation: $\mathbf{w}_k^t$ is an n-vector as well, containing, for example, the true gridpoint values or spectral coefficients. The *forecast error* is then simply

$$\boldsymbol{\epsilon}_k^f \equiv \mathbf{w}_k^f - \mathbf{w}_k^t. \tag{1}$$

For a p_k-vector $\mathbf{w}_k^0$ of measurements generated by a particular instrument at time t_k, the *observation error* is defined by

$$\boldsymbol{\epsilon}_k^0 \equiv \mathbf{w}_k^0 - \mathbf{h}_k(\mathbf{w}_k^t). \tag{2}$$

The nonlinear p_k-vector function $\mathbf{h}_k$ is the *discrete forward observation operator* (e.g., Cohn 1997), mapping model variables to the data type associated with the instrument.

We introduce the following notation for the forecast error mean and covariance

$$\mathbf{b}_k^f \equiv \langle \boldsymbol{\epsilon}_k^f \rangle, \qquad \mathbf{P}_k^f \equiv \langle (\boldsymbol{\epsilon}_k^f - \mathbf{b}_k^f)(\boldsymbol{\epsilon}_k^f - \mathbf{b}_k^f)^{\mathrm{T}} \rangle, \tag{3}$$

for the observation error mean and covariance

$$\mathbf{b}_k^0 \equiv \langle \boldsymbol{\epsilon}_k^0 \rangle, \qquad \mathbf{R}_k \equiv \langle (\boldsymbol{\epsilon}_k^0 - \mathbf{b}_k^0)(\boldsymbol{\epsilon}_k^0 - \mathbf{b}_k^0)^{\mathrm{T}} \rangle, \tag{4}$$

and for the cross-covariance between observation and forecast errors

$$\mathbf{X}_k \equiv \langle (\boldsymbol{\epsilon}_k^0 - \mathbf{b}_k^0)(\boldsymbol{\epsilon}_k^f - \mathbf{b}_k^f)^{\mathrm{T}} \rangle. \tag{5}$$

The notation $\langle \cdot \rangle$ used in (3)–(5) denotes the *ensemble averaging* or *expectation operator*, whose proper definition involves the (typically unknown) joint probability distribution of forecast and observation errors.

The observation operator $\mathbf{h}_k$ and its associated observation error distribution are different for each data type. It is of course possible, and sometimes convenient, to combine all available observations at a time t_k into the observation vector $\mathbf{w}_k^0$. Throughout this paper the vector $\mathbf{w}_k^0$ will always denote a specific subset of the observations, obtained by restriction to a single observing system and to a limited region in space. It will be clear from the context which restriction is implied. We will also have occasion to consider simultaneous observations obtained from two different instruments, and in that case the notation will be suitably generalized.

All operational data assimilation systems rely on approximate information about error means and covariances. In practice (3)–(5) are modeled by introducing various simplifying assumptions about the underlying error distributions. For example, the means $\mathbf{b}_k^f$ and $\mathbf{b}_k^0$ are often disregarded, amounting to the assumption that the forecast model as well as the observing instruments are unbiased. Also, for most data types it is assumed that observation errors and forecast errors are mutually independent, which is equivalent to taking $\mathbf{X}_k \equiv 0$. This assumption ignores the fact that both types of error depend on the local state of the atmosphere and must therefore be correlated with each other. The most familiar example of state-dependent observation error is representativeness error (Daley 1993); model error is also easily shown to be state dependent [Dee 1995, Eqs. (4) and (5)]. We will mention other assumptions about errors associated with specific data types below and in Part II. Some are not necessarily realistic, but in practice the information required to remove them is usually lacking.

In general, theoretical statements about the error distributions combined with practical considerations lead to specific *covariance models* for the forecast and observation errors. Typically such models involve unknown parameters, which must then be estimated from actual atmospheric data.

For example, quality-controlled rawinsonde observations are usually regarded as unbiased measurements of the true atmospheric state. Measurement errors associated with separate vertical soundings are assumed independent, and the errors for the different measurement variables (temperature, relative humidity, and wind components) are assumed to be independent as well. The statistical properties of the errors in all individual, univariate soundings are generally taken to be identical; that is, independent of time and station.

For spatially distributed univariate observations, this set of assumptions corresponds to an observation error covariance model $\mathbf{R}$, with

$$[\mathbf{R}]_{ij}^{(mn)} = \sigma^{(m)}\sigma^{(n)}\nu^{(mn)}\delta(r_{ij}). \tag{6}$$

The notation $[\mathbf{R}]_{ij}^{(mn)}$ means the covariance of the error at station i, level m, with the error at station j, level n. The parameter $\sigma^{(m)}$ is the observation error standard deviation at pressure level m, and $\nu^{(mn)}$ is the vertical correlation between errors at levels m and n. The quantity r_{ij} is the (horizontal) distance between stations i and j, and

$$\delta(r) = \begin{cases} 1, & \text{if } r = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Thus, a complete univariate rawinsonde observation error covariance model for each measurement variable is determined by the set of parameters $\{\sigma^{(m)}, \nu^{(mn)}\}$.

The nature of forecast errors is more complicated, primarily because model errors are inherently multivariate and correlated in space and time. By expressing these properties in a forecast error covariance model, the information contained in a set of localized, univariate observations can be exploited to estimate multivariate atmospheric fields, even in regions where no observations exist. In order for such estimates to be meaningful, it is of course necessary that the covariance model formulations be sufficiently realistic. Forecast error covariance modeling is an active field of research that we will not attempt to review here. Rather, given a particular formulation of a forecast error covariance model, our concern in this work is to determine the best set of parameters for the model based on the available observations.

The applications described in Part II all involve forecast error covariance models that can be expressed in terms of the simple univariate model given by

$$[\mathbf{P}]_{ij}^{(mn)} = \sigma^{(m)}\sigma^{(n)}\nu^{(mn)}\rho(r_{ij}; L^{(m)}, L^{(n)}). \tag{8}$$

The function $\rho(r_{ij}; L^{(m)}, L^{(n)})$ represents spatial correlations while the quantity $\nu^{(mn)}\rho(0; L^{(m)}, L^{(n)})$ is the correlation between errors at locations at levels m and n

but in the same vertical column. Such models may be called *quasi-separable* because the vertical correlations are invariant with respect to translation along levels of constant pressure.

First consider the restriction of the model (8) to a fixed pressure level m, where it depends on two parameters only: the error standard deviation $\sigma^{(m)}$ and a decorrelation length scale $L^{(m)}$. The function $\rho(r; L^{(m)}) \equiv \rho(r; L^{(m)}, L^{(m)})$ represents the (horizontal) isotropic correlations between errors at any two locations on a fixed pressure level. Only a special class of functions of r gives rise to a legitimate (i.e., positive-semidefinite) fixed-level covariance model on a spherical surface; appendix A describes a few such functions as well as a definition of the decorrelation length-scale parameter L. Isotropic models derive from an assumption that the correlation between errors at any two locations depends only on the distance between the two locations: the isolines of the correlation functions are circular, and the parameter L controls the distance between the contours. The isotropic assumption is clearly not valid for actual forecast errors, which generally depend on local properties of the flow. The widespread use of isotropic univariate covariance models in atmospheric data assimilation systems can be explained by the fact that error correlations have traditionally been calculated by averaging data over relatively long periods of time, for example, 1–3 months (Rutherford 1972; Hollingsworth and Lönnberg 1986; Lönnberg and Hollingsworth 1986; Bartello and Mitchell 1992).

