Statistical Design for Adaptive Weather Observations

L. MARK BERLINER

Department of Statistics, Ohio State University, Columbus, Ohio, and
National Institute of Statistical Sciences, Research Triangle Park, North Carolina

ZHAN-QIAN LU AND CHRIS SNYDER

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 15 August 1997, in final form 19 November 1998)

ABSTRACT

Suppose that one has the freedom to adapt the observational network by choosing the times and locations of observations. Which choices would yield the best analysis of the atmospheric state or the best subsequent forecast? Here, this problem of ''adaptive observations'' is formulated as a problem in statistical design. The statistical framework provides a rigorous mathematical statement of the adaptive observations problem and indicates where the uncertainty of the current analysis, the dynamics of error evolution, the form and errors of observations, and data assimilation each enter the calculation. The statistical formulation of the problem also makes clear the importance of the optimality criteria (for instance, one might choose to minimize the total error variance in a given forecast) and identifies approximations that make calculation of optimal solutions feasible in principle. Optimal solutions are discussed and interpreted for a variety of cases. Selected approaches to the adaptive observations problem found in the literature are reviewed and interpreted from the optimal statistical design viewpoint. In addition, a numerical example, using the 40-variable model of Lorenz and Emanuel, suggests that some other proposed approaches may often be close to the optimal solution, at least in this highly idealized model.

1. Introduction

Data used in meteorological forecasting currently consist mainly of routine observations from dedicated platforms, such as radiosonde stations and satellites, and observations of opportunity, such as those from commercial aircraft. There is increasing interest, however, in adaptively modifying and supplementing the existing observational network according to the key forecast problems of a given day. This interest is motivated both by the scarcity of resources available for meteorological observations and by the now routine availability of information concerning the growth and propagation of errors within forecasts. We will refer to the notion of changing and supplementing the observational network in order to optimize the quality of a specific forecast, as adaptive observations.

Since forecast errors have a random component (because analysis errors arise in part from random errors in observations), it is natural to cast the problem in a statistical framework. Thus, our topic is the statistical design of data collection processes in order to optimize statistical measures of the quality of prediction. Though our statistical formulation is appropriate in general design problems, the focus here is to take advantage of opportunities for choosing specific geographical areas in which specialized observational data will be obtained. We also present an example of optimal design for an idealized low-order model of the atmosphere (following Lorenz and Emanuel 1998) and compare that design against other proposed adaptive observation strategies.

This work was motivated in part by the opportunity to test adaptive strategies during the Fronts and Atlantic Storm Tracks Experiment (FASTEX; see Joly et al. 1997). FASTEX included two long-range jet aircraft capable of providing between 10 and 50 additional drop soundings over the North Atlantic, upstream of the main observational area centered on Shannon, Ireland. The design problem was to regularly and optimally devise flight paths for the aircraft. Future experiments will further test the notion of adaptive data collection.

To best communicate the ideas the following idealized problem is the focus of this article: let $\mathbf{X}_0$, $\mathbf{X}_1$, and $\mathbf{X}_2$ be n-dimensional vectors representing the state of the atmosphere at times t_0, t_1, and t_2, respectively, based on some finite-dimensional representation, such as mod-

Corresponding author address: Dr. Mark Berliner, Department of Statistics, Ohio State University, Cockins Hall, 1958 Neil Ave., Columbus, OH 43210-1247.
E-mail: mb@stat.ohio-state.edu

el gridpoint values or spectral coefficients. Given all available information at $t = t_0$ (the decision or design time), we wish to decide how to collect additional observations at $t = t_1$ (the targeted or observation time) in order to optimize statistical properties of a forecast from $t = t_1$ valid at $t = t_2$ (the forecast or validation time). The exposition here will primarily assume that all observations at $t = t_1$ are our adaptive observations; more general situations are discussed in section 2c. Also, note that, in general, the quantities represented in $\mathbf{X}$ may include a variety of different physical variables, or in idealized settings (e.g., section 4), the values of a single variable at different physical locations or sites. The structure of our formulation is not dependent on these differences; hence, we generically use ''site'' to indicate an element of $\mathbf{X}$.

The choice of data collection design is to be guided by trying to obtain the most efficient forecast of $\mathbf{X}_2$. (The accuracy of a prediction of course depends on how the prediction is made. The forecast model considered here is described in section 2.) The task involves a tradeoff between regions in which one expects to be very uncertain about $\mathbf{X}_1$ versus regions of $\mathbf{X}_1$ that are extremely important in terms of their role in the evolution of the dynamics. We used the word ''expects'' because the entire issue is statistical. That is, our forecast of $\mathbf{X}_2$ will be based on our best assessment of $\mathbf{X}_1$, but this assessment depends on the (unknown) future data whose collection is being designed.

To frame the problem statistically, we first note that, quite generally, forecasts (estimators of random variables) are chosen to be the *conditional* expected values of those random variables. The conditioning is based on all information available at the time of prediction. The corresponding measure of predictive accuracy is then the *conditional* covariance matrix of the variables. These notions have been the basis for data assimilation. In particular, ''objective analysis'' procedures typically produce conditional expectations. See Lorenc (1986) for clarifications.

We use the variable $\boldsymbol{\nu}$, subscripted by time, to represent an analysis; A, again subscripted by time, represents the corresponding analysis errors covariance matrix. We use the variable $\boldsymbol{\mu}$, subscripted by time, to represent a forecast, and B, appropriately subscripted, represents the corresponding forecast errors covariance matrix. The procedure described here focuses on the quantities described in Table 1. The indicated formulas for these quantities are developed in section 2. With this notation, the adaptive observation problem is essentially, at time t_0, to decide where to observe the system at time t_1, with the intent of making B_2 ''small.''

