ADAPTIVE TUNING, 4D-VAR AND REPRESENTERS IN RKHS

Grace Wahba [1]
Department of Statistics
University of Wisconsin
Madison WI USA

Summary: We (abstractly) generalize the 'toy' weak 4D-Var model in Gong, Wahba, Johnson & Tribbia (1998) to include adaptive tuning of a variety of parameters throughout the 4D-Var variational problem, and note issues of sensitivity and identifiability. We discuss 'models' for model errors which include systematic, short memory and long memory errors. Finally we remark on the role of the theory of representers in reproducing kernel Hilbert spaces in the weak 4D-Var setting.

1 INTRODUCTION

We first consider the general setup in the experiment in Gong et al. (1998), which is a toy weak 4D-Var model (actually one time and one space variable) with five unknown smoothing, weighting and distributed parameters, which were simultaneously adaptively tuned using generalized cross validation (GCV) calculated via the randomized trace technique. In that setup 'model error' was generated as the difference between a 'nature' model and the 'computer' model, but white noise model errors were assumed in the weak 4D-Var variational problem. In this paper we then (i) review the use of model errors as dual variables, (ii) review the GCV and generalized maximum likelihood (GML) tuning methods, and pinpoint sensitivity issues as tunable parameters are sprinkled liberally throughout the weak 4D-Var problem, noting that they can be studied in the influence matrix (or influence operator in the nonlinear case). Then (iii) we describe some simple models for correlated model errors and the simultaneous consideration of systematic (bias), short memory and long memory correlation. We end with (iv) a summary of some representer theory in reproducing kernel Hilbert space (RKHS) relevant to the weak 4D-Var setting.

Let $t = 1, \cdots, T$ denote discrete time and let $\mathbf{\Psi}_t, t = 1, \cdots T$ be a sequence of state vectors representing (some part of) nature that evolves according to

$$\mathbf{\Psi}_{t+1} = \mathsf{M}_t \mathbf{\Psi}_t + \mathbf{N}_t + \xi_t, \quad t = 1, \cdots T-1, \tag{1}$$

where M_t is the model evolution operator, $\mathbf{N}_t$ is a forcing function, and the ξ_t represent model errors, which we will discuss in more detail later. Here $\mathbf{\Psi}_*$ is the forecast for $t = 1$ assumed to satisfy

$$\mathbf{\Psi}_1 = \mathbf{\Psi}_* + \epsilon_* \tag{2}$$

with $\epsilon_* \sim \mathcal{N}(0, \sigma_f^2 \mathsf{Q}_*)$. $\mathsf{M}_t = \mathsf{M}_t(\theta_\mathsf{M}), \mathbf{N}_t = \mathbf{N}_t(\theta_\mathsf{M})$ are assumed to contain some tunable distributed parameters θ_M. The observations are

$$\mathbf{y}_t = \mathsf{K}_t \mathbf{\Psi}_t + \epsilon_t, \quad t \in \Lambda, \tag{3}$$

with $\epsilon_t \sim \mathcal{N}(0, \sigma_o^2 \mathsf{S}_t)$. K_t is a map from state vector space to observation space at time $t \in \Lambda$, and $\mathsf{K}_t = \mathsf{K}_t(\theta_\mathsf{K})$ may also contain some tunable parameters, for example calibration coefficients or bias corrections. Here Λ is the set of observation times, which are assumed to be a subset of the model update times $t = 1, \cdots, T$.

In Gong et al. (1998) a toy weak 4D-Var problem was formulated as: Find $\hat{\mathbf{\Psi}} = (\hat{\mathbf{\Psi}}_1', \cdots, \hat{\mathbf{\Psi}}_T)'$, to minimize

$$\frac{1}{\sigma_o^2} \sum_{t \in \Lambda} \|\mathbf{y}_t - \mathsf{K}_t \mathbf{\Psi}_t\|^2_{\mathsf{S}_t^{-1}} + \frac{1}{\sigma_m^2} \sum_{t=1}^{T-1} \|\mathbf{\Psi}_{t+1} - \mathsf{M}_t(\theta_\mathsf{M}) \mathbf{\Psi}_t - \mathbf{N}_t(\theta_\mathsf{M})\|^2_{\mathsf{Q}_t^{-1}} + \frac{1}{\sigma_f^2} \|\mathbf{\Psi}_* - \mathbf{\Psi}_1\|^2_{\mathsf{Q}_*^{-1}} + \frac{1}{b} \|\mathbf{\Psi}_T\|^2_\mathsf{J}. \tag{4}$$

