Adaptive Tuning of Numerical Weather Prediction Models: Randomized GCV in Three- and Four-Dimensional Data Assimilation

GRACE WAHBA

Department of Statistics, University of Wisconsin—Madison, Madison, Wisconsin

DONALD R. JOHNSON

Space Science and Engineering Center, University of Wisconsin—Madison, Madison, Wisconsin

FENG GAO

Battelle Institute, Richland, Washington

JIANJIAN GONG

Department of Statistics, University of Wisconsin—Madison, Madison, Wisconsin

(Manuscript received 23 May 1994, in final form 1 May 1995)

ABSTRACT

In variational data assimilation, optimal ingestion of the observational data, and optimal use of prior physical and statistical information involve the choice of numerous weighting, smoothing, and tuning parameters that control the filtering and merging of diverse sources of information. Generally these weights must be obtained from a partial and imperfect understanding of various sources of errors and are frequently chosen by a combination of historical information, physical reasoning, and trial and error.

Generalized cross validation (GCV) has long been one of the methods of choice for choosing certain tuning, smoothing, regularization parameters in ill-posed inverse problems, smoothing, and filtering problems. In theory, it is well suited for the adaptive choice of certain parameters that occur in variational objective analysis and for data assimilation problems that are mathematically equivalent to variational problems. The main drawback of the use of GCV in data assimilation problems was that matrix decompositions were apparently needed to compute the GCV estimates. This limited the application of GCV to datasets of the order of less than about 1000. Recently, the randomized trace technique for computing the GCV estimates has been developed, and this makes the use of GCV feasible in essentially any variational problem that has an operating algorithm to produce estimates, given data. In this paper the authors demonstrate that the answers given by the randomized trace estimate are indistinguishable in a practical sense from those computed more exactly by traditional methods. Then the authors carry out an experiment to choose one of the main smoothing parameters (λ) in the context of a variational objective analysis problem that is approximately solved by k iterations of a conjugate gradient algorithm. The authors show how the randomized trace technique can be used to obtain good values of both λ and k in this context. Finally, the authors describe how the method can be applied in operational-sized three- and four-dimensional variational data assimilation schemes, as well as in conjunction with a Kalman filter.

1. Introduction

In modern global-scale numerical weather prediction models, it is common to update the state vector ($\mathbf{x}$), which will serve as initial conditions to integrate forward the primitive equations, by combining observations, forecast, and possibly other physical constraints in a manner that is (some approximation to) the solution to a minimization problem of the following general form: find $\mathbf{x}$ to minimize

$$J(\mathbf{x}) = [\mathbf{y} - \mathbf{K}(\mathbf{x})]'\mathbf{S}^{-1}[\mathbf{y} - \mathbf{K}(\mathbf{x})] + (\mathbf{x} - \mathbf{x}^*)'\Sigma^{-1}(\mathbf{x} - \mathbf{x}^*). \quad (1.1)$$

Here $\mathbf{y}$ is a vector of data that are related to the state vector $\mathbf{x}$ by the possibly partly nonlinear operator $\mathbf{K}$, $\mathbf{x}^*$ is a known vector, and the matrices $\mathbf{S}$ and Σ embody information concerning how close $\mathbf{y}$ is expected to be to $\mathbf{K}(\mathbf{x})$ and how close $\mathbf{x}$ is expected to be to $\mathbf{x}^*$. At this point we are being deliberately vague about $\mathbf{y}$ because it may include forecast "data" as well as observational data; alternatively, forecast data may be in-

Corresponding author address: Prof. Grace Wahba, Department of Statistics, University of Wisconsin—Madison, 1210 W. Dayton St., Madison, WI 53706.

corporated into $\mathbf{x}^*$. Formally, the matrices $\mathbf{S}$ and Σ may be derived as covariance matrices under certain statistical assumptions, as in Parrish and Derber (1992), Lorenc et al. (1991), Lorenc (1986), Wahba (1990b, 1985b, 1982b), Wahba and Wendelberger (1980), and Kimeldorf and Wahba (1970, 1971). Four-dimensional assimilation with the model as a strong constraint can be put in this framework by, for example, letting $\mathbf{x}$ be the state at an initial time and including model integrations in $\mathbf{K}$. Physically based penalties—for example, energy in gravity waves—may be incorporated in the penalty term $(\mathbf{x} - \mathbf{x}^*)' \Sigma^{-1} (\mathbf{x} - \mathbf{x}^*)$ (see Lewis and Derber 1985; Courtier and Talagrand 1987, 1990; Zou et al. 1992; Zou et al. 1993; Rabier et al. 1993; Li et al. 1993; and references cited therein). Other relevant references are Hoffman (1984, 1985), Hoffman and Louis (1990), and Bennett and Budgell (1987).

The problem is to choose certain unknown parameters in Σ, $\mathbf{K}$, and $\mathbf{S}$. This fairly simple statement of the problem conceals many choices that must be made in practice. The entries in Σ and $\mathbf{S}$ contain numerous smoothing, tuning, and weighting parameters that in practice are obtained from postulated error covariances that aim to take into account measurement error, forecast error, errors of representativeness, and model error from prior physical and statistical information about the atmosphere, from physical intuition, and from trial and error. Operator $\mathbf{K}$ may contain instrument calibration constants, physical parameters, and so forth. Some of the discrepancies between $\mathbf{y}$ and $\mathbf{K}(\mathbf{x})$ are fairly well understood (e.g., radiosonde measurements), while others, particularly satellite radiances, are not, due in part to the difficulty of modeling the forward problem accurately. This problem is exacerbated in four-dimensional assimilation where $\mathbf{K}$ includes model integrations. Forecast error covariances can, to a certain extent, be studied empirically from historical data by comparing forecast and observation over a period of time (see Hollingsworth and Lonnberg 1986; Lonnberg and Hollingsworth 1986; Bartello and Mitchell 1993; Mitchell et al. 1990; Goerss and Phoebus 1993).

It is the purpose of this paper to initiate the development of a general theory of adaptive estimation of smoothing, weighting, and tuning parameters based on generalized cross validation (GCV) (Wahba and Wendelberger 1980; Wahba 1990b, and references therein) and related methods, for parameters that are hidden in Σ and in $\mathbf{K}$ and also to some extent in $\mathbf{S}$, that is applicable to the tuning of three- and four-dimensional numerical weather prediction models and to other data assimilation problems solved via variational problems that can be put in the general form (1.1).

There is, of course, much interest in developing objective methods of obtaining these parameters. Aside from work in the spirit of Hollingsworth and Lonnberg, which involves directly fitting parametric or semiparametric models of covariances to large historical datasets of forecasts minus observations, there are several other trains of research with similar goals, based on Kalman filter theory. The Kalman filter theory shows how the forecast error covariance evolves, given past data patterns and past observational error and "plant noise" (usually considered to be model errors in this context). Cohn (1993) looks at stochastic dynamic equations for the growth and propagation of forecast errors. Dee et al. (1985) and later Dee (1990) use a simplified model for the evolution of the forecast error covariance and fit the simplified model using historical data. Recently, Daley (1992a,b,c,d) developed some ingenious methods for estimating the stationary isotropic part of certain required covariances in the context of Kalman filtering, after having parameterized them with a small number of unknowns.

The GCV estimates that we discuss here are "adaptive" or "dynamic," in the sense that they are carried out *simultaneously* with the estimation of $\mathbf{x}$, unlike the methods described in the references given above that use historical "after the fact" data. By "after the fact," we mean that a historical sequence of estimates of $\mathbf{x}$'s are obtained from a model using whatever parameters exist in the model. Then new parameters are estimated given this series of $\mathbf{x}$'s along with their associated series of forecasts and data. Once these new parameters are obtained they are then substituted in the model. After this substitution, the model error properties may also change, so that this procedure needs to be iterated see, for example, Daley (1992b, section 3b). The proposed adaptive methods can be used to monitor or fine-tune certain parameters dynamically that have been obtained from historical data, from a Kalman filter method, or from other methods. This paper will focus on estimates of parameters primarily in Σ, and $\mathbf{K}$ above, although we will briefly mention other estimates. Historically, the use of GCV in data assimilation problems was limited by the fact that matrix decompositions were apparently needed to compute the GCV estimates. This limited its use to datasets very much smaller than those occurring in operational numerical weather prediction (NWP). Recently, the randomized trace technique for computing the GCV estimates has been developed (Girard 1987, 1989, 1991; Deshpande and Girard 1991; Hutchinson 1989), and the method does not require matrix decompositions. It is the purpose of this paper to demonstrate some of the properties of this technique in the context of variational data assimilation and methods (such as optimum interpolation—OI) that are mathematically equivalent to variational problems and to show how the technique may be used in essentially any size variational problem that has an operating algorithm to produce estimates given data, provided only that the algorithm can be run several times.

Dee (1995) has used maximum likelihood (ML) estimates to tune parameters in covariances occurring in a Kalman filter applied to a shallow-water equation, and D. Dee and G. Cats (1994, personal communication) have applied ML estimates to tune error covari-

ances in the High-Resolution Limited-Area Model (HIRLAM). This important work demonstrates the feasibility and potential value of adaptive on-line parameter estimation.

In section 2 we briefly review the properties of GCV, which suggest the approximate range of validity of this method, and identify the key role of the so-called influence matrix in adaptive tuning of NWP models.

In section 3 we describe the randomized trace technique and demonstrate that the answers given by the randomized trace technique are essentially indistinguishable from those calculated by more traditional methods, for datasets as small as 400. Then, we carry out an experiment to demonstrate the efficacy of the method in the context of a variational objective analysis problem that is approximately solved by k iterations of a conjugate gradient algorithm. Running the experiment on simulated data where the ground truth is known, we demonstrate that a good value of a smoothing parameter as well as a good value of k can be chosen by this technique, without using any matrix decompositions.

