# An Entropy Approach to Tuning Weights and Smoothing in the Generalized Inversion

GENNADY A. KIVMAN, ALEXANDRE L. KURAPOV, AND ALINA V. GUESSEN

*St. Petersburg Branch, P. P. Shirshov Institute of Oceanology, Russian Academy of Sciences, St. Petersburg, Russia*

(Manuscript received 20 April 1999, in final form 17 April 2000)

ABSTRACT

Weak constraint data assimilation involves a certain number of weighting and smoothing parameters. The authors present an approach to estimate them based on maximizing the entropy. Because application of this rigorous scheme to large-dimensional data assimilation problems is a tedious task, the authors also consider a simplified version of the entropy method, which assumes maximizing a data cost as a function of relative data weights. It is proven to be equivalent to maximizing the entropy under certain assumptions. In the scope of this method, the authors have also proposed a smoothing procedure necessary for very fine grids. The schemes have been checked using a tidal channel model for Tatarsky Strait.

## 1. Introduction

The generalized inversion (GI) is the most flexible tool for combining information on the state of the ocean provided by physical models and observations (Egbert and Bennett 1996; Kivman 1997a). The approach explicitly takes into account uncertainties inherent in both model equations and measurements; this is achieved by the proper weighting of corresponding cost terms. An assumption that the model equations are perfect has been widely used in data assimilation. Though accounting for the model errors allows one to improve accuracy of the prediction (see, e.g., Evensen 1997; Gong et al. 1998), the question of what the model errors are remains an open scientific issue (Courtier 1997). The main goal of this paper is to show that as soon as we combine information from different sources into a unique solution, methods of the information theory are applicable and may be of help in selecting a number of error statistics parameters.

From a probabilistic standpoint on data assimilation (van Leeuwen and Evensen 1996), a system state $\psi$ and measurement results $d_1, \ldots, d_M$ are viewed as random quantities from a set $\Psi$ of admissible solutions and a data space $D$, respectively. Given a prior probability density function (PDF) $p(\psi)$ and the data $\mathbf{d} = (d_1, \ldots, d_M)$, with errors described by a conditional PDF $p(\mathbf{d}\,|\,\psi)$, we update our knowledge about $\psi$ using the Bayes theorem

$$p(\psi\,|\,\mathbf{d}) = \frac{p(\mathbf{d}\,|\,\psi)p(\psi)}{\int_{\Psi} p(\mathbf{d}\,|\,\psi)p(\psi)\Pi d\psi}. \tag{1}$$

Then, the objective of data assimilation is to seek a maximum probable state $\psi_i$ when $\mathbf{d}$ is observed.

The problem is that we know little about the prior $p(\psi)$. As soon as $\psi_i$ becomes heavily dependent on $p(\psi)$, tuning statistical parameters of the prior raises considerable interest. A prerequisite in prescribing the prior is the mathematical expectation or the mean

$$E(\psi) = \int_{\Psi} \psi \, dp(\psi). \tag{2}$$

In oceanography, the amount of data is typically much less than the number of unknown variables, and we must look for additional prior information to construct $E(\psi)$. A useful way is to invoke a physical model for $\psi$; however, model uncertainties are rarely estimated a priori. Since errors in model equations are never observed directly, their statistical properties could be derived only from the data. To minimize the effects of measurement noise, the number of observations should be at least an order of magnitude larger than the dimension of the system state space, which is rarely the case. To close an inverse problem, we should adopt assumptions about the model uncertainties. Further, physical laws underlying models are usually expressed as differential equations. Initial and/or boundary conditions for them can

*Corresponding author address:* Dr. Gennady Kivman, Alfred Wegener Institute for Polar and Marine Research, Columbusstrasse Postfach 120161, 27515 Bremerhaven, Germany.
E-mail: gkivman@awi_bremerhaven.de

come only from observations. Thus, even if the probability distribution of the model errors was known, it would not be enough for defining the prior. We have to guess the statistics of $\psi$ at the initial moment and at the boundary of the modeled domain.

Though there have been some attempts to estimate systematic model errors (Dee and da Silva 1998; Griffith and Nichols 1996), in most studies, the errors in the model equations and initial and boundary conditions have been taken to be unbiased and Gaussian distributed. That is, in the linear case, the mean $E(\psi)$ is assumed to be given by the model solution satisfying the initial and boundary data exactly (Bennett 1992). The initial and boundary conditions must be obtained from observations. Thus, the data obtained at the initial time and/or at the boundary of the domain are considered to be equal to the mean of the prior $E(\psi)$, while the data assimilated inside the domain do not possess that exclusive property. It is a rather general situation that there are not enough observations at the intial moment or at the boundary; then, we should introduce bogus data to get a well-posed problem (Bennett and Miller 1991). We work with the two datasets that get different interpretations: the data at the initial time and/or at the boundary of the domain are considered to be equal to the mean $E(\psi)$, while the data assimilated inside the domain are not. Partitioning the data into the two subsets is arbitrary to some extent; thus, essential asymmetry in the interpretation of these two sets is introduced.

Having few observations, we cannot estimate everything and should restrict our attention to tuning a certain number of statistical parameters, such as error variances and error decorrelation scales, while keeping other inputs fixed. Two methods have been widely used in data assimilation community, namely, maximum likelihood (ML; Dee 1995) and general cross-validation (GCV; Egbert et al. 1994). As the statistical theory suggests (Cramer 1954), if the true prior PDF is indeed of an assumed functional shape, then ML is the most powerful tool for estimating parameters of the prior PDF from the data in the sense of minimizing variances of the estimates. However, if the assumptions about $p(\psi)$ are violated, this approach may be less effective than some other tuning criteria (Wahba 1985). Gao (1994) demonstrated that in a particular example, the ML substantially overestimated a parameter of smoothing (error decorrelation length scale) in comparison to the GCV. When even the mean $E(\psi)$ is not known, the ML estimates of the statistical parameters are suspect.

Another approach to building the prior PDF may be based on the principle of maximum entropy (PME). The PME and Bayesian viewpoints complement each other (Jaynes 1988a). Indeed, the former assesses the prior probabilities in the most cautious way in the sense that the PDF obtained assumes nothing beyond what was actually known and used as input. Once the prior PDF is chosen, the Bayes method updates it when new data become available. It is worth emphasizing that probabilities derived from the PME should be considered as ''degrees of belief'' rather than ''relative frequencies'' (Jaynes 1988b); it has been precisely the interpretation of probabilities used by Tarantola (1987) in the statistical inverse problems theory. Thacker and Holloway (1990) pointed out that the PME might be of help in choosing the prior in oceanographic inverse problems.

We outline the fundamentals of the PME in section 2. Then, we show that in the linear case, any choice of the prior generates a measure in the space of possible outcomes, which may be viewed as a generalization of the probability measures. Thus, one is able to use the PME to select a certain number of statistical parameters (see section 3 for the theory and section 4 for a numerical example). Any method for tuning the statistical parameters substantially complicates data assimilation and is very difficult to employ in realistic large-dimensional problems. The maximum data cost criterion (MDC), which is easier to implement, is shown to be an approximation to the PME. Though the MDC established on heuristic grounds has worked well (Kivman 1997c; Kurapov and Kivman 1999), its statistical background is clarified only now (see section 5).

The cost functional generally includes certain smoothing terms, which are introduced to provide well-posedness. It was demonstrated by Gong et al. (1998) that they could be as important as the weight controlling the trade-off between the model and the data. We will show in section 6 how to tune a smoothing parameter in the scope of the PME and the MDC.

## 2. Maximum entropy formalism

What can one conclude about $p(\psi|\mathbf{d})$ having a number of constraints on it? This question is as old as statistical mechanics, and an answer was provided by Boltzmann (1964) and Gibbs (1902). Namely, one should maximize entropy

$$H(p) = -\int_{\Psi} p(\psi|\mathbf{d}) \ln\frac{p(\psi|\mathbf{d})}{\mu(\psi)}\Pi \, d\psi, \qquad (3)$$

subject to the constraints imposed. Here, $\mu(\psi)$ is the so-called noninformative PDF that expresses the state of ''lowest'' information about $\psi$ (Tarantola 1987).