[1]Research supported in part by NASA Grant NAG5-3769 and NSF Grant DMS9121003

Here $\|\mathbf{v}\|_{\mathsf{C}}^2 = \mathbf{v}'\mathsf{C}\mathbf{v}$ for C a non-negative definite matrix. This formulation corresponds to $\xi_t \sim \mathcal{N}(0, \sigma_m^2 \mathsf{Q}_t)$, independent from time to time, and a prior belief that $\|\mathbf{\Psi}_T\|_{\mathsf{J}}^2$ is 'small' where J is a quadratic penalty representing a toy version of e. g. a penalty for lack of balance. Letting $\gamma = \sigma_o^2/\sigma_f^2, \alpha = \sigma_o^2/\sigma_m^2$ and $\eta = \sigma_o^2/b$, then the minimizer of (4) is the same as the minimizer of

$$\begin{aligned} & \sum_{t\in\Lambda} \|\mathbf{y}_t - \mathsf{K}_t\mathbf{\Psi}_t\|^2_{\mathsf{S}_t^{-1}} + \alpha \sum_{t=1}^{T-1} \|\mathbf{\Psi}_{t+1} - \mathsf{M}_t(\theta_{\mathsf{M}})\mathbf{\Psi}_t - \mathbf{N}_t(\theta_{\mathsf{M}})\|^2_{\mathsf{Q}_t^{-1}} + \gamma\|\mathbf{\Psi}_* - \mathbf{\Psi}_1\|^2_{\mathsf{Q}_*^{-1}} + \eta\|\mathbf{\Psi}_T\|^2_{\mathsf{J}} \\ = \; & J_o + J_m + J_b + J_c, \end{aligned} \tag{5}$$

say. The J terms have there usual meaning as observation, model, forecast and constraint except that the coefficients in front have been scaled relative to observations. The major tuning parameters $\{\gamma, \alpha, \eta\}$ and a two coefficient distributed parameter θ_{M} in the model were simultaneously tuned by the GCV method, using the randomized trace technique, which allows the computation of the cross-validation function to be carried out by rerunning the model with perturbed data. 'Nature' was simulated using the barotropic vorticity equation on a latitude circle, solved with a high order method on a fine time and space grid, and noisy observations were generated from nature using a random number generator. S_t was taken as I both in generating the data and in analyzing it via (5), and ϵ_* was generated as a zero mean random Gaussian vector with covariance a multiple of Q_*. 'The model' was based on a cruder integration of the barotropic vorticity equation, so that 'model error' may be thought of as the difference between 'nature' and 'the model'. In this case, as in nature, model error is not readily describable in terms of means and covariances, nevertheless, they may be a convenient, although crude way of dealing with model error that is not well understood. In the experiment Q_t was taken as the identity matrix. The five tuning parameters were selected for adaptive tuning via the GCV method because it was believed, as a result of some preliminary experiments, (plus guesswork) that the solution and the 'predicted' observations computed from the solution, were sensitive to them. It turned out that the predictive mean square error, based on $\mathbf{y}_{true} - \mathbf{y}_{fitted}$, where $\mathbf{y}_{true}$ is what would have been observed if there were no errors anywhere, (known only in a simulation, of course), was sensitive to all five parameters, but, on comparable scales, α, the parameter relating observational to model error, had a much broader, flat minimum. One possible explanation is that the 'white noise' assumption for model error was not a very good representation for model error. Although the sensitivity to α was not great, the weak constraint estimate gave better results than the strong constraint ($\alpha \to \infty$). Dealing with model error is an open scientific issue, according to Courtier (1997) and others. Before going on to some speculative discussion of approaches to model error in the weak 4D-Var problem, we note that it was clear in the experiments in Gong et al. (1998) that the five parameters being tuned interacted with one another. For example the optimal value of one of the physical parameters in θ_{M} was systematically larger than the 'true' or 'nature' value but it depended on the choice of (γ, α, η).

