DAO Office Note 97-12

Office Note Series on Global Modeling and Data Assimilation

Richard B. Rood, Head
Data Assimilation Office
Goddard Space Flight Center
Greenbelt, Maryland

Maximum-likelihood estimation of forecast and observation error covariance parameters

Dick Dee*
Arlindo da Silva†
Chris Redder*
Greg Gaspari*
Leonid Rukhovets*

Data Assimilation Office, Goddard Laboratory for Atmospheres
* *General Sciences Corporation, Laurel, Maryland*
† *Goddard Space Flight Center, Greenbelt, Maryland*

Goddard Space Flight Center
Greenbelt, Maryland 20771
August 1 1997

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 many practical aspects of implementation. Issues such as bias estimation and correction, parameter identifiability, estimation accuracy, and robustness of the method, are discussed from both a theoretical and a practical point of view.

Three different applications are presented, intended to illustrate both the generality and the limitations of the method. Different observation and forecast error parameters are estimated from a February 1995 time series of observed-minus-forecast residuals, using rawinsonde height data, ship-based sea-level pressure reports, and aircraft wind observations.

It is shown that many statistical parameters usually specified to be constant in operational data assimilation systems in fact vary significantly in both space and time. It is also found that systematic errors in both forecasts and observations cannot be ignored. Finally, the results reported here demonstrate both the necessity and feasibility of on-line tuning of covariance models.

Contents

List of Figures

1 Introduction

This report describes the tools developed at the DAO for estimating unknown parameters of the PSAS observation and forecast error covariance models. The methodology is based on maximum-likelihood covariance parameter estimation as described by Dee (1995). During the last two years it has been applied at the DAO to estimate:

- rawinsonde observation error standard deviations for height, wind components, and relative humidity at different pressure levels; vertical correlation coefficients for height and relative humidity;
- TOVS height retrieval error standard deviations at different pressure levels, both for spatially uncorrelated and correlated error components; horizontal de-correlation length scales and vertical correlation coefficients;
- parameters of the cross-covariances between TOVS height retrieval errors and forecast height errors;
- ship-based sea-level pressure observation error standard deviations;
- aircraft-based wind observation error standard deviations at various pressure levels;
- forecast height and wind error standard deviations; horizontal de-correlation length scales and vertical correlation coefficients for forecast height errors; and various parameters describing the multivariate wind-mass error covariances.

Error covariance models for convential observations as well as for TOVS retrievals are described in DAO 1996, Chapter 5. The initial formulation and implementation of

the PSAS forecast error covariance model will be described in full detail in a separate report. Here we focus on parameter estimation; i.e., we are not primarily concerned with the formulation of a covariance model, but rather with the problem of estimating unknown parameters of a model from observations.

The PSAS covariance model development has been guided by the following considerations:

- maximum use should be made of available observations;
- there should be no restriction to specific categories of covariance models (e.g., isotropic, univariate) or data (e.g., station data, single-level data);
- model formulations and estimation methods must be consistent with the general estimation-theoretical framework of PSAS.

The covariance models will be re-tuned and re-formulated as often as necessary, using the methods described in this report. We developed a a collection of MATLAB programs for solving the following parameter estimation problem:

> Given a multivariate data set $\{\mathbf{v}_k : k = 1,2,\ldots,K\}$ and a covariance model $\mathbf{S}_k(\boldsymbol{\alpha}) \approx \left\langle \mathbf{v}_k \mathbf{v}_k^T \right\rangle$ depending on unknown parameters $\boldsymbol{\alpha}$, find the maximum-likelihood estimate $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}$.

The data set $\{\mathbf{v}_k\}$ typically consists a time series of observed-minus-forecast residuals, restricted to a certain spatial region and time period. The covariance model $\mathbf{S}_k(\boldsymbol{\alpha})$ can involve any number of unknown parameters $\boldsymbol{\alpha}$ (e.g., variance parameters,

de-correlation length scales, vertical correlation coefficients), subject only to the condition that it be positive semi-definite on the model domain. If the parameters are identifiable from the data then the software will produce parameter estimates, as well as error estimates for the parameter values.

This report consists of two parts. In the first we present the method, in general terms but with particular emphasis on many practical aspects of implementation. The second part describes three different applications, intended to illustrate both the generality and the limitations of the method. We include the standard example of estimating covariance parameters from rawinsonde height data, but also discuss the estimation of parameters of univariate and multivariate covariance models, using data from moving observers.

2 Methodology

In this section we present the general methodology for tuning forecast and observation error covariance models to observational data. We first consider forecast and observation errors and their covariances. We then discuss the available observational data and the relationship between the data and the covariance models. Finally we present the maximum-likelihood method for tuning the models to the data, and discuss some important properties of the method.

2.1 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 grid-point values or spectral coefficients. The *forecast error* is then simply

$$\boldsymbol{\varepsilon}_k^f \equiv \mathbf{w}_k^f - \mathbf{w}_k^t. \tag{2.1}$$

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

$$\boldsymbol{\varepsilon}_k^o \equiv \mathbf{w}_k^o - \mathbf{h}_k(\mathbf{w}_k^t). \tag{2.2}$$

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

$$\mathbf{b}_k^f \equiv \langle \boldsymbol{\varepsilon}_k^f \rangle, \qquad \mathbf{P}_k^f \equiv \left\langle (\boldsymbol{\varepsilon}_k^f - \mathbf{b}_k^f)(\boldsymbol{\varepsilon}_k^f - \mathbf{b}_k^f)^T \right\rangle, \tag{2.3}$$

for the observation error mean and covariance

$$\mathbf{b}_k^o \equiv \langle \boldsymbol{\varepsilon}_k^o \rangle, \qquad \mathbf{R}_k \equiv \left\langle (\boldsymbol{\varepsilon}_k^o - \mathbf{b}_k^o)(\boldsymbol{\varepsilon}_k^o - \mathbf{b}_k^o)^T \right\rangle \tag{2.4}$$

and for the cross-covariance between observation and forecast errors

$$\mathbf{X}_k \equiv \left\langle (\boldsymbol{\varepsilon}_k^o - \mathbf{b}_k^o)(\boldsymbol{\varepsilon}_k^f - \mathbf{b}_k^f)^T \right\rangle. \tag{2.5}$$

Here $\langle \cdot \rangle$ denotes the *ensemble averaging* or *expectation operator*, whose proper definition involves the (typically unknown) joint probability distribution of forecast and observation errors.

Atmospheric data assimilation systems involve many different observing systems, and the observation operator $\mathbf{h}_k$ and its associated observation error must be considered separately 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 $\mathbf{w}_k^o$. In this paper the vector $\mathbf{w}_k^o$ will always denote a specific subset of the observations, obtained by restricting to a single data type 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 (2.3–2.5) are modeled by introducing various simplifying assumptions about the underlying error distributions. For example, the means $\mathbf{b}_k^f$ and $\mathbf{b}_k^o$ 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 independent, so that $\mathbf{X}_k \equiv 0$. We will mention several additional assumptions for errors associated with

specific data types below. Some are not necessarily realistic, but often the information required to remove them is 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 generally 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 often taken to be identical; i.e., independent of time and station. For spatially distributed univariate observations, this set of assumptions leads to an observation error covariance model $\mathbf{R}$, with

$$[\mathbf{R}]_{ij}^{(mn)} = \sigma^{(m)}\sigma^{(n)}\nu^{(mn)}\delta(r_{ij}). \tag{2.6}$$

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

$$\delta(r) = \begin{cases} 1, & \text{if } r = 0 \\ 0, & \text{otherwise} \end{cases} . \tag{2.7}$$

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

Understanding 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 which 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 Section 3 all involve simple representations of forecast error covariances, based on univariate single-level isotropic models of the form

$$[\mathrm{Cov}]_{ij} = \sigma^2 \rho(r_{ij}; L). \tag{2.8}$$

These models depend on two parameters only: the error standard deviation σ and a de-correlation length scale L. Appendix A describes a few choices for the function $\rho(r; L)$ as well as a definition of the parameter L. Isotropic models are based on the 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 is due to the fact that error correlations have traditionally been calculated by averaging data over relatively long periods of time, e.g. one to three months (Hollingsworth and Lönnberg 1986;

Lönnberg and Hollingsworth 1986).

Models of the form (2.8) will be used in this study to describe univariate forecast height errors at a fixed pressure level within a limited spatial region, so that the parameter σ represents the forecast height error standard deviation for that particular level and region. Estimation of this parameter based on regional time series data will, at best, produce a spatial and temporal average of the actual forecast error standard deviations. We will consider different choices for the horizontal correlation function $\rho(r; L)$ in (2.8), only in order to examine the effect on covariance parameter estimation of some of the uncertainties inherent in the description of forecast errors.

Spatially correlated multivariate wind error covariances can be modeled based on (2.8) as follows. Let $\varepsilon^u, \varepsilon^v$ denote the wind error components at a location (x, y) on a fixed pressure level. Define an *error stream function* ψ and *error velocity potential* χ on that pressure level, and write

$$\begin{bmatrix} \varepsilon^u \\ \varepsilon^v \end{bmatrix} = \begin{bmatrix} -\frac{\partial \psi}{\partial y} + \frac{\partial \chi}{\partial x} \\ \frac{\partial \psi}{\partial x} + \frac{\partial \chi}{\partial y} \end{bmatrix}. \tag{2.9}$$

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. See, for example, Daley (1991, Section 5.2) for details. 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 (2.8).

For the applications in section 3 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 among wind errors and height errors. The

coupling between wind and height errors is known to be strong in mid-latitudes; 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 object is only to estimate wind observation error covariance parameters, then a model derived from (2.9) will be adequate.

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 will write

$$\mathbf{P}_k^f \approx \mathbf{P}_k^f(\boldsymbol{\alpha}^f), \qquad \mathbf{R}_k \approx \mathbf{R}_k(\boldsymbol{\alpha}^o), \qquad \mathbf{X}_k \approx \mathbf{X}_k(\boldsymbol{\alpha}^x) \tag{2.10}$$

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

2.2 Observational residuals

The observation operator introduced in (2.2) is a device for comparing forecasts with observations. The *observed-minus-forecast residuals* defined by

$$\mathbf{v}_k \equiv \mathbf{w}_k^o - \mathbf{h}_k(\mathbf{w}_k^f) \tag{2.11}$$

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

$$\mathbf{v}_k \approx \boldsymbol{\varepsilon}_k^o - \mathbf{H}_k \boldsymbol{\varepsilon}_k^f, \tag{2.12}$$

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

$$\mathbf{H}_k \equiv \left. \frac{\partial \mathbf{h}_k}{\partial \mathbf{w}} \right|_{\mathbf{w}=\mathbf{w}_k^f}. \tag{2.13}$$

Equation (2.12) is obtained by linearizing (2.11) about the forecast state and using (2.1) and (2.2). The accuracy of the approximation (2.12) 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, one can also define the *observed-minus-observed residuals*

$$\mathbf{r}_k \equiv \mathbf{w}_k^{o1} - \mathbf{w}_k^{o2}, \tag{2.14}$$

where the second superscript refers to the instrument. For (2.14) to make sense the observation operators $\mathbf{h}_k^1$ and $\mathbf{h}_k^2$ associated with the two separate instruments must be compatible, in the sense that they both map to the same observation space; see (2.2). They need not be identical, however, and so it follows from (2.2) applied to each data type that

$$\mathbf{r}_k \approx \boldsymbol{\varepsilon}_k^{o1} - \boldsymbol{\varepsilon}_k^{o2}. \tag{2.15}$$

As an example, consider a set of temperatures retrieved from remotely sensed radiances, valid at a particular time t_k. If the retrievals are co-located with a set of rawinsonde temperature observations, then the residuals (2.14) can be computed. In this case the observation operators, although very different, are compatible. Some kind of interpolation will generally be required in order to co-locate the retrievals

with the rawinsonde observations, and therefore (2.15) 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 (2.11) are easily obtained from (2.12):

$$\langle \mathbf{v}_k \rangle \approx \mathbf{b}_k^o - \mathbf{H}_k \mathbf{b}_k^f, \tag{2.16}$$

$$\left\langle (\mathbf{v}_k - \langle \mathbf{v}_k \rangle)(\mathbf{v}_k - \langle \mathbf{v}_k \rangle)^T \right\rangle \approx \mathbf{R}_k - \mathbf{X}_k \mathbf{H}_k^T - \mathbf{H}_k \mathbf{X}_k^T + \mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^T. \tag{2.17}$$

We used the additional approximation $\langle \mathbf{H}_k \cdot \rangle \approx \mathbf{H}_k \langle \cdot \rangle$; both (2.16) and (2.17) are exact for linear observation operators.

Dee and da Silva (1997) show how the mean equation (2.16) 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; i.e.,

$$\langle \mathbf{v}_k \rangle = \boldsymbol{\mu}_k. \tag{2.18}$$

For now we regard the mean $\boldsymbol{\mu}_k$ as known; see however Section 2.3.2 below.