In section 4 we describe how the method can be implemented in the context of objective analysis, the Kalman filter, and four-dimensional variational data assimilation. Section 5 is a summary. Appendix A discusses theoretical conditions for the range of applicability of GCV estimates, and appendix B discusses relationships between GCV and ML.

2. The GCV estimate

We first review the well-known statistical assumptions that relate OI and variational methods (see Kimeldorf and Wahba 1970, 1971; Lorenc 1986, 1988; Wahba 1982b, 1985b, 1990b).

First, let $\mathbf{x}$ be the state vector of an NWP system. We suppose that $\mathbf{x}$ has some climatological mean that has already been subtracted out, and we suppose that the mean values of the components of $\mathbf{x}$ can be treated as though they are zero. We let $\mathbf{y}$ be a vector of observations. Later we will let $\mathbf{x}$ be an analysis increment and $\mathbf{y}$ an observation increment; the mathematics will be essentially the same.

If $\mathbf{u}$ is a vector, we will use the notation $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ to mean we will treat $\mathbf{u}$ as though it has a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$. First we suppose $\mathbf{x} \sim \mathcal{N}(0, b\boldsymbol{\Sigma})$, where b is a (free) positive constant and $\boldsymbol{\Sigma}$ is some nonnegative-definite matrix. We next suppose that $\mathbf{y}$ is related to $\mathbf{x}$ by

$$\mathbf{y} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{S})$, where σ^2 is a (free) positive constant and $\mathbf{S}$ is some positive-definite matrix. The components of $\mathbf{K}\mathbf{x}$ represent functionals of $\mathbf{x}$. If they are linear functionals, then $\mathbf{K}$ is simply a matrix, which we will assume for now. In the case of satellite radiances the functionals in $\mathbf{K}$ are mildly nonlinear integrals (see O'Sullivan and Wahba 1985). It is generally desirable to linearize as late as is practicable. Given the statistical assumptions on $\mathbf{x}$ and $\boldsymbol{\epsilon}$ above, and letting $\lambda = \sigma^2/b$, the conditional expectation $\mathbf{x}_\lambda$ of $\mathbf{x}$, given the data $\mathbf{y}$, is given by the minimizer of

$$(\mathbf{y} - \mathbf{K}\mathbf{x})'\mathbf{S}^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}) + \lambda\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}, \tag{2.2}$$

and $\mathbf{x}_\lambda$ is given by

$$\mathbf{x}_\lambda = (\mathbf{K}'\mathbf{S}^{-1}\mathbf{K} + \lambda\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{K}'\mathbf{S}^{-1}\mathbf{y}. \tag{2.3}$$

The identity

$$(\mathbf{K}'\mathbf{S}^{-1}\mathbf{K} + \boldsymbol{\Sigma}^{-1})^{-1} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{K}'(\mathbf{K}\boldsymbol{\Sigma}\mathbf{K}' + \mathbf{S})^{-1}\mathbf{K}\boldsymbol{\Sigma} \tag{2.4}$$

can be used to give a formula for $\mathbf{x}_\lambda$ in another, possibly more familiar, form.

However, it is not required to make any statistical assumptions on $\mathbf{x}$ in order for the variational problem of (2.2) to be sensible. The quantity $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ may, for example, be some quadratic functional of the state vector that penalizes derivatives or penalizes some physical quantity that it is desired to partially suppress, say, gravity wave tendencies.

To define the GCV estimate of λ and any other (identifiable) parameters $\boldsymbol{\theta}$ in $\mathbf{K}$ and $\boldsymbol{\Sigma}$ we need to define the (standardized) influence matrix $\mathbf{A}$ for this variational problem. Let $\mathbf{y}$ be standardized as

$$\tilde{\mathbf{y}} = \mathbf{S}^{-1/2}\mathbf{y}, \tag{2.5}$$

where $\mathbf{S}^{-1/2}$ is the symmetric square root of $\mathbf{S}$. If the observation errors are independent, then $\mathbf{S}^{-1/2}$ is just the diagonal matrix with inverse standard deviations (scaled by σ) down the diagonal. We have

$$\tilde{\mathbf{y}} = \mathbf{S}^{-1/2}\mathbf{K}\mathbf{x} + \tilde{\boldsymbol{\epsilon}}, \tag{2.6}$$

where $\tilde{\boldsymbol{\epsilon}}$ has the standardized distribution

$$\tilde{\boldsymbol{\epsilon}} \sim \mathcal{N}(0, \sigma^2\mathbf{I}). \tag{2.7}$$

The influence matrix $\mathbf{A}(\lambda, \boldsymbol{\theta})$ is defined as the matrix that relates the (standardized) data $\tilde{\mathbf{y}}$ to the predicted (standardized) data

$$\hat{\tilde{\mathbf{y}}} = \mathbf{S}^{-1/2}\mathbf{K}\hat{\mathbf{x}}. \tag{2.8}$$

That is,

$$\hat{\tilde{\mathbf{y}}} \equiv \mathbf{A}\tilde{\mathbf{y}}, \tag{2.9}$$

and it can be checked that $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{S}^{-1/2}\mathbf{K}(\mathbf{K}'\mathbf{S}^{-1}\mathbf{K} + \lambda\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{K}'\mathbf{S}^{-1/2}. \tag{2.10}$$

The influence matrix $\mathbf{A}$ is a so-called smoother matrix—that is, it is symmetric, nonnegative definite and all its eigenvalues are in the interval $[0, 1]$. This fact will play an important role in the randomized trace calculations to be described later. The GCV estimate $\hat{\lambda}$ of λ and other (identifiable) parameters in $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ and

$\mathbf{K} = \mathbf{K}(\theta)$ is the minimizer of $V(\lambda)$ [or $V(\lambda, \theta)$] given by

$$V(\lambda) = \frac{\mathbf{y}'\mathbf{S}^{-1/2}[\mathbf{I} - \mathbf{A}(\lambda)]^2\mathbf{S}^{-1/2}\mathbf{y}}{\{\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\lambda)]\}^2} \quad (2.11)$$

$$= \frac{\|[\mathbf{I} - \mathbf{A}(\lambda)]\mathbf{S}^{-1/2}\mathbf{y}\|^2}{\{\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\lambda)]\}^2}, \quad (2.12)$$

where Tr is trace (see Wahba and Wendelberger 1980; Wahba 1990b). Setting $\mathbf{S} = \mathbf{I}$ wherever it occurs results in the familiar form of V in the literature, where it is assumed that the data vector $\mathbf{y}$ has already been rescaled by $\mathbf{S}^{-1/2}$.

The GCV estimate of λ, θ is based on a predictive mean-square error criterion that attempts to obtain a "best" estimate of $\mathbf{x}$ within the family of possible minimizers of (2.2), parameterized by λ, θ. It will do this under rather general conditions, independent of whether Σ represents a covariance matrix, a "smoothness" penalty, or a physical quantity suppressing, say, some form of energy. The "predictive mean-square error" is with respect to data with ϵ of (2.1) being white-noise errors, or data normalized by a covariance matrix so that the normalized errors $\tilde{\epsilon}$ of (2.6) are at least roughly "white." To be specific, let the predictive mean-square error $R(\lambda)$ be defined by

$$R(\lambda) = \frac{1}{n}\|\mathbf{S}^{-1/2}(\mathbf{K}\mathbf{x}_{\mathrm{true}} - \mathbf{K}\mathbf{x}_\lambda)\|^2. \quad (2.13)$$

Here $\mathbf{x}_{\mathrm{true}}$ is the source of the data in (2.1), and we consider it fixed (not a random vector). Then the GCV estimate $\hat{\lambda}$ of λ is a good estimate of the λ *that minimizes* $R(\lambda)$, under fairly general conditions irrespective of whether $\mathbf{x}_{\mathrm{true}}$ is considered to be a fixed vector satisfying certain conditions, or is considered as a random vector with covariance matrix $b\Sigma$ (see Craven and Wahba 1979; Wahba and Wendelberger 1980; Speckman 1985; Li 1986; Wahba 1990, section 4.4). Some of these mathematical conditions are described for the reader's convenience in appendix A. Generally $\hat{\lambda}$ is also a good estimate of the minimizer of $D(\lambda) = \|\mathbf{x}_\lambda - \mathbf{x}_{\mathrm{true}}\|^2$ under some fairly but not completely general conditions (see Wahba and Wang 1990). A cross-validation-based estimate for σ^2 that has been shown to work well in examples is

$$\sigma^2_{\mathrm{GCV}} = \frac{\mathrm{RSS}(\hat{\lambda})}{\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\hat{\lambda})]}. \quad (2.14)$$

Here $\mathrm{RSS}(\hat{\lambda}) = \|[\mathbf{I} - \mathbf{A}(\hat{\lambda})]\mathbf{S}^{-1/2}\mathbf{y}\|^2$ is the (scaled) residual sum of squares when $\hat{\lambda}$ is used, and $\hat{\lambda}$ is the GCV estimate of λ, that is, the minimizer of $V(\lambda)$. Viewing $\mathrm{Tr}\mathbf{A}(\hat{\lambda})$ as the degrees of freedom for signal, this estimate is the analog of the usual estimate of the variance after linear regression (see Wahba 1983). Note that, due to the presence of $\mathbf{S}$ in the theoretical loss function R of (2.13), GCV in this form is not in general appropriate to estimate unknown parameters in $\mathbf{S}$. Partial GCV (see section 4b) may be used when a sufficiently large submatrix of S is known, and other GCV-related methods are discussed in Gao (1993, 1994) and Gao et al. 1995 (unpublished manuscript).