With regard to θ_{M}, in the present study, θ_{M} contributed essentially two degrees of freedom to the fit. In practice, θ_{M} may be widely distributed. If it contributes many degrees of freedom, then in general it will be appropriate to include a (tunable) penalty term, say $\delta\|\theta_{\mathsf{M}}\|^2_{\mathsf{D}}$ to the variational problem, see Wahba (1990*a*), O'Sullivan (1991), Navon (1998), Evensen, Dee & Schroter (1998).

2 DUAL VARIABLES, NONLINEAR FORWARD OPERATORS AND MODELS, CORRELATED MODEL ERRORS

Following Bennett (1997), Courtier (1997), notice that, assuming that $\mathbf{\Psi}_*$ and the $\mathbf{N}_t$ known, that $\mathbf{\Psi}_1, \cdots, \mathbf{\Psi}_T$ are determined by ϵ_* and $\xi = (\xi_1', \cdots, \xi_{T-1})'$ and vice versa. Thus one may change

variables from $\Psi_1, \cdots, \Psi_T$ in (5) to (ϵ_*, ξ) to get

$$\sum_{t\in\Lambda} \|\mathbf{y}_t - \mathsf{K}_t\Psi_t(\epsilon_*,\xi)\|^2_{\mathsf{S}_t^{-1}} + \alpha \sum_{t=1}^{T-1} \|\xi_t\|^2_{\mathsf{Q}_t^{-1}} + \gamma\|\epsilon_*\|^2_{\mathsf{Q}_*^{-1}} + \eta\|\Psi_T(\epsilon_*,\xi)\|^2_{\mathsf{J}}, \tag{6}$$

and solve the variational problem for $\epsilon*$ and ξ instead of for Ψ. To take a closer look at this problem, let $\tilde{\mathsf{K}}_t = \mathsf{K}_t\Psi_t$, $\mathbf{x} = (\epsilon_*', \xi')'$, and redesign J_c so that it is is quadratic in $\mathbf{x}$, letting $J_c(\mathbf{x}) = \|\mathbf{x}\|^2_{\mathsf{J}}$, $\mathsf{J} = \mathsf{J}(\theta_c)$. Concatenate the $\mathbf{y}_t, \tilde{\mathsf{K}}_t, \mathsf{S}_t$ and Q_t in an obvious way, and allow more tuning parameters to get $\tilde{\mathsf{K}}(\theta_{\mathsf{K}})$, $\mathsf{S} = \mathsf{S}(\theta_o)$ and $\mathsf{Q} = \mathsf{Q}(\theta_\xi)$, and let $\mathsf{Q}_* = \mathsf{Q}_*(\theta_b)$. Furthermore, we specifically do not want to restrict Q to be block diagonal, so that we can allow for model errors correlated from time to time. Let $\lambda^{-1}\Sigma = \lambda^{-1}\Sigma(\theta_\Sigma)$ be a quadratic form standing in collectively for $\alpha\mathsf{Q}, \gamma\mathsf{Q}_*$ and $\eta\mathsf{J}$. The result is the variational problem

$$(\mathbf{y} - \tilde{\mathsf{K}}\mathbf{x})'\mathsf{S}^{-1}(\mathbf{y} - \tilde{\mathsf{K}}\mathbf{x}) + \lambda\mathbf{x}'\Sigma^{-1}\mathbf{x}, \tag{7}$$

with tuning parameters $\theta = (\theta_{\mathsf{K}}, \theta_{\mathsf{M}}, \theta_o, \theta_\xi, \theta_b, \theta_c)$. We note that (7)is not changed if the K_t and M_t are nonlinear, in that case $\tilde{\mathsf{K}}$ is a nonlinear map from $\mathbf{x}$ to $\mathbf{y}$, but under the assumption about J_c, the second term is quadratic.