The covariance equation (2.17) can be used to tune parameters of the forecast and observation covariance models discussed in the previous section. Substitution of (2.10) and (2.18) gives

$$\left\langle (\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^T \right\rangle \approx \mathbf{S}_k(\boldsymbol{\alpha}), \tag{2.19}$$

where

$$\mathbf{S}_k(\boldsymbol{\alpha}) = \mathbf{S}_k(\boldsymbol{\alpha}^f, \boldsymbol{\alpha}^o, \boldsymbol{\alpha}^x) \tag{2.20}$$

$$= \mathbf{R}_k(\boldsymbol{\alpha}^o) - \mathbf{X}_k(\boldsymbol{\alpha}^x)\mathbf{H}_k^T - \mathbf{H}_k\mathbf{X}_k^T(\boldsymbol{\alpha}^x) + \mathbf{H}_k\mathbf{P}_k^f(\boldsymbol{\alpha}^f)\mathbf{H}_k^T. \tag{2.21}$$

Thus, models for forecast and observation error covariances imply a model for the observed-minus-forecast residuals, and (2.19) provides a relationship between the models and the data.

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

$$\langle \mathbf{r}_k \rangle \approx \mathbf{b}_k^{o1} - \mathbf{b}_k^{o2}, \tag{2.22}$$

$$\left\langle (\mathbf{r}_k - \langle \mathbf{r}_k \rangle)(\mathbf{r}_k - \langle \mathbf{r}_k \rangle)^T \right\rangle \approx \mathbf{R}_k^1 - \mathbf{Y}_k - \mathbf{Y}_k^T + \mathbf{R}_k^2, \tag{2.23}$$

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

$$\mathbf{Y}_k \equiv \left\langle (\boldsymbol{\varepsilon}_k^{o1} - \mathbf{b}_k^{o1})(\boldsymbol{\varepsilon}_k^{o2} - \mathbf{b}_k^{o2})^T \right\rangle. \tag{2.24}$$

The mean equation (2.22) can be used to estimate and correct the bias in one set of observations based on another unbiased (or bias-corrected) set; see, for example, **[refs]**. For now we assume

$$\langle \mathbf{r}_k \rangle = \boldsymbol{\zeta}_k \tag{2.25}$$

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

$$\left\langle (\mathbf{r}_k - \boldsymbol{\zeta}_k)(\mathbf{r}_k - \boldsymbol{\zeta}_k)^T \right\rangle \approx \mathbf{T}_k(\boldsymbol{\alpha}) \tag{2.26}$$

with

$$\mathbf{T}_k(\boldsymbol{\alpha}) = \mathbf{T}_k(\boldsymbol{\alpha}^{o1}, \boldsymbol{\alpha}^{o2}, \boldsymbol{\alpha}^y) \tag{2.27}$$

$$= \mathbf{R}_k^1(\boldsymbol{\alpha}^{o1}) - \mathbf{Y}_k(\boldsymbol{\alpha}^y) - \mathbf{Y}_k^T(\boldsymbol{\alpha}^y) + \mathbf{R}_k^2(\boldsymbol{\alpha}^{o2}). \tag{2.28}$$

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; equation (2.26) establishes a relationship between the models and the data analogous with (2.19).

2.3 Covariance tuning

In the previous section we developed relationships between forecast and observation error covariance models and residuals obtained from observed data. We now consider a general method for adjusting the free model parameters in order to improve the consistency between the models and any finite subset of the data. To simplify the presentation we will use the notation for observed-minus-forecast residuals (i.e., the data are $\mathbf{v}_k$ and the covariance model is $\mathbf{S}_k(\boldsymbol{\alpha})$); however the following applies equally well to observed-minus-observed residuals.

2.3.1 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* $\widehat{\boldsymbol{\alpha}}$ is obtained by finding the maximum (the mode) of the pdf.

To apply the maximum-likelihood method to our problem, we need to assume a

probability density for the underlying process $\{\mathbf{V}_k\}$ which has the covariances given by (2.21). 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),

$$\begin{aligned} p(\{\mathbf{v}_k\};\boldsymbol{\alpha}) &= \prod_{k=1}^{K} p(\mathbf{v}_k;\boldsymbol{\alpha}) \\ &\propto \prod_{k=1}^{K} (\det \mathbf{S}_k(\boldsymbol{\alpha}))^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{v}_k-\boldsymbol{\mu}_k)^T\mathbf{S}_k^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k-\boldsymbol{\mu}_k)\right]. \end{aligned} \tag{2.29}$$

The maximum-likelihood estimate $\widehat{\boldsymbol{\alpha}}$ is obtained by maximizing (2.29), or equivalently, by minimizing the *log-likelihood* function

$$f(\boldsymbol{\alpha}) = \frac{1}{K}\sum_{k=1}^{K}\left[\log\det \mathbf{S}_k(\boldsymbol{\alpha}) + (\mathbf{v}_k-\boldsymbol{\mu}_k)^T\mathbf{S}_k^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k-\boldsymbol{\mu}_k)\right]. \tag{2.30}$$

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

For a fixed data set $\{\mathbf{v}_k\}$ 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 (2.6) and (2.8) therefore depends primarily on the size of the data set. As we show below, the error variance in the parameter estimates is proportional to $1/\nu$, with ν the size of the data set. 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 data items to estimate a single parameter with meaningful accuracy.

The log-likelihood function formulation (2.30) is appropriate for estimating parameters associated with a time-dependent covariance model $\mathbf{S}_k(\boldsymbol{\alpha})$. Referring to (2.21), it is clear that time-dependence arises when either the forecast and/or the observation error covariance model formulations involve time as an independent variable, or when the observation operator itself depends on time. This is the case, for example, for observation operators associated with moving platforms such as ships or aircraft.

In the special case when the covariance model (2.21) is stationary; i.e.,

$$\mathbf{S}_k(\boldsymbol{\alpha}) = \mathbf{S}(\boldsymbol{\alpha}), \tag{2.31}$$

the log-likelihood function simplifies to

$$f(\boldsymbol{\alpha}) = \log\det \mathbf{S}(\boldsymbol{\alpha}) + \text{trace}\left[\mathbf{S}^{-1}(\boldsymbol{\alpha})\overline{\mathbf{S}}\right] \tag{2.32}$$

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

$$\overline{\mathbf{S}} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^T. \tag{2.33}$$

Equations (2.32) and (2.30) are equivalent when (2.31) holds because

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}(\mathbf{v}_k - \boldsymbol{\mu}_k)^T\mathbf{S}^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k) \\
&\qquad = \frac{1}{K}\sum_{k=1}^{K}\text{trace}\left[\mathbf{S}^{-1}(\boldsymbol{\alpha})(\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^T\right] \\
&\qquad = \text{trace}\left[\mathbf{S}^{-1}(\boldsymbol{\alpha})\frac{1}{K}\sum_{k=1}^{K}(\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^T\right] \\
&\qquad = \text{trace}\left[\mathbf{S}^{-1}(\boldsymbol{\alpha})\overline{\mathbf{S}}\right].
\end{aligned} \tag{2.34}$$

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 (2.32) 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 (2.32) of the log-likelihood function is appropriate (and computationally convenient) when neither the covariance models nor the observation operator depend on time. The procedure in this case is first to compute the sample covariance of the data (2.33), and then to minimize (2.32) with respect to $\boldsymbol{\alpha}$. Many of the covariance models currently implemented in operational data assimilation systems do not depend on time. Time-independent observation operators arise from stationary observing systems, such as rawinsonde networks. Truly stationary observation operators do not occur in practice, however, because of occasional missing reports or data rejected by quality control. Still, in case of a stationary observing network 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

$$\overline{\mathbf{S}}_{ij} = \frac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} ([\mathbf{v}_k]_i - [\boldsymbol{\mu}_k]_i)([\mathbf{v}_k]_j - [\boldsymbol{\mu}_k]_j), \tag{2.35}$$

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 insure that all K_{ij} exceed a certain threshold; however, our experience indicates that the parameter estimates are not very sensitive to this.

2.3.2 Bias estimation

In order to implement the parameter estimation procedure outlined above, it is necessary to specify the residual means $\boldsymbol{\mu}_k$ in (2.30) or in (2.33). These depend on the mean observation errors $\mathbf{b}_k^o$ and on the mean forecast errors $\mathbf{b}_k^f$ (see (2.16)), neither of which are accurately known in practice.

For the purpose of tuning covariance models there are two choices. The first is to simply ignore the bias by assuming it to be zero; i.e., to set $\boldsymbol{\mu}_k = 0$. 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 system which does not explicitly account for bias. 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 1997).

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 should naturally 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

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_k(\boldsymbol{\beta}) \tag{2.36}$$

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 (2.30), after substitution of (2.36), 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 think it very 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 data assimilation are generally defined in terms of time averages, because true ensemble averaging is not possible with only a single available realization. Accordingly one can estimate $\boldsymbol{\mu}_k$ by calculating the time average of the residuals. In the stationary case (i.e., for stationary observing systems) we then have

$$[\boldsymbol{\mu}_k]_i = [\overline{\boldsymbol{\mu}}]_i \,, \tag{2.37}$$

where

$$[\overline{\boldsymbol{\mu}}]_i = \frac{1}{K_i} \sum_{k=1}^{K_i} [\mathbf{v}_k]_i \tag{2.38}$$

with K_i the number of reports from station i. In the general (non-stationary) 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 will experiment with such a technique in Section 3.

The presence of bias which 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 analyses, independently of the covariance models used for the analysis. Although it is not the subject of this report, bias estimation and correction should take precedence over covariance modeling and estimation.

2.3.3 Identifiability of the parameters

The maximum-likelihood method is appealing because of 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 (2.30) must have a unique global minimum with probability one; i.e., 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, provided the residuals 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 the next section, in which the data 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 data set. However, 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:

$$\mathbf{A}_{ij} \equiv \left. \frac{\partial^2 f}{\partial \alpha_i \partial \alpha_j} \right|_{\alpha=\widehat{\alpha}} . \tag{2.39}$$

The Hessian matrix can be approximated by finite differences, or it may in fact 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 $\widehat{\boldsymbol{\alpha}}$,

$$f(\boldsymbol{\alpha}) \approx f(\widehat{\boldsymbol{\alpha}}) + \frac{1}{2}(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^T \mathbf{A} (\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}), \tag{2.40}$$

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

Equation (2.40) 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 $\widehat{\boldsymbol{\alpha}}$ along the direction of an eigenvector of $\mathbf{A}$ produces a change in f by an amount which is proportional to the corresponding eigenvalue. 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 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 which is not specifically connected with the maximum-likelihood method; it is simply a practical requirement for any parameter estimation method which is based on minimizing a cost function.

2.3.4 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 which produces more accurate parameter estimates. In practice we only have finite data sets 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.

Here we suppose, for the moment, that the *modeling assumption* holds: all assumptions about the data as expressed by the likelihood function are actually valid. In that case the parameter estimates produced by minimizing (2.30) are truly the maximum-likelihood estimates. Then it can be shown (e.g., Sorenson 1980, Theorem 5.4) that in the limit of infinite data the estimates approach a normal distribution with

$$\langle \widehat{\boldsymbol{\alpha}} \rangle = \boldsymbol{\alpha}^\star, \qquad \left\langle (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star)^T \right\rangle = \frac{2}{\nu} \mathbf{A}^{-1}. \tag{2.41}$$

Here $\mathbf{A}$ is the Hessian of the log-likelihood function (see (2.39)), and ν is the number of degrees of freedom associated with the estimation problem. In the general (instationary) case corresponding to (2.30) we have

$$\nu = \sum_{k=1}^{K} n_k, \qquad n_k = \dim \mathbf{v}_k. \tag{2.42}$$

In the truly stationary case where $n_k = n = \text{const}$, we would have $\nu = nK$; however, (2.42) should be applied to account for missing data (see the discussion following (2.34)). 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 (2.41) is the lower bound of the Cramér-Rao inequality (Sorenson 1980, Section 3B). 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 pdf of the data 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 data set increases (see also Lupton 1993, Chapter 10).

We routinely use (2.41) to estimate the standard errors of the parameter estimates under the modeling assumption. For any given data set and covariance model formulation, the validity of (2.41) for finite ν can be checked using a Monte-Carlo experiment with synthetic data; it turns out to be quite accurate for the applications described in Section 3.

It is important to keep in mind that the maximum-likelihood standard errors represent idealized accuracy estimates; in practice they should be regarded as a lower bound 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 modeling assumption. 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.

2.3.5 Robustness of the parameter estimates

The fact that parameters associated with a particular covariance model can be estimated from a data set (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: there may be a strong state-dependent component of error which 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. We will address these issues and some practical tools for dealing with them in a separate report.

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 does depend on the 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 Section 3.

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 the 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 (2.30) (or (2.32), 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 data set. Traditional methods for determining covariance parameters from observed data are generally based on a least-squares criterion (Rutherford 1972; Thiébaux *et al.* 1986; Bartello and Mitchell 1991).

As stated earlier, an advantage of the maximum-likelihood criterion is that it can be made entirely consistent with the complete set of assumptions that have been invoked in the actual data assimilation procedure. However, if these assumptions 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

$$\begin{aligned} \mathbf{S}(\boldsymbol{\alpha}) &= \mathbf{S}(\sigma_1, \sigma_2, \boldsymbol{\theta}) \\ &= \sigma_1^2 \mathbf{S}_1 + \sigma_2^2 \mathbf{S}_2(\boldsymbol{\theta}), \end{aligned} \tag{2.43}$$

where $\mathbf{S}_1$ is constant, and both $\mathbf{S}_1$ and $\mathbf{S}_2$ are positive definite. Such a model is sufficiently general for most applications considered in this report: 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

$$\lambda = \left(\frac{\sigma_1}{\sigma_2} \right)^2, \tag{2.44}$$

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 (2.43), 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 on-line in a data assimilation system, for the adaptive tuning of system parameters. However, it can also be applied off-line 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 (B.10) with (2.30) (or, for stationary models, (B.15) with (2.32)). It is the case for both methods that if either the covariance model formulation or the data change, so will the parameter estimates produced by both methods 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 data set can be compared by simply interchanging the cost functions. Our experiments, some of which are reported in Section 3, indicate that the differences are insignificant.