Evaluating the model (8) for locations at two different pressure levels m and n that are not in the same vertical column requires the specification of the function $\rho(r_{ij}; L^{(m)}, L^{(n)})$, $m \neq n$. Special care must be taken in constructing such a function so that its restriction to any fixed pressure level has the prescribed decorrelation length scale, while it still gives rise to a model that is positive-semidefinite on the domain of interest. This problem and its solution are addressed in detail by Gaspari and Cohn (1999). For estimating the vertical correlations between the errors at any two levels m and n we use the approximation

$$\rho(r_{ij}; L^{(m)}, L^{(n)}) = \tfrac{1}{2}[\rho(r_{ij}; L^{(m)}) + \rho(r_{ij}; L^{(n)})], \quad (9)$$

which is sufficiently accurate for our purpose, although it can be shown to give rise to negative eigenvalues in (8) when either the vertical correlations are large or when the decorrelation length scales vary greatly between levels.

Models of the form (8) will be used in Part II of this study to describe univariate forecast error covariances. With the approximation (9), such a model is completely determined by the parameters $\{\sigma^{(m)}, L^{(m)}, \nu^{(mn)}\}$ and by the choice of the function $\rho(r; L)$. We will consider a number of alternatives for this function, primarily in order to examine the effect on the parameter estimates of some of the uncertainties inherent in the description of forecast errors. In any case, estimation from regional time series data will, at best, produce parameter estimates that are representative of the actual forecast errors averaged over the space and time domain of the data.

Spatially correlated multivariate wind error covariances can be modeled using (8) as well. Following Daley (1991, his section 5.2), let ε^u, ε^v denote the wind error components and define an *error stream function* ψ and *error velocity potential* χ. Then one may write

$$\begin{bmatrix} \boldsymbol{\epsilon}^u \\ \boldsymbol{\epsilon}^v \end{bmatrix} = \begin{bmatrix} -\dfrac{\partial \psi}{\partial y} + \dfrac{\partial \chi}{\partial x} \\ \dfrac{\partial \psi}{\partial x} + \dfrac{\partial \chi}{\partial y} \end{bmatrix}. \quad (10)$$

Note that the stream function and velocity potential, in the present context, are associated with the error fields rather than with the flow itself. A multivariate wind error covariance model can then be constructed based on separate univariate covariance models for ψ and χ. The simplest such model results from the assumption that ψ and χ are statistically independent, and that the covariance of each can be modeled by (8).

For the applications in Part II we will use this simple approach to represent forecast wind error covariances at fixed pressure levels. This model does not provide any information about the cross covariances between wind errors and height errors. The coupling between wind and height errors is known to be strong in midlatitudes; this information must be incorporated in a multivariate forecast error covariance model in order to take full advantage of the observations in a data assimilation system (Hollingsworth and Lönnberg 1986). However, if the goal is only to estimate wind observation error covariance parameters, then a forecast error covariance model based on (10) will serve the purpose.

We have just discussed some examples of forecast and observation error covariance models, all of which involve several unknown parameters. In the following section we consider the general relationship between the covariance models on the one hand, and the actual observed data on the other. Without referring to specific models we may write

$$\mathbf{P}_k^f \approx \mathbf{P}_k^f(\boldsymbol{\alpha}^f), \qquad \mathbf{R}_k \approx \mathbf{R}_k(\boldsymbol{\alpha}^0), \qquad \mathbf{X}_k \approx \mathbf{X}_k(\boldsymbol{\alpha}^x), \quad (11)$$

with $\boldsymbol{\alpha}^f$, $\boldsymbol{\alpha}^0$, and $\boldsymbol{\alpha}^x$ unknown parameters whose definition depends on the particular modeling assumptions. Our goal will be to determine values for these parameters that, in a sense to be made precise, are most compatible with the data.

3. Observational residuals

The observation operator introduced in (2) is a device for comparing forecasts with observations. The *observed-minus-forecast residuals* defined by

$$\mathbf{v}_k \equiv \mathbf{w}_k^0 - \mathbf{h}_k(\mathbf{w}_k^f) \quad (12)$$

are routinely computed in operational data assimilation

systems. The residual p_k-vector time series $\{\mathbf{v}_k\}$ depends on actual observation and forecast errors, since

$$\mathbf{v}_k \approx \boldsymbol{\epsilon}_k^0 - \mathbf{H}_k\boldsymbol{\epsilon}_k^f, \tag{13}$$

where the linearized observation operator $\mathbf{H}_k$, a $p_k \times n$-matrix, is defined by

$$\mathbf{H}_k \equiv \left.\frac{\partial \mathbf{h}_k}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_k^f}. \tag{14}$$

Equation (13) is obtained by linearizing (12) about the forecast state and using (1) and (2). The accuracy of (13) depends on the size of the forecast errors; it is exact for linear observation operators.

If two distinct observing systems simultaneously measure the same quantity, then one can also define the *observed-minus-observed residuals*

$$\mathbf{r}_k \equiv \mathbf{w}_k^{o1} - \mathbf{w}_k^{o2}, \tag{15}$$

where the second superscript refers to the instrument. For (15) to make sense the observation operators $\mathbf{h}_k^1$ and $\mathbf{h}_k^2$ associated with the separate instruments must be compatible, in the sense that they both map to the same observation space; see (2). They need not be identical, however, and so it follows from (2) applied to each data type that

$$\mathbf{r}_k \approx \boldsymbol{\epsilon}_k^{o1} - \boldsymbol{\epsilon}_k^{o2}. \tag{16}$$

As an example, consider a set of temperatures retrieved from remotely sensed radiances, valid at a particular time t_k. If the retrievals are collocated with a set of rawinsonde temperature observations, then the residuals (15) can be computed. In this case the observation operators, although very different, are compatible. Some kind of interpolation will be required in order to collocate the retrievals with the rawinsonde observations, and therefore (16) is not exact. Provided the interpolation errors are small compared with the observation errors themselves, the observed-minus-observed residuals contain useful information about the observation errors associated with the two data types.

The mean and covariance of the observed-minus-forecast residuals defined by (12) are easily obtained from (13):

$$\langle \mathbf{v}_k \rangle \approx \mathbf{b}_k^o - \mathbf{H}_k\mathbf{b}_k^f, \tag{17}$$

$$\langle (\mathbf{v}_k - \langle \mathbf{v}_k \rangle)(\mathbf{v}_k - \langle \mathbf{v}_k \rangle)^{\mathrm{T}} \rangle \approx \mathbf{R}_k - \mathbf{X}_k\mathbf{H}_k^{\mathrm{T}} - \mathbf{H}_k\mathbf{X}_k^{\mathrm{T}} + \mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^{\mathrm{T}}. \tag{18}$$

We used the additional approximation $\langle \mathbf{H}_k \cdot \rangle \approx \mathbf{H}_k \langle \cdot \rangle$; both (17) and (18) are exact for linear observation operators.

Dee and da Silva (1998) show how the mean equation (17) can be used to estimate forecast bias in a statistical data assimilation system, using unbiased (or bias-corrected) observations. They also discuss, in general terms, the implications of using biased forecasts and/or biased observations in an analysis system. For the purpose of covariance estimation based on data residuals we will need to specify the mean of the residuals; that is,

$$\langle \mathbf{v}_k \rangle = \boldsymbol{\mu}_k. \tag{19}$$

For the moment we regard the mean $\boldsymbol{\mu}_k$ as known; see, however, section 4b below.