In section 2 selected principles of statistical design of experiments are introduced. A more general overview of the subject is given in appendix A. Applications of statistical design in the context of prediction of linearizable dynamical systems are described. Section 3 develops the essence of some other approaches to the adaptive observation problem leading to comparisons and new interpretations vis-à-vis our approach. In particular, we explicitly compute examples of our optimal statistical designs in the case of negligible observational noise. This permits new insights, interpretations, and comparisons with other approaches.

TABLE 1. Guide to key definitions.

	Expected value (analysis or forecast)	Covariance
At time t_0		
To estimate $\mathbf{X}_0$	$\boldsymbol{\nu}_0$ (input)	A_0 (input)
To forecast $\mathbf{X}_1$	$\boldsymbol{\mu}_1$ (2.8)	B_1 (2.9)
To forecast $\mathbf{X}_2$	$\boldsymbol{\mu}_2^o$ (3.15)	B_2^o (3.16)
At time t_1, after observation		
To estimate $\mathbf{X}_1$	$\boldsymbol{\nu}_1$ (2.15)	A_1 (2.13)
To forecast $\mathbf{X}_2$	$\boldsymbol{\mu}_2$ (2.19)	B_2 (2.20)

Section 4 is devoted to example derivations of our results and some comparisons to other approaches. Following Lorenz and Emanuel (1998), we use a toy adaptive observation problem based on a 40-variable model. Section 5 is devoted to comments and a summary.

2. Statistical formulation of design problems

The statistical design of experiments is a fundamental subdiscipline of statistics. The approach taken in this article is known as optimal experimental design. Though the literature is rich and complex, the guiding principle can be stated succinctly:

> For a given formulation of a problem, choose a procedure for collecting future data with the intent of optimizing mathematical criteria that reflect statistical accuracy of the conclusion to be made.

If our problem is to predict (forecast) a random vector, say $\mathbf{X}$, we wish to obtain data that is informative regarding the probability distribution of that vector. A very common assumption that well motivates our approach is to control predictive mean-squared error (MSE). Namely, we wish to minimize the expected squared difference between our predictor and $\mathbf{X}$; formally, let $p(\mathbf{Y})$ denote a prediction procedure based on data $\mathbf{Y}$ whose collection we are designing. We are to minimize the expectation of the Euclidean norm squared prediction errors:

$$\text{MSE} = E(|\mathbf{X} - p(\mathbf{Y})|^2), \tag{2.1}$$

where the indicated expectation is taken with respect to *both* $\mathbf{X}$ and $\mathbf{Y}$. A result from the theory of prediction (e.g., Aitchison and Dunsmore 1975, pp. 47–50) tells us that we should choose the predictor to be the conditional expectation of $\mathbf{X}$, given the observed data $\mathbf{y}$:

$$E(\mathbf{X} \,|\, \mathbf{Y} = \mathbf{y}). \tag{2.2}$$

This result is derived by noting that

$$\text{MSE} = E[|\mathbf{X} - E(\mathbf{X}|\mathbf{Y}) + E(\mathbf{X}|\mathbf{Y}) - p(\mathbf{Y})|^2] \quad (2.3)$$

$$= E\{\text{tr}[\text{Cov}(\mathbf{X}|\mathbf{Y})]\} + E[|E(\mathbf{X}|\mathbf{Y}) - p(\mathbf{Y})|^2]. \quad (2.4)$$

By choosing $p(\mathbf{Y}) = E(\mathbf{X}|\mathbf{Y})$, the second term of (2.4) vanishes. For design purposes our goal then becomes optimization of the expected (with respect to $\mathbf{Y}$) trace of the conditional covariance matrix of $\mathbf{X}$.

A message in this derivation is that we formulate a criterion by first anticipating what predictive procedure we plan to use. This procedure's accuracy is unknown, since both $\mathbf{Y}$ and $\mathbf{X}$ are unknown (''random'') at the moment of design. Hence, we can only control *expected*, not actual, behavior.

We present additional review of the literature and the development of criteria for optimal design in appendix A.

a. Primary statistical formulation

1) MODEL AT TIME t_0

We assume the following statistical and dynamical formulations. At time t_0 the information available to us concerning $\mathbf{X}_0$ is summarized in the probabilistic summary, or ''prior,''

$$\mathbf{X}_0 \sim N(\boldsymbol{\nu}_0, \mathbf{A}_0),$$

where this notation is read as ''$\mathbf{X}_0$ has a multivariate normal (Gaussian) distribution with expected value (mean) $\boldsymbol{\nu}_0$ and covariance matrix $\mathbf{A}_0$.'' Some readers may find this counterintuitive: the state of the atmosphere at an instant is some fixed vector. The statistical view is to summarize our uncertainty via a probability distribution, thereby treating $\mathbf{X}_0$ as a random vector. Typically, this distribution will be the result of some data assimilation process. In particular, $\boldsymbol{\nu}_0$ is the *analyzed* state and $\mathbf{A}_0$ is the corresponding analysis error covariance matrix.