3 Applications

We now discuss three different applications of the methodology described in the previous section. The quality-controlled data used in this study were obtained from a February 1995 time series of observed-minus-forecast residuals produced by the Goddard EOS Data Assimilation System, Version 1.2 (GEOS-1.2 DAS) (Pfaendtner *et al.* 1995).

3.1 Rawinsonde height residuals

We first consider the estimation of rawinsonde height error standard deviations and forecast height error covariance parameters from observed-minus-forecast height residuals over North America for the month of February 1995. This represents a classic example of covariance parameter estimation for atmospheric data assimilation; see e.g., Gandin (1963, Section 2.3); Daley (1991, Section 4.3). Our purpose here is to test the performance of the maximum-likelihood method, and to assess the accuracy of the parameter estimates in light of various uncertainties inherent in both the models and the data.

We will use (2.6) for modeling rawinsonde height error covariances. At each fixed pressure level this model reduces to

$$[\mathbf{R}]_{ij} = (\sigma_h^o)^2 \delta(r_{ij}), \tag{3.1}$$

where i, j are station indices, r_{ij} is the chordal distance between the stations and σ_h^o is the observation height error standard deviation. As stated in Section 2, this model assumes that observation errors associated with different soundings are statistically independent, and that their standard deviations are identical. The latter assump-

tion may not be adequate if instruments from more than a single manufacturer are involved. Moreover, the model (3.1) cannot properly account for representativeness error, which is state-dependent and therefore spatially correlated (Daley 1993).

Single-level forecast height error covariances will be modeled by (2.8), using the *spline-windowed powerlaw function* ρ_w (see Appendix A) to represent horizontal correlations. For our experiments we set $r_* = 6000km$, which is the distance beyond which the modeled correlations are identically zero. Later we will consider alternative correlation models as well. Thus, the covariance between forecast height errors at locations i and j is modeled by

$$\left[\mathbf{P}^f\right]_{ij} = (\sigma_h^f)^2 \rho_w(r_{ij}; L_h) \tag{3.2}$$

with σ_h^f the forecast height error standard deviation and L_h the *de-correlation length scale*; see (A.3).

Combining (3.1) and (3.2) with the assumption that observation errors are uncorrelated with forecast errors, the covariance of the observed-minus-forecast height residuals is

$$[\mathbf{S}]_{ij} = (\sigma_h^o)^2 \delta(r_{ij}) + (\sigma_h^f)^2 \rho_w(r_{ij}; L_h). \tag{3.3}$$

This expression completely specifies the residual covariances, except for the three parameters σ_h^o, σ_h^f, and L_h which will be estimated from the data.

We now turn our attention to the data. Day-time rawinsonde temperature measurements are affected by sunlight, which can cause systematic errors in reported heights. Corrections may be applied to the data in order to reduce the effects of solar radiation; the method of correction generally depends on the manufacturer of the rawinsonde equipment (Mitchell *et al.* 1996). In some cases corrections are applied at the source;

i.e., prior to communicating the reports to the operational weather centers. Other than that, no corrections were applied to the data used for this study. In order to eliminate the possible contaminating effect of this aspect of quality control, we use night-time observations only for this study.

We consider a report to be a night-time report when the sun is below the horizon at the time and location of the observation. When selecting a set of stations for covariance parameter estimation we also require that all stations in the set produce a certain minimum number (usually 10) of simultaneous night-time reports during the period in question. Figure 1 shows, for example, a subset of the North American rawinsonde stations which produced at least 10 simultaneous night-time 500mb reports during the month of February 1995.

Having selected a subset of stations in this manner, we computed the mean residuals by averaging all night-time residuals at each station (see (2.37, 2.38)). The results are shown at four different pressure levels in Figure 2. Closed disks indicate positive mean residuals; circles correspond to negative values. The diameter of each disk or circle is proportional to the absolute value of the mean; the minimum, median, and maximum values at each level are indicated in each panel.

It is likely that the occasionally large monthly mean residuals are primarily due to systematic errors in the forecast model. For example, the means at 100mb clearly show a large-scale spatial pattern; it is difficult to imagine that this would be due to observational bias. The magnitudes of monthly mean height residuals computed for different periods are similar, although the detailed spatial distributions generally depend on the prevailing large-scale circulation. By introducing the assumption that the observations are unbiased, it is possible to estimate forecast bias from the time

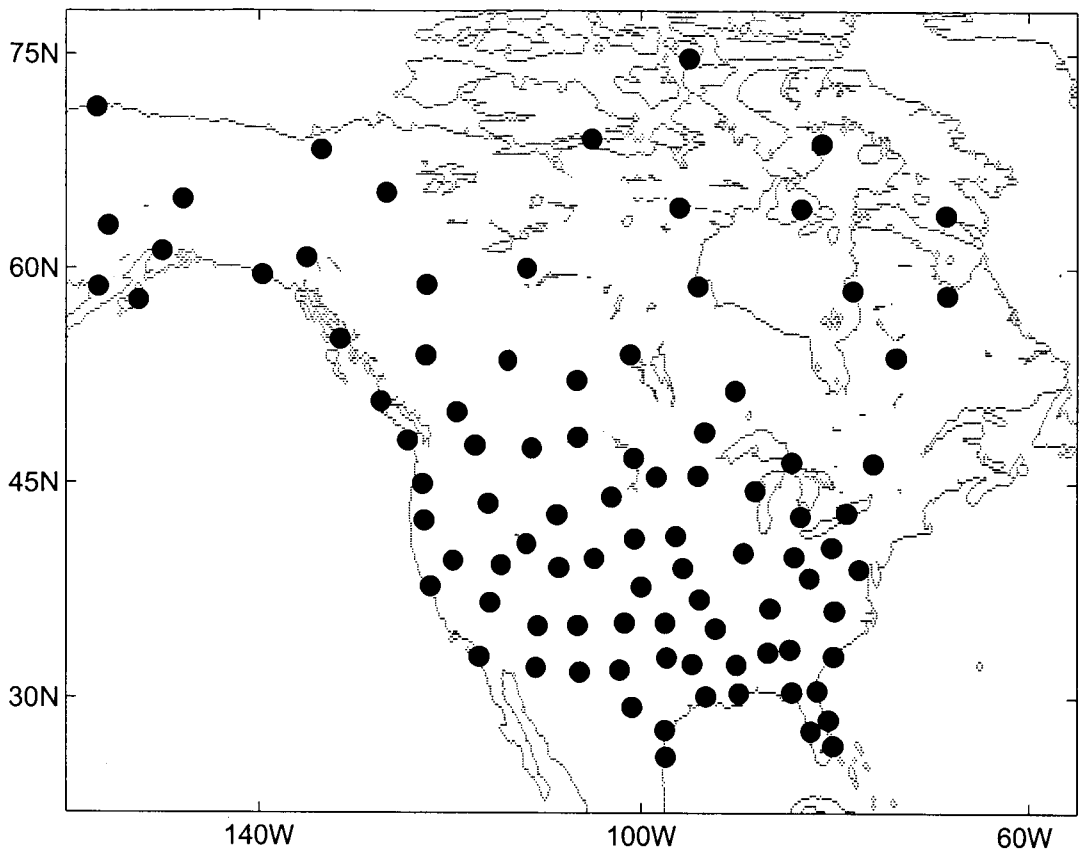

Figure 1: Locations of North-American rawinsonde stations which produced at least 10 simultaneous 500mb night-time reports during the month of February 1995.

series of residuals. Dee and da Silva (1997) have developed a sequential forecast bias estimation algorithm that can be incorporated into existing statistical data assimilation systems. The algorithm will produce multivariate forecast bias estimates that can be used to reduce climate errors in assimilated data sets.

The data set consists of 3447 night-time reports from 86 North-American stations at 850mb, 3684 reports from 95 stations at 500mb, 3628 reports from 95 stations at 250mb, and 3195 reports from 84 stations at 100mb. The covariance model (3.3) does not depend on time so that the stationary form (2.32) of the maximum-likelihood cost

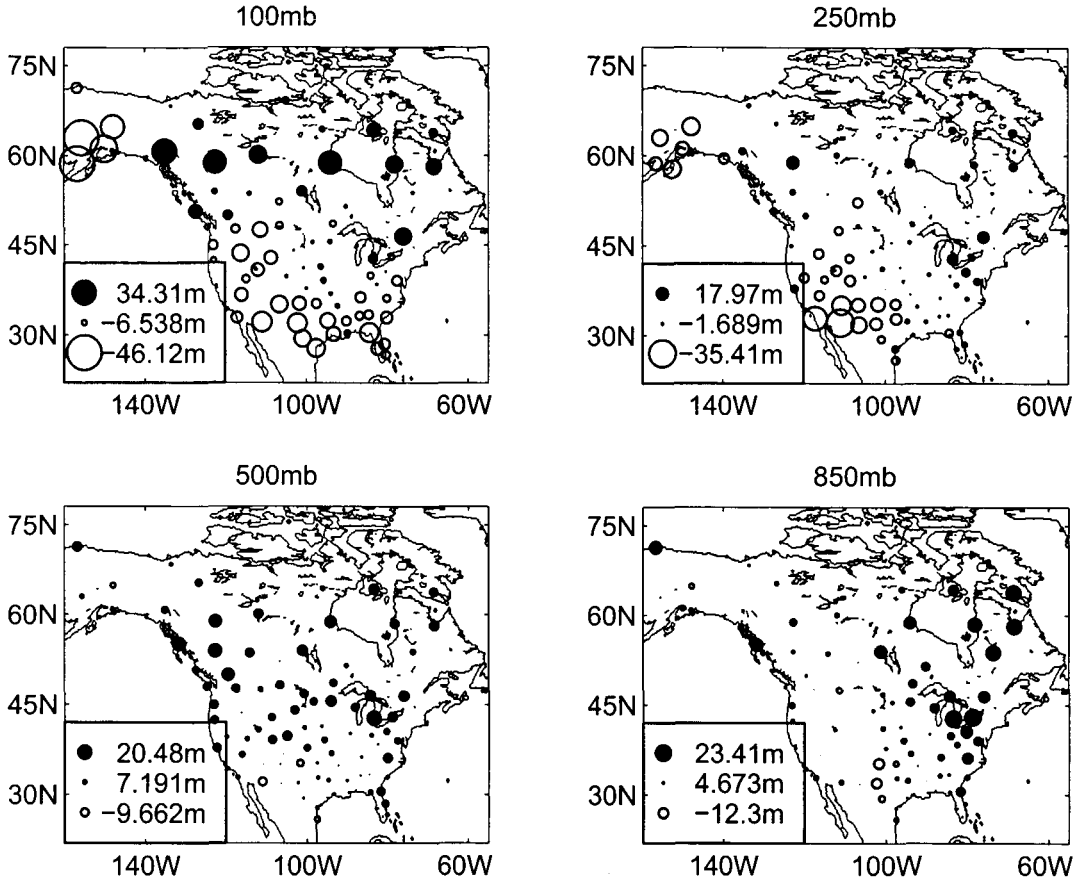

Figure 2: Mean observed-minus-forecast night-time height residuals for February 1995, at stations with at least 10 simultaneous night-time reports, at 100mb, 250mb, 500mb, and 850mb. These are station averages of all night-time residuals during the month. Closed disks correspond to positive values; open circles to negative values. The diameter of each disk or circle is proportional to the absolute value of the mean. Maximum, median, and minimum values are indicated in each panel.

function may be used in this case. We therefore first estimated the sample covariance $\overline{\mathbf{S}}$ from the data at each pressure level by applying (2.35), using the threshhold $K_{ij} \geq 10$. Changing the threshold value to 1 (the lowest possible) or to 40 (the maximum number of reports per station being 56) did not have a significant effect on any of the parameter estimates. The reason is that the total number of data is not significantly

reduced by removing those stations which report infrequently.

Given the sample covariance $\overline{\mathbf{S}}$ at a fixed pressure level, the cost function (2.32) can be minimized as a function of the covariance parameters σ_h^o, σ_h^f, and L_h. We use a quasi-Newton method with a BFGS update (Gill *et al.* 1981) for this purpose, allowing the scheme to approximate cost function gradients as needed by finite differences. Occasionally the scheme has trouble converging when the initial guess for the parameters is poor; then a simplex search method (Nelder and Mead 1965) is used instead. Optimization is considered complete when an iteration results in a relative change of less than 10^{-4} in each parameter estimate as well as in the value of the cost function. We also computed the maximum-likelihood parameter accuracies as in (2.41), using finite differences to approximate the Hessian matrix $\mathbf{A}$. The entire procedure requires repeated evaluation of the cost function—typically on the order of 50–100 times in this three-parameter example, depending on the initial guess for the parameters—and takes only a few seconds on a moderately powerful desktop computer.

The maximum-likelihood parameter estimates as well as their estimated standard errors are shown in Figure 3. As expected, observation and forecast error standard deviations increase with height, although not monotonically. The estimated standard errors for the observation error standard deviations are roughly $2-3\%$ at all levels, indicating that this is the most easily identifiable parameter. The standard errors for the forecast error standard deviations increase to about 6% at higher levels. The forecast error de-correlation length scale estimates vary between 530±110km at 1000mb, 520±20km at 500mb, and 1250±110km at 20mb. The relatively large uncertainties in the length scale estimates at 1000mb and at 20mb are due to the fact that there are fewer data available there.