The covariance Eq. (18) can be used to tune parameters of the forecast and observation error covariance models discussed in the previous section. Substitution of (11) and (19) gives

$$\langle (\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \rangle \approx \mathbf{S}_k(\boldsymbol{\alpha}), \tag{20}$$

where

$$\mathbf{S}_k(\boldsymbol{\alpha}) = \mathbf{S}_k(\boldsymbol{\alpha}^f, \boldsymbol{\alpha}^o, \boldsymbol{\alpha}^x) \tag{21}$$

$$= \mathbf{R}_k(\boldsymbol{\alpha}^o) - \mathbf{X}_k(\boldsymbol{\alpha}^x)\mathbf{H}_k^{\mathrm{T}} - \mathbf{H}_k\mathbf{X}_k^{\mathrm{T}}(\boldsymbol{\alpha}^x) + \mathbf{H}_k\mathbf{P}_k^f(\boldsymbol{\alpha}^f)\mathbf{H}_k^{\mathrm{T}}. \tag{22}$$

Models for forecast and observation error covariances imply a model for the observed-minus-forecast residuals, and (20) provides a relationship between the models and the data.

Similarly, the mean and covariance of the observed-minus-observed residuals for two sets of collocated observations are

$$\langle \mathbf{r}_k \rangle \approx \mathbf{b}_k^{o1} - \mathbf{b}_k^{o2} \tag{23}$$

and

$$\langle (\mathbf{r}_k - \langle \mathbf{r}_k \rangle)(\mathbf{r}_k - \langle \mathbf{r}_k \rangle)^{\mathrm{T}} \rangle \approx \mathbf{R}_k^1 - \mathbf{Y}_k - \mathbf{Y}_k^{\mathrm{T}} + \mathbf{R}_k^2, \tag{24}$$

where $\mathbf{Y}_k$ is the cross-covariance between the observation errors,

$$\mathbf{Y}_k \equiv \langle (\boldsymbol{\epsilon}_k^{o1} - \mathbf{b}_k^{o1})(\boldsymbol{\epsilon}_k^{o2} - \mathbf{b}_k^{o2})^{\mathrm{T}} \rangle. \tag{25}$$

The mean Eq. (23) can be used to estimate and correct the bias in one set of observations based on another unbiased (or bias-corrected) set. For now we assume

$$\langle \mathbf{r}_k \rangle - \boldsymbol{\zeta}_k, \tag{26}$$

with $\boldsymbol{\zeta}_k$ known. We then have

$$\langle (\mathbf{r}_k - \boldsymbol{\zeta}_k)(\mathbf{r}_k - \boldsymbol{\zeta}_k)^{\mathrm{T}} \rangle \approx \mathbf{T}_k(\boldsymbol{\alpha}) \tag{27}$$

with

$$\mathbf{T}_k(\boldsymbol{\alpha}) = \mathbf{T}_k(\boldsymbol{\alpha}^{o1}, \boldsymbol{\alpha}^{o2}, \boldsymbol{\alpha}^y) \tag{28}$$

$$= \mathbf{R}_k^1(\boldsymbol{\alpha}^{o1}) - \mathbf{Y}_k(\boldsymbol{\alpha}^y) - \mathbf{Y}_k^{\mathrm{T}}(\boldsymbol{\alpha}^y) + \mathbf{R}_k^2(\boldsymbol{\alpha}^{o2}). \tag{29}$$

Here, we have introduced the possibility of parameterizing the cross-covariance $\mathbf{Y}_k$ as well. In this case, models for observation error covariances imply a model for the observed-minus-observed residuals; Eq. (27) establishes a relationship between the models and the data analogous with (20).

4. Covariance parameter estimation

The previous section describes the relationship between error covariance models and data residuals. We now consider a general method for adjusting the free model parameters in order to improve the consistency between the models and a given finite subset of the data. In order to keep the presentation simple we will use the notation for observed-minus-forecast residuals: the data are $\mathbf{v}_k$ and the covariance model is $\mathbf{S}_k(\boldsymbol{\alpha})$, although the method applies equally well to observed-minus-observed residuals.

a. Maximum-likelihood estimation

One way to fit a model to a dataset is by maximizing the likelihood that the actual observed data did, in fact, arise from the model. To be precise, suppose that the actual sequence of residuals $\{\mathbf{v}_k\}$ is a realization of a multivariate stochastic process $\{\mathbf{V}_k\}$, whose joint probability density function (pdf) is $p(\{\mathbf{v}_k\}; \boldsymbol{\alpha})$. If the functional form of the pdf is known, then its value for a fixed dataset $\{\mathbf{v}_k, k = 1, \ldots, K\}$ depends on $\boldsymbol{\alpha}$ only: the function of $\boldsymbol{\alpha}$ thus defined is called the *likelihood function* (Fisher 1922). The *maximum-likelihood estimate* $\hat{\boldsymbol{\alpha}}$ is obtained by finding the maximum of the likelihood function.

To apply the maximum-likelihood method to our problem, we need to assume a probability density for the underlying process $\{\mathbf{V}_k\}$ with covariances given by (22). To this end we postulate that the process is white and Gaussian, with covariances at times t_k given by $\mathbf{S}_k(\boldsymbol{\alpha})$ for some $\boldsymbol{\alpha}$. We also assume that the means $\boldsymbol{\mu}_k$ are known, or that they can be estimated independently. Using the familiar expression for the multivariate Gaussian pdf (Jazwinski 1970, section 2.4),

$$\begin{aligned} p(\{\mathbf{v}_k\}; \boldsymbol{\alpha}) &= \prod_{k=1}^{K} p(\mathbf{v}_k; \boldsymbol{\alpha}) \\ &\propto \prod_{k=1}^{K} (\det \mathbf{S}_k(\boldsymbol{\alpha}))^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathbf{S}_k^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k)\right]. \end{aligned} \quad (30)$$

The maximum-likelihood estimate $\hat{\boldsymbol{\alpha}}$ is obtained by maximizing (30) with respect to $\boldsymbol{\alpha}$, or equivalently, by minimizing the *log-likelihood function*

$$f(\boldsymbol{\alpha}) = \frac{1}{K} \sum_{k=1}^{K} [\log \det \mathbf{S}_k(\boldsymbol{\alpha}) + (\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathbf{S}_k^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k)]. \quad (31)$$

Note that this expression depends on the data, is therefore random, and that there is no guarantee of a unique minimum.

The assumption that the process $\{\mathbf{V}_k\}$ is white and Gaussian is not essential; the pdf (30) can be replaced by any other. For example, it is straightforward to incorporate a description of the serial correlations between successive residuals. For some data types a lognormal distribution may be more appropriate, although in the applications we have studied so far the Gaussian assumption appears to be adequate. In any case, the sensitivity of the parameter estimates to assumptions about the pdf of the data should be addressed experimentally; we will return to this issue in section 4e below and in Part II.