The concept of entropy originated about a century ago within the context of a fundamental but specific physical problem; it has recently found widespread application in dynamical systems theory, ergodicity theory, inverse problems, communication theory, numerical analysis, theory of functions, decision theory, etc., that is, to problems staying far from the classical thermodynamics. Thanks to Shannon (1948), all these applications were made possible only after recognizing the fact that entropy has much more wider content than had been previously thought. To be concise, Shannon proved that any measure of uncertainty in a probability distribution is proportional to the Boltzmann–Gibbs entropy if that measure is

(i) a continuous function of probabilities;
(ii) a decreasing function of the number of possible events in the case when they are equally probable; and
(iii) consistent, that is, if there is more than one way of evaluating its value, every possible way yields the same result.

Thus, the PME is in fact a refined version of Bernoulli's principle of insufficient reason for assigning probabilities to possible events in the light of all prior knowledge (see Thacker and Holloway 1990, and references therein). It should be emphasized that the PME is an extremely conservative procedure in the sense that its application yields a probability distribution that is as uncertain or ''spread'' as allowed by properties we wish the distribution to have and that possesses a minimal information content among all distributions satisfying the constraints. In other words, the maximum entropy distribution is free from any information additional to that we actually knew and used as input.

## 3. An entropy approach to selecting priors in weak constraint data assimilation

We will treat the data as outputs of certain observational functionals $\{L_1, \ldots, L_M\} = \mathbf{L}$,

$$\mathbf{L}(\psi) = \mathbf{d}, \tag{4}$$

acting on the vector of the system state variables $\psi$. If $\dim\Psi > \dim D$, which is typical for oceanographic inverse problems, a probability distribution over $D$ cannot be uniquely transferred onto $\Psi$. A physical model,

$$\mathbf{A}(\psi) = \mathbf{f}, \tag{5}$$

provides additional information and may be of help. Here, $\mathbf{A}$ is an operator mapping $\Psi$ onto a Hilbert space $R$ of admissible forcing vectors $\mathbf{f} \in R$. Usually, (5) represents a set of differential equations and thus $\Psi$ is infinite-dimensional. To avoid consideration of very sophisticated aspects of the measure theory in infinite-dimensional spaces, from here on, we will assume that a discretization was applied and $\dim\Psi = N < \infty$.

An additional assumption is that the discrete set (4), (5) is overdetermined. If it is not the case, one should augment the dataset with bogus data at boundary and/or initial points to provide well-posedness of weak constraint data assimilation (Bennett and Miller 1991). Thus, our assumptions are valid in all practical applications. If (4), (5) constitute the overdetermined set, we have to admit that at least a part of information about $\psi$ contains errors. It is natural to assume that the errors inherent in the data and model equations have finite variances $\sigma_d$ and $\sigma_m$, respectively, that is,

$$\int_\Psi [\mathbf{L}(\psi) - \mathbf{d}, \mathbf{L}(\psi) - \mathbf{d}]_d p(\psi|\mathbf{d})\Pi\, d\psi = \sigma_d^2, \tag{6}$$

$$\int_\Psi (\mathbf{A}(\psi) - \mathbf{f}, \mathbf{A}(\psi) - \mathbf{f})_r p(\psi|\mathbf{d})\Pi\, d\psi = \sigma_m^2. \tag{7}$$

Here, the PDF $p(\psi|\mathbf{d})$ is defined with respect to the standard Lebesgue measure $\Pi d\psi$ on $\Psi$, and $(\cdot)_d$ and $(\cdot)_r$ are inner products in the spaces $D$ and $R$, respectively. They are not necessarily Euclidean inner products but may involve any positively defined matrices $\mathbf{C}_d$ and $\mathbf{C}_m$:

$$(\mathbf{d}_1, \mathbf{d}_2)_d = \mathbf{d}_1 \cdot \mathbf{C}_d^{-1}\mathbf{d}_2,$$
$$(\mathbf{f}_1, \mathbf{f}_2)_r = \mathbf{f}_1 \cdot \mathbf{C}_m^{-1}\mathbf{f}_2. \tag{8}$$

The rationale behind (6) and (7) is that at first, our measurement instruments indeed have finite error variance; at second, the desired smoothness of physical fields may be expressed as the boundedness of a norm (usually a Hilbert norm) of uncertainties in (5). The matrices $\mathbf{C}_d$ and $\mathbf{C}_m$ are usually interpreted as covariances of the data and model errors. As it was mentioned in section 1, there is a lot of uncertainty in the assessment of the model error covariances, at least in the oceanographic context. The choice will always be a hypothesis, which is difficult, if ever possible, to verify a posteriori.

An exercise in the calculus of variations leads to the PDF

$$p(\psi|\mathbf{d}) = C\mu(\psi)\exp\{-a_m\|\mathbf{A}(\psi) - \mathbf{f}\|_r^2 - a_d\|\mathbf{L}(\psi) - \mathbf{d}\|_d^2\}, \tag{9}$$

which maximizes (3) subject to (6), (7). Here,

$$\|\mathbf{A}(\psi) - \mathbf{f}\|_r^2 = [\mathbf{A}(\psi) - \mathbf{f}, \mathbf{A}(\psi) - \mathbf{f}]_r,$$
$$\|\mathbf{L}(\psi) - \mathbf{d}\|_d^2 = [\mathbf{L}(\psi) - \mathbf{d}, \mathbf{L}(\psi) - \mathbf{d}]_d, \tag{10}$$

and $a_m$, $a_d$ are some positive weights appearing as Lagrangian multipliers. It must be emphasized that $a_m \neq (1/2)\sigma_m^{-2}$ and $a_d \neq (1/2)\sigma_d^{-2}$. Together with a normalization constant $C$, they should be determined from (6), (7) and normalizability

$$\int_\Psi p(\psi|\mathbf{d})\Pi d\psi = 1, \tag{11}$$

If $\mu(\psi)$ is chosen as a constant, that is, we consider a priori all the states $\psi$ as equally probable, the most likely state minimizes a cost function

$$J(\psi) = a_m\|\mathbf{A}(\psi) - \mathbf{f}\|_r^2 + a_d\|\mathbf{L}(\psi) - \mathbf{d}\|_d^2. \tag{12}$$

Hence, the PME leads to a Gaussian distribution if $\mathbf{A}$ and $\mathbf{L}$ are linear. Based on the PME, we obtain the cost functional commonly used in weak constraint data assimilation when information about higher moments is not available. Though, if we had possessed the information about the higher moments, we would have been

able to account for it in the framework of the PME by imposing additional constraints on $p(\psi|\mathbf{d})$.

Calculating the weights $a_m$ and $a_d$ from the error variances with (6), (7), (11) is a tough routine (Urban 1996). We will not be able to determine the weights unless both $\sigma_d$ and $\sigma_m$ are known. Luckily, we are rarely interested in the whole PDF and would be pleased to obtain a reliable estimate of its maximizer or the mean. To calculate it, we only need the ratio $w = a_d/a_m$. Surprisingly, if there is no model solution fitting the data exactly (i.e., the model and the data are in contradiction), the PME makes it possible to proceed a bit further.