If $\tilde{\mathsf{K}} = \mathsf{K}$, meaning $\tilde{\mathsf{K}}$ is linear, then the minimizer $\mathbf{x}_\lambda$ of (7) is[2]

$$\mathbf{x}_\lambda = (\mathsf{K}'\mathsf{S}^{-1}\mathsf{K} + \lambda\Sigma^{-1})^{-1}\mathsf{K}'\mathsf{S}^{-1}y = \Sigma\mathsf{K}'(\mathsf{K}\Sigma\mathsf{K}' + \lambda\mathsf{S})^{-1}\mathbf{y} = \sum c_i\eta_i, \tag{8}$$

where η_i is the *ith* column of $\Sigma\mathsf{K}'$, and, letting $\mathbf{c} = (c_1, \cdots, c_n)'$, $(\mathsf{K}\Sigma\mathsf{K}' + \lambda\mathsf{S})\mathbf{c} = \mathbf{y}$. This is a trivial example of representer theory (implemented , for example in PSAS, Cohn, daSilva, Guo, Sienkiewicz & Lamich (1998)), where a system of size the dimension of $\mathbf{y}$ is to be solved even if the dimension of $\mathbf{x}_\lambda$ is much bigger than the dimension of $\mathbf{y}$.

3 TUNING METHODS

With the dimension of $\mathbf{x}$ orders of magnitude greater than the dimension of $\mathbf{y}$, it is clear that the number of tunable parameters in $\tilde{\mathsf{K}}, \mathsf{S}$ and Σ is limited by the amount of information in $\mathbf{y}$, and by the possibility of aliasing/identifiability. (Of course to the extent that desirable values of these parameters do not vary over long periods of time, historical information may be collected). In any case, it is neither possible, nor even desirable to have models for the model error covariance matrix Q, the forecast error matrix Q_* or the constraint functional J_c to have an overabundance of free parameters. Furthermore, the solution should be sensitive to any parameters considered for adaptive tuning.

Equation (7) may be interpreted as the variational problem associated with the statistical assumptions

$$\mathbf{y} = \mathsf{K}\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_o^2\mathsf{S}), \quad \mathbf{x} \sim \mathcal{N}(0, b\Sigma), \lambda = \sigma_o^2/b. \tag{9}$$

The influence matrix $\mathsf{A}(\lambda, \theta)$, which maps $\mathsf{S}^{-1/2}\mathbf{y}$ into $\mathsf{S}^{-1/2}\mathsf{K}\mathbf{x}_\lambda$ [3] is given by

$$\mathsf{A} = \mathsf{S}^{-1/2}\mathsf{K}(\mathsf{K}'\mathsf{S}^{-1}\mathsf{K} + \lambda\Sigma^{-1})^{-1}\mathsf{K}'\mathsf{S}^{-1/2} = \mathsf{B}(\mathsf{B} + I)^{-1}, \tag{10}$$

[2]This formula assumes that the quadratic form $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ is of full rank. If it is not, see Kimeldorf & Wahba (1971), Wahba (1990*b*). Having already abused notation, replace $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ by $\mathbf{x}'\mathsf{P}\mathbf{x}$. For (7) to have a unique minimizer in the linear case it is necessary and sufficient that $\mathbf{x}'P\mathbf{x} = 0$ and $\mathsf{K}\mathbf{x} = 0$ imply that $\mathbf{x} = 0$. In practice users should make sure that the null space of P is not too big!

[3]In the nonlinear case use this to define the influence operator.

where $\mathsf{B} = \mathsf{B}(\lambda,\theta) = \frac{1}{\lambda}\mathsf{S}^{-1/2}\mathsf{K}\Sigma\mathsf{K}'\mathsf{S}^{-1/2}$. Its clear that B must be sensitive to a component of θ in order for estimation of that component to make sense and aliasing of parameters inside B is to be avoided [4]. The GML estimate for λ,θ is the minimizer of

$$M(\lambda,\theta) = \frac{y'\mathsf{S}^{-1/2}(\mathsf{I}-\mathsf{A}(\lambda,\theta))\mathsf{S}^{-1/2}y}{[det\mathsf{S}^{-1}]^{1/n}[det(\mathsf{I}-\mathsf{A}(\lambda,\theta))]^{1/n}}. \tag{11}$$