Various parameters θ in $\Sigma = \Sigma(\theta)$ are known to be amenable to estimation by GCV by replacing Σ by $\Sigma(\theta)$ in (2.2) and by minimizing V of (2.11) with respect to both λ and θ (see, for example, Wahba and Wendelberger 1980; Hutchinson et al. 1984; Wahba 1990, chapter 3). Certain parameters θ in $\mathbb{K} = \mathbb{K}(\theta)$ may also be estimated this way (O'Sullivan 1991; Wahba 1990a; Wahba et al. 1995, unpublished manuscript), by setting $\mathbb{K} = \mathbb{K}(\theta)$ in the ingredients of V. Then the minimizer of V should be a good estimate of the minimizer of R given by

$$R(\lambda, \theta) = \frac{1}{n}\|\mathbb{S}^{-1/2}[\mathbb{K}(\theta_{\mathrm{true}})\mathbb{x}_{\mathrm{true}} - \mathbb{K}(\theta)\hat{\mathbb{x}}_{\lambda,\theta}]\|^2, \quad (2.15)$$

where θ_{true} contains the true (but unknown) components of θ in $\mathbb{K}$ and $\mathbb{x}_{\mathrm{true}}$ is the true (but unknown) state vector. The establishment of which parameters can be/should be tuned in this way is an important separate subject that we will treat elsewhere; however, it is clear that the Hessians of R and V with respect to the parameters being estimated should be well conditioned.

The minimization of V can be carried out for medium-sized datasets via the algorithm of Gu and Wahba (1991), which uses truncated matrix decomposition methods. The code RKPACK (Gu 1989), which implements this algorithm, is available over the Internet through the public library netlib in the gcv directory there.[1] RKPACK will actually minimize V with respect to multiple smoothing parameters $\lambda_1, \ldots, \lambda_p$ that arise in problems when $\lambda\mathbb{x}'\Sigma^{-1}\mathbb{x}$ of (2.2) is replaced by $\Sigma_{\alpha=1}^{p}\lambda_\alpha\mathbb{x}'J_\alpha\mathbb{x}$. The code GCVPACK (Bates et al. 1986) will minimize V in the context of the thin plate spline described in Wahba and Wendelberger (1980), as well as in the general context of (2.2), again using matrix decomposition methods. GCVPACK and other computer code containing GCV estimates can also be found in the gcv directory of netlib. Matrix decomposition methods, however, are not at present suitable for datasets of the size that occur in global-scale numerical weather prediction.

3. The randomized computation of V

a. *Exact and randomized GCV*

In this section we describe a method that may be used to calculate V in the context of operational global-scale NWP, whenever the means are available to solve

[1] Write netlib@ornl.gov with the words "send index" in the body of the message, and the netlib robot mailserver will respond with instructions for using the system.

the variational problem for the state vector with a (single) random perturbation of the data, along with the original data. The idea is to estimate the required trace by Monte Carlo, or randomized methods, and was proposed in connection with the calculation of GCV functions like those of (2.11) by Deshpande and Girard (1991), Girard (1987, 1989, 1991), and Hutchinson (1989). Girard (1991) proved that the error due to the randomization part was generally negligible in the context of estimation of certain parameters by GCV. Let $\mathbf{B}(\boldsymbol{\theta})$ be any $n \times n$ matrix depending on some parameter vector $\boldsymbol{\theta}$, with i, jth entry $b_{ij}(\boldsymbol{\theta})$, and let $\boldsymbol{\xi}$ be an n-dimensional random vector with components $\{\xi_i\}$ satisfying $E\xi_i = 0$, $E\xi_i\xi_j = 1$, $i = j$, $= 0$, $i \neq j$, where E is expectation. Then $n^{-1}\boldsymbol{\xi}'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{\xi} = n^{-1}\sum_{i,j=1}^{n} \xi_i\xi_j b_{ij}(\boldsymbol{\theta})$ and $En^{-1}\boldsymbol{\xi}'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{\xi} = n^{-1}\sum_{i=1}^{n} b_{ii}(\boldsymbol{\theta}) = n^{-1} \times \mathrm{Tr}\mathbf{B}(\boldsymbol{\theta})$. The randomized trace estimate of $n^{-1}\mathrm{Tr}\mathbf{B}(\boldsymbol{\theta})$ is then given by $n^{-1}\boldsymbol{\xi}'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ comes from a random number generator. If $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I})$, then the standard deviation of this randomized trace estimate is $(2n^{-1})^{1/2}[n^{-1}\mathrm{Tr}\,\mathbf{B}^2(\boldsymbol{\theta})]^{1/2}$ (see Girard 1989, 1991). If $\mathbf{B}(\boldsymbol{\theta})$ is a smoother matrix, that is, it is symmetric, nonnegative definite, with all its eigenvalues between 0 and 1, as is the case for any matrix $\mathbf{A}$ of the form of (2.10), then $0 \leq n^{-1}\mathrm{Tr}\mathbf{B} \leq 1$, and the standard deviation of $n^{-1}\boldsymbol{\xi}'\mathbf{B}(\boldsymbol{\theta})\boldsymbol{\xi}$ is no greater than $(2n^{-1})^{1/2}[n^{-1} \times \mathrm{Tr}\mathbf{B}(\boldsymbol{\theta})]^{1/2}$.

We have run a small toy problem that demonstrates that an estimate of a smoothing parameter λ in this problem calculated via randomized GCV gives, for all practical purposes, just as good a value for λ as one calculated more exactly using matrix decompositions. We generated data from the model

$$y_i = f[\mathbf{x}(i)] + \epsilon_i, \quad i = 1, \cdots, n, \qquad (3.1)$$

where $\mathbf{x}(i) = [x_1(i), x_2(i)]$ is a point in the unit square and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)' \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We took $f(\mathbf{x})$ as Franke's principal test function. A formula and plot of f appear in Wahba (1983) and Wahba (1990b, Fig. 5.1), where f was used to test Bayesian confidence intervals; f is a smooth function with two round peaks and a rounded valley with minimum height near 0 and maximum height at approximately 1.2. The standard deviation σ of the noise ϵ was 0.1, and $676 = 26^2$ values of $\mathbf{x}(i) = [x_1(i), x_2(i)]$ were chosen on a regular 26×26 grid on the unit square. GCVPACK was used to estimate f given $\mathbf{y}$ by a thin-plate smoothing spline (Wahba and Wendelberger 1980). The thin-plate spline used here is the solution to the minimization problem: find f_λ (in an appropriate function space) to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f[\mathbf{x}(i)]\}^2 + \lambda\int_{\infty}^{\infty}\int_{\infty}^{\infty}(f_{x_1,x_1}^2 + 2f_{x_1,x_2}^2 + f_{x_2,x_2}^2)dx_1dx_2, \qquad (3.2)$$

where $\mathbf{x} = (x_1, x_2)$. An analytical representation for f_λ and for the influence matrix $\mathbf{A}(\lambda)$ that satisfies $\{f_\lambda[\mathbf{x}(1)], f_\lambda[\mathbf{x}(2)], \cdots, f_\lambda[\mathbf{x}(n)]\}' = \mathbf{A}(\lambda)\mathbf{y}$ is part of GCVPACK. GCVPACK uses matrix decomposition methods to compute

$$V(\lambda) = \frac{\mathbf{y}'[\mathbf{I} - \mathbf{A}(\lambda)]^2\mathbf{y}}{\{\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\lambda)]\}^2} \qquad (3.3)$$

to a large number of significant figures. The solid line in Fig. 1 is a plot of $V(\lambda)$ as a function of λ as computed by GCVPACK. The dotted line is $R(\lambda)$, the predictive mean-square error (PMSE) function, defined by

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{f[\mathbf{x}(i)] - f_\lambda[\mathbf{x}(i)]\}^2. \qquad (3.4)$$

Function $R(\lambda)$ can be plotted only in a synthetic experiment when the truth is known and it is used to check the performance of the GCV in synthetic experiments. Note that the minimizer of $V(\lambda)$, indicated by a diamond in the figure, is very close to the minimizer of $R(\lambda)$, indicated by the circle. This is as predicted by the theory. [See Wahba (1990) or Wahba and Wendelberger (1980) for further discussion and references.] The dashed line in Fig. 1 is a plot of $\mathrm{Ran}V(\lambda)$, given by

$$\mathrm{Ran}V(\lambda) = \frac{n^{-1}\mathbf{y}'[\mathbf{I} - \mathbf{A}(\lambda)]^2\mathbf{y}}{\{n^{-1}\boldsymbol{\xi}'[\mathbf{I} - \mathbf{A}(\lambda)]\boldsymbol{\xi}\}^2}, \qquad (3.5)$$

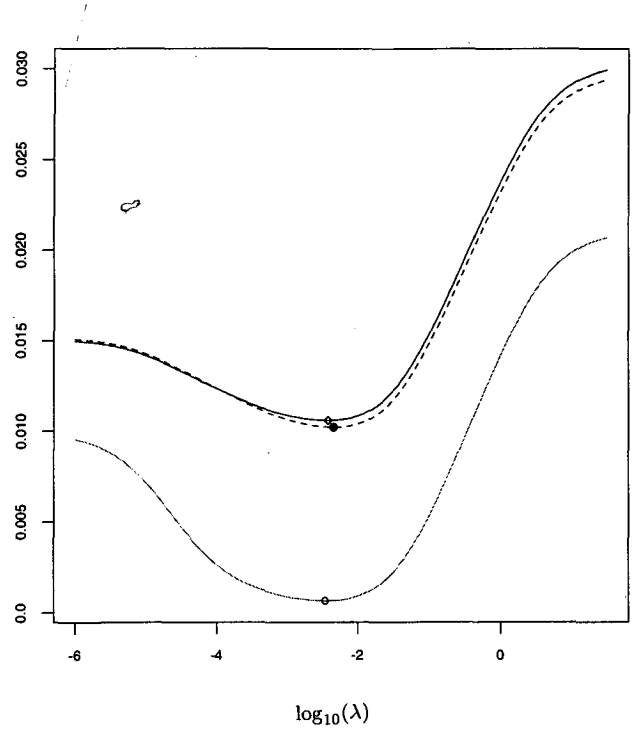

FIG. 1. The exact GCV function and one replicate of the randomized GCV function with the PMSE function. Solid line, $V(\lambda)$; dotted line, $R(\lambda)$; dashed line, $\mathrm{Ran}V(\lambda)$.

where ξ came from a random number generator $\xi \sim \mathcal{N}(0, \mathbf{I})$ and $\mathrm{Tr}[\mathbf{I} - \mathbf{A}(\lambda)]$ in $V(\lambda)$ has been replaced by a randomized estimate of it. It is important to note that the *same* ξ is used for all values of λ. Figure 2 contains the exact GCV function and PMSE function of Fig. 1 on an expanded scale as well as 10 replicates of Ran$V$. Each replicate represents a different ξ obtained from a random number generator. It can be seen that the minimizers of all 10 replicates do an excellent job of estimating the minimizer of $R(\lambda)$, even though the heights of the replicates vary.