For a fixed dataset $\{\mathbf{v}_k\}$ and given formulations of the covariance models $\mathbf{S}_k(\boldsymbol{\alpha})$, the function $f(\boldsymbol{\alpha})$ can be minimized using standard optimization software (e.g., Press et al. 1992, chapter 10). Forecast and observation error covariance models implemented in current atmospheric data assimilation systems are relatively simple to evaluate; they have to be in order for the assimilation of large volumes of data to be computationally viable. The effort involved in tuning models such as (6) and (8) therefore depends primarily on the size of the dataset. As we show below, the error variance in the parameter estimates is proportional to $1/\nu$, where ν is the total number of data. The constant of proportionality varies from case to case (it depends on the identifiability of the parameters), but as a rule it requires on the order of a hundred observations to estimate a single parameter with meaningful accuracy.

The log-likelihood function (31) is formulated for the general case of time-dependent forecast and/or observation error covariances; see (22). This is necessary, for example, when the observation operator depends on time, which is the case for observations originating from moving platforms such as ships, aircraft, or satellites. Time-dependent covariance models $\mathbf{S}_k(\boldsymbol{\alpha})$ can also result from expressing forecast error covariances in terms of local atmospheric flow and/or thermodynamic conditions; see Riishøjgaard (1998) for an example of such a model.

In the special case when the covariance (22) is stationary, that is, when

$$\mathbf{S}_k(\boldsymbol{\alpha}) = \mathbf{S}(\boldsymbol{\alpha}), \quad (32)$$

the log-likelihood function simplifies to

$$f(\boldsymbol{\alpha}) = \log \det \mathbf{S}(\boldsymbol{\alpha}) + \mathrm{trace}[\mathbf{S}^{-1}(\boldsymbol{\alpha})\overline{\mathbf{S}}], \quad (33)$$

where $\overline{\mathbf{S}}$ is the sample covariance of the data defined by

$$\overline{\mathbf{S}} = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}. \quad (34)$$

Equations (33) and (31) are equivalent when (32) holds because

$$\frac{1}{K}\sum_{k=1}^{K}(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{S}^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k)$$

$$= \frac{1}{K}\sum_{k=1}^{K}\mathrm{trace}[\mathbf{S}^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}]$$

$$= \mathrm{trace}\left[\mathbf{S}^{-1}(\boldsymbol{\alpha})\frac{1}{K}\sum_{k=1}^{K}(\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}\right]$$

$$= \mathrm{trace}[\mathbf{S}^{-1}(\boldsymbol{\alpha})\overline{\mathbf{S}}]. \quad (35)$$

Note that $\overline{\mathbf{S}}$ is the *unconstrained* maximum-likelihood covariance estimate for a white, stationary Gaussian time series (e.g., Muirhead 1982). Minimization of (33) provides instead the *constrained* maximum-likelihood covariance estimate—constrained to be of the form $\mathbf{S}(\boldsymbol{\alpha})$ for some $\boldsymbol{\alpha}$. This approach to the estimation of structured covariance matrices from stationary time series was first proposed by Burg et al. (1982).

The stationary form (33) of the log-likelihood function is appropriate for estimating covariance parameters from station data. Most covariance models currently implemented in operational data assimilation systems are independent of the state of the system and generally vary slowly with time, if at all. In this case it is convenient to first compute the sample covariance matrix (34), and then to minimize (33) with respect to $\boldsymbol{\alpha}$. This procedure is similar to the traditional method described in the introduction, except that the parameter estimates are based on the maximum-likelihood criterion rather than on a least squares fit. The two norms are different in general since the least squares procedure does not take into account any a priori information about the probability distribution of the data.

Time-independent observation operators arise from stationary observing systems such as rawinsonde networks. However, truly stationary observation operators do not exist in practice, because data are occasionally missing or rejected by quality-control procedures. In that case the matrix $\overline{\mathbf{S}}$ can be constructed element by element, by considering, for each pair of stations, the set of all simultaneous quality-controlled reports. The sample covariance between data from stations i and j can then be estimated from this set by

$$\overline{\mathbf{S}}_{ij} = \frac{1}{K_{ij}}\sum_{k=1}^{K_{ij}}([\mathbf{v}_k]_i - [\boldsymbol{\mu}_k]_i)([\mathbf{v}_k]_j - [\boldsymbol{\mu}_k]_j), \quad (36)$$

where $[\,\cdot\,]_i$ denotes the element associated with station i and K_{ij} is the number of simultaneous reports at stations i and j. If K_{ij} is small then $\overline{\mathbf{S}}_{ij}$ is generally not an accurate estimate of the covariance between stations i and j. One might exclude a number of stations in order to ensure that all K_{ij} exceed a certain threshold; however, our experience with the maximum-likelihood method is that the parameter estimates are rather insensitive to this.

b. Bias estimation

In order to implement the estimation procedure, it is necessary to specify the residual means $\boldsymbol{\mu}_k$ in (31) or in (34). These depend on the mean observation errors $\mathbf{b}_k^0$ and on the mean forecast errors $\mathbf{b}_k^f$ [see (17)], neither of which is accurately known in practice.

There are two choices for the purpose of tuning covariance models. The first is to simply ignore the bias by taking $\boldsymbol{\mu}_k = 0$. This choice will, of course, affect the parameter estimates. For example, variance parameters will tend to be overestimated when bias is ignored. This approach is not unreasonable if the tuned covariance models are to be used for a statistical analysis where bias is not explicitly accounted for. In that case the total (systematic plus random) root-mean-square analysis error will actually be smallest when the forecast and observation error variances are suitably inflated in order to account for the bias (Dee and da Silva 1998).

The alternative is to estimate the residual mean $\boldsymbol{\mu}_k$ prior to, or concurrent with, the estimation of covariance parameters. If independent information about forecast and/or observation bias is available, then this information should obviously be used. In practice this is unlikely to be the case and therefore $\boldsymbol{\mu}_k$ must be estimated from the data.

One approach would be to generalize the maximum-likelihood estimation procedure by formulating a parameterized bias model

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_k(\boldsymbol{\beta}), \quad (37)$$

and then to produce maximum-likelihood estimates of the bias parameters as well. For example, bias over a fixed domain might be modeled by a truncated spectral expansion. Parameter estimation would involve minimizing the log-likelihood function (31), after substitution of (37), with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Estimating bias parameters in this fashion amounts to a weighted least squares bias estimation procedure in which the weights (determined by the covariances) are adjusted adaptively. Although the generality of this approach is appealing, we do not expect it to be practical. The difficulty in bias estimation lies not so much in the techniques as in the ability to formulate sensible bias models.

Error statistics used in atmospheric data assimilation are usually defined in terms of time averages, because true ensemble averaging is not possible; after all, only a single realization of the physical atmosphere is available for observation. Accordingly one can estimate $\boldsymbol{\mu}_k$ by calculating the time-mean residuals. In the stationary case (i.e., for stationary observing systems) we then have

$$[\boldsymbol{\mu}_k]_i = [\overline{\boldsymbol{\mu}}]_i, \quad (38)$$

where

$$[\overline{\boldsymbol{\mu}}]_i = \frac{1}{K_i}\sum_{k=1}^{K_i}[\mathbf{v}_k]_i \quad (39)$$

with K_i the number of reports from station i. In the general (nonstationary) case when observation locations vary with time, one might define a spatially varying bias estimate $\overline{\boldsymbol{\mu}}$ on some arbitrary grid by means of a successive correction scheme. We report on experiments with such a technique in Part II.