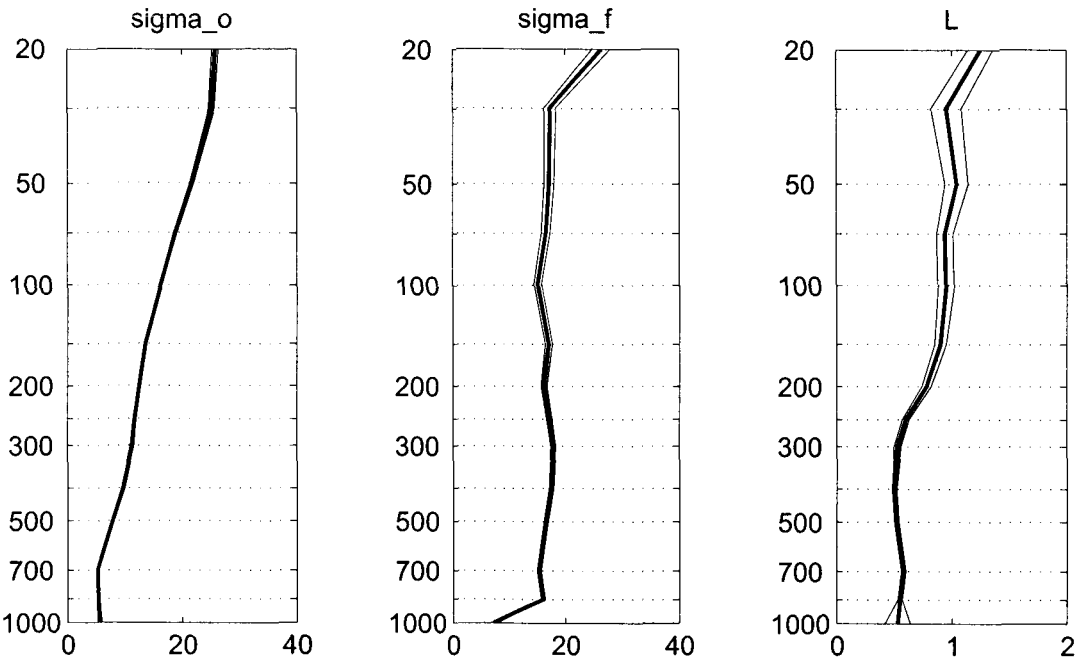

Figure 3: Maximum-likelihood estimates, based on night-time data only, of rawinsonde height error standard deviations (in meters), forecast height error standard deviations (in meters), and forecast height error de-correlation length scales (in thousands of kilometers), shown as a function of pressure (thick curves). Also shown are the estimated standard errors (thin curves), although these are barely visible in the left-most panel.

The estimated standard errors, obtained from the asymptotic theory described in Section 2.3.4, do not necessarily provide a realistic measure of the actual uncertainties in these monthly parameter estimates. The reason for this is that the estimates are not truly maximum-likelihood estimates, since many of the assumptions about the data that enter into the maximum-likelihood formulation are in fact violated. Let us introduce, for the sake of discussion, the *model hypothesis* which states that all assumptions made in modeling the data are in fact satisfied. Under the model hypothesis the uncertainty in the parameter estimates is due only to sampling error: the estimates depend on a finite number of noisy data. The effect of sampling error

is different for each parameter and depends on the nature of the model. For example, the error bars on the de-correlation length scale estimates in Figure 3 are relatively large compared to those on the estimates of the standard deviations: the data contain more useful information about the latter. Ultimately, the standard errors obtained from the asymptotic theory are useful because they indicate (i) whether a parameter can be actually identified from the available data and (ii) whether the parameter uncertainty due to sampling error is acceptable. The answer to both questions, for the three-parameter case presented here, is affirmative.

A better indication of the actual uncertainty in the monthly parameter estimates can be obtained by changing the data selection in various ways. The standard error estimates obtained for the data selection presented above suggest that the sampling error would still be acceptable even if the number of data were significantly smaller. Surely it does not require thousands of observations to estimate only three identifiable covariance parameters; this insight provided the basis for on-line method proposed by Dee (1995). This opens up a number of interesting possibilities for studying the dependence of the parameter estimates on the spatial and temporal selection of data.

As an example, we emulated an on-line estimation procedure by cycling through the month of February 1995 and re-estimating covariance parameters each day, based on the most recent 10 days of data. We used night-time rawinsonde reports only, so that the daily parameter estimates are always based on a subset (a sliding 10-day window) of the one-month data set described earlier. The procedure starts on February 10, using night-time reports from the period February 1–10 only. The resulting parameter estimates at four pressure levels are shown as a function of time in Figure 4.

The variability in the estimates is remarkable in several respects. The estimated

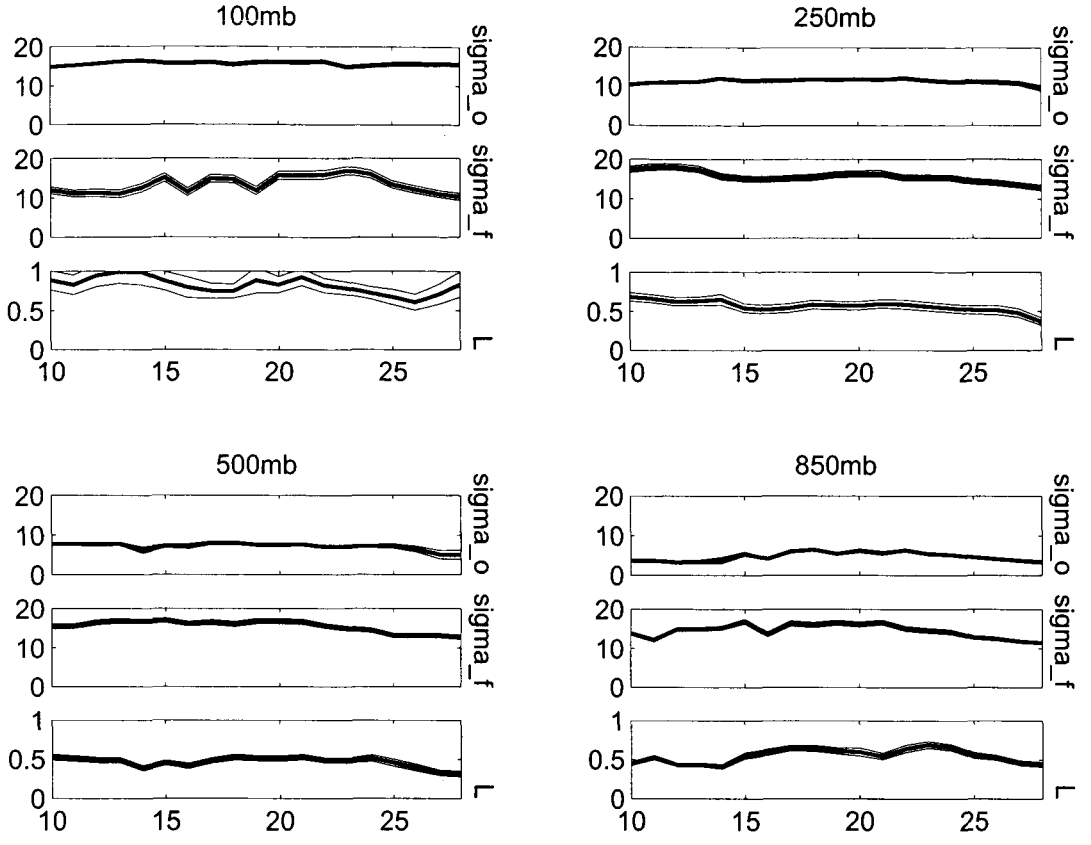

Figure 4: Maximum-likelihood covariance parameter estimates as a function of time (in days). Shown are the estimated rawinsonde height error standard deviations (in meters), forecast height error standard deviations (in meters), and forecast height error de-correlation length scales (in thousands of kilometers) at 100mb, 250mb, 500mb, 850mb. The estimates are produced once a day and are based on the latest 10 days of available night-time reports. The thin curves indicate the estimated standard errors.

observation error standard deviations range between 14.7m and 16.4m at 100mb, between 9.6m and 12.1m at 250mb, between 5.0m and 8.0m at 500mb, and between 3.3m and 6.4m at 850mb. Those are rather large variations for a parameter which is usually presumed to be a function of pressure only! The sampling error, even with a 10-day data set, is too small to be of influence (the standard error curves for the estimates are barely visible in the figure). The variability of the estimated forecast

error standard deviations is not unexpected, since forecast errors are state-dependent. Note that the estimated forecast error standard deviations at 500mb actually exceed those at 100mb during most of the month. The length scale estimates at 500mb, for example, change from 527km on February 21 to 313km on February 28.

Current operational data assimilation systems use constant (in time) parameters to describe most observation error statistics; these values are usually tuned on the basis of at least a month of data. The variability observed in the 10-daily parameter estimates can be interpreted as an uncertainty estimate for the monthly parameter estimates: our results indicate, for example, that this uncertainty for the rawinsonde height error parameter at 500mb is at least 40%. This result is disturbing since this parameter, among all parameters used to describe observation error statistics, is the easiest to estimate, and probably has the largest impact on analysis accuracy.

Another way of assessing parameter uncertainty is by changing the model hypothesis and examining the effect of such a change on the parameter estimates. There are many dubious aspects to the model hypothesis, and a complete sensitivity analysis would not be practical. On the other hand, even a limited analysis can be valuable if it helps to identify particular components of the model formulation that can have a significant impact on the results.

Consider, for example, the specific form of the function used to represent forecast height correlations in the present application. Simultaneous estimation of observation and forecast error standard deviations from residuals is possible only by virtue of the fact that the forecast errors are spatially correlated. The precise nature of the correlations is highly uncertain and the choice of the function used to model them must have an effect on the estimated standard deviations, because the values of the

likelihood function depend on this choice. To illustrate this, we repeated the parameter estimation procedure for night-time data, using the same residual covariance model (3.3), but now with the *powerlaw* and the *compactly supported spline* (see Appendix A) representing the isotropic forecast error correlations. Figure 5 shows the results. The impact on the estimates of the error standard deviations is not much larger than that of sampling error; compare Figure 3. The de-correlation length scale estimates do change significantly depending on the choice of correlation model, but this is not surprising since this parameter describes only the behavior of the model near the origin (see Figure 14). We conclude that, in this case at least, the choice of forecast error correlation model does not significantly affect the estimates of forecast and observation error standard deviations.

An intriguing question is whether any of the isotropic correlation models considered so far actually describe the forecast errors well. The minimum values of the maximum-likelihood functional obtained for each choice of correlation model provides some information about this: since each of the models depends on the same number (one, in this case) of parameters, the value of the functional indicates which model provides the better fit. When the parameters are tuned to the complete month of night-time data it is found that, of the three candidate models considered, the spline-windowed powerlaw consistently provides the best fit at all levels. However, this is not true when only a 10-day window of data are used. In the cycling experiment described above, there was no clear preference toward any of the three candidate models: the smallest cost function value was obtained with different models depending on the particular time and pressure level.

It is probably the case that none of the models considered here fit the data well.

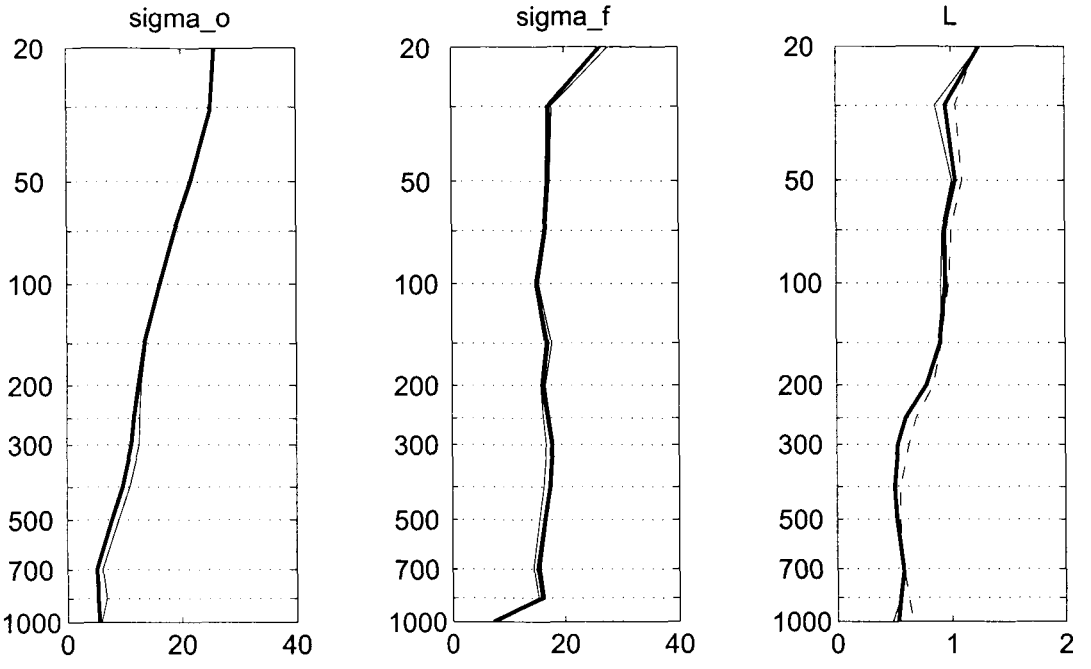

Figure 5: The effect of the choice of forecast error correlation model on the maximum-likelihood parameter estimates. Estimates of rawinsonde height error standard deviations (in meters), forecast height error standard deviations (in meters) and de-correlation length scales (in thousands of kilometers) are shown as a function of pressure. The thick solid curves were obtained using the spline-windowed powerlaw function (identical to Figure 3), the thin solid curves with the powerlaw function, and the dashed curves with the compactly supported spline function. Parameter estimates are based on North-American night-time February 1995 data.