The dynamics of the evolution of the process are represented as

$$\mathbf{X}_1 = f(\mathbf{X}_0), \quad (2.5)$$

where f is a known, nonlinear function. (We discuss formulations that include model uncertainty in section 5.) Note that despite the determinism implicit in (2.5), if $\mathbf{X}_0$ is random, then so is $\mathbf{X}_1$. Next, consider the tangent linear approximation:

$$\mathbf{X}_1 \approx f(\boldsymbol{\nu}_0) + \mathbf{F}(\boldsymbol{\nu}_0)(\mathbf{X}_0 - \boldsymbol{\nu}_0), \quad (2.6)$$

where $\mathbf{F}(\boldsymbol{\nu}_0)$ is the $n \times n$ Jacobian matrix of the transformation f, evaluated at $\boldsymbol{\nu}_0$. Based on this approximation, it follows from standard statistical theory that our implied, approximate distribution for $\mathbf{X}_1$ is

$$\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \mathbf{B}_1), \quad (2.7)$$

where

$$\boldsymbol{\mu}_1 = f(\boldsymbol{\nu}_0) \quad (2.8)$$

and

$$\mathbf{B}_1 = \mathbf{F}(\boldsymbol{\nu}_0)\mathbf{A}_0\mathbf{F}(\boldsymbol{\nu}_0)^{\mathrm{T}}. \quad (2.9)$$

(Also, see Ehrendorfer and Tribbia 1997.)

At time $t = t_0$ we have the opportunity to design an experiment that will take place at time $t = t_1$. We will observe a function of the state variables $\mathbf{X}_1$, with measurement error. In the adaptive observation problem of FASTEX, the problem was the selection of a limited *subarea* of the region of interest for data collection. We assume that the random data to be observed, denoted by $\mathbf{Y}$, follows a model,

$$\mathbf{Y} = \mathbf{K}\mathbf{X}_1 + \boldsymbol{\epsilon}, \quad (2.10)$$

where $\mathbf{K}$ is a $d \times n$ matrix. We assume that the measurement error vector $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$. (Here $\boldsymbol{\Sigma}$ is a $d \times d$ matrix that implicitly depends on $\mathbf{K}$.) The best choice of $\mathbf{K}$, under appropriate restrictions, is our problem. In a simple adaptive observation context, $\mathbf{K}$ could be viewed as an ''incidence matrix,'' consisting of d rows, each containing $n - 1$ zeroes and a single 1, to signify that we are to take d total observations at the indicated regions. Restrictions include the specification of d, typically very small compared to n, and the condition that the regions used be geographically contiguous. Finally, note that the linearity assumption implicit in (2.10) is often itself an approximation, typically justified via additional linearization arguments.

2) UPDATING AT TIME t_1

A standard calculation provides the following updating of knowledge about $\mathbf{X}_1$, combining the actual observed data $\mathbf{y}$ and the previous data and dynamical relationships summarized by (2.7). Under the above specifications, it can be shown that *conditional* on $\mathbf{Y} = \mathbf{y}$, the distribution of $\mathbf{X}_1$ is

$$\mathbf{X}_1|\mathbf{y} \sim N(\boldsymbol{\nu}_1, \mathbf{A}_1), \quad (2.11)$$

where

$$\mathbf{A}_1 = (\mathbf{B}_1^{-1} + \mathbf{K}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{K})^{-1} \quad (2.12)$$

$$= \mathbf{B}_1 - \mathbf{B}_1\mathbf{K}^{\mathrm{T}}(\boldsymbol{\Sigma} + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1}\mathbf{K}\mathbf{B}_1, \quad (2.13)$$

$$\boldsymbol{\nu}_1 = \mathbf{A}_1(\mathbf{B}_1^{-1}\boldsymbol{\mu}_1 + \mathbf{K}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}) \quad (2.14)$$

$$= \boldsymbol{\mu}_1 - \mathbf{B}_1\mathbf{K}^{\mathrm{T}}(\boldsymbol{\Sigma} + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1}(\mathbf{K}\boldsymbol{\mu}_1 - \mathbf{y}). \quad (2.15)$$

The results in (2.13) and (2.15) were obtained via application of Bayes' theorem. [See West and Harrison (1989) for details particularly relevant to the formulation here.] It is well known that they coincide with familiar results from the data assimilation literature, often known as the extended Kalman filter (Lorenc 1986; Tarantola 1987; Courtier 1997). In particular, the estimate given in (2.15) is obtainable as the solution to an optimization problem: find the minimizer of

$$J(\mathbf{x}_1) = (\mathbf{x}_1 - \boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{B}_1^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{y} - \mathbf{K}\mathbf{x}_1)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{K}\mathbf{x}_1). \quad (2.16)$$

The fact that the two approaches agree involves a bit of calculus and the observation that for Gaussian distributions the most likely (maximum likelihood) or modal value coincides with the mean of the distribution. For our purposes it is convenient to present the analyses in standard statistical notation. This permits a clear parallel to statistical design theory and enables a clear tracking of the impact on uncertainties resulting from the approximations used.

3) DESIGN FOR PREDICTION OF $\mathbf{X}_2$

The above notions are next applied to the basic problem described in section 1. Throughout this formulation, we act as if all means and covariance matrices derived using tangent linear approximations are exact, rather than approximate. Our goal is to predict

$$\mathbf{X}_2 = g(\mathbf{X}_1), \tag{2.17}$$

where g is a known, nonlinear function. The notation g allows for the possibility that the time lags $t_1 - t_0$ and $t_2 - t_1$ may be different.