The presence of bias that is not properly accounted for will generally result in inaccurate covariance parameter estimates. More importantly, biased data and/or forecasts will result in a biased analysis, independently of the covariance models used. Although it is not the subject of this study, bias estimation and correction must take precedence over covariance modeling and estimation.

c. Identifiability of the parameters

The maximum-likelihood method is appealing for its generality; it can be used to estimate any set of parameters of the pdf, provided those parameters are *jointly identifiable*. This is a fundamental requirement for the estimation problem to be well posed. There exist different technical definitions of this notion (e.g., Chavent 1979), but we will take it to mean that the log-likelihood function (31) must have a unique global minimum with probability one; that is, for almost all realizations of the process $\{\mathbf{V}_k\}$. In practice this imposes requirements on the model formulation as well as on the data: there must be no dependency among the model parameters, and the data must provide an adequate sampling.

Consider, for example, two independent *scalar* random variables w^1 and w^2 with identical means but different variances, and suppose that only the residual $w^1 - w^2$ is observed. It is clearly impossible to estimate the variances of w^1 and w^2 separately, no matter how many sample residuals are available: only the sum of the variances can be estimated. (The means of w^1 and w^2 are not identifiable either in this case.) Suppose now that $\mathbf{w}^1$ and $\mathbf{w}^2$ are independent *vector* random variables, representing, for example, two spatially distributed random fields. If $\mathbf{w}^2$ is spatially correlated with constant variance but $\mathbf{w}^1$ is spatially uncorrelated with constant variance, then it is, in fact, possible to estimate both variances from the residuals $\boldsymbol{w}^1 - \boldsymbol{w}^2$, provided they are sampled at more than a single location. Thus, there is a data requirement as well as a model requirement. This example is prototypical for the applications described in Part II, where the data residuals contain both a spatially correlated and a spatially uncorrelated error component.

It is not possible in general to prove identifiability for a given parameter estimation problem, although simple examples such as the above can lead to useful insights about the types of parameters one can hope to estimate from a given dataset. On the other hand, it is easy to check numerically for indications that identifiability might be a problem by evaluating the Hessian of the log-likelihood function at its minimum,

$$\mathbf{A}_{ij} = \left.\frac{\partial^2 f}{\partial\alpha_i \partial\alpha_j}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}}. \tag{40}$$

The Hessian matrix can be approximated by finite differences; depending on the optimization method used, it may even be available as a by-product of the optimization of the function f. At the minimum the gradient of the log-likelihood function vanishes, so for $\boldsymbol{\alpha}$ near $\hat{\boldsymbol{\alpha}}$,

$$f(\boldsymbol{\alpha}) \approx f(\hat{\boldsymbol{\alpha}}) + \tfrac{1}{2}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^{\mathrm{T}}\mathbf{A}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}}), \tag{41}$$

provided f is a sufficiently smooth function of $\boldsymbol{\alpha}$.

Equation (41) shows that the sensitivity of the log-likelihood function f to the parameters $\boldsymbol{\alpha}$ near its minimum is controlled by the Hessian. A small perturbation of $\hat{\boldsymbol{\alpha}}$ along the direction of an eigenvector of $\mathbf{A}$ produces a change in f by an amount proportional to the corresponding eigenvalue (e.g., Strang 1988, section 7.2). If the Hessian has a large condition number then the identifiability of the parameters is poor along the directions associated with the smallest eigenvalues.

Since the nonlinearity of the function f is generally not quadratic, the Hessian matrix $\mathbf{A}$ only describes the *local* identifiability of the parameters. Note, however, that the analysis does not depend on any properties of f other than its differentiability with respect to the parameters. Identifiability is a notion that is not specifically connected with the maximum-likelihood method; it is simply a practical requirement for any parameter estimation method that is based on minimizing a cost function.

d. Accuracy of maximum-likelihood parameter estimates

The maximum-likelihood method has many appealing theoretical properties (Cramér 1946); in particular it is *asymptotically efficient.* This means that in the limit of infinite data there is no other unbiased estimator that produces more accurate parameter estimates. In practice, we have only finite datasets at our disposal, and more importantly, many of the assumptions required to implement the method are, in fact, violated. The parameter estimates produced in any realistic application therefore will *not* be true maximum-likelihood estimates. Nevertheless, it is useful to compute the asymptotic accuracy of the maximum-likelihood estimates along with the parameter estimates themselves; as we explain below, this provides important information about the uncertainty of the estimates due to sampling error.

Suppose, for the moment, that the *model hypothesis* holds. By this we mean that all assumptions about the data as expressed by the likelihood function are actually valid. In that case the parameter estimates produced by minimizing (31) are truly the maximum-likelihood estimates. Then it can be shown (e.g., Sorenson 1980, his Theorem 5.4) that, if $\boldsymbol{\alpha}^*$ is the vector of true parameter values, then in the limit of infinite data the estimates approach a normal distribution with

$$\langle \hat{\boldsymbol{\alpha}} \rangle = \boldsymbol{\alpha}^*, \qquad \langle (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^{\mathrm{T}} \rangle = \frac{2}{\nu} \mathbf{A}^{-1}. \quad (42)$$

Here $\mathbf{A}$ is the Hessian of the log-likelihood function [see (40)], and ν is the number of degrees of freedom associated with the estimation problem. In the general (nonstationary) case corresponding to (31) we have

$$\nu = \sum_{k=1}^{K} n_k, \qquad n_k = \dim \mathbf{v}_k. \quad (43)$$

In the truly stationary case where $n_k = n = \text{const}$, we would have $\nu = nK$; however, (43) should be applied to account for missing data [see the discussion following (35)]. If bias parameters are estimated from the same data used for covariance estimation, then the number of degrees of freedom ν should be reduced accordingly.

The estimation error covariance in (42) is the lower bound of the Cramér–Rao inequality (Sorenson 1980). The Cramér–Rao inequality can be regarded as an *uncertainty principle* for parameter estimation: it expresses the fact that the random nature of the data imposes a fundamental limitation on the accuracy with which parameters of the pdf can be estimated from the data. The Hessian matrix $\mathbf{A}$ is related to the curvature of the likelihood function at its mode; the broader the mode, the harder it is to estimate parameters within a certain accuracy. The theory states that, under rather general conditions, the error covariance of the maximum-likelihood estimates tends to the Cramér–Rao bound as the size of the dataset increases (see also Lupton 1993, chapter 10).

We routinely use (42) to estimate the standard errors of the parameter estimates under the modeling hypothesis. For any given dataset and covariance model formulation, the validity of (42) for finite ν can be checked using a Monte Carlo experiment with synthetic data. The procedure is simply to compute parameter estimates from the output of a random-number generator, making sure that the generator produces normally distributed vector samples whose covariances are given by the covariance model with specified parameters. By repeating this many times, the actual estimation errors due to sampling error can be computed. We performed this procedure for all applications described in Part II; although not reported there, the standard error estimates given by (42) turned out to be quite accurate in all cases.

The maximum-likelihood standard errors represent idealized accuracy estimates; in practice they should be regarded as lower bounds on the true accuracy. These error estimates are useful in practice because they quantify the effect of sampling error, which is the only source of error under the model hypothesis. Thus, the standard error estimates indicate whether a given set of covariance parameters can be actually identified from the available data, and whether the parameter uncertainty due to sampling error is acceptable. Making sure that sampling error is small allows one to investigate other sources of uncertainty by modifying the assumptions underlying the model hypothesis. We include several examples of this type of uncertainty analysis in Part II.

e. Robustness of the parameter estimates

The fact that parameters associated with a particular covariance model can be estimated from a dataset (i.e., that they are identifiable) does not imply that the estimates are actually meaningful. There are many possible reasons why a tuned covariance model may not in fact provide a good fit to the actual data. First of all, the covariance model may be incorrect for *any* set of parameter values. For example, the model might be isotropic while actual covariances are highly anisotropic: a strong state-dependent component of error cannot be accounted for in an isotropic model. To some extent the validity of the assumptions that enter into the formulation of a covariance model can be examined using standard statistical techniques. This requires long-term monitoring of the actual residuals produced by an operational data assimilation system.