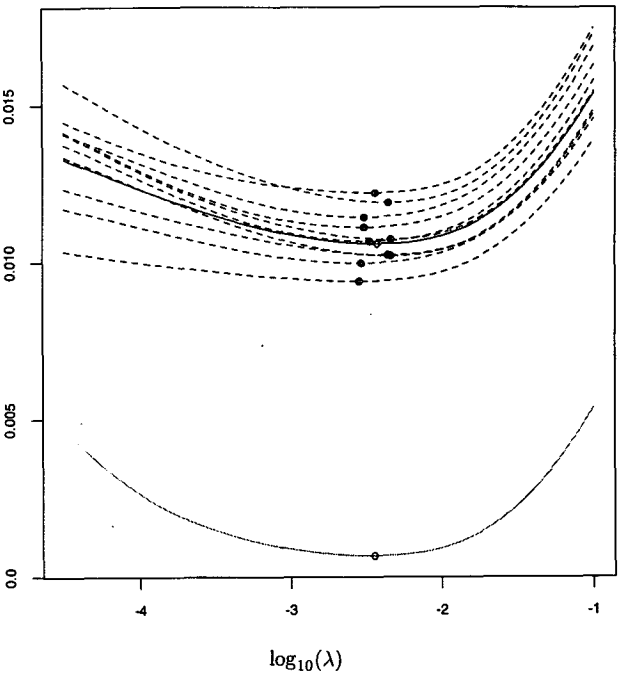

FIG. 2. Ten replicates of the randomized GCV function along with the exact GCV function and the PMSE function. Solid line, $V(\lambda)$; dotted line, $R(\lambda)$; dashed lines, 10 replicates of Ran$V(\lambda)$.

b. *Randomized GCV and the iterative solution of extremely large variational problems*

In modern operational global-scale numerical weather prediction models, some iterative method, with k iterations, is used to obtain an approximate solution $\hat{\mathbf{x}}^k$ to the minimizer of

$$[\mathbf{y} - \mathbf{K}(\mathbf{x})]'\mathbf{S}^{-1}[\mathbf{y} - \mathbf{K}(\mathbf{x})] + \mathbf{x}'\Sigma^{-1}\mathbf{x}, \quad (3.6)$$

given $\mathbf{K}$, $\mathbf{S}^{-1}$, and Σ^{-1}. See for example Parrish and Derber (1992),[2] who use a conjugate gradient method, and Lorenc et al. (1991) and Lorenc (1992), who discuss successive correction and other methods. The L-BFGS algorithm (Liu and Nocedal 1989) is also popular. The particular method used is not important in what follows. We assume only that there is an operational code that we call the "black box" that, given $\mathbf{K}$, $\mathbf{S}^{-1}$, Σ^{-1}, and $\mathbf{y}$, returns (after k iterations) $\hat{\mathbf{x}}^k = \hat{\mathbf{x}}^k(\mathbf{y})$, which is an approximation to the minimizer $\hat{\mathbf{x}}$ of (3.6). This black box can be augmented to return $\mathbf{K}(\hat{\mathbf{x}}^k)$. In what follows, we assume that $\mathbf{S}$ is known and incorporated into $\tilde{\mathbf{y}} = \mathbf{S}^{-1/2}\mathbf{y}$ and $\tilde{\mathbf{K}} = \mathbf{S}^{-1/2}\mathbf{K}$, and we drop the tilde on $\mathbf{y}$ and $\mathbf{K}$ in the rest of this section. The black box may now be used to obtain a randomized estimate of the trace of the matrix that plays the role of $\mathbf{A}(\theta)$, where now θ represents the unknown parameters (including λ). Even if $\mathbf{K}$ is linear in $\mathbf{x}$, the relationship between $\mathbf{y}$ and $\mathbf{K}[\hat{\mathbf{x}}^k(\mathbf{y})]$ is not necessarily linear in $\mathbf{y}$ if the iteration is stopped before the exact minimizer $\hat{\mathbf{x}}$ has been found. This happens in, for example, the conjugate gradient algorithm if $k < n$. Thus, we no longer have an influence matrix $\mathbf{A}(\theta)$ that satisfies $\mathbf{K}\hat{\mathbf{x}}_\theta = \mathbf{A}(\theta)\mathbf{y}$ but an influence operator $\mathbf{A}^k(\theta, \mathbf{y})$ defined by $\mathbf{K}[\hat{\mathbf{x}}^k_\theta(\mathbf{y})] = \mathbf{A}^k(\theta, \mathbf{y})$. Define the matrix $\mathbf{A}^k_{\mathbf{y}}(\theta)$ by

$$\mathbf{K}(\hat{\mathbf{x}}^k_\theta) \approx \mathbf{A}^k_{\mathbf{y}}(\theta)\mathbf{y}; \quad (3.7)$$

that is, $\mathbf{A}^k_{\mathbf{y}}(\theta)$ is the linearized version of the influence operator implicitly defined by the black box, evaluated at $\mathbf{y}$. If the black box is always used to find the approximate minimizer of (3.6), then it is actually $\mathrm{Tr}\mathbf{A}^k_{\mathbf{y}}(\theta)$ and not $\mathrm{Tr}\mathbf{A}(\theta)$ that should be used in computing the GCV function $V(\theta)$ of (2.11). In any case, let ξ come from a random number generator with $\xi \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{I})$ and let $\hat{\mathbf{x}}^k_\theta(\mathbf{y} + \xi)$ be the estimate for $\mathbf{x}$ obtained by substituting $(\mathbf{y} + \xi)$ for $\mathbf{y}$ in (3.6) and by using the black box to obtain the approximate minimizer. Then

$$\mathbf{K}[\hat{\mathbf{x}}^k_\theta(\mathbf{y} + \xi)] \approx \mathbf{A}^k_{\mathbf{y}+\xi}(\theta)(\mathbf{y} + \xi). \quad (3.8)$$

A randomized estimate of $\mathrm{Tr}\mathbf{A}^k_{\mathbf{y}}(\theta)$ is given by

$$\frac{1}{\sigma_\xi^2}[\xi'\mathbf{A}^k_{\mathbf{y}+\xi}(\theta)(\mathbf{y} + \xi) - \xi'\mathbf{A}^k_{\mathbf{y}}(\theta)\mathbf{y}] \quad (3.9)$$

$$\approx \frac{1}{\sigma_\xi^2}\xi'\{\mathbf{K}[\hat{\mathbf{x}}^k_\theta(\mathbf{y} + \xi)] - \mathbf{K}[\hat{\mathbf{x}}^k_\theta(\mathbf{y})]\}, \quad (3.10)$$

where we have written $\hat{\mathbf{x}}^k_\theta(\mathbf{y} + \xi)$ for $\hat{\mathbf{x}}^k_\theta$ based on the data $\mathbf{y} + \xi$.

In the case of linear iterative methods, such as the Richardson–Landweber–Fridman–Pickard–Cimino (RLFPC) iteration (with $\mathbf{K}$ linear), simple exact formulas can be obtained for $\mathbf{A}^k_{\mathbf{y}}(\theta)$, not depending on $\mathbf{y}$ (see Wahba 1987). It is shown there and in Fleming (1990) that early stopping of the iteration with such a linear method is a form of regularization, or low-pass filtering. Roughly speaking, a stopped iteration tends to project the exact solution toward eigenvectors corresponding to large eigenvalues, and these eigenvectors tend to be smooth. Thus, both k, the number of iterations, and λ, a multiplier on Σ^{-1}, can be thought of as

[2] In Parrish and Derber and elsewhere $\mathbf{y}$ represents a forecast increment; see section 4a.

smoothing or regularization parameters. It was suggested that GCV could be used to choose both k and λ simultaneously in Wahba (1987); however, no procedure for carrying out the calculations with large datasets was provided there.