How the principle works may be illustrated with the example of multiple measurements of a scalar random variable $y$ (Ghill and Malanotte-Rizzoli 1991). If two measurements $y_1$ and $y_2$ are available, an estimate $\tilde{y}$ of $y$ may be obtained by minimizing a cost function

$$J(y) = a_1(y - y_1)^2 + a_2(y - y_2)^2, \tag{13}$$

with certain positive weights $a_i$, $i = 1, 2$. Then,

$$\tilde{y} = \alpha_1 y_1 + \alpha_2 y_2, \tag{14}$$

where

$$\alpha_1 = \frac{a_2}{a_1 + a_2}, \qquad \alpha_2 = \frac{a_1}{a_1 + a_2}. \tag{15}$$

Since

$$\alpha_1 + \alpha_2 = 1, \tag{16}$$

$\alpha_1$, $\alpha_2$ can be viewed as probabilities of the realizations $y_1$ and $y_2$. Thus, any choice of the weights in (13) generates a probabilistic measure $d\alpha$ with support in the set of the observed states. The mean with respect to this measure coincides with the minimizer (14) of the objective function (13). If we do not know the error variance of either measurement, the PME is of help. Maximizing the entropy

$$H(d\alpha) = -\sum_{i=1}^{2} \alpha_i \ln\alpha_i \tag{17}$$

subject to (16) yields $\alpha_1 = \alpha_2$ and $a_1 = a_2$. That is, not having enough constraints to maximize the entropy over continuous PDFs, we ''quantize'' the system state space $\Psi$ and pick up the maximum-entropy singular PDF.

Let us return to the multidimensional case. The information presented in (12) is excessive, and we will try to construct a singular measure over a quantized space, which would consist of the observed system states. Then, the PME would be a guide to choosing the weight ratio. At first, in the case of the linear model and observational operators, an inner product on $\Psi$ may be defined as

$$\langle \psi_1, \psi_2 \rangle = a_m[\mathbf{A}(\psi_1), \mathbf{A}(\psi_2)]_r + a_d[\mathbf{L}(\psi_1), \mathbf{L}(\psi_2)]_d, \tag{18}$$

which allows us to rewrite (9) as

$$p(\psi|\mathbf{d}) = C_1 \exp\{-\|\psi\|_\Psi^2 + 2a_m \mathrm{Re}[\mathbf{A}(\psi), \mathbf{f}]_r + 2a_d \mathrm{Re}[\mathbf{L}(\psi), \mathbf{d}]_d\}, \tag{19}$$

with a new normalization constant $C_1$. The last two terms are bounded functionals in $\Psi$. According to the Riesz theorem (Yosida 1980), for any $\mathbf{f} \in R$ and $\mathbf{d} \in D$ there exist unique vectors $\boldsymbol{\psi}_A$ and $\boldsymbol{\psi}_L$ such that

$$a_m[\mathbf{A}(\psi), \mathbf{f}]_r = \langle \psi, \boldsymbol{\psi}_A \rangle,$$
$$a_d[\mathbf{L}(\psi), \mathbf{d}]_d = \langle \psi, \boldsymbol{\psi}_L \rangle, \tag{20}$$

for all $\psi \in \Psi$. Consequently, we may define operators $\mathbf{A}_*: R \to \Psi$ and $\mathbf{L}_*: D \to \Psi$ such that

$$\mathbf{A}_*\mathbf{f} = \psi_A, \qquad \mathbf{L}_*\mathbf{d} = \psi_L. \tag{21}$$

Note that these operators are not adjoint to $\mathbf{A}$ and $\mathbf{L}$, though they have very close meaning.

Taking derivatives of the log-likelihood functions $S = -\ln p(\psi|\mathbf{d})$ with respect to $\psi$ and keeping in mind the definitions of the operators $\mathbf{A}_*$ and $\mathbf{L}_*$ [see (20) and (21)], we immediately arrive at an expression for the maximum likelihood estimate of $\psi$ or equivalently the mean with respect to the probability distribution (19):

$$\psi_i = \mathbf{M}_m\psi_m + \mathbf{M}_d\psi_d, \tag{22}$$

where $\psi_d$ and $\psi_m$ are any system states satisfying (4) and (5), respectively, and

$$\mathbf{M}_d = \mathbf{L}_*\mathbf{L}, \qquad \mathbf{M}_m = \mathbf{A}_*\mathbf{A}. \tag{23}$$

These operators act in $\Psi$ and have three attractive properties: they are nonnegative, self-adjoint, and

$$\mathbf{M}_m + \mathbf{M}_d = \mathbf{I}, \tag{24}$$

where $\mathbf{I}$ is the unit operator (see appendix A for a proof). Hence, these operators generate an operator-valued measure (OVM) $\mathbf{M}$ (Davies and Lewis 1970) in the system state space $\Psi$ with support in two disjoint subsets of $\Psi$: the set $\Psi_m$ of the solutions to the model (5) and the set $\Psi_d$ of the states fitting the data exactly.

Now we need to define the entropy $H(\mathbf{M})$ of the OVM that would meet (i)–(iii) (see section 2). A possible candidate is

$$H(\mathbf{M}) = -\mathrm{trace}(\mathbf{M}_d \ln\mathbf{M}_d + \mathbf{M}_m \ln\mathbf{M}_m) = -\sum_{i=1}^{N} [\lambda_i \ln\lambda_i + (1 - \lambda_i)\ln(1 - \lambda_i)], \tag{25}$$

where $\lambda_i$, $1 - \lambda_i$ are eigenvalues of the operators $\mathbf{M}_d$ and $\mathbf{M}_m$, respectively. Since we have more freedom in the assessment of the OVM than an ordinary probability measure (PM), conditions (i)–(iii) are not enough to determine the entropy uniquely. However, if we wish to retain additiveness of Shannon's entropy, a property intensively used in thermodynamics, (25) is the only choice. A rigorous mathematical proof of this fact is out of the scope of the present paper and will be given elsewhere. Thus, (25) can be considered as a natural generalization of (17).

Two properties of $H(\mathbf{M})$ are easy to establish. First, $H(\mathbf{M})$ is maximum when $\mathbf{M}_m = \mathbf{M}_d = 0.5\mathbf{I}$. Second, if the mean eigenvalue,

$$\lambda = N^{-1} \sum_{i=1}^{N} \lambda_i,$$

is fixed, $H(\mathbf{M})$ is maximum when $\lambda_i = \lambda$ for all $i$. In the latter case, the OVM becomes the PM. Thus, OVMs generally contain more information than PMs. This conclusion is very natural, since the OVM appeared due to our knowledge of the inner structure of random events in the multidimensional space $\Psi$; by assigning ordinary probabilities to the events observed ($\psi_m$ and $\psi_d$), we would ignore that structure.

Note that only nonzero and nonunity eigenvalues make a contribution to $H(\mathrm{M})$, and there are not more than

$$N_+ = \dim D - \dim \ker \mathbf{M}_m$$

such eigenvalues. Here, ker $\mathbf{M}_m$ denotes the null space of $\mathbf{M}_m$. As each term under the summation symbol in (25) reaches its upper bound at $\lambda_i = 0.5$, it immediately follows that

$$H(M) \leq N_+ \frac{1}{2} \ln \frac{1}{2}. \tag{26}$$

We emphasize that the operators $\mathbf{M}_m$ and $\mathbf{M}_d$ depend not only on weights $a_m$, $a_d$ and the operators $\mathbf{A}$ and $\mathbf{L}$, but also on the specific form of the inner products $(\cdot)_r$ and $(\cdot)_d$. Involved in the definitions of the operators $\mathbf{A}_*$ and $\mathbf{L}_*$, they reflect our a priori assumptions about the data and model error covariances, $\mathbf{C}_d$ and $\mathbf{C}_m$. In the standard practice of data assimilation, we know the approximate $\mathbf{C}_d$ and have few ideas about what $\mathbf{C}_m$ looks like. Hence, we might try to maximize $H(\mathbf{M})$ as a guide to selecting the priors. However, $\dim D \ll \dim R$ as a rule, and maximizing the entropy (25) will not determine $\mathbf{C}_m$ uniquely. In this instance, if we could estimate the relative weight $w$ and the decorrelation length of the model errors, that would be quite a satisfactory result.