Applying the tangent linear approximation on (2.17) yields the following approximation, *conditional* on the data $\mathbf{y}$:

$$\mathbf{X}_2 | \mathbf{y} \sim N(\boldsymbol{\mu}_2, \mathbf{B}_2), \tag{2.18}$$

where

$$\boldsymbol{\mu}_2 = g(\boldsymbol{\nu}_1) \tag{2.19}$$

and

$$\mathbf{B}_2 = \mathbf{G}(\boldsymbol{\nu}_1)\mathbf{A}_1\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}. \tag{2.20}$$

The matrix $\mathbf{G}$ is the collection of first partials of the transformation g; in (2.20) $\mathbf{G}$ is evaluated at $\boldsymbol{\nu}_1$. Note that $\mathbf{B}_2$ depends on the values of the very data $\mathbf{y}$ whose design is being formulated. Following the paradigm outlined earlier, we should seek optima of expectations of functions of the predictive covariance matrix $\mathbf{B}_2$; namely, optimize

$$E_0[\mathcal{F}(\mathbf{B}_2)]. \tag{2.21}$$

The subscript 0 on the expectation operator reminds us that this calculation is done at the present time.

b. Some issues

From a theoretical view the problem of adaptive observations is now well posed, statistically. However, serious difficulties can arise in practical implementation.

The key difficulty is that the computation and subsequent optimization of the criterion given in (2.21) are virtually intractable in very high-dimensional problems. Hence, additional simplifications are needed. These simplifications involve additional statistical approximations as well as numerical simplifications and dimension reduction.

1) DESIGN FOR ESTIMATION OF $\mathbf{X}_1$

Rather than attempting to arrive at the most efficient prediction of $\mathbf{X}_2$, consider the problem of efficient estimation of $\mathbf{X}_1$. The hope, parallel to much of the reasoning in the data assimilation literature, is that a very good estimate of $\mathbf{X}_1$, say $\boldsymbol{\nu}_1$, can be used to obtain reasonable predictions of $\mathbf{X}_2$. The resulting design problem is then one of finding optima of functions of $\mathbf{A}_1$. This problem is well posed, since $\mathbf{A}_1$ does not depend on the actual value of the data $\mathbf{y}$. (This feature is quite dependent on the Gaussian assumptions used here.) This approach also seems natural in settings in which the analyst wishes to predict at a variety of future time points. This notion seems to be a motivation of the discussion presented in Lorenz and Emanuel (1998).

2) APPROXIMATING THE DESIGN CRITERION

The primary criterion (2.21) is complex. Referring to (2.20), the matrix $\mathbf{G}$ is a complicated function of the data through the quantity $\boldsymbol{\nu}_1$. Furthermore, $\mathbf{G}(\boldsymbol{\nu}_1)$ itself enters the criterion function (2.21) nonlinearly. These complexities and the size of the problem combine to mandate a significant simplification.

Recall that $E_0(\boldsymbol{\nu}_1) = \boldsymbol{\mu}_1 = f(\boldsymbol{\nu}_0)$. [See (2.8) and (2.15).] The suggestion is that we simply ''plug in'' the expected value of $\boldsymbol{\nu}_1$, that is, replace $\mathbf{G}(\boldsymbol{\nu}_1)$ in $E_0\{\mathcal{F}[\mathbf{G}(\boldsymbol{\nu}_1)\mathbf{A}_1\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}]\}$ by $\mathbf{G}(\boldsymbol{\mu}_1)$. This eliminates the need for computing an expectation, since, as noted earlier, $\mathbf{A}_1$ is independent of the observed data. Analyses based on the resulting criterion,

$$\mathcal{F}[\mathbf{G}(\boldsymbol{\mu}_1)\mathbf{A}_1\mathbf{G}(\boldsymbol{\mu}_1)^{\mathrm{T}}], \tag{2.22}$$

are considered in section 3.

To clarify the simplification accruing from the plug-in approximation, consider the following special cases (see appendix A for discussion of these example criteria).

1) Considering the determinant function (D optimality) leads to

$$\begin{aligned} E_0[\mathcal{F}(\mathbf{B}_2)] &= E_0\{\det[\mathbf{G}(\boldsymbol{\nu}_1)\mathbf{A}_1\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}]\} \\ &= \det(\mathbf{A}_1)E_0\{\det[\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}\mathbf{G}(\boldsymbol{\nu}_1)]\}. \end{aligned} \tag{2.23}$$

   Applying the approximation reduces to optimizing $\det(\mathbf{A}_1)$. Note that the dynamics beyond t_1 would play no role and we are in the mode of section 2b(1).
2) The basic A-optimality criterion corresponds to optimization of

$$\begin{aligned} E_0[\mathcal{F}(\mathbf{B}_2)] &= E_0\{\mathrm{tr}[\mathbf{G}(\boldsymbol{\nu}_1)\mathbf{A}_1\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}]\} \\ &= \mathrm{tr}\{\mathbf{A}_1E_0[\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}}\mathbf{G}(\boldsymbol{\nu}_1)]\}. \end{aligned} \tag{2.24}$$

   Applying the approximation implies that we optimize $\mathrm{tr}[\mathbf{A}_1\mathbf{G}(\boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{G}(\boldsymbol{\mu}_1)]$. [Interest in features of the matrix $\mathbf{A}_1\mathbf{G}\,(\boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{G}(\boldsymbol{\mu}_1)$ will arise again in section 3.]
3) For E optimality the approximation implies that we

are to minimize the largest eigenvalue of the matrix in (2.22).

3) DIMENSION REDUCTION

In most settings the dimension of the state variable of interest is very high, on the order of millions. The dimensions of the corresponding covariance matrices lead to severe limitations on the numerical search for statistically optimal designs. A natural suggestion is to seek designs based on low-dimensional collection of variables that are themselves functions of the larger state vector. Once a small number of variables are agreed upon, the design approach is that outlined here, with the reduced set of variables simply replacing the original state variables, $\mathbf{X}$. In some cases this strategy may remove the need for the plug-in approximation described in section 2b(2).