We have constructed a semirealistic toy problem to test and demonstrate the feasibility and efficiency of choosing both k and λ via GCV, in conjunction with the randomized trace estimation of (3.10). Rather than use an RLFPC algorithm for this demonstration, we have chosen to use a preconditioned conjugate gradient algorithm, since conjugate gradients are used operationally and are well known to have favorable properties. The experimental setup we use here is a part of the experimental setup in Gao (1993) and Gao et al. (1995, unpublished manuscript). European Centre for Medium-Range Weather Forecasts (ECMWF) Gridded Level IIIB First GARP (Global Atmospheric Research Program) Global Experiment data for the 500-mb height for 2 January 1979 was used to obtain a spherical harmonic representation for this 500-mb height field of the form

$$f(P) = \sum_{l=0}^{30} \sum_{s=-l}^{l} x_{ls} Y_{ls}(P), \tag{3.11}$$

where P is a point on the sphere and the Y_{ls} are spherical harmonics. This representation was obtained by solving a variational problem given the gridded data. The amount of smoothing was chosen to make the resulting contour plots match the ECMWF plots visually (see Gao 1993 for details). Simulated observational data at $n = 600$ North American radiosonde stations was generated by

$$y_i = f(P_i) + \epsilon_i, \tag{3.12}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_{600})' \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and the P_i are station locations. We chose $\sigma = 9$ m to represent observational error. A spline on the sphere analysis analogous to Wahba and Wendelberger (1980) is obtained by letting $\hat{\mathbf{x}}_\lambda = (\hat{x}_{00,\lambda}, \hat{x}_{10,\lambda}, \cdots)$ be the minimizer of

$$\sum_{i=1}^{n} [y_i - \sum_{l=0}^{30} \sum_{s=-l}^{l} x_{ls} Y_{ls}(P_i)]^2 + \lambda \sum_{l=0}^{30} \sum_{s=-l}^{l} [(l)(l+1)]^2 x_{ls}^2. \tag{3.13}$$

The penalty functional $J(f) = \Sigma_{ls} [(l)(l+1)]^2 x_{ls}^2$ is a multiple of $J(f) = \int_S (\Delta f)^2$, where Δ is the Laplacian on the sphere (see Wahba 1981, 1982a). Letting $\mathbf{K}$ be the 600×960 matrix with entries $Y_{ls}(P_i)$ and letting $\mathbf{D}$ be the diagonal matrix with ls, ls entries $[l(l+1)]^2$, then the minimizer $\hat{\mathbf{x}}_\lambda$ satisfies

$$(\mathbf{K}'\mathbf{K} + \lambda\mathbf{D})\hat{\mathbf{x}}_\lambda = \mathbf{K}'\mathbf{y}. \tag{3.14}$$

A preconditioned conjugate gradient algorithm with (symmetric, invertible) preconditioner $\mathbf{C}$ replaces $\mathbf{x}_\lambda$ by $\mathbf{C}^{-1}\mathbf{w}$ in (3.14) and solves for $\mathbf{w}$ in

$$\mathbf{C}^{-1}(\mathbf{K}'\mathbf{K} + \lambda\mathbf{D})\mathbf{C}^{-1}\mathbf{w} = \mathbf{C}^{-1}\mathbf{K}'\mathbf{y} \tag{3.15}$$

(see Golub and van Loan 1989, section 10.3). In the experiment below $\mathbf{C}$ was taken as $[\text{diag}(\mathbf{K}'\mathbf{K} + \lambda\mathbf{D})]^{1/2}$. The starting point $\mathbf{x}^0$ of the iteration was taken as $\mathbf{x}^0 = \tilde{\mathbf{D}}^{-1}\mathbf{K}'\mathbf{y}$, where $\tilde{\mathbf{D}}^{-1}$ is a diagonal approximation to $(\mathbf{K}'\mathbf{K} + \lambda\mathbf{D})^{-1}$ obtained by replacing the lower right 959×959 dimensional block of $(\mathbf{K}'\mathbf{K} + \lambda\mathbf{D})$ by its diagonal and by inverting analytically.

Following (2.13) we define the predictive mean-square error as

$$R(\lambda, k) = \frac{1}{n} \sum_{i=1}^{n} [f_\lambda^k(P_i) - f(P_i)]^2, \tag{3.16}$$

where now

$$f_\lambda^k(P) = \sum_{ls} \hat{x}_{ls,\lambda}^k Y_{ls}(P), \tag{3.17}$$

$$\hat{\mathbf{x}}_\lambda^k = \{\hat{x}_{ls,\lambda}^k\}, \tag{3.18}$$

and $\hat{\mathbf{x}}_\lambda^k$ is the approximate solution after k iterations. Figure 3 gives a plot of $R^{1/2}(\lambda, k)$ as a function of λ and k, where k is the number of iterations in the conjugate gradient iterative solution of (3.15).

This kind of plot is, of course, available only in a simulation study where the ground truth is known; $R(\lambda, k)$ is minimized at around $-\log_{10}(\lambda) = 4.5$, and $k = 75$. Note that the value of $R^{1/2}(\lambda, k)$ at the minimum is about 6 m. Assuming that good approximation to these optimal values of λ and k can be found, the smoothing procedure has resulted in a smoothed minus true standard deviation that is about one-third less than the observational standard deviation. (This would be reduced further if unbiased forecast data were also available.)

Figure 4 gives a plot of the randomized version of the GCV function of (2.11) as a function of k and λ. The randomized GCV function is computed as

$$\text{Ran}V(\lambda, k) = \frac{n^{-1}\|\mathbf{y} - \mathbf{K}\hat{\mathbf{x}}_\lambda^k\|^2}{(n^{-1}\langle\sigma_\xi^{-2}\boldsymbol{\xi}'\{\boldsymbol{\xi} - [\mathbf{K}\hat{\mathbf{x}}_\lambda^k(\mathbf{y} + \boldsymbol{\xi}) - \mathbf{K}\hat{\mathbf{x}}_\lambda^k(\mathbf{y})]\}\rangle)^2}, \tag{3.19}$$

where $\boldsymbol{\xi}$ came from a random number generator, $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2\mathbf{I})$. The numerator in (3.19) is the mean residual sum of squares, and the expression in the denominator is the randomized trace estimate; compare (2.11). Ran$V$ is based only on the data. The same $\boldsymbol{\xi}$ was used for the entire plot. The standard deviation σ_ξ for the random vector $\boldsymbol{\xi}$ should be chosen carefully if $\mathbf{A}_y^k(\boldsymbol{\theta})\mathbf{y}$ is not linear in $\mathbf{y}$. If σ_ξ is too small, then the calculation of the difference in (3.10) may be unstable, while if σ_ξ is too large, the behavior at $\mathbf{A}_y^k(\boldsymbol{\theta})$ may not be captured. After a little trial and error we found that a σ_ξ of the order of but smaller than the presumed σ of the noise in $\mathbf{y}$ worked well. In these experiments we took $\sigma_\xi = 3$ m $= 1/3\sigma$. The value of Ran$V^{1/2}$ at the minimum (11.35) is roughly an estimate of $[\min_{\lambda,k} R(\lambda, k) + \sigma^2]^{1/2} = 10.8$,

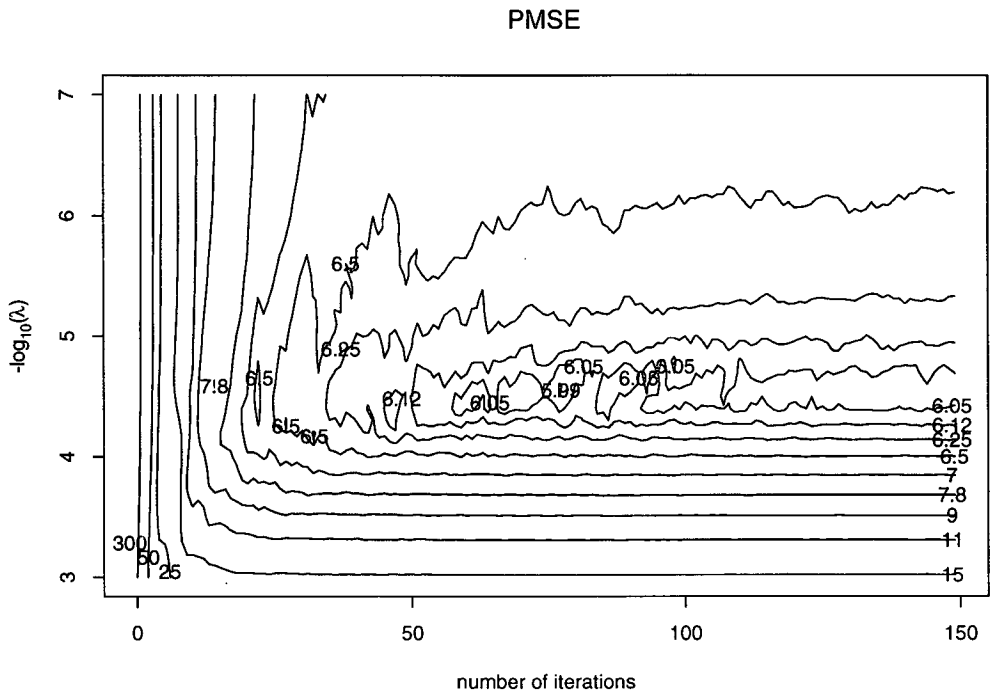

FIG. 3. The root predictive mean-square error $R^{1/2}$ as a function of $\log_{10}(\lambda)$ and k.

as predicted by the theory (see Wahba 1990). The striking thing to observe about Ran$V$ is how well it actually estimates $R(\lambda, k) + \sigma^2$. The minimum GCV score is located in a region for which the PMSE score $R^{1/2}$ is less than or equal to 6.25, so that if a value of λ and a stopping rule k based on minimizing Ran$V$ were