## 4. Case study: Weak constraint data assimilation into a 1D model for tides in Tatarsky Strait

### a. Model

To test what the PME can give in practice, we have applied it to a weak constraint formulation of a linear channel model for the $M_2$ tide in Tatarsky Strait (Fig. 1). If the strait width $W$ is much smaller than a characteristic wavelength and varies smoothly, the cross-channel velocity is negligible. Then, complex amplitudes of surface elevation $\zeta$ and alongstrait mass transport $q$ averaged over the strait width are related by the continuity and by alongchannel momentum equations

$$i\omega\zeta - (hq)_x = 0, \tag{27}$$

$$i\omega q - g\zeta_x = 0. \tag{28}$$

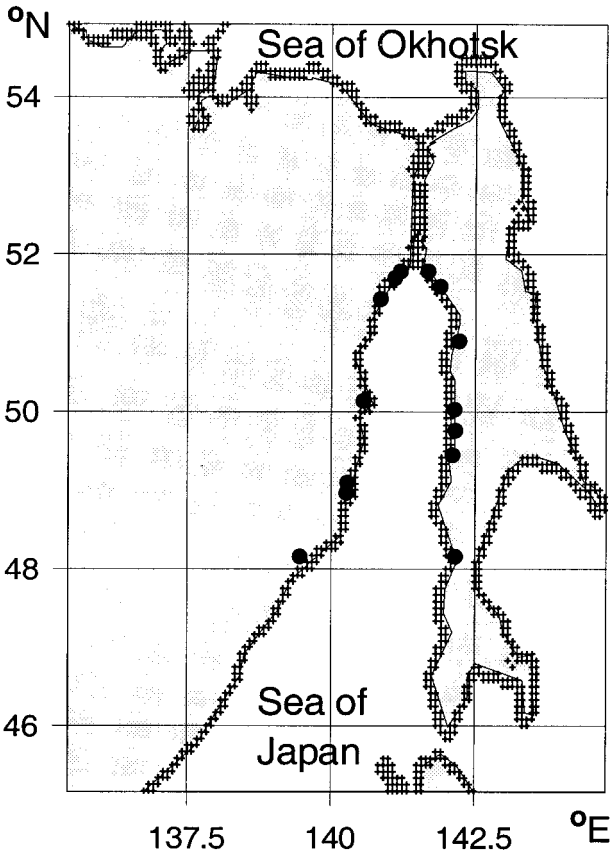


FIG. 1. Tatarsky Strait. Dots indicate measurement locations.

Here, $\omega = 1.405 \times 10^{-4}$ s$^{-1}$ is the frequency of the $M_2$ tide constituent considered, $g$ is the gravity acceleration, $h(x)$ is the mean depth across the strait, and $x$ is the alongstrait coordinate. Set (27), (28) was reduced to a second-order equation for $q$:

$$(hq)_{xx} + \frac{\omega^2}{g} q = 0. \tag{29}$$

The data $d(x)$ available are complex amplitudes of the surface elevation at certain coastal points (see Fig. 1). They are related to the model variable $q$ via the cross-strait momentum equation, which is reduced to geostrophy under the assumptions adopted (Toulany and Garrett 1984):

$$\zeta_+ - \zeta_- = -\frac{f}{g} qW. \tag{30}$$

Here, $f$ is the Coriolis parameter and $\zeta_+$ and $\zeta_-$ are complex amplitudes of the surface elevation at the west and east coast, respectively. If we assume surface elevation to vary linearly in the cross-strait direction and take approximately

$$\zeta = \frac{1}{2}(\zeta_+ + \zeta_-), \tag{31}$$

Eqs. (27), (30), and (31) relate $q$ to the data $d(x)$,

$$(hq)_x \pm \frac{i\omega f W}{2g} q = i\omega d(x), \tag{32}$$

where + (−) is held for the west (east) coast.

We assimilate the east coast data, while those located at the west side of the strait are left for verification of the inverse solution. The cost function $J(q)$ is

$$J(q) = J_m + wJ_d. \tag{33}$$

Here, the first term represents the mean-square misfit in the weak finite-element approximation of (29) (Kurapov and Kivman 1999), the second one refers to the square misfit in the data equation (32), and $w$ is the weight to be tuned. The cost function is minimized by the conjugate gradient.

We would like the generalized inverse of our 1D model to reproduce features essential for weak constraint data assimilation into 2D and 3D models. One of them is ill-posedness of the problem in the continuum limit (see section 6). To allow for it, the dynamical errors are supposed to be uncorrelated. The same are assumed for the data errors. If we penalized the mean-square norm of the residuals referred to the model equations (27), (28), we would obtain $\zeta, q \in W_2^1$ in the continuum limit, that is, $\zeta$, $q$, $\zeta_X$, $q_x$ would belong to the space $L_2$ of square-integrable functions. Since every function from $W_2^1$ in the 1D case is continuous (Adams 1975), the $\zeta$ data at isolated points would generate observational functionals bounded in the continuum limit; the problem would be well-posed, even if a diagonal dynamical error covariance matrix was used. This circumstance would not allow us to test the ability of the entropy approach to recognize ill-possedness and tell us when smoothing is necessary. Therefore, to introduce ill-posedness, we have used the second-order equation (29) for the tidal transport $q$. Then, the observations are related to values of $q_x$ at the isolated points via (32). We can guarantee $q_x \in L_2$ only, and the observational functionals are unbounded in the continuum limit.

### b. Maximum entropy and the quality of the solution

Inverse solutions have been computed at four different grids containing $N = 20$, 38, 75, or 149 nodes. At each grid, a series of calculations have been made to obtain the entropy $H$ as a function of the weight ratio $w$. Along with it, we have estimated root-mean-square (rms) errors of the solution with respect to the data constraining the solution ($\text{rms}_{as}$) and to the data left for verification ($\text{rms}_{v}$) (Fig. 2).

In principle, we wish to obtain a solution that would have as small $\text{rms}_{as}$ and $\text{rms}_{v}$ as possible. By increasing the data weight $w$, we can achieve an arbitrary small value of the former. However, the latter does not generally diminish with $w \to \infty$ and usually has a positive lower bound. Closeness between $\text{rms}_{as}$ and $\text{rms}_{v}$ points to the fact that the inverse solution experiences minor sensitivity to the choice of the subset of the data taken from the whole set of observations and assimilated (presuming the data are all of the same quality). Alternatively, large differences between $\text{rms}_{as}$ and $\text{rms}_{v}$ mean that the inverse solution varies notably depending on whether it is constrained by certain data; in the latter case, the inverse solution will be far from the truth. Therefore, in speaking about the quality of the solution, we impose two basic requirements. First, the solution should yield $\text{rms}_{v}$ as small as possible. Second, we would want $\text{rms}_{as}$ to be close to $\text{rms}_{v}$ if the two datasets are of the same accuracy.

Let us check the quality of the solution corresponding to the weight $w = w_{\text{me}}$ at which it yields the maximum entropy. At each ''coarse'' grid (grids 20, 38, and 75), we will be quite satisfied if $\text{rms}_{v}$ is close to its minimum (see Fig. 2). Underdetermining the weight leads to a much bigger value. The second quality criterion is also fulfilled at $w = w_{\text{me}}$: $\text{rms}_{v}$ is close to $\text{rms}_{as}$ (Table 1). It is worth noting that the mean data amplitude is about 47 cm. Keeping in mind the crudeness of the model, the approximate 20% relative error seems satisfactory.

Overestimating $w$ does not increase $\text{rms}_{v}$ significantly. However, $\text{rms}_{as}$ becomes close to zero, and thus the inverse solution is sensitive to the input information. The other limit case ($w \to 0$ or $a_m \to \infty$) corresponds to the so-called strong constraint data assimilation. This approach has been widely used in the oceanographic inverse modeling and is based on the assumption that the model equations are error free. An observation made in several studies (see Gong et al. 1998) was that the weak constraint estimates ($w > 0$) were better if compared with the strong constraint solutions. Figure 2 clearly indicates the advantage of weak constraint data assimilation.