4) MODEL UNCERTAINTY

To this point, we have have not attempted to adjust for model uncertainty or unmodeled forcings. Suppose that we extend the formulation in (2.5) and (2.17) to

$$\mathbf{X}_1 = f(\mathbf{X}_0) + \mathbf{S}_1 \tag{2.25}$$

and

$$\mathbf{X}_2 = g(\mathbf{X}_1) + \mathbf{S}_2, \tag{2.26}$$

respectively, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are independent vectors of mean-zero stochastic elements. We also assume that these shocks are independent of the state of the system at the instant before they are added (i.e., $\mathbf{X}_0$ and $\mathbf{S}_1$ are independent). Finally, let $\boldsymbol{\Xi}_1$ and $\boldsymbol{\Xi}_2$ be the covariance matrices of $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively.

If we are willing to assert the validity of the tangent linear approximations, we can develop a simple analog of the design problem. First, the analogs of (2.7)–(2.9) are that the approximate prior for $\mathbf{X}_1$ is

$$\mathbf{X}_1 \sim N(\boldsymbol{\mu}_1, \mathbf{B}_1^{(s)}), \tag{2.27}$$

where $\boldsymbol{\mu}_1$ is as given in (2.8) and

$$\mathbf{B}_1^{(s)} = \mathbf{B}_1 + \boldsymbol{\Xi}_1, \tag{2.28}$$

where $\mathbf{B}_1$ is defined in (2.9).

The assimilation of the adaptive observations is performed as in (2.11). One simply replaces $\mathbf{B}_1$ by $\mathbf{B}_1^{(s)}$ everywhere in (2.13)–(2.15). Let the resulting analog of $\mathbf{A}_1$ be denoted by $\mathbf{A}_1^{(s)}$. To go forward to $t = t_2$, we could consider approximations as in section 2b. In particular, a natural suggestion is to simply replace $\mathbf{B}_2$ given in (2.20) by

$$\mathbf{B}_2{}^{(s)} = \mathbf{G}(\boldsymbol{\nu}_1)\mathbf{A}_1{}^{(s)}\mathbf{G}(\boldsymbol{\nu}_1)^{\mathrm{T}} + \boldsymbol{\Xi}_2, \tag{2.29}$$

and proceed with the optimization.

c. Alternative data collection procedures

In most practical weather forecasting contexts, a variety of additional, sometimes termed ''routine,'' observations are available *after* time t_0. Adaptive design should adjust for such data. We consider three cases: routine observations are 1) before, 2) (essentially) simultaneous with, or 3) after the adaptive observations. We briefly describe adjustments to the analyses so far discussed to account for such data. In designing adaptive observation collection for prediction, the recipe for the statistical approach is to find a formula by the predictive covariance matrix for the state $\mathbf{X}_2$ at time t_2, as a function of the representation of the adaptive design (in our notation, $\mathbf{K}$), and the unobserved data (now both the adaptive and routine observations). We then find $\mathbf{K}$ to optimize the expectation of a function of the covariance matrix where the expectation is taken over the unobserved data; in practice we typically approximate the expectation. Hence, there is virtually no new conceptual baggage to introduce beyond that already described. However, technical problems in deriving formulas for covariances and approximate expectations may not be obvious. We formulate the setup in this section, but defer the calculations to appendix B.

Case 1. Suppose that at some time t_α, $t_0 < t_\alpha < t_1$, we will observe a dataset, denoted by $\mathbf{Y}_\alpha$, that is directly informative about the state of the system $\mathbf{X}_\alpha$ at time t_α.

At time t_0 we can compute a linearization-based approximation [analogous to (2.3)] for the distribution of $\mathbf{X}_\alpha$. Updating based on the observed data $\mathbf{y}_\alpha$ leads to an analog of (2.11). Next, our predictive distribution, conditional on the data $\mathbf{y}_\alpha$, for $\mathbf{X}_1$ can be approximated by employing another linearization (from time t_α). The final step is to form the conditional distribution of $\mathbf{X}_2$ given both $\mathbf{y}_\alpha$ and $\mathbf{y}$. With these definitions, one can proceed as in section 2a.

Case 2. Now suppose that the routine and adaptive observations are to be (essentially) simultaneously observed ($t_\alpha = t_1$). We then find new formulas corresponding to (2.13) and (2.15) based on both sets of data and again proceed as in section 2a.

Case 3. Finally, suppose that the routine data is collected at time t_β, $t_1 < t_\beta < t_2$. The analysis proceeds identically to that in section 2a up through the definition of the conditional distribution $\mathbf{X}_1$ given $\mathbf{y}$. Hence, if one seeks optimal designs for estimating $\mathbf{X}_1$ the analysis is unchanged.

For optimal prediction of $\mathbf{X}_2$, two linearization arguments (one from t_1 to t_β and then from t_β to t_2) would be used. We propagate this information forward in time to t_2.

Extensions of these analyses can be pursued. First, we may envision *collections* of both future times for adaptive observations as well as times at which we will

make predictions. Those predictions will be modified as we collect future observations. The adaptive observations problem is then a *sequential* statistical design problem. (In this article we have used a ''greedy'' or one-step approximation in which we only have controlled prediction variance at t_2, but no control downstream.) Further, selection of times as well as locations for adaptive observation may be considered in the optimization. Indeed, statistical design can, in principle, take a unified approach to the design of both routine and adaptive data collection. Though such fully sequential statistical designs have been studied and applied in other contexts (Chernoff 1972), the complexities and size of weather forecasting applications may moderate the richness of the approach.