and the GML estimate for σ_o^2 is $\frac{1}{n}\mathbf{y}'\mathsf{S}^{-1/2}(\mathsf{I}-\mathsf{A})\mathsf{S}^{-1/2}\mathbf{y}$. The GML estimate has certain optimality properties when the stochastic model (9) is correct up to the unknown parameters. Little is known concerning its robustness to this assumption, see Wahba (1985). Note that there are other forms of maximum likelihood estimates , depending on which unknowns are included and how they enter into the formulas. The GCV estimate is the minimizer of

$$V(\lambda,\theta) = \frac{\|(\mathsf{I}-\mathsf{A})\mathsf{S}^{-1/2}\mathbf{y}\|^2}{(\frac{1}{n}trace(\mathsf{I}-\mathsf{A}))^2} \equiv \frac{\|y-\mathsf{K}\mathbf{x}_\lambda\|^2_{\mathsf{S}^{-1}}}{(\frac{1}{n}trace(\mathsf{I}-\mathsf{A}))^2}. \tag{12}$$

In theory it may not be suitable for estimating sensitive parameters inside S (i.e. θ_o), since it is theoretically based on assuming that the problem is being scaled so that $\mathsf{S}^{-1/2}\epsilon \sim \mathcal{N}(0,\sigma_0^2\mathsf{I})$, where S is assumed reasonably correct. Subsets of observations where this is not true, (for example radiance data), may be excluded from V by partial GCV, see Wahba, Johnson, Gao & Gong (1994), there $(\mathsf{I}-\mathsf{A})$ is replaced by $\mathsf{E}(\mathsf{I}-\mathsf{A})$ where E is a possibly weighted indicator matrix for the observations to be included. However the GCV estimate is robust to various assumptions about K,Σ and $\mathbf{x}$. We remark that both the GCV and GML estimate can be defined when K is nonlinear. In either the linear or nonlinear case the trace of A may be estimated by the randomized trace method without having A explicitly, given a 'black box' which produces $\mathsf{K}\mathbf{x}_\lambda$ given y, see Wahba, Johnson, Gao & Gong (1995). Dee & daSilva (1998) and Dee, Gaspari, Redder, Rukhovets & daSilva (1998) have used maximum likelihood methods to estimate parameters in forecast error covariances in several practical examples, and have compared some of the results with GCV estimates, obtaining generally similar results in the examples tried. It may be possible to combine the strengths of both methods by iterating back and forth, using likelihood methods for parameters in S and GCV for parameters in K and Σ, this is speculative at the moment. The ordinary cross-validation 'warhorse' of leaving out a subset of the observations may also be used (with care).

4 MODELS FOR MODEL ERROR, DISCRETE TIME CASE

We outline some classes of models for time dependent model error, $t = 1,\cdots,T$. Let g be a generic index, $g = (lat,\ long,\ z,\ type)$ where *type* indexes the analysis variables, i. e. *type* = surface temperature, *type* = vorticity, etc. See Wahba (1992) for more on generic indices. We list some stochastic models for $\xi_t(g)$ which remove the restriction that $E\xi_t = 0, E\xi_s'\xi_t = 0, s \neq t$. A fairly general class of models is

$$\xi_t(g) = \mu_t(g) + \sum_{k=1}^{\infty} z_k(t)\sqrt{\lambda_k(t)}\Phi_k(g) \tag{13}$$

where $\mu_t(g) = \sum_{\nu=1}^{M} d_\nu F_\nu(t,g)$ is a mean function (bias term) specified except for a modest number of coefficients $d_\nu, \nu = 1,\cdots,M$, to be found, the λ_k and Φ_k are specified up to some parameters θ_ξ, and $Ez_k(t) = 0, Ez_k(s)z_l(t) = r_{k,l}(s,t)$. Then

$$cov\xi_s(g)\xi_t(h) = \sum_{k,l} r_{k,l}(s,t)\sqrt{\lambda_k(s)}\sqrt{\lambda_l(t)}\Phi_k(g)\Phi_l(h). \tag{14}$$

[4] e. g. In principle at least, the Hessian of B with respect to (λ,θ) should be well conditioned.