At the very fine resolution (grid 149), the situation is not as good as that at the coarser grids. The PME overestimates the data weight, which leads to a large discrepancy between $\text{rms}_{as}$ and $\text{rms}_{v}$. In the case study, the entropy peak value is attained at grid 38, and the maximum entropy value decreases when a finer or coarser grid is taken (Table 1). Numerical discretization acts as a regularizor to the problem, which is ill-posed in the continuum limit (see section 6). Consequently, following the PME we are to adopt grid 38 as optimal. Numerical results support our choice: $\text{rms}_{v}(w_{\text{me}})$ obtained at grid 149 is not improved in comparison with the value obtained at grid 75 (see Table 1). Thus, choosing the grid much finer than that having the biggest entropy does not improve the accuracy of the inverse solution.

## 5. Maximum data cost criterion

To compute the entropy, one needs a calculation of eigenvalues of the operator $\mathbf{M}_d$ or $\mathbf{M}_m$, which is not feasible to obtain for large $N$. We would like to find an approximation to the entropy approach and to avoid explicit calculation of the entropy. It is easy to check

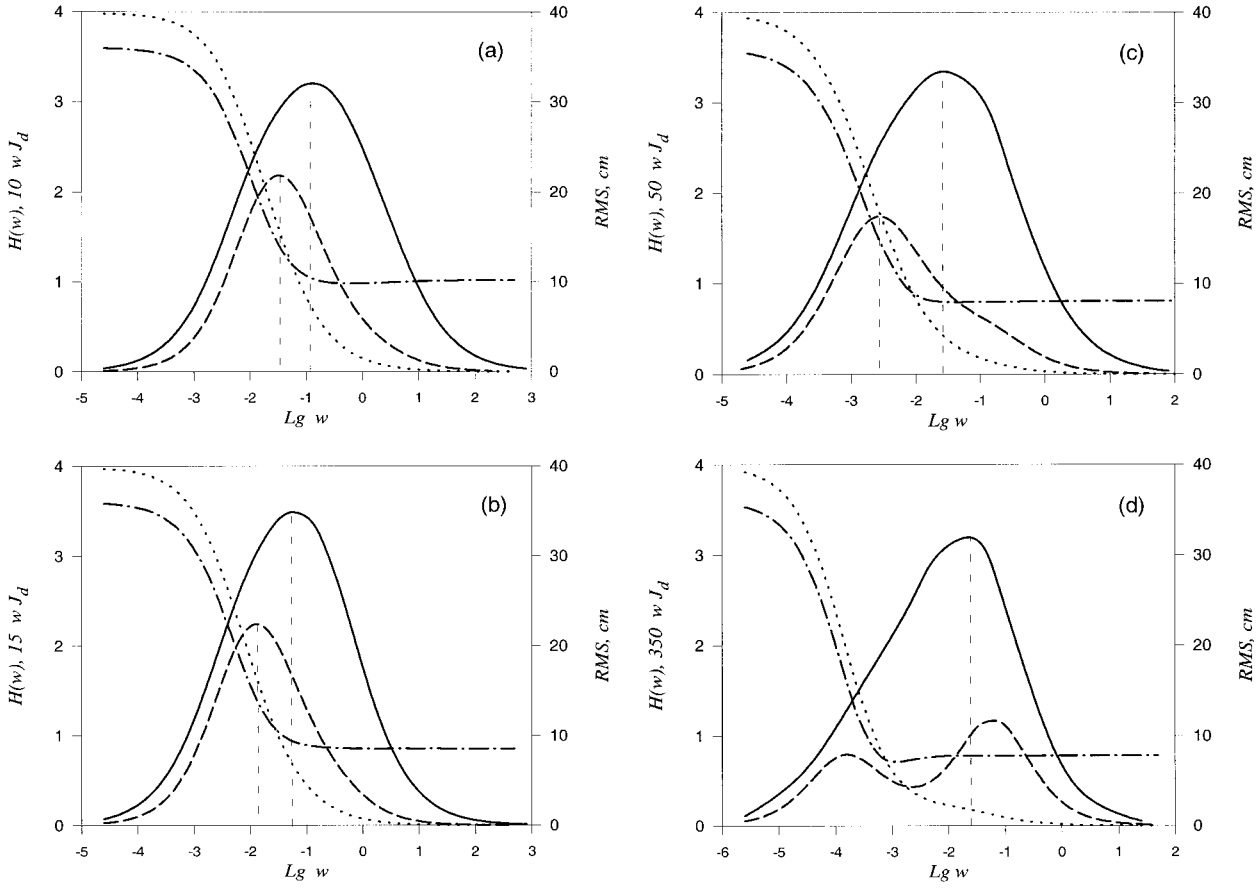


FIG. 2. The entropy (solid line), the data cost (dashed line), $\mathrm{rms}_{as}$ (dotted line), and $\mathrm{rms}_{v}$ (dashed–dotted line) as functions of the data weight $w$ obtained on different grids: (a) 20 nodes, (b) 38 nodes, (c) 75 nodes, (d) 149 nodes. The units for $\mathrm{rms}_{as}$ and $\mathrm{rms}_{v}$ are in centimeters.

the multiple measurements example discussed in section 2 so that the maximum-entropy measure $d\alpha$ yields the weight maximizing the cost $a_1(\tilde{y} - y_1)^2$ when $a_2$ is fixed, and vice versa. Does the same fact hold for operator-valued measures appearing in multidimensional spaces? A relevant theorem is proven in appendix B. The MDC coincides with the PME under the assumption that the shape of the error covariance matrices allows the entropy to attain its upper bound (26) at some $w$. Fixing $\mathbf{C}_m$ and $\mathbf{C}_d$ a priori, we cannot guarantee that the upper bound of the entropy will be attained. However, we may still select $w$ in accordance with the MDC criterion, considering it as an approximation to the PME.

TABLE 1. Comparison of the PME and the MDC.

| $N$ | Method | $w$ | rms $v$ (incm) | $\mathrm{rms}_{as}$ (incm) | $H(w)$ |
|---|---|---|---|---|---|
| 20 | PME | 0.1250 | 10.0 | 7.1 | 3.19 |
| | MDC | 0.0375 | 14.0 | 15.8 | 2.93 |
| 38 | PME | 0.0550 | 9.4 | 6.9 | 3.49 |
| | MDC | 0.0125 | 14.2 | 16.7 | 3.02 |
| 75 | PME | 0.0250 | 7.9 | 4.4 | 3.35 |
| | MDC | 0.0025 | 14.7 | 17.8 | 2.51 |
| 149 | PME | 0.0250 | 7.8 | 1.7 | 3.19 |
| | MDC | | Two maxima | | |

Let us return to the case study. The MDC predicts weights $w_{\mathrm{mdc}} = 15$ and $w_{\mathrm{mdc}} = 5$ at grids 20 and 38, correspondingly. Though these weights are underestimated in comparison with the maximum-entropy weights $w_{\mathrm{me}} = 45$ and $w_{\mathrm{me}} = 22$, the MDC solutions are still of good quality (see Fig. 2). At grid 75, the MDC weight is already an order of magnitude less than the PME weight ($w_{\mathrm{mdc}} = 1$ against $w_{\mathrm{me}} = 10$). As we know from above, a sort of smoothing should be applied at this resolution. Further grid refinement leads to splitting the extremum of the data cost curve. Hence, the necessary condition for the MDC to coincide with the PME is violated (see appendix B).

## 6. Smoothing

What lies behind difficulties rising at the fine resolution? Since the model errors are assumed to be uncorrelated, the Hilbert structure in $\Psi$, introduced by the inner product (18), does not guarantee smoothness of $\psi$ in the continuum limit. Although the inverse solution

is bounded at any discrete grid, it is very sensitive to the data noise and to the spatial resolution if the grid becomes too fine. As an example, if the data are values of model variables at isolated points, the inverse solution has pathological behavior at the data sites with holes and spikes in their vicinity, and it does not converge to a continuous function as numerical lattice tends to zero (Bennett and McIntosh 1982). Kivman (1997b) has demonstrated that ill-posedness is not just a mathematical technicality; refining the grid can in fact degrade the inverse solution compared with that obtained at a coarser numerical lattice. In principle, well-posedness may be restored by reducing the relative data weight with grid refinement. However, this scheme relies on the assumption that the continuous model equations are error free, which is difficult to judge.