3. Results and comparisons

In this section we derive statistical design results for the A-optimality criterion. These results are based on the plug-in approximation idea of section 2b(2). Further, under the assumptions that (i) adaptive observations have negligible errors and (ii) the design $\mathbf{K}$ is of an arbitrary form (i.e., the observation may be any linear combination of the state variables), we obtain formulas for optimal designs that carry substantial intuitive value. Our results also lead to new interpretations of previously proposed strategies for adaptive observations (Langland and Rohaly 1996; Bishop and Toth 1996; Lorenz and Emanuel 1998; Palmer et al. 1998; see also Snyder 1996).

a. Computations for A-optimal designs

1) SINGLE OBSERVATION SETTING

Suppose that a single adaptive observation is to be taken. That is, $\mathbf{K}$ is a single row vector. If that observation is to represent one coordinate of $\mathbf{X}_1$, we restrict the admissible $\mathbf{K}$ to be row vectors whose elements are all zero, except for one element equal to 1; that is, $\mathbf{K}$ is an incidence vector [see (2.10)]. Our data $Y = \mathbf{K}\mathbf{X}_1 + \boldsymbol{\epsilon}$ is a scalar. We assume that the measurement error $\boldsymbol{\epsilon}$ has variance σ_i^2, for $\mathbf{K}$ containing its 1 at index i. As in section 2b(2), all calculations will be based on the plug-in approximation

$$\mathbf{G}_0 \equiv \mathbf{G}(\boldsymbol{\mu}_1). \tag{3.1}$$

Under these assumptions (2.13) reduces to

$$\mathbf{A}_1 = \mathbf{B}_1 - \frac{\mathbf{B}_1\mathbf{K}^{\mathrm{T}}\mathbf{K}\mathbf{B}_1}{\sigma_i^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}.$$

An (approximate) A-optimal design minimizes $\mathrm{tr}(\mathbf{G}_0\mathbf{A}_1\mathbf{G}_0^{\mathrm{T}})$ or, equivalently, maximizes

$$\frac{\mathrm{tr}(\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}})}{\sigma^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}} = \frac{\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}{\sigma_i^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}. \tag{3.2}$$

The solution is well defined: namely, we should observe at that site that maximizes (over i) the quantity

$$\frac{(gb)_{ii}}{\sigma_i^2 + b_{ii}}, \tag{3.3}$$

where $(gb)_{ii}$ is the ith diagonal element of $\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1$ and b_{ii} is the ith diagonal element of $\mathbf{B}_1$.

Finding the optimal site requires further knowledge of $\mathbf{B}_1$ and $\mathbf{G}_0$ to make progress. To provide some intuition, note that in the unlikely situation that both $\mathbf{G}_0$ and $\mathbf{B}_1$ are diagonal matrices, (3.3) implies that we should observe at that site that maximizes $(b_{ii}g_{ii})^2/(\sigma_i^2 + b_{ii})$, where g_{ii} is the ith element of $\mathbf{G}_0$. The trade-offs between the magnitudes of b_{ii}, σ_i^2, and g_{ii}^2 are interesting. For example, if the b_{ii} and σ_i^2 vary very little compared to the g_{ii}, then the dynamics reflected in $\mathbf{G}_0$ dominate the design. Alternatively, if all g_{ii} are essentially equal, the optimal site is that which has the largest total variance $\sigma_i^2 + b_{ii}$.

2) MULTIPLE OBSERVATIONS SETTING

Here, $\mathbf{K}$ is a $d \times n$ matrix, where d is the number of observations in the design. An (approximate) A-optimal design minimizes $a = \mathrm{tr}(\mathbf{G}_0\mathbf{A}_1\,\mathbf{G}_0^{\mathrm{T}})$ where

$$\mathrm{tr}(\mathbf{G}_0\mathbf{A}_1\mathbf{G}_0^{\mathrm{T}}) = \mathrm{tr}(\mathbf{G}_0\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}) - \mathrm{tr}[(\boldsymbol{\Sigma} + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1}\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}]. \tag{3.4}$$

Hence, A optimality is equivalent to maximizing the quantity

$$\tilde{a} = \mathrm{tr}[(\boldsymbol{\Sigma} + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1}\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}], \tag{3.5}$$

with respect to $\mathbf{K}$ {recall [see (2.10)] that $\boldsymbol{\Sigma}$ depends on $\mathbf{K}$}.

Computation of the optimal design in this case is challenging in general. If $\mathbf{K}$ is an incidence matrix, we must calculate $\tilde{a}\ n!/[d!(n-d)!]$ times; this is certainly prohibitive for n huge and d moderate. However, in practical adaptive observation problems, not all designs are feasible. For example, if the elements of the state vector $\mathbf{X}_1$ represent sites, it is often the case that the d sites to be selected must be contiguous neighbors. This yields a major reduction in the computational overhead. Also, note that the matrix inversion indicated in (3.5) is of a $d \times d$ matrix; typically, we expect d to be comparatively small.

3) UNRESTRICTED DESIGNS FOR SMALL OBSERVATION ERRORS

Further approximations to A-optimal solutions can be obtained if the measurement error variances are essentially zero and the elements of $\mathbf{K}$ are unrestricted. That is, we are asking which linear combinations of $\mathbf{X}_1$ would be best to observe. Formally, we should impose some normalization side conditions on $\mathbf{K}$. Here, we are only

interested in the direction of solutions. Hence, normalization may be ignored. [Note that observing an unrestricted linear combination of $\mathbf{X}_1$ is implausible in the context of real adaptive observations. We pursue this case for its intuitive value; also, this case enables an interesting comparison to an analysis in Palmer et al. (1998).]