The simplest generalization over model error independent from time to time is the tensor product case, $r_{k,l}(s,t) = 0, k \neq l, r_{k,k} = r(s,t)$, independent of k, and $\lambda_k(s) = \lambda_k$ independent of s, which gives

$$E\xi_s(g)\xi_t(h) = r(s,t)\mathsf{R}(g,h), \tag{15}$$

where $\mathsf{R}(g,h) = \sum_k \lambda_k \Phi_k(g)\Phi_k(h)$. This gives the penalty term for model error as

$$\sum_{s,t=1}^{T-1} r^{st}(\xi_s - \mu_s)'\mathsf{R}^{-1}(\xi_t - \mu_t), \tag{16}$$

where r^{st} is the (s,t)th entry of the inverse of the matrix with (s,t) entry $r(s,t)$. Here the coefficients in μ_t become part of the variational problem. Simple examples include $z(\cdot)$ an autoregressive scheme or moving average where the correlation structure of the $z(\cdot)$ process can be defined as short or long memory. A model like

$$E\xi_s(g)\xi_t(h) = \sum_\alpha r_\alpha(s,t)\mathsf{R}_\alpha(g,h) \tag{17}$$

would allow for different time scales in the r_α. Luo, Wahba & Johnson (1998) consider (15) in a simple situation with g on the sphere, R an isotropic covariance on the sphere and $z(\cdot)$ a second order difference scheme forced by white noise. They used this model to estimate the linear time trend as a function of space given historical data irregular in time and space, by using the fact that the time trend is obtained as an orthogonal projection of the fit onto the relevant subspace of the implied RKHS. It may be possible to use similar techniques to diagnose model error. Griffith & Nichols (1998) have recently examined some simplified dynamical models for model error.

5 ELEMENTS OF REPRESENTER THEORY

In this section, time is continuous, $t \in [0,1]$, and some components of g (e. g. space variables) are also to be thought of as continuous. Continuous time representer theory in RKHS has recently been applied in a number of places, see Bennett (1992), Bennett (1997), Bennett, Chua & Leslie (1996), Eknes & Evensen (1997), Evensen et al. (1998), Amodei (1997), Wahba (1992). We remind the reader that for every positive definite function R on $\mathcal{T} \otimes \mathcal{T}$, where $\mathcal{T}$ is an abstract index set, there exists a unique RKHS and vice versa (The Moore-Aronszajn Theorem). There also exists a well defined zero mean Gaussian stochastic process with R as its covariance, however, sample functions of the stochastic process are not, with probability 1 in the RKHS if the RKHS is infinite dimensional. See Wahba (1990*b*), Weinert (1982) for more on reproducing kernel Hilbert spaces.

Let

$$\xi_t(g) = \xi_t^0(g) + \xi_t^1(g), \tag{18}$$

where

$$\mathcal{L}\xi_t^0(g) = 0, \quad \mathcal{B}\xi_t^1(g) = 0, \tag{19}$$

where $\mathcal{L}_t$ is a (linear) evolution operator, and $\mathcal{B}$ are initial/boundary conditions which serve to make the solution of differential equation unique, so that

$$\xi_t(g) = \xi_t^0(g) + \int G(t,g;\tilde{t},\tilde{g})u(\tilde{t},\tilde{g})d\tilde{t}d\tilde{g} \tag{20}$$

for some $u(\cdot)$ where G is the Green's function for $\mathcal{L}$ and $\mathcal{B}$. If u is treated as though is is a zero mean Gaussian stochastic process with covariance $R_u(s,g;t,h)$ then

$$E\xi_s^1(g)\xi_t^1(h) = \int\int G(s,g;\tilde{s},\tilde{g})G(t,h;\tilde{t},\tilde{h})R_u(\tilde{s},\tilde{g};\tilde{t},\tilde{h})d\tilde{s}d\tilde{g}d\tilde{t}d\tilde{h} = R^1(s,g;t,h), \quad [say]. \tag{21}$$

Let

$$E\xi_s^0(g)\xi_t^0(h) = R^0(s,g;t,h), \tag{22}$$

and suppose that ξ_s^0 and ξ_s^1 are independent, then

$$E\xi_s(g)\xi_t(h) \equiv R(s,g;t,h) = R^0(s,g;t,h) + R^1(s,g;t,h) \equiv R^0_{\theta_b}(s,g;t,h) + R^1_{\theta_M}(s,g;t,h). \tag{23}$$