Isotropic models are perhaps appropriate for describing forecast error correlations averaged over a sufficiently long (seasonal) time period, but on a shorter time scale the forecast errors are state-dependent and their spatial correlations must therefore be anisotropic. See Riishojgaard (1997) for a promising approach toward covariance modeling for state-dependent forecast errors. In any case, since we question the model hypothesis it would be more meaningful to study goodness-of-fit of various candidate models based on independent data sets and on a variety of parametric and

nonparameteric statistical tests. We will report on such a study in a subsequent report.

As a final experiment with this data set we produced parameter estimates using the *Generalized Cross-Validation* (GCV) method (Wahba and Wendelberger 1980), briefly described in Section 2.3.5 and summarized in Appendix B. Figure 6 shows the GCV estimates superimposed on the maximum-likelihood estimates. The only difference between the two sets of estimates is that they are based on the minimization of two different cost functions. The estimates are not significantly different in light of the parameter uncertainties alluded to earlier, except perhaps near the surface where the GCV estimates of the observation error standard deviations are consistently smaller. The GCV estimate at 1000mb is zero, which is suspicious; yet it is not possible in general to determine which method is more accurate.

As discussed in Section 2.3.5, the likelihood cost function leads to asymptotically (i.e., for large number of data) optimal parameter estimates under the model hypothesis. This property is academic if the model hypothesis is violated, as it invariably is in practice. A practical advantage of the maximum-likelihood method is that it produces standard errors of the parameter estimates, which can be interpreted as estimates of the parameter uncertainty due to sampling error. Conceptually we prefer the maximum-likelihood formulation because it is consistent with current implementations of statistical analysis systems. However, we have found the GCV method to be computationally more robust in some cases when the initial parameter estimates were very poor. In those cases the initial phase of the optimization process (the bracketing or approximate localization of the minimum) was more rapidly achieved for the GCV cost function than for the log-likelihood function. This is probably due to the fact

that the GCV method first estimates the ratio of the observation and forecast error variances (see (2.44)); this ratio is generally more easily identifiable from the data than each of the variances separately.

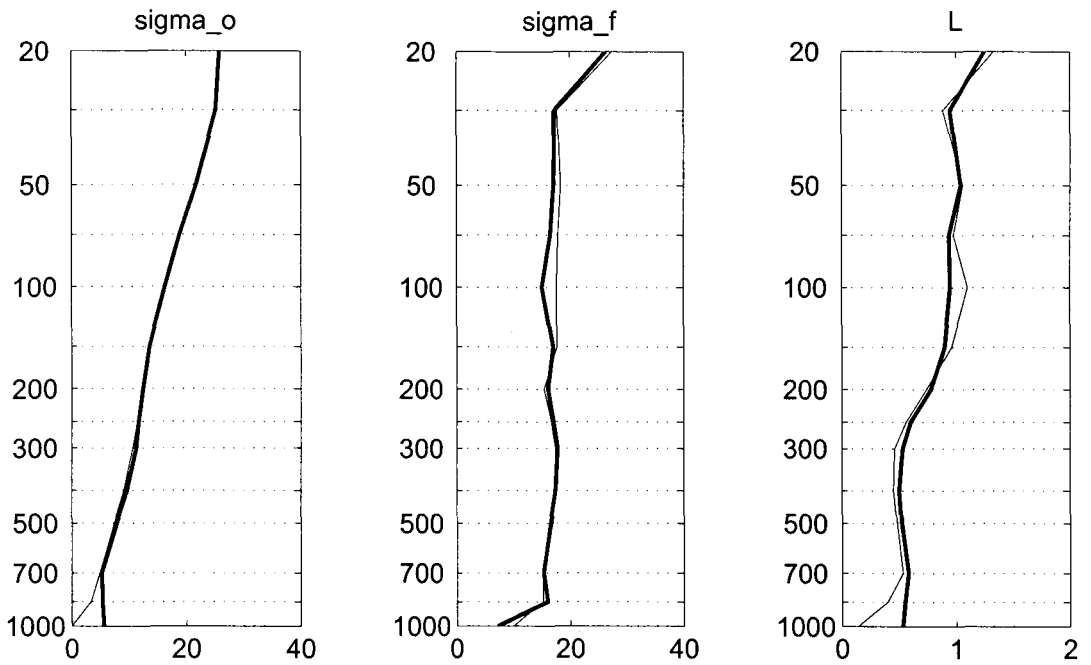

Figure 6: The effect of the choice of estimation method on the parameter estimates. Estimates of rawinsonde height error standard deviations (in meters), forecast height error standard deviations (in meters) and de-correlation length scales (in thousands of kilometers) are shown as a function of pressure. The thick curves were obtained with the maximum-likelihood method (identical to Figure 3); the thin curves are the Generalized Cross-Validation parameter estimates. The spline-windowed powerlaw function was used to model forecast error correlations in both cases, and the parameter estimates are based on the same set of North-American night-time February 1995 data.

3.2 Sealevel pressure residuals from ship reports

In this section we apply the maximum-likelihood method to the estimation of sea-level pressure observation and forecast error parameters. We use ship reports for this purpose, obtained during February 1995 in a section of the North Atlantic situated off the East Coast of the United States. Figure 7 shows the locations of each of the 3573 reports included in the data set. Superimposed is an estimate of the monthly mean residuals, which will be discussed below. The data distribution in this area is fairly uniform, and some major shipping routes are clearly visible.

An interesting aspect of this application is that the observing system is not stationary. Consequently the general, time-dependent formulation (2.30) of the maximum-likelihood cost function must be used in this case. This does not present any serious difficulties as long as the covariance between residuals at any two locations can be evaluated as a function of the parameters to be estimated. Computations are slower than in the stationary case, typically by a factor of 10 or so, depending on the size of the data set. Still, the examples in this section were easily calculated on a desktop computer.

The covariance model for this case is similar to that for rawinsonde height residuals; i.e., observation errors are assumed to be uncorrelated in space, and forecast errors are modeled using a simple univariate isotropic model. The covariance of the observed-minus-forecast sea-level pressure residuals is therefore represented by

$$[\mathbf{S}]_{ij} = (\sigma_p^o)^2 \delta(r_{ij}) + (\sigma_p^f)^2 \rho_w(r_{ij}; L_p), \tag{3.4}$$

where σ_p^o and σ_p^f are the observation and forecast sea-level pressure error standard deviations, respectively, and L_p is the de-correlation length scale associated with the

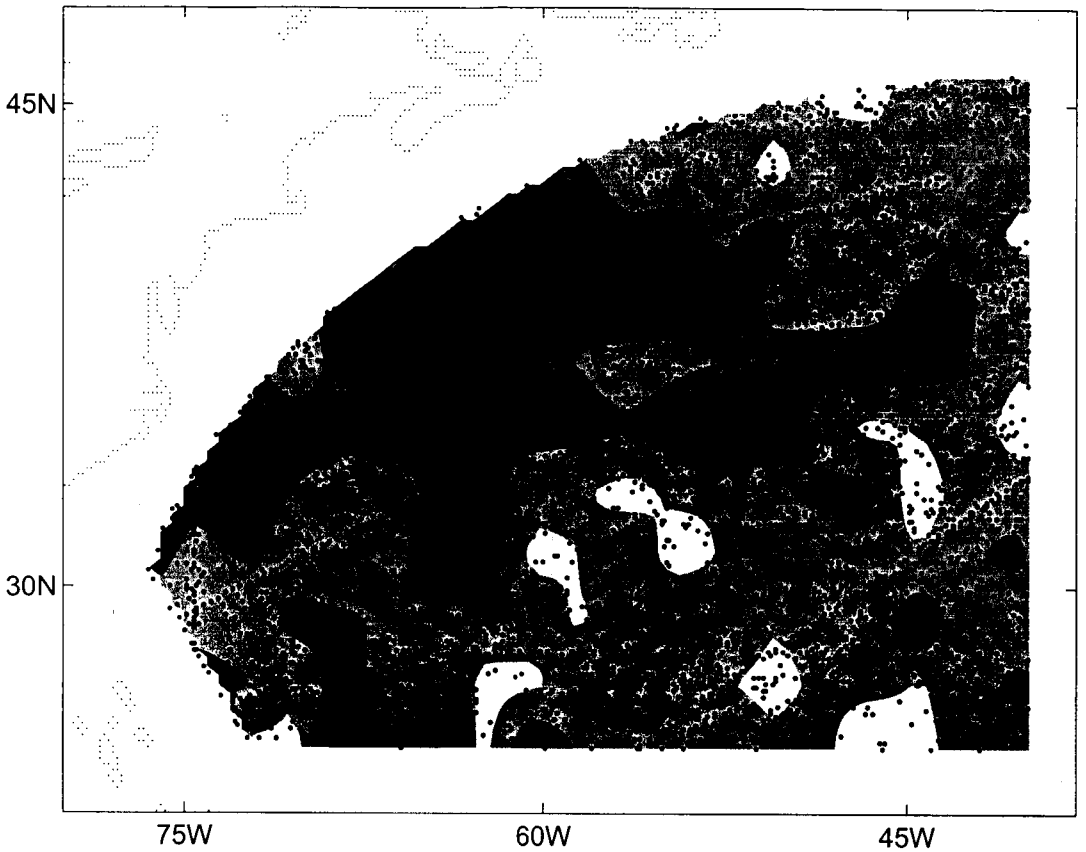

Figure 7: Locations of February 1995 ship reports, and mean observed-minus-forecast sea-level pressure residuals for that month. The mean field was computed using a successive correction method; see text for details. The four different shades of gray, from darkest to lightest, correspond to mean values in the intervals (-3mb,-2mb], (-2mb,-1mb], (-1mb,0mb], and (0mb,1mb].

forecast sea-level pressure errors. Given the coordinates of any two locations, this expression completely specifies the residual error covariance, except for the three parameters σ_p^o, σ_p^f, and L_p which will be estimated from the data.

Bias estimation is more complicated in this case, since there are no station locations that can be used to define the bias estimates. As discussed in Section 2.3.2, several possibilities present themselves, and we will experiment with a few of them here. The

obvious approach is to construct a grid which covers the data locations, and then to define and estimate the bias on the grid. (The grid may or may not coincide with the forecast model grid, but this is not relevant here). For example, the mean field shown in Figure 7 was computed by first constructing a $2° \times 2°$ grid and then, for each grid location, averaging all nearest residuals. Subsequently the estimate was smoothed by applying two iterations of a successive correction method, using a Gaussian weighting function with a length scale of $200km$. The gridded bias estimates are ultimately interpolated back to the data locations for the purpose of covariance parameter estimation, since the likelihood functional (2.30) involves the specification of the bias at the data locations.

The bias estimation and correction procedure just outlined is simple to implement, but involves a number of choices regarding the technical details: the definition of the grid, the method of estimation, and the interpolation scheme. Accurate bias estimates may be of interest for reasons other than covariance estimation, but our working assumption here is that *the covariance parameter estimates are not greatly sensitive to the details of the bias estimation procedure.* This assumption needs, of course, to be tested, and we will do so below. If in fact a small change in the bias estimation procedure causes a great change in the covariance parameter estimates, then the latter are not very meaningful.

Figure 8 shows, as a function of time, the maximum-likelihood estimates of the parameters σ_p^o, σ_p^f, and L_p based on a sliding 10-day window of data. The estimated standard errors are included as well. Bias was estimated at each time step from the data themselves (i.e., using the same 10 days of data) on a $2° \times 2°$ grid. The gridded bias estimates were calculated by applying two iterations of a successive correction

method, using a Gaussian weighting function with a length scale of $200km$. Bias estimates at the data locations were obtained by means of bilinear interpolation.