First, consider the single-observation case. Suppose that all σ_i^2 are very small, so that σ_i^2 may be neglected in the denominators in (3.2) for approximation purposes. To roughly assess the validity of this approximation, note that algebra yields

$$\frac{\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}{\sigma_i^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}} = \frac{\mathbf{K}\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}{\mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}\left[1 - \frac{\sigma_i^2}{\sigma_i^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}}\right]. \tag{3.6}$$

Hence, σ_i^2 should be at least an order smaller than the smallest feasible value of $\mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}$ for the approximation to be of value.

Setting $\sigma_i^2 = 0$ and assuming $\mathbf{K}$ is unrestricted, maximization of (3.2) is equivalent to finding the leading eigenvector of the eigenvalue problem,

$$\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{K}^{\mathrm{T}} = \lambda\mathbf{K}^{\mathrm{T}}. \tag{3.7}$$

Let

$$\mathbf{K}_*^{\mathrm{T}} \tag{3.8}$$

denote that eigenvector.

In the multiple-observation case, a corresponding analysis can be obtained when $\boldsymbol{\Sigma}$ is *small*. By small, we mean a condition analogous to that motivated for the single-observation case: namely, $\boldsymbol{\Sigma}(\mathbf{K}\mathbf{B}_1\ \mathbf{K}^{\mathrm{T}})^{-1}$ is ''small.'' Setting $\boldsymbol{\Sigma}$ equal to a matrix of zeros in (3.5), we have

$$\tilde{a} = \mathrm{tr}[(\mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1/2}\mathbf{K}\mathbf{B}_1^{1/2}(\mathbf{B}_1^{1/2}\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1^{1/2})\mathbf{B}_1^{1/2}\mathbf{K}^{\mathrm{T}}(\mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}})^{-1/2}]. \tag{3.9}$$

The maximizer of $\tilde{a}$ is

$$\mathbf{K}_*^{\mathrm{T}} = (\mathbf{k}_1^{\mathrm{T}}, \ldots, \mathbf{k}_d^{\mathrm{T}}), \tag{3.10}$$

where $\mathbf{k}_1^{\mathrm{T}}, \ldots, \mathbf{k}_d^{\mathrm{T}}$ denote leading d eigenvectors corresponding to the eigenproblem

$$\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1\mathbf{k}^{\mathrm{T}} = \lambda\mathbf{k}^{\mathrm{T}}. \tag{3.11}$$

(See Rao 1973, p. 74.)

As a closing comment, we note that if the unrestricted solutions $\mathbf{k}_j^{\mathrm{T}}$ above are highly localized, they can direct the search for the restricted, ''incidence'' solutions.

b. A simple principle for adaptive design

We reconsider the design for estimation of $\mathbf{X}_1$ problem as discussed in section 2b(1) in the context of small observation errors. Results for this case are most readily obtained by simply setting $\mathbf{G}_0$ to be the identity matrix in the A-optimality condition (3.5) as well as the eigenvalue problem (3.11). Assuming again that $\mathbf{K}$ is an incidence matrix, note that specification of $\mathbf{K}$ is equivalent to specification of a partition $\mathbf{X}_1 = (\mathbf{X}_o^{\mathrm{T}}, \mathbf{X}_u^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{X}_o$ is the d vector of variables to be directly observed and $\mathbf{X}_u$ correspond to the unobserved sites. It follows that (3.5) reduces to

$$\tilde{a} = \mathrm{tr}\{[\boldsymbol{\Sigma} + \mathrm{Cov}(\mathbf{X}_o)]^{-1} \times [(\mathrm{Cov}(\mathbf{X}_o))^2 + \mathrm{Cov}(\mathbf{X}_o, \mathbf{X}_u)(\mathrm{Cov}(\mathbf{X}_o, \mathbf{X}_u))^{\mathrm{T}}]\}, \tag{3.12}$$

where $\mathrm{Cov}(\mathbf{X}_o, \mathbf{X}_u)$ is the $d \times n$ matrix of pairwise covariances of elements of $\mathbf{X}_o$ with elements of $\mathbf{X}_u$. If $\boldsymbol{\Sigma}$ is small, $\tilde{a}$ is approximated by

$$\mathrm{tr}\{\mathrm{cov}(\mathbf{X}_o) + [\mathrm{cov}(\mathbf{X}_o)]^{-1}\ \mathrm{cov}(\mathbf{X}_o, \mathbf{X}_u)\ \mathrm{cov}(\mathbf{X}_o, \mathbf{X}_u)^{\mathrm{T}}\}. \tag{3.13}$$

Hence, A-optimal choices for the sites to be observed involve sites that not only have large variances but are also highly correlated with those sites that are unobserved. (We refer to variances and correlations given in $\mathbf{B}_1$.) The intuition seems clear; observing such sites both reduces our uncertainty at those sites and provides indirect information about the unobserved sites through the high correlation structure.

c. Other approaches

In the atmospheric science literature it is common to represent notions gauging quality of estimates in terms of ''analysis errors'' (here denoted by **a**) and ''forecast errors'' (here denoted by **e**). In particular, our expression (2.6) could be written as

$$\mathbf{e}_{01} \approx \mathbf{F}(\boldsymbol{\nu}_0)\mathbf{a}_0, \tag{3.14}$$

where the subscript 01 on **e** explicitly notes that we mean the forecast error in forecasting from t_0 to t_1; similarly, $\mathbf{a}_0$ is the analysis error at t_0. For the statistician, it is natural to describe formulations in terms of the probability distributions of these errors, rather than the errors themselves; that is, at t_0, we consider a probability model $\mathbf{a}_0 \sim N(0, \mathbf{A}_0)$ and $\mathbf{e}_{01} \sim N[(0, \mathbf{F}(\boldsymbol{\nu}_0)\mathbf{A}_0\mathbf{F}(\boldsymbol{\nu}_0)^{\mathrm{T}})]$.