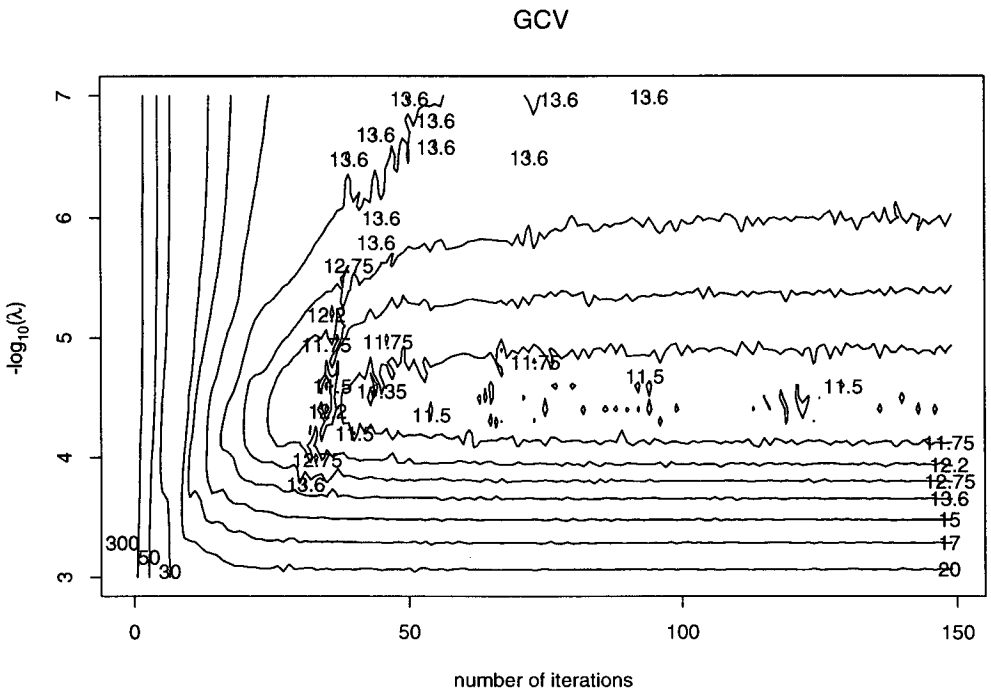

FIG. 4. Ran$V^{1/2}$ as function of $\log_{10}(\lambda)$ and k.

used, the ratio of the resulting predictive mean-square error to the minimum possible predictive mean-square error (known as the inefficiency) would be no larger than 6.25/5.99 = 1.04.

Before this experiment was conducted we had conjectured that there would be a minimum in $R(\lambda, k)$ and, consequently, in Ran$V(\lambda, k)$ at some value of $k \ll n$, but here the minimum is fairly shallow and the (λ, k) surfaces flatten out as a function of k as k gets larger. In practice, of course, one could stop as soon as the surface has flattened out. In another experiment, where the noise standard deviation σ was inadvertently set to the unrealistically large value of 120 m, there was a very distinct minimum in k, suggesting that k was an important regularization parameter. We also conjectured that the best λ might depend on k (that is, smaller λ might want a smaller k, due to a regularizing effect of the smaller number of iterations) but that is not evident in Fig. 3, since the minimizing λ appears not to depend on k. However, we do not rule out this phenomena in other experiments or in practice.

4. Other applications

a. Objective analysis and Kalman filtering

In an NWP analysis such as the spectral statistical interpolation (Parrish and Derber 1992), one may consider that the forecast $\mathbf{x}^f = \mathbf{x} + \boldsymbol{\eta}$ where $\mathbf{x}$ is the true state vector[3] and $\boldsymbol{\eta}$ is modeled as $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{S}_f)$, where $\mathbf{S}_f = \mathbf{S}_f(\boldsymbol{\theta})$ is the putative forecast error covariance, depending on some parameters $\boldsymbol{\theta}$, and $\mathbf{y} = \mathbf{K}\mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{S})$. Then the Gandin (Bayes) estimate $\mathbf{x}^a$ of $\mathbf{x}$ is the minimizer of

$$\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{K}\mathbf{x})'\mathbf{S}^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}) + (\mathbf{x} - \mathbf{x}^f)'\mathbf{S}_f^{-1}(\mathbf{x} - \mathbf{x}^f). \quad (4.1)$$

Let $\mathbf{y}^* = \mathbf{y} - \mathbf{K}\mathbf{x}^f$ be the so-called innovation vector and let $\boldsymbol{\delta}^a = \mathbf{x}^a - \mathbf{x}^f$ be the analysis increment. Reparameterizing $\mathbf{S}_f$ as $b\boldsymbol{\Sigma}(\boldsymbol{\theta})$, with b and $\boldsymbol{\theta}$ to be estimated and setting $\lambda = \sigma^2/b$, then $\boldsymbol{\delta}^a$ is the minimizer of

$$(\mathbf{y}^* - \mathbf{K}\boldsymbol{\delta})'\mathbf{S}^{-1}(\mathbf{y}^* - \mathbf{K}\boldsymbol{\delta}) + \lambda\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\delta}. \quad (4.2)$$

Then λ and (estimable) $\boldsymbol{\theta}$ may be obtained as the minimizers of

$$\text{Ran}V(\boldsymbol{\theta}, \lambda) = \frac{n^{-1}(\mathbf{y}^* - \mathbf{K}\boldsymbol{\delta}^a_{\theta,\lambda})'\mathbf{S}^{-1}(\mathbf{y}^* - \mathbf{K}\boldsymbol{\delta}^a_{\theta,\lambda})}{[n^{-1}(\sigma_\xi^{-2}\boldsymbol{\xi}'\langle\boldsymbol{\xi} - \{\mathbf{S}^{-1/2}[\mathbf{K}\boldsymbol{\delta}^a_{\theta,\lambda}(\mathbf{y}^* + \mathbf{S}^{1/2}\boldsymbol{\xi}) - \mathbf{K}\boldsymbol{\delta}^a_{\theta,\lambda}(\mathbf{y}^*)]\}\rangle)]^2}, \quad (4.3)$$

where $\boldsymbol{\delta}^a_{\theta,\lambda}$ is the minimizer of (4.2).

[3] This is a fiction of course. We omit discussion of the relationship between the true atmosphere and the best vector representation of it in a model.

In the context of the (linearized) Kalman filter the forecast error covariance at time t, call it $\mathbf{S}_{f,t}$ is modeled by

$$\mathbf{S}_{f,t} = \mathbf{M}_{t-1}\mathbf{P}^a_{t-1}\mathbf{M}'_{t-1} + \mathbf{Q}_{t-1}, \quad (4.4)$$

where $\mathbf{M}_t$ is the operation that produces $\mathbf{x}^f_t$, the forecast at time t, in terms of the analyses $\mathbf{x}^a_{t-1}$ at time $t-1$, by $\mathbf{x}^f_t = \mathbf{M}_{t-1}\mathbf{x}^a_{t-1}$, $\mathbf{Q}_{t-1}$ is the model error covariance matrix and $\mathbf{P}^a_{t-1}$ is the analysis error covariance at time $t-1$; $\mathbf{P}^a_{t-1}$ in theory satisfies well-known recursion relations that we will not repeat here. The (usual) Kalman filter produces $\mathbf{x}^a$, and hence $\mathbf{K}\boldsymbol{\delta}^a = \mathbf{K}\mathbf{x}^a - \mathbf{K}\mathbf{x}^f$, although not usually by solving the variational problem. If $\mathbf{Q}_{t-1} = \mathbf{Q}_{t-1}(\boldsymbol{\theta})$ and $\mathbf{P}^a_{t-1} = \mathbf{P}^a_{t-1}(\boldsymbol{\theta})$ depend on unknown (identifiable) parameters, then they may be estimated by minimizing Ran$V$ of (4.3) with $\mathbf{S}_f = \mathbf{S}_{f,t}(\boldsymbol{\theta})$ given by (4.4). Dee (1995) has estimated some of these parameters by maximum likelihood. He notes that "since the underlying assumptions are actually violated, the unknown quantities . . . should be regarded as calibration parameters, which do not necessarily have any physical meaning." This makes them candidates for estimation by GCV (see appendix B).

b. Four-dimensional variational objective analysis

Assuming that there is no model error, the state $\mathbf{x}_t$ is related to the state $\mathbf{x}_0$ by

$$\mathbf{x}_t = \mathbf{M}_{t-1}\{\mathbf{M}_{t-2}[\cdots\mathbf{M}_0(\mathbf{x}_0)]\} = \tilde{\mathbf{M}}_t(\mathbf{x}_0). \quad (4.5)$$

Let $\mathbf{y}_t$ be a vector of observations at time t related to $\mathbf{x}_t$ by $\mathbf{y}_t = \mathbf{K}_t(\mathbf{x}_t) + \boldsymbol{\epsilon}_t$, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{S}_t)$. A four-dimensional variational data assimilation with the model as a strong constraint, assuming that model errors are negligible, would find the initial state $\mathbf{x}_0$ to minimize[4]

$$\sum_{t=0}^{T}\{\mathbf{y}_t - \mathbf{K}_t[\tilde{\mathbf{M}}_t(\mathbf{x}_0)]\}'\mathbf{S}_t^{-1}\{\mathbf{y}_t - \mathbf{K}_t[\tilde{\mathbf{M}}_t(\mathbf{x}_0)]\} + (\mathbf{x}_0 - \mathbf{x}^*)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \mathbf{x}^*), \quad (4.6)$$

[4] Most (but not all) authors experimenting with four-dimensional variational data assimilation have found that some penalty term of the form $(\mathbf{x}_0 - \mathbf{x}^*)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \mathbf{x}^*)$($\mathbf{x}^*$ may be 0) based on balance or related physical considerations is necessary or at least improves the analysis. From a mathematical point of view, in order for a variational problem of the form $(\mathbf{y} - \mathbf{K}\mathbf{x})'\mathbf{S}^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}) + \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ to have a unique minimizer it is necessary that the intersection of the null spaces of $\mathbf{K}$ and $\boldsymbol{\Sigma}^{-1}$ are empty; that is, $\mathbf{K}\mathbf{x} = 0$ and $\boldsymbol{\Sigma}^{-1}\mathbf{x} = 0$ imply that $\mathbf{x} = 0$. If this penalty term is omitted and there are a large number of degrees of freedom in the model, this may not be true (see also Bennett and Miller 1991). Of course the larger T is the less important this penalty term may become. Balance or other penalty terms may also be imposed at some time other than t_0.