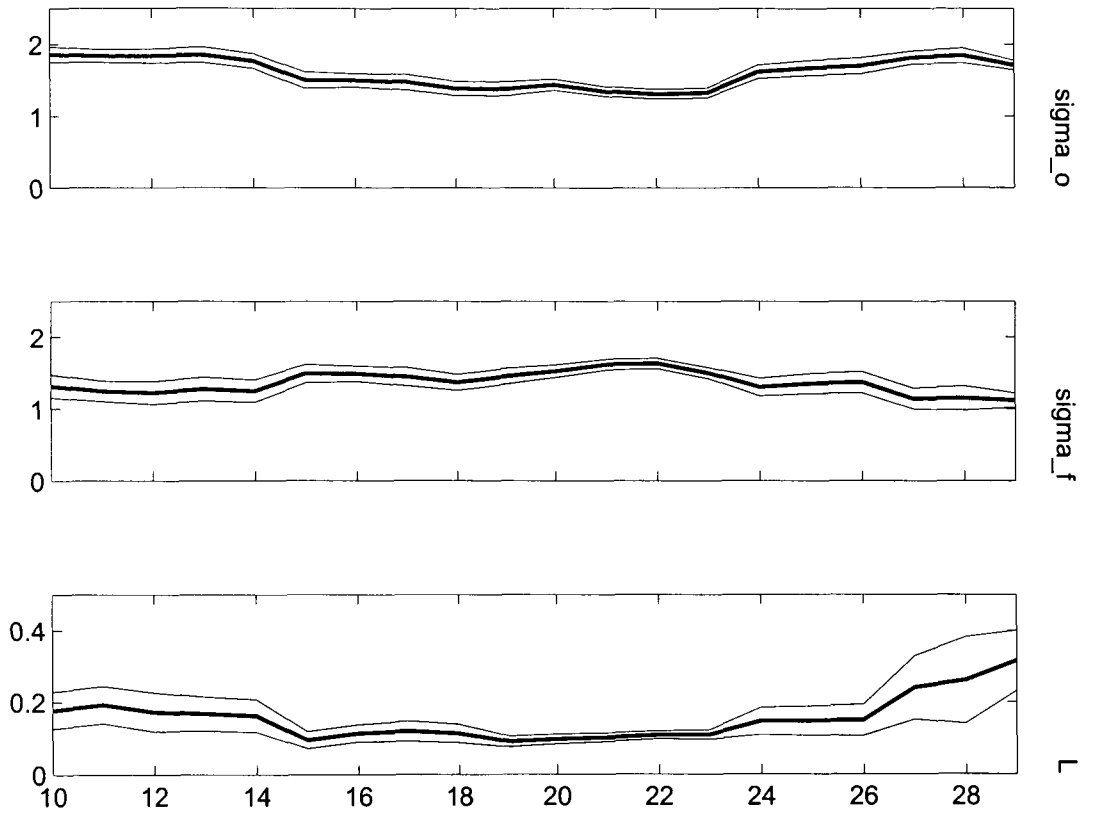

Figure 8: Maximum-likelihood parameter estimates as a function of time (in days). Shown are the estimates ship sea-level pressure error standard deviations (in millibars), forecast sea-level pressure error standard deviations (in millibars), and forecast sea-level pressure error de-correlation length scales (in thousands of kilometers). The estimates are produced once a day and are based on the latest 10 days of available reports. Also shown are the estimated standard errors (thin curves).

Figure 9 shows the results of introducing various modifications to the bias estimation and correction procedure. The thick curves in this figure are identical to those in Figure 8, while the dotted curves were obtained by not correcting for bias at all. Ignoring the bias altogether results in significantly larger de-correlation length scale

estimates and somewhat larger variance estimates. This is not surprising, since the bias here is mistaken for a spatially correlated random component of error. The thin solid curves are the result of using bias estimates which are less smooth: the estimation procedure involves only a single pass of the successive correction method with a more localized weighting function (using a length scale of $100km$). These curves are barely visible in the figure since they almost exactly coincide with the thick solid curves in most places. The thin dashed curves correspond to a radical simplification of the bias estimation procedure: the bias at each time was taken to be constant in space, with the constant obtained by simply averaging all the data within the 10-day period. This crude change in the procedure appears to mostly effect the estimates of the forecast error standard deviation, which turn out somewhat larger.

The differences among the parameter estimates which involve some form of spatially variable bias estimation and correction are generally small; in fact they appear to be comparable to the standard error estimates which are plotted in Figure 8. This implies that, in this case at least, the details of the bias correction procedure do not significantly affect the covariance parameter estimates.

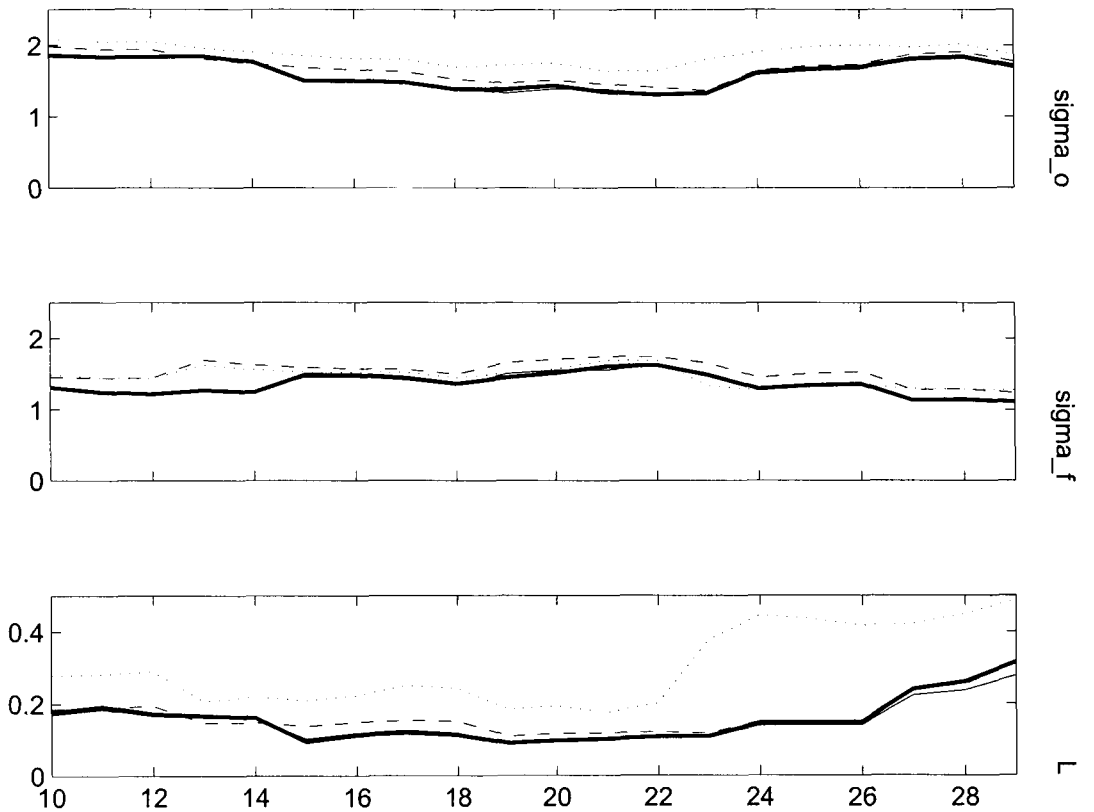

Figure 9: The effect of the bias estimation procedure on covariance parameter estimates. The thick solid curves are identical to those in Figure 8. The dotted curves were obtained by not correcting for bias at all. The thin solid curves (mostly hidden by the thick solid curves) correspond to a slight change in the bias estimation procedure—see text for details. The dashed curve was obtained by taking bias to be constant in space.

3.3 Wind residuals from aircraft reports

The final application we present here involves aircraft wind data, and the estimation of observation wind error standard deviations from these data. We use wind reports obtained from various flights over a North-Eastern portion of the North American continent during February 1995. Figure 10 shows the 1295 locations of all two-component wind observations reported at pressure levels between 225mb and 275mb during that month. The data distribution is highly irregular and mostly concentrated about fixed flight paths.

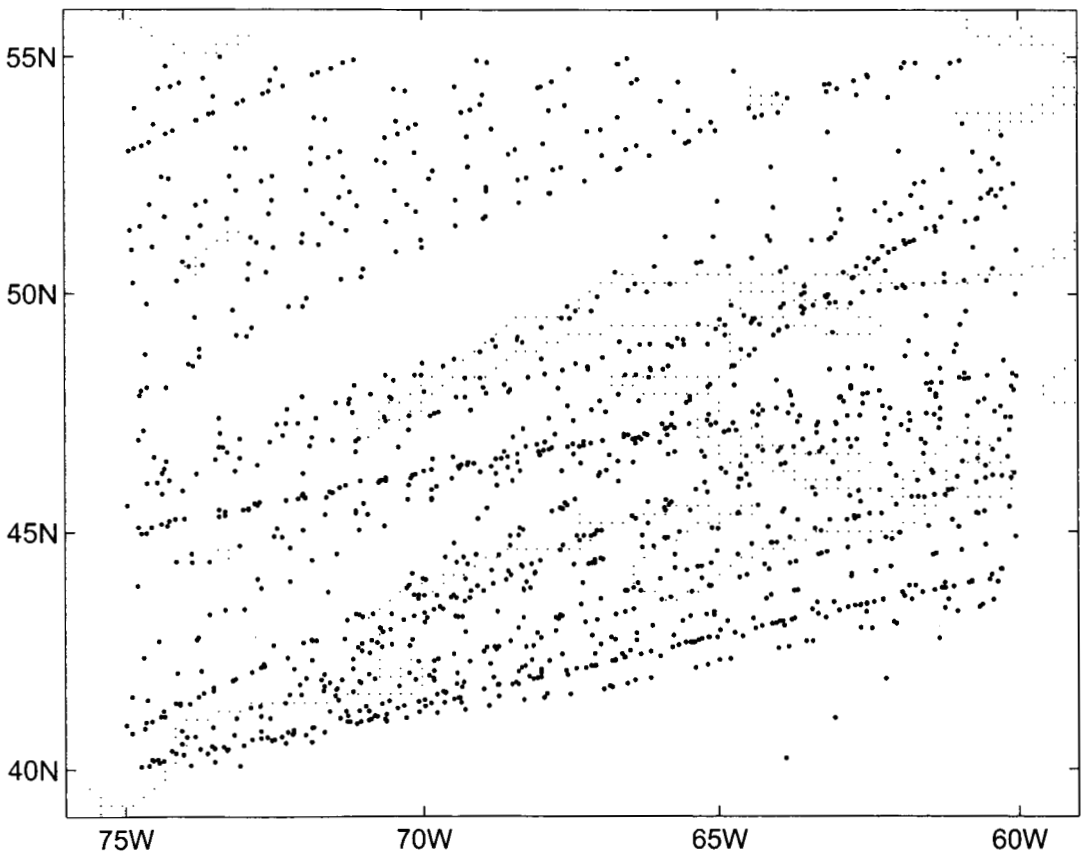

Figure 10: Locations of February 1995 aircraft wind observations with reported pressure levels between 225mb and 275mb, within a $15° \times 15°$ degree portion of the North American continent.

As before, we will check the sensitivity of our results to the treatment of bias. In Figure 11 we show the February 1995 mean observed-minus-forecast wind residuals computed on a $1° \times 1°$ grid. The mean was computed at each grid location by first averaging (over time) all nearest data and then applying two iterations of a successive correction method, using a Gaussian weighting function with a length scale of $200km$. The figure shows a consistent pattern in the residual wind directions along the major flight paths toward the north-east. The maximum residual wind speed is $8.4ms^{-1}$, while the median is $2.0ms^{-1}$.

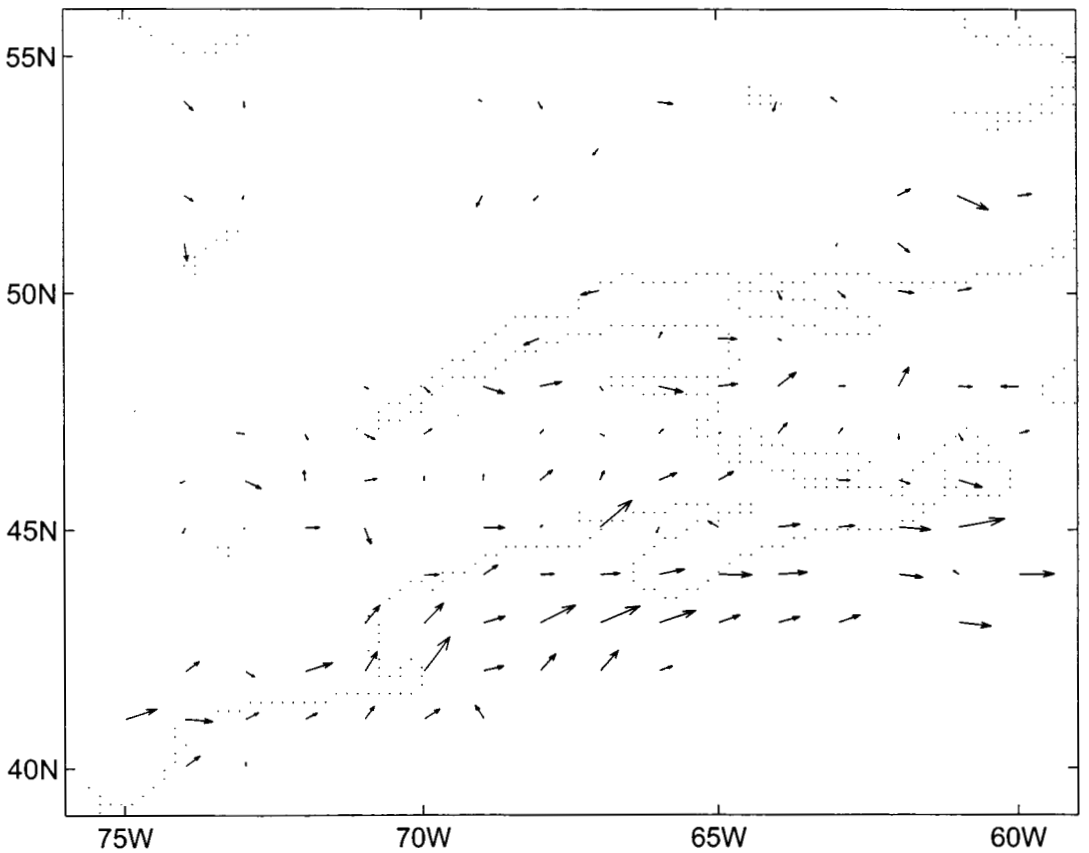

Figure 11: Mean observed-minus-forecast wind residuals from aircraft data, computed using February 1995 reports at levels between 225–275mb. The maximum residual wind speed is $8.4ms^{-1}$, the median is $2.0ms^{-1}$.