Smoothing at the fine resolution allows us to account for the model uncertainties that are not caused by discretization. Several strategies have been put forward to smooth the solution to the generalized inverse. Bennett and McIntosh (1982) used spatially variable dynamical weights with suitable singularities at the data sites. Provost and Salmon (1986) augmented the cost functional by a penalty on higher derivatives of the solution. Dynamical error covariances with nonzero off-diagonal elements were adopted in Egbert et al. (1994). Instead of penalizing the $L_2$ norm of the model equation residuals, Kivman (1996) imposed a penalty on higher norms.

The theory of partial differential equations says that if the inverse solution satisfies an elliptic equation, it is continuous everywhere except, possibly, at the data sites. Hence, smoothing is necessary only in the data neighborhood, not in the whole computational domain. Based on this fact, we propose to ''spread'' the data in order to regularize the problem. The procedure consists in the insertion of bogus data of the same value as the original datum in the vicinity of each data point. It is possible to show that the inverse solution converges to a smooth function in the continuum limit if the radius of spreading is fixed.

From the physical standpoint, the decorrelation length of the field of interst is nonzero, and thus the values that the field takes in the vicinity of the data site is close to what is actually observed. Thus, inserting such bogus data around the measurement location within an area of strong correlations is plausible. Furthermore, we have more ideas about the decorrelation length of physical fields than about that of the model errors.

Because the number of observations $M$ is always finite, the norm generated by (18) becomes only a seminorm at some resolution in 2D and 3D applications. That is, $\langle \psi, \psi \rangle = 0$ does not necessarily mean that $\psi = 0$. Consequently, we will have to close the inverse problem. When we consider an open system exchanging the energy with its surroundings, the number $M_o$ of numerical grid nodes at the open boundary may exceed the amount of the data there, and maximizing (19) becomes an ill-posed problem, even if $M > M_o$. A standard regularization routine reduces to spreading the data along the open boundary (interpolating between available data). Here, we propose to apply this regularization methodology to the interior observations.

How can we embed a procedure for tuning the smoothing parameter into the MDC? We may try to determine a threshold spatial resolution where the general inverse starts working badly and additional regularization is already necessary. Smoothing makes the inverse solution more tolerant to the inputs. However, the solution must be insensitive to a smoothing parameter. The optimal weight $w_{\mathrm{mdc}}$ estimated by the MDC can play the role of a sensitivity measure of the inverse solution to the spreading length. If the data cost curve changes dramatically after the data are spread to the closest neighbors of the actual data locations, the spreading is not necessary at all.

Does the sensitivity study point to the same threshold resolution as that estimated by the PME ($N = 75$)? Yes, it does. As Fig. 3 depicts, spreading the data over adjacent nodes results in a dramatic increase of the MDC weight for $N = 20$. At grid $N = 38$, spreading the data causes the appearance of the second maximum. For $N = 75$, the MDC weight does not change if we spread the data just over two neighboring nodes, while further spreading drastically affects the MDC weight. An interesting point is that the optimal spreading length obtained at the two finer grids ($N = 75, 149$) remains the same in physical units. Thus, the smoothing length scale may be tuned at the coarser grid and used to obtain a solution at the finer resolution.

## 7. Summary

The principle of maximum entropy (PME) is a rigorous scheme for estimating weights and smoothing parameters involved in weak constraint data assimilation. Though the operator-valued measure generated by the prior seems to be an extravagant mathematical object, measures of that type arise naturally in the quantum probability theory. They appear in there instead of classical probability distributions because of the incomplete observability of quantum systems (Kholevo 1972). We face quite the same situation in data assimilation. The data are not enough to estimate even the mean $E(\psi)$, let alone the full statistics. In this instance, we invoke a hypothesis about the distribution of the model errors, which are unobservable; thus, the data space and the space of the model errors are generally nonisomorphic. Only in the simplest example of multiple measurements of one variable do we have $\Psi = D = R$; thus the classical probability measure $d\alpha$ can be used.

Application of the PME to weak constraint data assimilation in the 1D tide model has allowed us to obtain the inverse solution of good quality. In addition, the PME has been able to point at the threshold spatial resolution where the generalized inverse already requires smoothing.

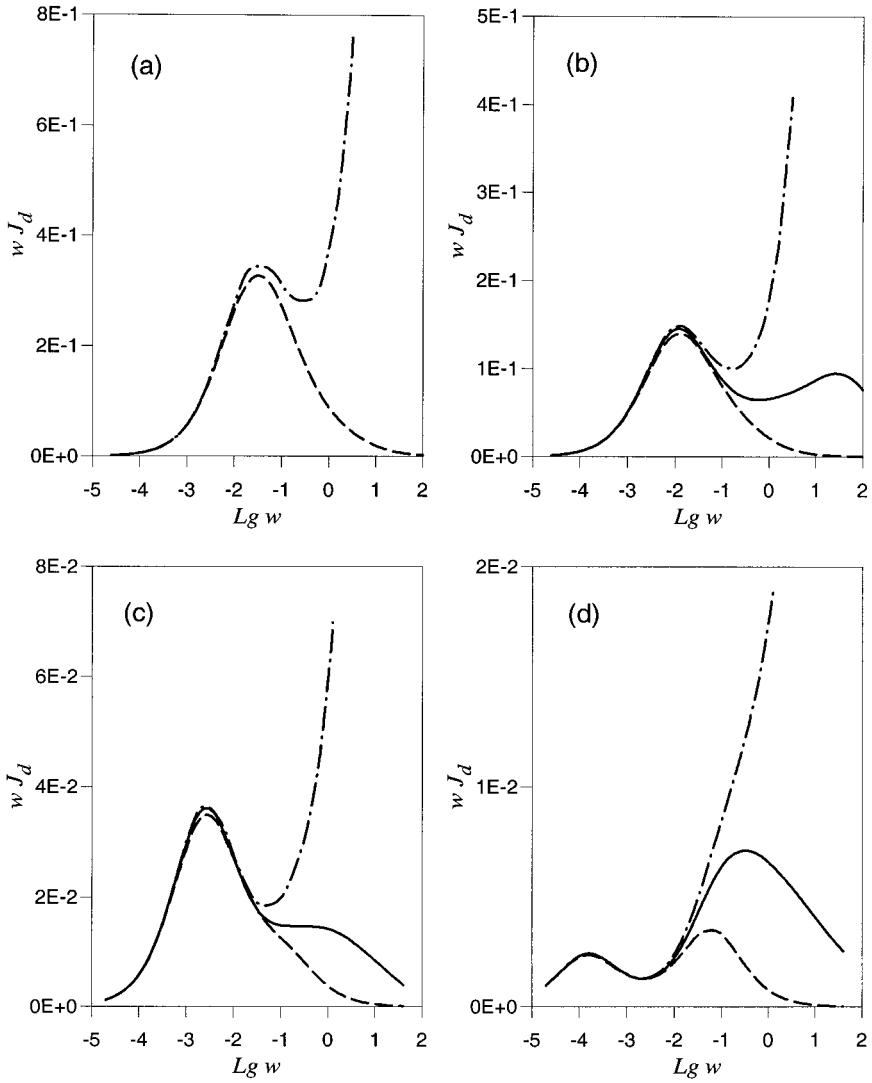


FIG. 3. The data cost for different spreading lengths. (a)–(d) as in Fig. 2. Dashed line, no spreading; solid line, optimal spreading: for 3 nodes in (b) and (c), for 7 nodes in (d); dashed–dotted line, spreading is overestimated when made: for 3 nodes in (a), 5 in (b) and (c), and 9 in (d).