where $\mathbf{x}^*$ is a starting guess, perhaps a forecast. Sophisticated methods for finding $\mathbf{x}_0$ to minimize (4.6) are a subject of great interest (see, for example, Zou et al. 1992). Letting $\mathbf{K} = \{\mathbf{K}_0[\tilde{\mathbf{M}}_0(\cdot)]', \mathbf{K}_1[\tilde{\mathbf{M}}_1(\cdot)]', \cdots, \mathbf{K}_T[\tilde{\mathbf{M}}_T(\cdot)]'\}'$, we may write (4.6) as

$$[\mathbf{y} - \mathbf{K}(\mathbf{x}_0)]'\mathbf{S}^{-1}[\mathbf{y} - \mathbf{K}(\mathbf{x}_0)] + (\mathbf{x}_0 - \mathbf{x}^*)'\Sigma^{-1}(\mathbf{x}_0 - \mathbf{x}^*). \quad (4.7)$$

Letting $\boldsymbol{\theta}$ be unknown parameters in $\mathbf{K}$ and Σ, and letting $\mathbf{x}^a_{0,\theta}(\mathbf{y})$ and $\mathbf{x}^a_{0,\theta}(\mathbf{y} + \mathbf{S}^{1/2}\boldsymbol{\xi})$ be the minimizers of (4.6) based on data $\mathbf{y}$ and $\mathbf{y} + \mathbf{S}^{1/2}\boldsymbol{\xi}$, respectively, the randomized GCV estimate of $\boldsymbol{\theta}$ is found by minimizing

$$\mathrm{Ran}V(\theta) = \frac{n^{-1}\{\mathbf{y} - \mathbb{K}[\mathbb{x}^a_{0,\theta}(\mathbf{y})]\}'\mathbf{S}^{-1}\{\mathbf{y} - \mathbb{K}[\mathbb{x}^a_{0,\theta}(\mathbf{y})]\}}{n^{-1}[\sigma_\xi^{-2}\boldsymbol{\xi}'(\boldsymbol{\xi} - \langle \mathbf{S}^{-1/2}\{\mathbb{K}[\mathbb{x}^a_{0,\theta}(\mathbf{y} + \mathbf{S}^{1/2}\boldsymbol{\xi})] - \mathbb{K}[\mathbb{x}^a_{0,\theta}(\mathbf{y})]\}\rangle)]^2}. \quad (4.8)$$

If model errors are important, then one may define a control variable $\mathbb{x}^c_0$ such that $\mathbb{x}^a_T = \tilde{\mathbb{M}}_T(\hat{\mathbb{x}}^c_0)$, where $\mathbb{x}^a_T$ is the analysis at time T and $\hat{\mathbb{x}}^c_0$ is an estimate of $\mathbb{x}^c_0$. Here, $\mathbb{x}^c_0$ is no longer necessarily considered to be the state vector at time 0 but rather a control variable that parametrizes trajectories. A good estimate $\hat{\mathbb{x}}^c_0$ of $\mathbb{x}^c_0$ is then one that provides a good analysis $\mathbb{x}^a_T$ and is not necessarily the best estimate of the state at time $t = 0$; $\hat{\mathbb{x}}^c_0 = \hat{\mathbb{x}}^c_{0,\theta}$ is taken as the minimizer of

$$\{\mathbf{y}_T - \mathbf{K}_T[\tilde{\mathbf{M}}_T(\mathbf{x}^c_0)]\}'\mathbf{S}_T^{-1}\{\mathbf{y}_T - \mathbf{K}_T[\tilde{\mathbf{M}}_T(\mathbf{x}^c_0)]\} + \sum_{t=0}^{T-1}\{\mathbf{y}_t - \mathbf{K}_t[\tilde{\mathbf{M}}_t(\mathbf{x}^c_0)]\}'[\mathbf{V}_t(\boldsymbol{\theta}) + \mathbf{S}_t]^{-1} \times \{\mathbf{y}_t - \mathbf{K}_t[\tilde{\mathbf{M}}_t(\mathbf{x}^c_0)]\} + (\mathbb{x}^c_0 - \mathbb{x}^*)'[\mathbb{V}_0(\boldsymbol{\theta}) + \Sigma]^{-1}(\mathbb{x}^c_0 - \mathbb{x}^*), \quad (4.9)$$

where the $\mathbf{V}_t(\boldsymbol{\theta})$ are included to compensate for model error. Note that this formulation is consistent with the desire to estimate $\mathbf{x}_T$ as well as possible, and then it would be expected that the $\mathbf{V}_t$ would increase going backward in time, consistent with the desire to rely more strongly on recent data when there is model error. Again, consistent with the desire to obtain a good estimate of $\mathbb{x}_T$, one may then do a partial GCV, based only on the data at time T. In that case $\boldsymbol{\theta}$ is chosen to minimize

$$\mathrm{Ran}_{\mathrm{partial}}V(\boldsymbol{\theta}) = \frac{\{\mathbf{y}_T - \mathbf{K}_T[\tilde{\mathbf{M}}_T(\hat{\mathbf{x}}^c_{0,\theta})]\}'\mathbf{S}_T^{-1}\{\mathbf{y}_T - \mathbf{K}_T[\tilde{\mathbf{M}}_T(\hat{\mathbf{x}}^c_{0,\theta})]\}}{\{\sigma_\xi^{-2}\boldsymbol{\xi}'[\boldsymbol{\xi} - (\mathbf{S}_T^{-1/2}\langle\mathbf{K}_T\{\tilde{\mathbf{M}}_T[\hat{\mathbf{x}}^c_{0,\theta}(\mathbf{y}_T + \mathbf{S}_T^{1/2}\boldsymbol{\xi})]\} - \mathbf{K}_T\{\tilde{\mathbf{M}}_T[\mathbf{x}^c_{0,\theta}(\mathbf{y}_T)]\}\rangle)]\}^2}. \quad (4.10)$$

A longer version of this section with further details may be found in Wahba et al. (1994).

5. Summary

We have demonstrated that the randomized trace technique for computing GCV estimates of a smoothing parameter gives essentially the same answer as traditional computational methods for datasets as small as 400. We have shown how this technique may be used to choose the number of iterations when solving a variational problem by an iterative method, while simultaneously choosing a smoothing parameter. Finally, we have described how the technique may potentially be used in the context of very large variational problems as occur in operational three- and four-dimensional data assimilation.

Acknowledgments. The first author wishes to thank a number of people for helpful discussions and early versions of their manuscripts: Phillipe Courtier, Roger Daley, John Derber, Andrew Lorenc, Dick Dee, Michael Ghil, Mike Navon, Dave Parrish, Jim Pfaendtner, Jim Purser, Florence Rabier, and Olivier Talagrand. Thanks also are due to Ken Bergman of NASA headquarters for his encouragement. This research was supported by NASA Grant NAGW-2961 and NSF Grant DMS 9121003. This paper is dedicated to the memory of Henry Fleming and his many contributions to meteorology and mathematics.

APPENDIX A

Conditions for the Validity of GCV

Recalling that the eigenvalues of $\mathbf{A}$ are all in the interval [0, 1], conditions on $\mathbf{A}$ given in the references in section 2 roughly translate into a requirement that for allowable values of $\lambda, \boldsymbol{\theta}$, $\mathbf{A}(\lambda, \boldsymbol{\theta})$ essentially defines a low-pass filter. This requires that a modest fraction

of the eigenvalues of $\mathbf{A}$ are close to 1, and most of the rest are close to 0. In addition, the signal $\mathbf{K}\mathbf{x}_{\text{true}}$ should have most of its energy concentrated in the passband defined by $\mathbf{A}$. Letting $\mu_1 = n^{-1}\,\text{Tr}\mathbf{A}(\lambda, \boldsymbol{\theta})$ and $\mu_2 = n^{-1} \times \text{Tr}\mathbf{A}^2(\lambda, \boldsymbol{\theta})$, the low-pass filter requirement is mathematically formulated as a requirement that μ_1 and μ_1^2/μ_2 be small for λ, $\boldsymbol{\theta}$ near the minimizer of R. The reader can translate these conditions on the eigenvalues of $\mathbf{A}$ to a condition on the eigenvalues of the matrix $\mathbf{S}^{-1/2}\mathbf{K}\Sigma\mathbf{K}'\mathbf{S}^{-1/2}$. Roughly speaking, it is required that a small fraction of these latter eigenvalues be large and most of the rest be small. It is cautioned that if the (rescaled) (observational) noise $\tilde{\boldsymbol{\epsilon}}$ is strongly positively correlated, the GCV cannot be expected to adequately separate it from the signal. This should not be a problem for data such as radiosonde observations that are spatially independent and whose vertical correlation structure is fairly well understood. However, care must be taken in cross-validating against data such as satellite radiance observations whose error structure may be highly correlated and poorly understood.

APPENDIX B

GCV and ML

Relationships between GCV and ML estimates, and a third estimate, the unbiased risk estimate, are discussed in Wahba (1990b). According to statistical theory (Cramer 1954), if the unknowns are parameters in distributions of random variables, and all of the assumptions concerning these random variables are true, then maximum-likelihood estimates will in general give the best estimates of the parameters, in the sense of minimizing the variance of the distributions of the estimates. If the statistical assumptions are sufficiently violated, however, then the maximum-likelihood estimates may fare relatively poorly with regard to other criteria, such as mean-square error of prediction (see Wahba 1985a). The GCV estimates have good properties irrespective of the nature of the second term in (1.1), and, for example, this term may represent an energy penalty and have nothing to do with a covariance. The primary requirement for the validity of the GCV estimates, aside from the condition described in appendix A, is that the observation error covariance matrix $\mathbf{S}$ is known sufficiently well that $\mathbf{S}^{-1/2}\boldsymbol{\epsilon}$ is close to white. If the components of this vector are strongly positively correlated, then the GCV may have difficulty telling signal from noise. The randomized trace technique may also be used to estimate some special parameters in maximum likelihood estimates (see Wahba et al. 1994).