In Figure 12 we show the weekly means, for each of the four weeks of the month, plotted at the same scale as in the previous figure. The number of reports during each week was 353, 408, 255, and 279, respectively. Although the predominant direction of the arrows is still visible in each of the four panels, there are obvious differences as well. One might suspect that covariance parameter estimates will be quite different, depending on the manner in which bias is estimated and removed from the data.

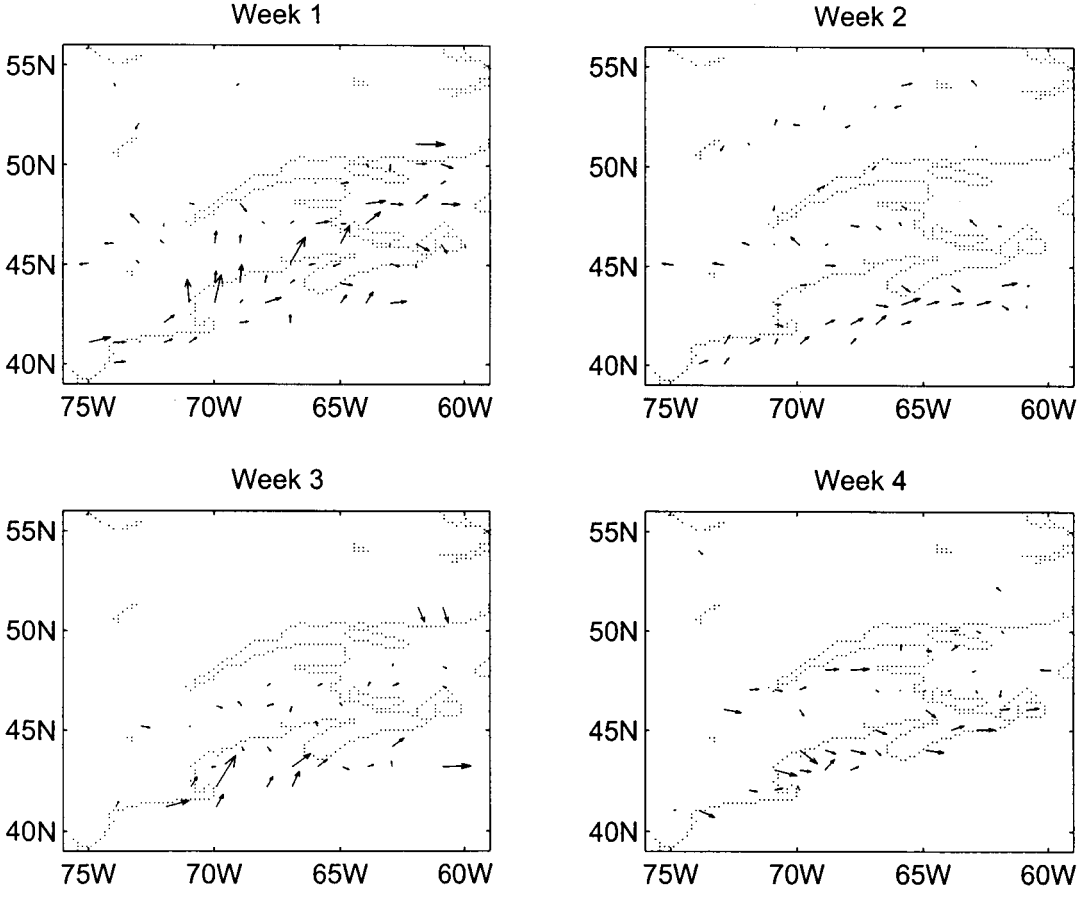

Figure 12: As in Figure 11, but computed for each week. The scale of the residual wind arrows in all four panels is identical to that in Figure 11.

Aircraft wind data comprise an important source of information about upper level atmospheric flow, yet the error characteristics associated with these data are not very

well known. Here we attempt to estimate only the standard deviation of the (spatially and temporally) uncorrelated component of observation error. This does not properly account for the contribution of representativeness error, which may very well be highly significant in this case. Since both forecast errors and representativeness errors are likely to be state-dependent and spatially correlated, it is not clear that the two can be statistically separated. This is a good example of an identifiability problem; i.e., a fundamental limitation of the approach of estimating observation and forecast error covariance parameters from residuals.

Aircraft wind observation errors are modeled simply by

$$R = \begin{bmatrix} R^u & 0 \\ 0 & R^v \end{bmatrix} \tag{3.5}$$

with

$$[\mathbf{R}^u]_{ij} = (\sigma^{u^o})^2 \delta(r_{ij}), \tag{3.6}$$

$$[\mathbf{R}^v]_{ij} = (\sigma^{v^o})^2 \delta(r_{ij}). \tag{3.7}$$

Here σ^{u^o} and σ^{v^o} are the observation error standard deviations for the $u-$ and $v-$components, respectively. There is no obvious reason to expect that these two quantities are greatly different, but for the moment we will retain the extra degree of freedom.

Wind forecast errors are modeled as described in Section 2.1, by introducing an error streamfunction ψ and error velocity potential χ and postulating simple univariate covariance models for each of these scalar fields. Here we assume that ψ and χ are independent with covariances

$$[\mathbf{P}^\psi]_{ij} = (\sigma^\psi)^2 \rho_w(r_{ij}; L^\psi), \tag{3.8}$$

$$[\mathbf{P}^\chi]_{ij} = (\sigma^\chi)^2 \rho_w(r_{ij}; L^\chi), \tag{3.9}$$

depending on the four parameters σ^{ψ}, L^{ψ} and σ^{χ}, L^{χ}. Combining (3.8, 3.9) with (2.9) results in an anisotropic forecast wind error covariance model (Daley 1991, Figures 5-2 and 5-4).

We estimated all six parameters from the entire month of data, using the general, time-dependent formulation (2.30) of the maximum-likelihood cost function. The monthly mean shown in Figure 11 was subtracted from the wind residuals prior to covariance tuning. The resulting parameter estimates and their estimated standard errors were

$$\sigma^{u^o} = (2.75 \pm 0.07) ms^{-1}, \tag{3.10}$$

$$\sigma^{v^o} = (2.77 \pm 0.07) ms^{-1}, \tag{3.11}$$

$$\sigma^{\psi} = (18.46 \pm 2.03) ms^{-1}, \tag{3.12}$$

$$\sigma^{\chi} = (19.77 \pm 2.59) ms^{-1}, \tag{3.13}$$

$$L^{\psi} = (0.51 \pm 0.04) \times 10^6 m, \tag{3.14}$$

$$L^{\chi} = (0.57 \pm 0.05) \times 10^6 m. \tag{3.15}$$

Note that the standard error estimates indicate that all six parameters are simultaneously identifiable from the data; i.e., the Hessian at the minimum of the cost function is well-conditioned.

The six-parameter estimation from a month of data took about an hour on a desktop workstation; the software we used for this purpose was designed for flexibility rather than efficiency. Still, it appears that the computation is unnecessarily expensive since (i) the estimates indicate that the observation error standard deviations can be represented by a single parameter and (ii) it is questionable whether four parameters are really needed to describe the forecast wind error covariance. We therefore repeated

the calculation after reducing the number of parameters to three, by setting

$$\sigma^{u^o} = \sigma^{v^o} \equiv \sigma^o, \tag{3.16}$$

$$\sigma^{\psi} = \sigma^{\chi} \equiv \sigma^f, \tag{3.17}$$

$$L^{\psi} = L^{\chi} \equiv L. \tag{3.18}$$

The resulting parameter estimates and their estimated standard errors were

$$\sigma^o = (2.77 \pm 0.05) ms^{-1}, \tag{3.19}$$

$$\sigma^f = (19.09 \pm 1.14) ms^{-1}, \tag{3.20}$$

$$L = (0.54 \pm 0.03) \times 10^6 m. \tag{3.21}$$

It appears that in this case the observation wind error standard deviation can be estimated well using three parameters only.

We repeated this procedure for each week of data separately, and obtained similar results—that is, the estimated $u-$ and $v-$observation error standard deviations are not significantly different, whether six or three parameters are used to describe the residual covariances. The estimates do vary from week to week, as shown in Figure 13. The horizontal lines in each panel correspond to (3.19–3.21): these are the estimates and their standard errors obtained from the entire month of data. The circles and plus signs in each panel mark weekly parameter estimates, using two different methods for bias correction. In case of the circles the estimates truly depend on one week of data only: the bias was estimated from the same week of data (see Figure 12). When, instead, the monthly mean (Figure 11) was used as a bias estimate for each week of data, slightly different estimates were obtained (marked by the plus signs). The discrepancy between the two sets of estimates is indicative of the uncertainty due to the treatment of bias.

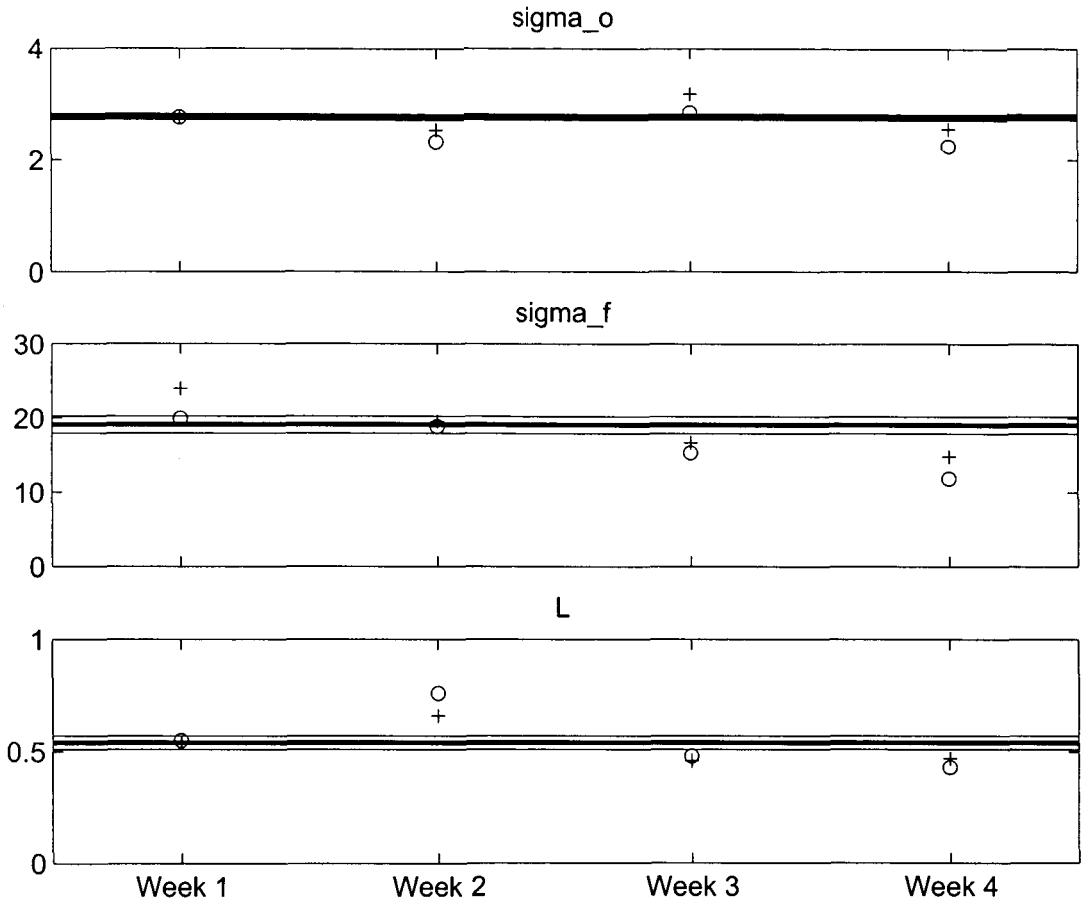

Figure 13: Wind error covariance parameters estimated from February 1995 near-250mb aircraft data. Shown are the estimated aircraft observation wind error standard deviation σ^o (in meters per second), the forecast wind error standard deviation parameter σ^f (in meters per second), and the forecast wind error de-correlation length scale parameter L (in thousands of kilometers). The horizontal lines in each panel correspond to the estimates and their standard errors obtained from the entire month of data. The circles and plus signs in each panel mark weekly estimates, using two different methods for bias correction; see text for details.

4 Summary and conclusions

We presented a general method for estimating forecast and observation error covariance parameters, based on the maximum-likelihood principle. We discussed issues such as bias estimation and correction, parameter identifiability, estimation accuracy, and robustness of the method, both from a theoretical and from a practical point of view.

Three different applications were used to describe the flexibility and limitations of the method. The maximum-likelihood method produces 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, by changing some of the assumptions that enter into the maximum-likelihood criterion, one can gain insight into the actual parameter uncertainty.

Using this approach, we infer the following general conclusions from the reported experiments:

- Many statistical parameters usually specified to be constant in operational data assimilation systems in fact vary significantly in both space and time. For example, we found that rawinsonde height error standard deviations estimated from a sliding 10-day window of 500mb reports ranged between 5m and 8m within less than a month. Some of these differences may be explained by the effects of solar radiation and other quality control issues, but they are more likely due to representativeness error.

- Systematic errors in both forecasts and observations cannot be ignored and

can induce large uncertainties in the covariance parameter estimates. In some cases, monthly mean observed-minus-forecast residuals can be as large as the estimated standard deviations. Nevertheless, our results appear to be fairly robust with respect to the treatment of bias.