The PME is difficult to employ in large-dimensional problems. In this instance, the maximum data cost criterion (MDC) is a feasible alternative to the PME. The MDC was used earlier as a heuristic rule for estimating the relative data weight in weak constraint data assimilation (Kivman 1997c; Kurapov and Kivman 1999). Connection between the MDC and the PME sheds light on reasons for efficiency of the former. Namely, the MDC can be viewed as an approximation to the PME when the curve $\psi(w) \subset \Psi$ is close to a straight segment. It is also important that the MDC is computationally simpler than the cross-validation. Moreover, if the data are from different sources (say, those from pressure gauges and current meters; in situ and satellite altimetry measurements), the method can be applied to multiple weights tuning (Kurapov and Kivman 1999).

In the case study, the MDC underestimated the data weight in comparison with the maximum entropy value. Nevertheless, the inverse solution was fairly good. We should admit that the MDC still requires large computational efforts because the data cost value can be determined only after the inverse solution was obtained. Luckily, the weight can be and should be estimated at the coarse spatial resolution where no smoothing is required. Then we transfer it onto a finer grid.

The method also points to the fine grids where smoothing is necessary. We spread the data to provide well posedness. The spreading length obtained by the sensitivity criterion is in agreement with the PME value. It is worth noting that this smoothing parameter corresponds to the decorrelation length of the tidal fields and, consequently, is easier to be evaluated a priori than the decorrelation length of the model errors. In our numerical example, the quality of the inverse solution did

not appear to degrade with grid refinement, even if the smoothing routine was not applied. A reason may lie in that the data are of high quality and stay close to each other if compared with the model length scale. To check this hypothesis, we have unrealistically shallowed the channel and observed that smoothing was actually necessary.

*Acknowledgments.* This material is based upon the work supported by the U.S. Civilian Research & Development Foundation under Award RGI-245.

## APPENDIX A

### Properties of $\mathbf{M}_d$ and $\mathbf{M}_m$

Here we will prove some statements concerning the operators $\mathbf{M}_d$ and $\mathbf{M}_m$. At first, since for any $\psi_1, \psi_2 \in \Psi$,

$$\begin{aligned}\langle \psi_1, (\mathbf{L}_*\mathbf{L} + \mathbf{A}_*\mathbf{A})\psi_2\rangle &= a_d[\mathbf{L}(\psi_1), \mathbf{L}(\psi_2)]_d + a_m[\mathbf{A}(\psi_1, \mathbf{A}(\psi_2)]_r \\ &= \langle \psi_1, \psi_2\rangle, \end{aligned} \tag{A1}$$

then, immediately

$$\mathbf{M}_m + \mathbf{M}_d = \mathbf{I}, \tag{A2}$$

where $\mathbf{I}$ is the unit operator.

Next, the operator $\mathbf{M}_d$ is self-adjoint with respect to inner product (18). Indeed,

$$\begin{aligned}\langle \psi_1, \mathbf{L}_*\mathbf{L}\psi_2\rangle &= a_d[\mathbf{L}(\psi_1), \mathbf{L}(\psi_2)]_d \\ &= a_d\overline{[\mathbf{L}(\psi_2), \mathbf{L}(\psi_1)]_d} \\ &= \overline{\langle \psi_2, \mathbf{L}_*\mathbf{L}\psi_1\rangle} \\ &= \langle \mathbf{L}_*\mathbf{L}\psi_1, \psi_2\rangle. \end{aligned} \tag{A3}$$

Here, an upper bar denotes complex conjugation.

Finally, the $\mathbf{M}_d$ is nonnegative because

$$\langle \mathbf{M}_d\psi, \psi\rangle = a_d(\mathbf{L}\psi, \mathbf{L}\psi)_d \geq 0. \tag{A4}$$

Proofs for $\mathbf{M}_m$ are identical.

## APPENDIX B

### Equivalence between the MDC and the PME in a Special Case

LEMMA. If $v$ is an eigenfunction of $\mathbf{M}_d$ for some $\tilde{a}_d > 0$:

$$\mathbf{M}_d v = \lambda(\tilde{a}_d)v, \tag{B1}$$

then $v$ is the eigenfunction of $\mathbf{M}_d$ for any $a_d \in [0, \infty)$.

*Proof.* According to the definition of $\mathbf{M}_d$ [see (23)], (B1) is equivalent to the validity of

$$\begin{aligned}\tilde{a}_d(\mathbf{L}v, \mathbf{L}\psi)_d &= \lambda(\tilde{a}_d)[a_m(\mathbf{A}v, \mathbf{A}\psi)_r + \tilde{a}_d(\mathbf{L}v, \mathbf{L}\psi)_d] \end{aligned} \tag{B2}$$

for any $\psi \in \Psi$. Choosing $\psi = v$ in (B2), we obtain

$$\lambda(\tilde{a}_d) = \frac{\tilde{a}_d\|\mathbf{L}v\|_d^2}{a_m\|\mathbf{A}v\|_r^2 + \tilde{a}_d\|\mathbf{L}v\|_d^2}. \tag{B3}$$

Further, combining (B2) and (B3) yields, after some algebra,

$$\|\mathbf{A}v\|_r^2(\mathbf{L}v, \mathbf{L}\psi)_d = \|\mathbf{L}v\|_d^2(\mathbf{A}v, \mathbf{A}\psi)_r. \tag{B4}$$

It follows from (B4) that

$$\begin{aligned}a_d(\mathbf{L}v, \mathbf{L}\psi)_d = \frac{a_d\|\mathbf{L}v\|_d^2}{a_m\|\mathbf{A}v\|_r^2 + a_d\|\mathbf{L}v\|_d^2} \\ \times [a_m(\mathbf{A}v, \mathbf{A}\psi)_r + a_d(\mathbf{L}v, \mathbf{L}\psi)_d] \end{aligned} \tag{B5}$$

for any $a_d \in [0, \infty)$. Comparing (B5) with (B2) and (B3) completes the proof. Obviously, matrix $\mathbf{M}_m$ possesses the same quality.

THEOREM. If $H(\mathbf{M})$ reaches the least upper bound (26) at some $a_d = a_d^{\text{me}}$, the data cost $\phi(a_d) = a_d\|\mathbf{L}\psi_i - \mathbf{d}\|_d^2$ has a maximum at $a_d = a_d^{\text{me}}$, and this extreme point is unique.

*Proof.* It follows from (22) and (24) that

$$\psi_i - \psi_d = \mathbf{M}_m\,\Delta, \tag{B6}$$

where $\Delta = \psi_m - \psi_d$. By applying $\mathbf{L}$ to both sides of (B6), we will obtain

$$\phi(a_d) = a_d\|\mathbf{L}\mathbf{M}_m\Delta\|_d^2. \tag{B7}$$

Let us decompose $\Delta$ in a series of orthogonal eigenfunctions $v_k$ of $\mathbf{M}_m$:

$$\Delta = \sum_k b_k v_k, \tag{B8}$$

where coefficient $b_k$ depends on $a_d$. According to the lemma, the set of the eigenfunctions is the same for any $a_d$.