REFERENCES

Bartello, P., and H. Mitchell, 1993: A continuous three-dimensional model of short-range forecast error covariances. *Tellus,* **44A,** 217–235.

Bates, D., M. Lindstrom, G. Wahba, and B. Yandell, 1986: GCVPACK—Routines for generalized cross validation. Tech. Rep. 775 (rev.), Dept. of Statistics, University of Wisconsin, 263–297.

Bennett, A. F., and W. P. Budgell, 1987: Ocean data assimilation and the Kalman filter: Spatial reglarity. *J. Phys. Oceanogr.,* **17,** 1583–1601.

——, and R. N. Miller, 1991: Weighting initial conditions in variational assimilation schemes. *Mon. Wea. Rev.,* **119,** 1098–1102.

Cohn, S., 1993: Dynamics of short-term univariate forecast error covariances. *Mon. Wea. Rev.,* **121,** 3123–3149.

Courtier, P., and O. Talagrand, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. Part II: Numerical results. *Quart. J. Roy. Meteor. Soc.,* **113,** 1329–1347.

——, and ——, 1990: Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus,* **42A,** 531–549.

Cramer, H., 1954: *Mathematical Methods of Statistics.* Princeton University Press, 575 pp.

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

Daley, R., 1992a: The effect of serially correlated observation and model error on atmospheric data assimilation. *Mon. Wea. Rev.,* **120,** 164–177.

——, 1992b: The lagged innovation covariance: A performance diagnostic for atmospheric data assimilation. *Mon. Wea. Rev.,* **120,** 178–196.

——, 1992c: Forecast error statistics for homogeneous and inhomogeneous observation networks. *Mon. Wea. Rev.,* **120,** 627–643.

——, 1992d: Estimating model-error covariances for application to atmospheric data assimilation. *Mon. Wea. Rev.,* **120,** 1735–1746.

Dee, D., 1990: Simplified adaptive Kalman filtering for large-scale geophysical models. *Realization and Modelling in System Theory: Proceedings of the International Symposium MTNS-89, v1,* M. Kaashoek, J. van Schuppen and A. Ram, Eds., Birkhauser, 567–574.

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

——, S. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Autom. Control,* **AC-30,** 1057–1065.

Deshpande, L., and D. Girard, 1991: Fast computation of cross-validated robust splines and other non-linear smoothing splines. *Curves and Surfaces,* P.-J. Laurent, A. LeMehaute, and L. Schumaker, Eds., Academic Press, 143–148.

Fleming, H., 1990: Equivalence of regularization and truncated iteration in the solution of ill-posed image reconstruction problems. *Linear Alg. Appl.,* **130,** 133–150.

Gao, F., 1993: On combining data from multiple sources with unknown relative weights (thesis), Tech. Rep. 902, Dept. of Statistics, University of Wisconsin, Madison, WI, 178 pp.

——, 1994: Fitting smoothing splines to data from multiple sources. *Commun. Statist.-Theory Meth.,* **23,** 1665–1698.

Girard, D., 1987: A fast 'Monte Carlo cross validation' procedure for large least squares problems with noisy data, Tech. Rep. RR 687-M, IMAG, Grenoble, France, 22 pp.

——, 1989: A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data. *Numer. Math.,* **56,** 1–23.

——, 1991: Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *Ann. Stat.,* **19,** 1950–1963.

Goerss, J., and P. Phoebus, 1993: The multivariate optimum interpolation analysis of meteorological data at the fleet numerical oceanography center, Tech. Rep. NRL/FR/7531-92-9413, Naval Research Laboratory, Monterey, CA, 58 pp.

Golub, G., and C. VanLoan, 1989: *Matrix Computations.* 2d ed. Johns Hopkins University Press, 642 pp.

Gu, C., 1989: RKPACK and its applications: Fitting smoothing spline models. *Proceedings of the Statistical Computing Section,* American Statistical Association, 42–51. [Code available through netlib.]

——, and G. Wahba, 1991: Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comput.,* **12,** 383–398.

Hoffman, R., 1984: SASS wind ambiguity removal by direct minimization. Part II. Effect of smoothing and dynamical constraints. *Mon. Wea. Rev.,* **112,** 1829–1852.

——, 1985: Using smoothness constraints in retrievals. *Advances in Remote Sensing Retrieval Methods,* A. Deepak, H. Fleming, and M. Chahine, Eds., A. Deepak Publishing, 411–436.

——, and J.-F. Louis, 1990: The influence of atmospheric stratification on scatterometer winds. *J. Geophys. Res.,* **95,** 9723–9730.

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

Hutchinson, M., 1989: A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Stat.-Simula.,* **18,** 1059–1076.

——, J. Kalma, and M. Johnson, 1984: Monthly estimates of windspeed and wind run for Australia. *J. Climatol.,* **4,** 311–324.

Kimeldorf, G., and G. Wahba, 1970: A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Stat.,* **41,** 495–502.

——, and ——, 1971: Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.,* **33,** 82–95.

Lewis, J., and J. Derber, 1985: The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus,* **37A,** 309–322.

Li, K. C., 1986: Asymptotic optimality of C_L and generalized cross validation in ridge regression with application to spline smoothing. *Ann. Stat.,* **14,** 1101–1112.

Li, Y., M. Navon, P. Courtier, and P. Gauthier, 1993: Variational data assimilation with a semi-Lagrangian semi-implicit global shallow-water equation model and its adjoint. *Mon. Wea. Rev.,* **121,** 1759–1769.

Liu, D., and J. Nocedal, 1989: On the limited memory BFGS method for large scale optimization. *Math. Progr.,* **45,** 503–528.

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

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

——, 1988: Optimal nonlinear objective analysis. *Quart. J. Roy. Meteor. Soc.,* **114,** 205–240.

——, 1992: Iterative analysis using covariance functions and filters. *Quart. J. Roy. Meteor. Soc.,* **118,** 569–591.

——, R. S. Bell, and B. MacPherson, 1991: The Meteorological Office analysis correction data assimilation scheme. *Quart. J. Roy. Meteor. Soc.,* **117,** 59–89.

Mitchell, H., C. Charette, C. Chouinard, and B. Brasnett, 1990: Revised interpolation statistics for the Canadian data assimilation procedure: Their derivation and application. *Mon. Wea. Rev.,* **118,** 1591–1614.

O'Sullivan, F., 1991: Sensitivity analysis for regularized estimation in some system identification problems. *SIAM J. Sci. Stat. Comput.,* **12,** 1266–1283.

——, and G. Wahba, 1985: A cross validated Bayesian retrieval algorithm for non-linear remote sensing. *J. Comput. Phys.,* **59,** 441–455.

Parrish, D., and J. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Rabier, F., P. Courtier, J. Pailleux, O. Talagrand, J. Thepaut, and D. Vasiljevic, 1993: A comparison between four dimensional variational assimilation and simplified sequential assimilation relying on three dimensional analysis. *Quart. J. Roy. Meteor. Soc.,* **119,** 845–880.

Speckman, P., 1985: Spline smoothing and optimal rates of convergence in nonparametric regression. *Ann. Stat.,* **13,** 970–983.

Wahba, G., 1981: Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Statist. Comput.,* **2,** 5–16.

——, 1982a: Erratum: Spline interpolation and smoothing on the sphere. *SIAM J. Sci. Stat. Comput.,* **3,** 385–386.

——, 1982b: Variational methods in simultaneous optimum interpolation and initialization. *The Interaction Between Objective Analysis and Initialization,* D. Williamson, Ed., Atmospheric Analysis and Prediction Division, National Center for Atmospheric Research, 178–185.

——, 1983: Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Stat. Soc. Ser. B,* **45,** 133–150.

——, 1985a: A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Stat.,* **13,** 1378–1402.

——, 1985b: Variational methods for multidimensional inverse problems. *Remote Sensing Retrieval Methods,* A. Deepak, H. Fleming, and M. Chahine, Eds., A. Deepak Publishing, 385–408.

——, 1987: Three topics in ill posed problems. *Proceedings of the Alpine-U.S. Seminar on Inverse and Ill Posed Problems,* H. Engl and C. Groetsch, Eds., Academic Press, 37–51.

——, 1990a: Regularization and cross validation methods for nonlinear, implicit, ill-posed inverse problems. *Geophysical Data Inversion Methods and Applications,* A. Vogel, C. Ofoegbu, R. Gorenflo, and B. Ursin, Eds., Vieweg, 3–13.

——, 1990b: *Spline Models for Observational Data,* SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, 165 pp.

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

——, and Y. Wang, 1990: When is the optimal regularization parameter insensitive to the choice of the loss function? *Commun. Stat.-Theory Meth.,* **19,** 1685–1700.

——, D. Johnson, F. Gao, and J. Gong, 1994: Adaptive tuning of numerical weather prediction models. Part I: Randomized GCV and related methods in three and four dimensional data assimilation, Tech. Rep. 920, Dept. of Statistics, University of Wisconsin, Madison, WI. [Available by ftp in the public directory ftp.stat.wisc.edu/pub/wahba in the file tuning-nwp.ps.gz, or, via the URL http://www.stat.wisc.edu/~wahba/.]

Zou, X., I. Navon, and F.-X. LeDimet, 1992: Incomplete observations and control of gravity waves in variational data assimilation. Part II: Applications and numerical results. *Tellus,* **44A,** 297–313.

——, ——, and J. Sela, 1993: Control of gravitational oscillations in variational data assimilation. *Mon. Wea. Rev.,* **121,** 272–289.