- Forecast error parameter estimates vary greatly, depending on the selection of data. This is true for both variance and correlation parameters. The assumption that forecast error correlations are isotropic may be appropriate when interpreted in a time-average (rather than ensemble-average) sense, but only on a seasonal time scale. Different isotropic models fit equally well (or equally badly) on shorter time scales.

- On-line estimation of covariance parameters is both necessary and feasible. Estimates of parameter uncertainty due to sampling error confirm that it requires on the order of a hundred data items to estimate each parameter, and the computational effort involved in doing so is small. Whether to use highly localized data spaced in time, or nearly instantaneous but spatially distributed data, is a matter of modeling strategy.

- A fundamental limitation of this and other estimation methods is identifiability. 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 they have distinguishable characteristics.

Work is currently underway to automate some aspects of the estimation procedures

described in this report, so that certain parameters of the PSAS covariance specifications can be updated in real time based on recent observations. The applications presented here show that this type of automation is feasible. The result will be an *adaptive PSAS* in which, for example, time- and space-dependent changes in forecast accuracy will be automatically reflected in updated forecast error statistics used by PSAS. This should improve analysis accuracy, which, in turn, will improve forecast accuracy, which, in turn, will improve analysis accuracy, which, in turn, will improve forecast accuracy, which, in turn, will improve analysis accuracy, which, in turn, will improve forecast accuracy, ...

A Correlation models

The general univariate isotropic covariance model is of the form

$$\mathrm{Cov}(\mathbf{x},\mathbf{y}) = \sigma(\mathbf{x})\sigma(\mathbf{y})\rho(r(\mathbf{x},\mathbf{y})), \tag{A.1}$$

with $\sigma(\mathbf{x})$ a positive real-valued function and $r(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||$ the Euclidean distance between locations $\mathbf{x}$ and $\mathbf{y}$. The *representing function* ρ must satisfy certain conditions for (A.1) to be a legitimate (*i.e.*, positive semi-definite) covariance model; see Gaspari and Cohn (1997) for details.

In this study we consider the following alternatives for the representing function ρ.

The *powerlaw*:

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

The parameter L is the de-correlation length scale defined by

$$L = \sqrt{\frac{-1}{\rho''(0)}}, \tag{A.3}$$

see Daley (1991, Section 4.3).

The *compactly supported spline* (Gaspari and Cohn 1997, section 4.3):

$$\rho(r) = \rho_c(r;L) \tag{A.4}$$

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

with

$$c = L\sqrt{\frac{10}{3}}. \tag{A.6}$$

and L is the de-correlation length scale defined by (A.3).

Covariances modeled by the compactly supported spline are identically zero whenever the distance between two locations exceeds the threshold $r = r_* = 2c$:

$$\rho_c(r; L) = 0 \qquad \text{for} \qquad r > 2c \approx 3.65L \tag{A.7}$$

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 the compactly supported spline is quite different from that of the powerlaw; see figure 14.

The *spline-windowed powerlaw*:

$$\rho(r) = \rho_w(r; L) = \rho_p(r; L_1) \times \rho_c(r; L_2), \tag{A.8}$$

which also has compact suppport. Using (A.3) and the fact that $\rho(0) = 1, \rho'(0) = 0$ for each of the functions considered here, it is easy to show that

$$\frac{1}{L^2} = \frac{1}{L_1^2} + \frac{1}{L_2^2}. \tag{A.9}$$

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

$$L_2 = \frac{r_*}{2}\sqrt{\frac{3}{10}}. \tag{A.10}$$

If we consider the de-correlation length scale L as the single free (tunable) parameter in (A.8), one should take

$$L_1 = \frac{L}{\sqrt{1 - \frac{40}{3}(\frac{L}{r_*})^2}} \tag{A.11}$$

which follows by substituting (A.10) into (A.9).

Figure 14 shows plots of the three functions for identical values of the length scale parameter L, as well as their discrete Legendre spectra.

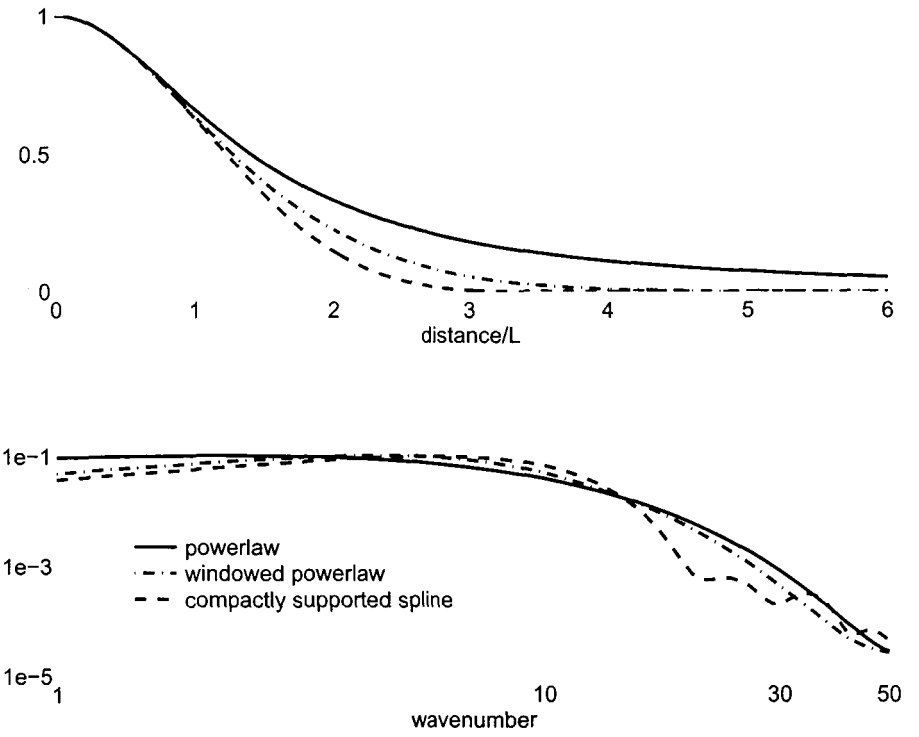

Figure 14: Correlation models and Legendre coefficients.

B Multiple-sample GCV

Wahba *et al.* (1995) show how to obtain GCV estimates of the covariance parameters $\sigma_1, \sigma_2, \boldsymbol{\theta}$ in (2.43) based on a single residual $\mathbf{v}$. It is assumed that

$$\langle \mathbf{v} \rangle = \boldsymbol{\mu}, \qquad \left\langle (\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})^T \right\rangle = \sigma_1^2 \mathbf{S}_1 + \sigma_2^2 \mathbf{S}_2(\boldsymbol{\theta}), \tag{B.1}$$

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

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

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

$$V(\lambda, \boldsymbol{\theta}) = \frac{|| \left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta}) \right] \mathbf{y} ||^2}{\{ \text{trace} \left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta}) \right] \}^2} \tag{B.3}$$

where

$$\mathbf{A}(\lambda,\boldsymbol{\theta}) = \left[\mathbf{I} + \lambda \mathbf{S}_1^{1/2}\mathbf{S}_2^{-1}(\boldsymbol{\theta})\mathbf{S}_1^{1/2}\right]^{-1}, \tag{B.4}$$

$$\mathbf{y} = \mathbf{S}_1^{-1/2}(\mathbf{v} - \boldsymbol{\mu}), \tag{B.5}$$

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

$$\widehat{\sigma}_1^2 = \{\text{trace}[\mathbf{I} - \mathbf{A}(\widehat{\lambda},\widehat{\boldsymbol{\theta}})]\} \times V(\widehat{\lambda},\widehat{\boldsymbol{\theta}}), \tag{B.6}$$

$$\widehat{\sigma}_2^2 = \frac{\widehat{\sigma}_1^2}{\widehat{\lambda}}. \tag{B.7}$$

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

$$\mathbf{v} = (\mathbf{v}_1^T \ldots \mathbf{v}_K^T)^T, \tag{B.8}$$

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

$$\langle \mathbf{v}_k \rangle = \boldsymbol{\mu}_k, \qquad \left\langle (\mathbf{v}_k - \boldsymbol{\mu}_k)(\mathbf{v}_k - \boldsymbol{\mu}_k)^T \right\rangle = \sigma_1^2 \mathbf{S}_{k1} + \sigma_2^2 \mathbf{S}_{k2}(\boldsymbol{\theta}). \tag{B.9}$$

The covariance model (B.1) for $\mathbf{v}$ is block-diagonal, with blocks given by (B.9). It is easily checked that

$$V(\lambda,\boldsymbol{\theta}) = \frac{\sum_{k=1}^K ||\left[\mathbf{I} - \mathbf{A}_k(\lambda,\boldsymbol{\theta})\right]\mathbf{y}_k||^2}{\left\{\sum_{k=1}^K \text{trace}\left[\mathbf{I} - \mathbf{A}_k(\lambda,\boldsymbol{\theta})\right]\right\}^2} \tag{B.10}$$

where

$$\mathbf{A}_k(\lambda,\boldsymbol{\theta}) = \left[\mathbf{I} + \lambda \mathbf{S}_{k1}^{1/2}\mathbf{S}_{k2}^{-1}(\boldsymbol{\theta})\mathbf{S}_{k1}^{1/2}\right]^{-1}, \tag{B.11}$$

$$\mathbf{y}_k = \mathbf{S}_{k1}^{-1/2}(\mathbf{v}_k - \boldsymbol{\mu}_k). \tag{B.12}$$

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,

$$\begin{aligned}
\sum_{k=1}^{K} || \left[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})\right] \mathbf{y}_k ||^2 &= \sum_{k=1}^{K} \mathbf{y}_k^T \left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \mathbf{y}_k \\
&= \sum_{k=1}^{K} \text{trace}\{\mathbf{y}_k^T \left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \mathbf{y}_k\} \\
&= \sum_{k=1}^{K} \text{trace}\{\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \mathbf{y}_k \mathbf{y}_k^T\} \\
&= \text{trace}\{\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \sum_{k=1}^{K} \mathbf{y}_k \mathbf{y}_k^T\} \\
&= K \times \text{trace}\{\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \mathbf{S}_1^{-1/2} \overline{\mathbf{S}} \mathbf{S}_1^{-1/2}\}, \qquad \text{(B.13)}
\end{aligned}$$

where $\overline{\mathbf{S}}$ is the sample covariance of the data defined in (2.33). For the denominator,

$$\sum_{k=1}^{K} \text{trace}\left[\mathbf{I} - \mathbf{A}_k(\lambda, \boldsymbol{\theta})\right] = K \times \text{trace}\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right] \qquad \text{(B.14)}$$

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

$$V(\lambda, \boldsymbol{\theta}) = \frac{1}{K} \times \frac{\text{trace}\{\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]^2 \mathbf{S}_1^{-1/2} \overline{\mathbf{S}} \mathbf{S}_1^{-1/2}\}}{\{\text{trace}\left[\mathbf{I} - \mathbf{A}(\lambda, \boldsymbol{\theta})\right]\}^2}. \qquad \text{(B.15)}$$

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. *Proceedings of the IEEE*, **70**, 963–974.

Chavent, G., 1979: Identification of distributed parameter systems: About the output least square method, its implementation, and identifiability. Pp. 85–97 in: R. Isermann (ed.), *Identification and System Parameter Estimation.* Proceedings of the Fifth IFAC Symposium, Darmstadt, Federal Republic of Germany, 24–28 September 1979, Vol. 1.

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

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

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

Daley, R., 1993: Estimating observation error statistics for atmospheric data assimilation. *Annales Geophysicae*, **11**, 634–647.

DAO, 1996: *Algorithm Theoretical Basis Document Version 1.01,* Data Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD 20771. Available on the internet 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.

Dee, D. P., and A. M. da Silva, 1997: Data assimilation in the presence of forecast bias. *Quart. J. R. Met. Soc.*, in press.

Fisher, R. A., 1922: On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A,* **222**, 309–368.

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

Gaspari, G., and S. Cohn, 1997: Construction of correlation functions in two and three dimensions. Submitted to *Quart. J. R. Met. Soc.*

Gill, P. E., W. Murray, and M. H. Wright, 1981: *Practical optimization.* Academic Press, 401pp.

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.

Jazwinski, A. H., 1970: *Stochastic Processes and Filtering Theory.* Academic Press, 376pp.

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. R. Met. Soc.*, **112**, 1177–1194.

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

Mitchell, H. L., C. Chounard, C. Charette, R. Hogue, and S. J. Lambert, 1996: Impact of a revised analysis algorithm on an operational data assimilation system. *Mon. Wea. Rev.*, **124**, 1243–1255.

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

Nelder, J. A., and R. Mead, 1965: A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Pfaendtner, J., S. Bloom, D. Lamich, M. Seablom, M. Sienkiewicz, J. Stobie, and A. da Silva, 1995: Documentation of the Goddard Earth Observing System (GEOS) Data Assimilation System–Version 1. *NASA Tech. Memo. No. 104606*, Vol. 4, Goddard Space Flight Center, Greenbelt, MD 20771. Available on the internet at `ftp://dao.gsfc.nasa.gov/pub/tech_memos/volume_4.ps.Z`

Press, W. H., S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The Art of Scientific Computing* (Second Edition). Cambridge University Press, 963pp.

Riishojgaard, L.-P., 1997: A direct way of specifying flow-dependent background error correlations for meteorological analysis systems. *Tellus,* in press.

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, 382pp.

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.

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.

Wahba, G., 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.