A second group of possible reasons for a poor fit concerns the additional assumptions involved in tuning the model. Even if the covariance model is appropriate, the parameter estimates may be far from optimal because, for example, the bias is handled incorrectly, or the data may be serially correlated. In fact it is very likely that some, if not all, of these violations apply in practice. Yet the maximum-likelihood method depends on a complete specification of the pdf of the data. In the absence of information, it is common practice to default to a standard set of assumptions. For example, lacking any specific indications to the contrary, it is almost always assumed that the data are Gaussian and white. This raises the issue of *robustness* of the maximum-likelihood method with respect to the information it requires; we will address this issue experimentally in Part II.

It is worth noting at this point that all currently operational atmospheric data assimilation systems can be regarded as particular applications of the maximum-likelihood method to the problem of estimating the state of the atmosphere from observations and model information (Lorenc 1986; Cohn 1997). Different assumptions about the underlying probability distributions lead to different solution methods, but in all cases that have been tried so far the errors (after quality control of observations) are assumed to be Gaussian and white. In the present work we try to be consistent in applying this same framework to the estimation of parameters of the covariance models, although the maximum-likelihood method is completely general in this respect.

Let us take the pragmatic point of view, then, that the majority of assumptions about error distributions are made primarily for practical reasons, and not necessarily because they are believed to be valid. Then the log-likelihood function (31) [or (33), in the stationary case] is simply one of many possible cost functions that could

be used for fitting a parameterized family of covariance models to a dataset. The traditional fitting procedure for station data described in the introduction, for example, is based on a least squares criterion. An advantage of the maximum-likelihood approach is that it incorporates the same statistical information about forecast and observation errors used in the data assimilation system. However, if the underlying assumptions on the error distributions are wrong, one might legitimately ask whether there are any other criteria that lead to a more robust parameter estimation procedure.

One candidate for such a criterion follows from the generalized cross-validation (GCV) method (Wahba and Wendelberger 1980). The cross-validation approach is based on maximizing the capability of a model to predict withheld data, and it does not require as many assumptions on the nature of the error distributions as does the maximum-likelihood method. Wahba et al. (1995) show how GCV can be applied to the estimation of covariance parameters and possibly other tuning parameters of an atmospheric assimilation system. Using our notation, they specifically consider covariance models of the form

$$\mathbf{S}(\boldsymbol{\alpha}) = \mathbf{S}(\sigma_1, \sigma_2, \boldsymbol{\theta}) = \sigma_1^2\mathbf{S}_1 + \sigma_2^2\mathbf{S}_2(\boldsymbol{\theta}), \quad (44)$$

with $\mathbf{S}_1$ a constant matrix, and both $\mathbf{S}_1$ and $\mathbf{S}_2$ positive definite. Such a model is sufficiently general for most applications of interest. The first term typically represents observation error covariances, while the second term can be used to model forecast error covariances. The GCV method estimates the parameter λ defined by

$$\lambda = \left(\frac{\sigma_1}{\sigma_2}\right)^2, \quad (45)$$

and possibly additional parameters $\boldsymbol{\theta}$ as well, from data residuals. We summarize the GCV estimation procedure in appendix B.

The scalar λ is actually the single most important parameter of the covariance model (44), being the ratio of the variances of the two signals present in the data residuals. Estimates of the separate variances σ_1 and σ_2 are obtained as a by-product of the GCV estimation procedure. Note that the identifiability requirement still holds; *no* method can produce meaningful estimates of poorly identifiable parameters. See appendix A of Wahba et al. (1995) for a discussion of identifiability in the context of GCV.

The GCV approach as formulated by Wahba et al. (1995) applies to the estimation of the parameters σ_1, σ_2, and $\boldsymbol{\theta}$ based on a single vector of residuals valid at a fixed time t_k. As in Dee (1995), their method was originally intended to be used online in a data assimilation system, for the adaptive tuning of system parameters. However, it can also be applied offline for covariance estimation based on data $\{\mathbf{v}_k\}$ collected over a finite interval, as we show in appendix B. From a practical point of view the GCV method and the maximum-likelihood method differ only in that they involve different cost functions; compare (B10) with (31) [or, for stationary models, (B15) with (33)]. It is true for both methods that if either the covariance model formulation or the data change, the parameter estimates produced by both methods will also change. Both methods are similar in terms of computational complexity, although we have found that the localization of a minimum of the GCV cost function is occasionally easier when starting from a poor initial guess. The parameter estimates obtained by the two methods from the same dataset can be compared by simply interchanging the cost functions. Experiments reported in Part II indicate that the differences are mostly insignificant.

5. Summary and conclusions

We described the application of the maximum-likelihood method to the estimation of unknown observation and forecast error covariance parameters from observational residuals. Issues such as bias estimation and correction, parameter identifiability, estimation accuracy, and robustness of the method were addressed with an eye toward practical application. We briefly discussed the relationship between the maximum-likelihood method and generalized cross-validation (Wahba and Wendelberger 1980). In Part II of this study we describe three different applications involving univariate and multivariate covariance models, and data from both stationary and moving observing systems.

Advantages of the maximum-likelihood method include the fact that it allows one to use a priori information about the error distributions, to the extent that such information is available. The method is consistent with current operational atmospheric data assimilation systems, all of which can be regarded as particular implementations of the maximum-likelihood method applied to the problem of estimating the state of the atmosphere from observations and from model information. We showed in section 4d how the maximum-likelihood method can be used to produce estimates of the effect of sampling error upon parameter uncertainty. By making sure that this effect is small, one can study the variability of the covariance parameters by changing the selection of data. In addition, it is possible to perform an uncertainty analysis on the parameter estimates by modifying the assumptions that enter into the maximum-likelihood formulation.

A fundamental limitation of this and other estimation methods is connected with the identifiability of the covariance parameters. Simultaneous estimation of multiple parameters is possible only when all parameters are jointly identifiable from the data. This imposes requirements on the model formulation as well as on the data. In practice, observation errors and forecast errors can be statistically separated only to the extent that 1) they have characteristics that are distinguishable in the observation space; and 2) these different characteristics are adequately modeled. For example, if the observation

error contains a spatially correlated component, then this component may be falsely attributed to forecast error. In a future paper we will present a strategy for dealing with such a situation.

Acknowledgments. Thanks to Steve Cohn, Roger Daley, Greg Gaspari, Peter Houtekamer, Herschel Mitchell, Chris Redder, Leonid Rukhovets, and Grace Wahba for many stimulating discussions about this work.

APPENDIX A

Correlation Models

The general univariate isotropic covariance model is of the form

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \sigma(\mathbf{x})\sigma(\mathbf{y})\rho(\|\mathbf{x} - \mathbf{y}\|), \qquad \text{(A1)}$$

with $\sigma(\mathbf{x})$ a positive real-valued function and $\|\mathbf{x} - \mathbf{y}\|$ the Euclidean distance between locations $\mathbf{x}$ and $\mathbf{y}$. The *representing function* ρ must satisfy certain conditions for (A1) to be a legitimate (i.e., positive semidefinite) covariance model; see Gaspari and Cohn (1999) for details.