Since

$$\begin{aligned}a_d(\mathbf{L}\mathbf{M}_m v_k, \mathbf{L}\mathbf{M}_m v_j)_d &= \langle \mathbf{M}_m v_k, \mathbf{M}_d\mathbf{M}_m v_j\rangle \\ &= \lambda_k\lambda_j(1 - \lambda_j)\delta_{kj}\langle v_k, v_j\rangle, \end{aligned} \tag{B9}$$

where $\lambda_k$ are eigenvalues of $\mathbf{M}_m$, substitution of (B8) for (B7) yields

$$\begin{aligned}\phi(a_d) = \sum_k b_k^2\lambda_k^2[1 - \lambda_k] \\ \times [a_m(\mathbf{A}v_k, \mathbf{A}v_k)_r + a_d(\mathbf{L}v_k, \mathbf{L}v_k)_d]. \end{aligned} \tag{B10}$$

Let us evaluate the least upper bound of the right-hand side of (B10). At first, the following obvious inequality holds:

$$\begin{aligned}\phi(a_d) \leq \max_k [\lambda_k(1 - \lambda_k)] \\ \times \sum_{\lambda_k \neq 1} b_k^2\lambda_k[a_m(\mathbf{A}v_k, \mathbf{A}v_k)_r + a_d(\mathbf{L}v_k, \mathbf{L}v_k)_d]. \end{aligned} \tag{B11}$$

It becomes an equality when all nonzero and nonunity eigenvalues are equal to each other. Next, since

$$a_m\|\mathbf{A}\Delta\|_r^2 = a_m \sum_{k,j} b_k b_j(\mathbf{A}\upsilon_k, \mathbf{A}\upsilon_j)_r$$
$$= \sum_k b_k^2\lambda_k[a_m(\mathbf{A}\upsilon_k, \mathbf{A}\upsilon_k)_r + a_d(\mathbf{L}\upsilon_k, \mathbf{L}\upsilon_k)_d], \quad \text{(B12)}$$

one can rewrite (B11) as

$$\phi(a_d) \le \max_k [\lambda_k(1 - \lambda_k)](a_m\|\mathbf{A}\Delta\|_r^2 - \|\mathbf{P}_1\Delta\|_\Psi^2), \quad \text{(B13)}$$

where $\mathbf{P}_1$ is the projector on to the subspace of eigenvectors corresponding to the unity eigenvalues.

Let us notice that if $\upsilon_l$ is an eigenfunction corresponding to $\lambda_l = 1$ for $a_d > 0$, then $\mathbf{L}\upsilon_l = 0$. This implies that neither $\langle\upsilon_l, \upsilon_l\rangle$ nor $\langle\Delta, \upsilon_l\rangle$ depend on $a_d$. Hence, the second multiplier at the right-hand side of (B13) does not depend on $a_d$. The least upper bound of the first multiplier is equal to 0.25, and it is reached when $\lambda_j = 0.5$ for some $j$.

It follows from (B3) that eigenvalues $\lambda_m$ are monotonic functions of $a_d$. According to the hypothesis of the theorem, the entropy attains its least upper bound. Thus, all nonzero and nonunity eigenvalues are equal to 0.5 when $a_d = a_d^{\text{me}}$. These imply that $a_d = a_d^{\text{me}}$ is an extreme point of the data cost $\phi(a_d)$, and this extreme point is unique. The theorem is proven.

Two points are worth noting in this connection. First, if the entropy attains its upper bound, all nonzero and nonunity eigenvalues of $\mathbf{M}_d$ are equal to each other, and thus the inverse solutions corresponding to different data weights all lie on a straight segment in the state space $\Psi$. Second, it is seen from the proof that except for uniqueness of the extreme point, the theorem remains valid for the case of several data weights referred to different multiple data terms (Kurapov and Kivman 1999).

## REFERENCES


Adams, R. A., 1975: *Sobolev Spaces.* Academic Press, 268 pp.

Bennett, A. F., 1992: *Inverse Methods in Physical Oceanography.* Cambridge University Press, 346 pp.

——, and P. C. McIntosh, 1982: Open ocean modeling as an inverse problem: Tidal theory. *J. Phys. Oceanogr.,* **12,** 1004–1018.

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

Boltzmann, L., 1964: *Lectures on Gas Theory.* Cambridge University Press, 490 pp. [First published as *Vorlesungen über Gastheorie,* Barth, 1896.]

Courtier, P., 1997: Dual formulation of four-dimensional variational assimilation. *Quart. J. Roy. Meter. Soc.,* **123,** 2449–2461.

Cramer, H., 1954: *Mathematical Methods of Statistics.* 6th ed. Princeton University Press, 575 pp.

Davies, E. B., and J. T. Lewis, 1970: An operational approach to quantum probability. *Commun. Math. Phys.,* **17,** 239–260.

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

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

Egbert, G. D., and A. F. Bennett, 1996: Data assimilation methods for ocean tides. *Modern Approaches to Data Assimilation in Ocean Modeling,* P. Malanotte-Rizzoli, Ed., Elsevier Press, 147–179.

——, ——, and M. G. G. Foreman, 1994: TOPEX/POSEIDON tides estimated using a global inverse model. *J. Geophys. Res.,* **99,** 24 821–24 852.

Evensen, G., 1997: Advanced data assimilation for strongly nonlinear dynamics. *Mon. Wea. Rev.,* **125,** 1342–1354.

Gao, F., 1994: Fitting smoothing splines to data from multiple sources with unknown relative weights. *Commun. Stat. Theor. Methods,* **23,** 1665–1698.

Ghil, M., and P. Malanotte-Rizzoli, 1991: Data assimilation in meteorology and oceanography. *Advances in Geophysics,* Vol. 33, Academic Press, 141–266.

Gibbs, J. W., 1902: *Elementary Principles in Statistical Mechanics.* Yale University Press, 207 pp.

Gong, J., G. Wahba, D. Johnson, and J. Tribbia, 1998: Adaptive tuning of numerical weather prediction models: Simultaneous estimation of weighting, smoothing, and physical parameters. *Mon. Wea. Rev.,* **126,** 210–231.

Griffith, A. K., and N. K. Nichols, 1996: Accounting for model error in data assimilation using the adjoint methods. *Proc. Second Int. SIAM Workshop on Computational Differentiation,* Santa Fe, NM, SIAM, 195–204.

Jaynes, E. T., 1988a: The relation of Bayesian and maximum entropy methods. *Maximum-Entropy and Bayesian Methods in Science and Engineering,* G. J. Erickson and C. R. Smith, Eds., Kluwer, Reidel, 25–29.

——, 1988b: How does the brain do plausible reasoning? *Maximum-Entropy and Bayesian Methods in Science and Engineering,* G. J. Erickson and C. R. Smith, Eds., Kluwer, Reidel, 1–24.

Kholevo, A. C., 1972: Statistical problems in quantum physics. *Proc. Second Japan–USSR Symp. on Probability Theory,* Kyoto, Japan, 22–40.

Kivman, G. A., 1996: Assimilating sea-level data into a global tidal model by means of the generalized method of residuals (in Russian). *Okeanologia,* **36,** 835–841.

——, 1997a: Assimilating data into open ocean models. *Surv. Geophys.,* **18,** 621–645.

——, 1997b: Weak constraint data assimilation for tides in the the Arctic Ocean. *Progress in Oceanography,* Vol. 40, Pergamon Press, 179–196.

——, 1997c: Weak constraint data assimilation for tides in the the Arctic Ocean. *Proc. Workshop on Data Assimilation in Large-scale Models,* Delft, Netherlands, Delft University of Technology, 53–55.

Kurapov, A. L., and G. A. Kivman, 1999: Data assimilation in a finite element model of $M_2$ tides in the Barents Sea (in Russian). *Okeanology,* **39,** 306–313.

Provost, C., and R. Salmon, 1986: A variational method for inverting hydrographic data. *J. Mar. Res.,* **44,** 1–34.

Shannon, C. E., 1948: A mathematical theory of communication. *Bell Syst. Tech. J.,* **27,** 379–423, 623–655.

Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, 613 pp.

Thacker, W. C., and G. Holloway, 1990: Inverse problems and the principle of maximum entropy. *Int. Symp. on the Assimilation of Observations in Meteorology and Oceanography,* Clermont-Ferrand, France, 6–11.

Toulany, B., and C. Garrett, 1984: Geostrophic control of fluctuating barotropic flow through straits. *J. Phys. Oceanogr.,* **14,** 649–655.

Urban, B., 1996: Retrieval of atmospheric thermodynamical parameters using satellite measurements with a maximum entropy method. *Inverse Probl.,* **12,** 779–796.

van Leeuwen, P. J., and G. Evensen, 1996: Data assimilation and inverse methods in terms of a probabilistic formulation. *Mon. Wea. Rev.,* **124,** 2898–2913.

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

Yosida, K., 1980: *Functional Analysis.* Springer-Verlag, 500 pp.