Under general circumstances $\mathcal{L}R^0(s,g;,\cdot,\cdot) = 0$, where $\mathcal{L}$ is applied to R^0 considered as a function of $(\cdot,\cdot)$ for each fixed (s,g). Furthermore, the RKHS $\mathcal{H}_R$ with reproducing kernel (RK) given by $R = R^0 + R^1$ of (23) consists of the direct sum of the orthogonal subspaces $\mathcal{H}_R = \mathcal{H}_{R^0} \oplus \mathcal{H}_{R^1}$, respectively containing solutions of the homogeneous equation and solutions to the differential equation satisfying homogeneous boundary conditions. Changing notation from $\xi_t(g)$ to $f_t(g)$ to indicate that we are now letting f be an element of $\mathcal{H}_R$, we have that if $f^1 \in \mathcal{H}_{R^1}$, then $\|f^1\|^2_{\mathcal{H}_{R^1}} = \|\mathcal{L}f^1\|^2_{\mathcal{H}_{R_u}}$ where $\|\cdot\|^2_{\mathcal{H}_{R_u}}$ is the square norm in $\mathcal{H}_{R_u}$. If u had instead been taken as 'white noise' then R_u would not appear in (21) and the $\mathcal{H}_{R_u}$ norm would be replaced with the usual L_2 norm. Decomposing f into $f^0$and f^1 analogous to (18,19) gives (the obvious) $\|f\|^2_{\mathcal{H}_R} = \|f^0\|^2_{\mathcal{H}_{R^0}} + \|f^1\|^2_{\mathcal{H}_{R^1}}$

Let $L_1, \cdots, L_n$ be n bounded linear functionals on $\mathcal{H}_R$. Basic representer theory (see Kimeldorf & Wahba (1971), Wahba (1990*b*)) in RKHS tells us that the solution to the problem: find $f \in \mathcal{H}_R$ to minimize

$$\sum_{i=1}^{n}(y_i - L_i f)^2 + \lambda\|f\|^2_{\mathcal{H}_R} \tag{24}$$

is in the span of the n representers η_i of L_i in $\mathcal{H}_R$, where

$$\eta_i(s,g) = L_{i(t,h)}R(s,g;t,h), \tag{25}$$

where $L_{i(t,h)}$ means L_i applied to what follows considered as a function of (t,h). We may replace $\|f\|^2_{\mathcal{H}_R}$ in (24)by e. g. $\|f^0\|^2_{\mathcal{H}_{R^0}} + w\|f^1\|^2_{\mathcal{H}_{R^1}}$, the new problem is in theory solved with the aid of the RK $R^0(s,g;t,h) + w^{-1}R^1(s,g;t,h)$; $w \to \infty$ corresponds to the 'perfect model' assumption. Simple prototypes appear in Kimeldorf & Wahba (1971) and Wahba (1990*b*) where a seminorm penalty is also allowed.

Tuning for model error may have potential to provide diagnostic information concerning model error. Tony Weaver (personal communication) has remarked on the necessity of considering extrapolating (forecasting) the correlated part of any fit to model error. Most desirable, of course is to eliminate model error to the extent possible.

Recently Lin, Wahba, Xiang, Gao, Klein & Klein (1998), in a different (and much simpler) context, but with a relatively large, irregularly spaced data set, solved the variational problem under consideration in the span of a selected subset of the representers $\eta_i, i = 1, \cdots, n$, with excellent results.

6 ACKNOWLEDGMENTS

We thank Andrew Bennett for generously providing advance copies of much of his recent work, and Phillipe Courtier for providing talk notes. We thank Tony Hollingsworth, Francois Bouttier, Heiki Jarvinen, Mike Fisher, Tony Weaver, Olivier Talagrand, Nancy Nichols and Gene Golub for helpful conversations and remarks.

References

Amodei, L. (1997), Reproducing kernels of vector-valued function spaces, *in* A. LeMehaute, C. Rabut & L. Schumaker, eds, 'Surface Fitting and Multiresolution Methods', Vanderbilt University Press, Nashville TN, pp. 17–26.