In this study and in Part II we consider three alternatives for the representing function ρ. In each case the Euclidean distance $\|\mathbf{x} - \mathbf{y}\|$ is replaced by the so-called *horizontal distance* $r = r(\mathbf{x}, \mathbf{y})$, defined as the chordal distance between the orthogonal projections of the locations $\mathbf{x}$ and $\mathbf{y}$ onto the earth's surface. Each alternative generates a legitimate covariance model on the sphere.

The *power law*:

$$\rho(r) = \rho_p(r; L) = \left[1 + \frac{1}{2}\left(\frac{r}{L}\right)^2\right]^{-1}. \qquad \text{(A2)}$$

The parameter L is the decorrelation length scale defined by

$$L = \sqrt{\frac{-1}{\rho''(0)}}, \qquad \text{(A3)}$$

see Daley (1991, section 4.3).

The *compactly supported fifth-order piecewise rational function* (Gaspari and Cohn 1999, section 4.3):

$$\rho(r) = \rho_c(r; L) \qquad \text{(A4)}$$

$$= \begin{cases} -\frac{1}{4}\left(\frac{r}{c}\right)^5 + \frac{1}{2}\left(\frac{r}{c}\right)^4 + \frac{5}{8}\left(\frac{r}{c}\right)^3 - \frac{5}{3}\left(\frac{r}{c}\right)^2 + 1, & \text{if } 0 \le r \le c, \\ \frac{1}{12}\left(\frac{r}{c}\right)^5 - \frac{1}{2}\left(\frac{r}{c}\right)^4 + \frac{5}{8}\left(\frac{r}{c}\right)^3 + \frac{5}{3}\left(\frac{r}{c}\right)^2 - 5\left(\frac{r}{c}\right) + 4 - \frac{2}{3}\left(\frac{r}{c}\right)^{-1}, & \text{if } c \le r \le 2c, \\ 0 & \text{otherwise} \end{cases} \qquad \text{(A5)}$$

with

$$c = L\sqrt{\frac{10}{3}}, \qquad \text{(A6)}$$

L being the decorrelation length scale defined by (A3).

Correlation models represented by (A4) are identically zero whenever the distance between two locations exceeds the threshold $r = r_* = 2c$,

$$\rho_c(r; L) = 0 \quad \text{for } r > 2c \approx 3.65L. \qquad \text{(A7)}$$

Taking advantage of this property can result in significant computational savings in the context of a global statistical analysis system (DAO 1996). However, the Legendre spectrum of this compactly supported correlation model is quite different from that of the power law; see Fig. A1.

The *windowed power law*:

$$\rho(r) = \rho_w(r; L) = \rho_p(r; L_1) \times \rho_c(r; L_2), \qquad \text{(A8)}$$

also with compact suppport. Using (A3) and the fact that $\rho(0) = 1$, $\rho'(0) = 0$ for each of the functions considered here, it is easy to show that

$$\frac{1}{L^2} = \frac{1}{L_1^2} + \frac{1}{L_2^2}. \qquad \text{(A9)}$$

The support of the windowed power law can be controlled by means of the parameter L_2: the function is identically zero for $r > r_*$ when

$$L_2 = \frac{r_*}{2}\sqrt{\frac{3}{10}}. \qquad \text{(A10)}$$

If we consider the decorrelation length-scale L as the single free (tunable) parameter in (A8), one should take

$$L_1 = \frac{L}{\sqrt{1 - \frac{40}{3}\left(\frac{L}{r_*}\right)^2}}, \qquad \text{(A11)}$$

which follows by substituting (A10) into (A9).

Figure A1 shows plots of the three functions for iden-

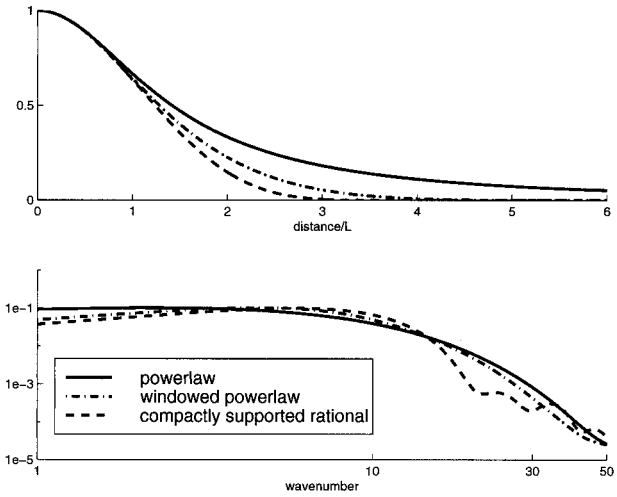

FIG. A1. Correlation models and Legendre coefficients, for three representing functions with identical length-scale parameters.

tical values of the length-scale parameter L, as well as their discrete Legendre spectra.

APPENDIX B

Multiple-Sample GCV

Wahba et al. (1995) show how to obtain GCV estimates of the covariance parameters σ_1, σ_2, $\boldsymbol{\theta}$ in (44) based on a single residual **v.** It is assumed that

$$\langle \mathbf{v} \rangle = \boldsymbol{\mu}, \qquad \langle (\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})^{\mathrm{T}} \rangle = \sigma_1^2 \mathbf{S}_1 + \sigma_2^2 \mathbf{S}_2(\boldsymbol{\theta}), \tag{B1}$$

with $\boldsymbol{\mu}$, $\mathbf{S}_1$, and $\mathbf{S}_2(\boldsymbol{\theta})$ known with the exception of the parameters $\boldsymbol{\theta}$. First, let

$$\lambda = \left(\frac{\sigma_1}{\sigma_2} \right)^2. \tag{B2}$$

Then find $\hat{\lambda}$, $\hat{\boldsymbol{\theta}}$ which minimize

$$V(\lambda, \boldsymbol{\theta}) = \frac{\|[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]\mathbf{y}\|^2}{\{\mathrm{trace}[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]\}^2}, \tag{B3}$$

where

$$\mathbf{A}(\lambda, \boldsymbol{\theta}) = [\mathbf{I} + \lambda \mathbf{S}_1^{1/2} \mathbf{S}_2^{-1}(\boldsymbol{\theta}) \mathbf{S}_1^{1/2}]^{-1} \tag{B4}$$

and

$$\mathbf{y} = \mathbf{S}_1^{-1/2}(\mathbf{v} - \boldsymbol{\mu}), \tag{B5}$$

and $\mathbf{S}_1^{1/2}$ is the symmetric square root of $\mathbf{S}_1$. This determines $\hat{\boldsymbol{\theta}}$, and then

$$\hat{\sigma}_1^2 = \{\mathrm{trace}[\mathbf{I} - \mathbf{A}(\hat{\lambda}, \hat{\boldsymbol{\theta}})]\} \times V(\hat{\lambda}, \hat{\boldsymbol{\theta}}) \tag{B6}$$

and

$$\hat{\sigma}_2^2 = \frac{\hat{\sigma}_1^2}{\hat{\lambda}}. \tag{B7}$$

In case the data consist of a time series $\{\mathbf{v}_k\}$ one can simply concatenate the $\mathbf{v}_k$ into a single random vector $\mathbf{v}$

$$\mathbf{v} = (\mathbf{v}_1^{\mathrm{T}} \cdots \mathbf{v}_K^{\mathrm{T}})^{\mathrm{T}}, \tag{B8}$$

formulate a covariance model for this concatenated vector, and apply the previous formulas. For simplicity we assume here that the $\mathbf{v}_k$ are independent. Suppose the mean and covariance models for the $\mathbf{v}_k$ are

$$\langle \mathbf{v}_k \rangle = \boldsymbol{\mu}_k,$$

$$\langle (\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \rangle = \sigma_1^2 \mathbf{S}_{k1} + \sigma_2^2 \mathbf{S}_{k2}(\boldsymbol{\theta}). \tag{B9}$$