Bennett, A. (1992), *Inverse methods in physical oceanography*, Cambridge University Press, 346pp.

Bennett, A. (1997), 'Inverse methods and data assimilation', College of Oceanic And Atmospheric Sciences, Corvallis OR. Summer School Lecture Notes.

Bennett, A., Chua, B. & Leslie, L. (1996), 'Generalized inversion of a global weather prediction model', *Meteorology and Atmospheric Physics* **60**, 165–178.

Cohn, S., daSilva, A., Guo, J., Sienkiewicz, M. & Lamich, D. (1998), 'Assessing the effects of data selection with the DAO physical-space statistical analysis system', *Monthly Weather Review* **126**, 2913–2926.

Courtier, P. (1997), 'Dual formulation of four-dimensional variational assimilation', *Q. J. R. Meterol. Soc.* **123**, 2449–2461.

Dee, D. & daSilva, A. (1998), 'Maximum-likelihood estimation of forecast and observation error covariance parameters. part i: Methodology', *Monthly Weather Review* **in press**, xx–xx.

Dee, D., Gaspari, G., Redder, C., Rukhovets, L. & daSilva, A. (1998), 'Maximum-likelihood estimation of forecast and observation error covariance parameters. part ii: Applications', *Monthly Weather Review* **in press**, xx–xx.

Eknes, M. & Evensen, G. (1997), 'Parameter estimation solving a weak constraint variational formulation for an Ekman model', *J. Geophysical Research* **102(C6)**, 479–491.

Evensen, G., Dee, D. & Schroter, J. (1998), Parameter estimation in dynamical models, *in* E. Chassignet & J. Verron, eds, 'Ocean Modeling and Parameterizations', NATO ASI, **in press**, xx-xx.

Gong, J., Wahba, G., Johnson, D. & Tribbia, J. (1998), 'Adaptive tuning of numerical weather prediction models: simultaneous estimation of weighting, smoothing and physical parameters', *Monthly Weather Review* **125**, 210–231.

Griffith, A. & Nichols, N. (1998), 'Adjoint methods for treating model error in data assimilation', *Proc. ICFD Conference on Numerical Methods in Fluid Dynamics* **in press**, xx-xx.

Kimeldorf, G. & Wahba, G. (1971), 'Some results on Tchebycheffian spline functions', *J. Math. Anal. Applic.* **33**, 82–95.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (1998), Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV, Technical Report 998, Department of Statistics, University of Wisconsin, Madison WI.

Luo, Z., Wahba, G. & Johnson, D. (1998), 'Spatial-temporal analysis of temperature using smoothing spline ANOVA', *J. Climate* **11**, 18–28.

Navon, I. (1998), 'Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography', *Dynamics of Atmospheres and Oceans* **27**, 55–79.

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

Wahba, G. (1985), 'A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem', *Ann. Statist.* **13**, 1378–1402.

Wahba, G. (1990*a*), Regularization and cross validation methods for nonlinear, implicit, ill-posed inverse problems, *in* A. Vogel, C. Ofoegbu, R. Gorenflo & B. Ursin, eds, 'Geophysical Data Inversion Methods and Applications', Vieweg, Weisbaden-Braunschweig, pp. 3–13.

Wahba, G. (1990*b*), *Spline Models for Observational Data*, SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

Wahba, G. (1992), Multivariate function and operator estimation, based on smoothing splines and reproducing kernels, *in* M. Casdagli & S. Eubank, eds, 'Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII', Addison-Wesley, pp. 95–112.

Wahba, G., Johnson, D., Gao, F. & Gong, J. (1994), Adaptive tuning of numerical weather prediction models: Part I, randomized GCV and related methods in three and four dimensional data assimilation, Technical Report 920, Department of Statistics, University of Wisconsin, Madison, WI.

Wahba, G., Johnson, D., Gao, F. & Gong, J. (1995), 'Adaptive tuning of numerical weather prediction models: randomized GCV in three and four dimensional data assimilation', *Mon. Wea. Rev.* **123**, 3358–3369.

Weinert, H., ed. (1982), *Reproducing kernel Hilbert spaces: Application in signal processing*, Hutchinson Ross, Stroudsburg, PA.