The covariance model (B1) for $\mathbf{v}$ is block-diagonal, with blocks given by (B9). It is easily checked that

$$V(\lambda, \boldsymbol{\theta}) = \frac{\sum_{k=1}^{K} \|[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})]\mathbf{y}_k\|^2}{\left\{ \sum_{k=1}^{K} \mathrm{trace}[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})] \right\}^2}, \tag{B10}$$

where

$$\mathbf{A}_k(\lambda, \boldsymbol{\theta}) = [\mathbf{I} + \lambda \mathbf{S}_{k1}^{1/2} \mathbf{S}_{k2}^{-1}(\boldsymbol{\theta}) \mathbf{S}_{k1}^{1/2}]^{-1} \tag{B11}$$

and

$$\mathbf{y}_k = \mathbf{S}_{k1}^{-1/2}(\mathbf{v}_k - \boldsymbol{\mu}_k). \tag{B12}$$

If we further assume that the covariance models are stationary with $\mathbf{S}_{k1} = \mathbf{S}_1$ and $\mathbf{S}_{k2} = \mathbf{S}_2$, then the function $V(\lambda, \boldsymbol{\theta})$ simplifies as follows. For the numerator,

$$\begin{aligned} &\sum_{k=1}^{K} \|[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})]\mathbf{y}_k\|^2 \\ &\quad = \sum_{k=1}^{K} \mathbf{y}_k^{\mathrm{T}} [\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]^2 \mathbf{y}_k \\ &\quad = \sum_{k=1}^{K} \mathrm{trace}\{[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]^2 \mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}\} \\ &\quad = \mathrm{trace}\{[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]^2 \sum_{k=1}^{K} \mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}\} \\ &\quad = K \times \mathrm{trace}\{[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]^2 \mathbf{S}_1^{-1/2} \overline{\mathbf{S}} \mathbf{S}_1^{-1/2}\}, \end{aligned} \tag{B13}$$

where $\overline{\mathbf{S}}$ is the sample covariance of the data defined in (34). For the denominator,

$$\sum_{k=1}^{K} \mathrm{trace}[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})] = K \times \mathrm{trace}[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})], \tag{B14}$$

so in case of multiple samples of a stationary time series the GCV criterion is

$$V(\lambda, \boldsymbol{\theta}) = \frac{1}{K} \times \frac{\mathrm{trace}\{[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]^2 \mathbf{S}_1^{-1/2} \overline{\mathbf{S}} \mathbf{S}_1^{-1/2}\}}{\{\mathrm{trace}[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})]\}^2}. \tag{B15}$$

REFERENCES

Bartello, P., and H. L. Mitchell, 1992: A continuous three-dimensional model of short-range forecast error covariances. *Tellus,* **44A,** 217–235.

Burg, J. P., D. G. Luenberger, and D. L. Wenger, 1982: Estimation of structured covariance matrices. *Proc. IEEE,* **70,** 963–974.

Chavent, G., 1979: Identification of distributed parameter systems: About the output least square method, its implementation, and identifiability. *Identification and System Parameter Estimation*: *Proceedings of the Fifth IFAC Symposium*, R. Iserman, Ed., Vol. 1, Pergamon Press, 85–97.

Cohn, S. E., 1997: Introduction to estimation theory. *J. Meteor. Soc. Japan,* **75,** 257–288.

Cramér, H., 1946: *Mathematical Methods of Statistics.* Princeton University Press, 575 pp.

Daley, R., 1985: The analysis of synoptic scale divergence by a statistical interpolation scheme. *Mon. Wea. Rev.*, **113,** 1066–1079.

——, 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

——, 1993: Estimating observation error statistics for atmospheric data assimilation. *Ann. Geophys.*, **11,** 634–647.

DAO, 1996: Algorithm theoretical basis document version 1.01. Data Assimilation Office, NASA/Goddard Space Flight Center, Greenbelt, MD. [Available online at http://dao.gsfc.nasa.gov/subpages/atbd.html.]

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.,* **123,** 1128–1145.

——, and A. M. da Silva, 1998: Data assimilation in the presence of forecast bias. *Quart. J. Roy. Meteor. Soc.*, **124,** 269–295.

——, G. Gaspari, C. Redder, L. Rukhovets, and A. M. da Silva, 1999: Maximum-likelihood estimation of forecast and observation error covariance parameters. Part II: Applications. *Mon. Wea. Rev.*, 1835–1849.

Devenyi, D., and T. W. Schlatter, 1994: Statistical properties of 3-hour prediction errors derived from the mesoscale analysis and prediction system. *Mon. Wea. Rev.*, **122,** 1263–1280.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London,* **222A,** 309–368.

Gandin, L. S., 1963: *Objective Analysis of Meteorological Fields* (in Russian). Israel Program for Scientific Translation, 242 pp.

Gaspari, G., and S. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125,** 723–757.

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from rawinsonde data. Part I: The wind field. *Tellus,* **38A,** 111–136.

Lönnberg, P., and A. Hollingsworth, 1986: The statistical structure of short-range forecast errors as determined from rawinsonde data. Part II: The covariance of height and wind errors. *Tellus,* **38A,** 137–161.

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

Lupton, R., 1993: *Statistics in Theory and Practice.* Princeton University Press, 188 pp.

Muirhead, R. J., 1982: *Aspects of Multivariate Statistical Theory.* Wiley, 673 pp.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The Art of Scientific Computing.* 2d ed. Cambridge University Press, 963 pp.

Riishøjgaard, L.-P., 1998: A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus,* **50A,** 42–57.

Rutherford, I., 1972: Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.*, **29,** 809–815.

Sorenson, H. W., 1980: *Parameter Estimation: Principles and Problems.* Marcel Dekker, 382 pp.

Strang, G., 1988: *Linear Algebra and its Applications.* 3d ed. Academic Press, 505 pp.

Thiébaux, H. J., H. L. Mitchell, and D. W. Shantz, 1986: Horizontal structure of hemispheric forecast error correlations for geopotential and temperature. *Mon. Wea. Rev.*, **114,** 1048–1066.

——, L. L. Morone, and R. L. Wobus, 1990: Global forecast error correlation. Part 1: Isobaric wind and geopotential. *Mon. Wea. Rev.*, **118,** 2117–2137.

Wahba, G., and J. Wendelberger, 1980: Some new mathematical methods for variational objective analysis using splines and cross-validation. *Mon. Wea. Rev.*, **108,** 1122–1145.

——, D. R. Johnson, F. Gao, and J. Gong, 1995: Adaptive tuning of numerical weather prediction models: Randomized GCV in three- and four-dimensional data assimilation. *Mon. Wea. Rev.*, **123,** 3358–3369.