At time t_0 the forecast error for $\mathbf{X}_2$ is $\mathbf{e}_{02}$. Based on the tangent linear approximation, the actual forecast is

$$\boldsymbol{\mu}_2^o = g(\boldsymbol{\mu}_1) = g[f(\boldsymbol{\nu}_0)]. \tag{3.15}$$

The corresponding forecast covariance is

$$\mathbf{B}_2^o = \mathbf{G}_0\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}, \tag{3.16}$$

where $\mathbf{G}_0$ is defined in (3.1). That is, we act as if

$$\mathbf{e}_{02} \sim N(0, \mathbf{B}_2^o). \tag{3.17}$$

1) Bishop and Toth (1996)

We will show that one suggestion of Bishop and Toth is to minimize (with the respect to $\mathbf{K}$) the largest ei-

genvalue of the (approximate) forecast covariance matrix $\mathbf{G}_0\mathbf{A}_1\mathbf{G}_0^{\mathrm{T}}$. In our terminology, they suggest an E-optimal design, for the approximate design as we described in section 2b(2).

Bishop and Toth propose minimizing, over all feasible designs, the quantity

$$\max_{\mathbf{a}_1}\left[\frac{\mathbf{e}_{12}^{\mathrm{T}}\mathbf{e}_{12}}{\mathbf{a}_1^{\mathrm{T}}\mathbf{A}_1^{-1}\mathbf{a}_1}\right]. \tag{3.18}$$

(Here $\mathbf{A}_1^{-1}$ is a function of $\mathbf{K}$.) Note that under the tangent linear approximation at $t = t_1$, we would act as if $\mathbf{e}_{12} = \mathbf{G}_0\mathbf{a}_1$. However, use of this fact creates a dilemma (we do not know $\boldsymbol{\nu}_1$ at t_0), as described in section 2b(2). Applying the approximation $\mathbf{e}_{12} \approx \mathbf{G}_0\mathbf{a}_1$, we can rewrite (3.18) as

$$\max_{\mathbf{a}_1}\left[\frac{\mathbf{a}_1^{\mathrm{T}}\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{a}_1}{\mathbf{a}_1^{\mathrm{T}}\mathbf{A}_1^{-1}\mathbf{a}_1}\right]. \tag{3.19}$$

Mathematically, the indicated maximization is equivalent to finding the largest solution $\lambda_1(\mathbf{K})$ to the eigenvalue problem

$$\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{a}_1 = \lambda\mathbf{A}_1^{-1}\mathbf{a}_1. \tag{3.20}$$

Setting $\mathbf{e}_{12} = \mathbf{G}_0\mathbf{a}_1$, (3.20) becomes

$$\mathbf{G}_0\mathbf{A}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{e}_{12} = \lambda\mathbf{e}_{12}. \tag{3.21}$$

In more recent work, Bishop and Toth also consider designs that minimize forecast variance, that is, A-optimal designs (C. Bishop 1998, personal communication).

2) PALMER ET AL. (1998)

In our notation, a simplified, though useful version of the formulation of Palmer et al. is as follows. They study features of the leading eigenvectors for the following problem:

$$\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{e}_{01} = \lambda\mathbf{B}_1^{-1}\mathbf{e}_{01}. \tag{3.22}$$

They advocate taking the adaptive observations at locations corresponding to the large absolute amplitudes of those leading eigenvectors.

Note that simple algebra implies that (3.22) is equivalent to

$$\mathbf{G}_0\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{e}_{01} = \lambda\mathbf{G}_0\mathbf{e}_{01}. \tag{3.23}$$

Next, we can rewrite (3.23) as

$$\mathbf{G}_0\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{e}_{02} = \lambda\mathbf{e}_{02}. \tag{3.24}$$

Hence, Palmer et al. focus on the forecast error $\mathbf{e}_{02}$ and its approximate covariance matrix $\mathbf{B}_2^0$ defined in (3.16). They suggest locating the leading eigenvectors (''worst forecast errors'') of $\mathbf{B}_2^0$. By reversing the above steps, this is related to finding the corresponding error $\mathbf{e}_{01}$ at time $t = t_1$. The idea is then that if one reduces the length of this ''error,'' one improves the structure of the forecast error covariance for $\mathbf{e}_{12}$. However, $\mathbf{B}_2^0$ is not the covariance for $\mathbf{e}_{12}$. Indeed, it is independent of the choice of adaptive observations design $\mathbf{K}$, whereas the statistical analysis yields criteria [see (2.20) and (2.22)] that depend on $\mathbf{K}$ through the forecast covariance matrix $\mathbf{B}_2$ [see (2.12) and (2.20)].

We note that Palmer et al. actually consider the case in which routine observations are also made available at or before time t_1. This coincides with our discussion in section 2c and appendix B. Indeed, to account for routine observations to be collected at time t_1, they would replace $\mathbf{B}_1$ in (3.21) by our $\mathbf{B}^*$ given in (B.13) of appendix B and then proceed as described above. Nevertheless, this enhancement still yields a criterion that does not depend on the adaptive design $\mathbf{K}$.

There is a second relationship between the statistical approach and that of Palmer et al. Let $\mathbf{v}^1$ be the leading eigenvector for (3.22). Algebra relates (3.22) and (3.7). Indeed, our small observation error solution [see (3.8)] is

$$\mathbf{K}_*^{\mathrm{T}} = \mathbf{B}_1^{-1}\mathbf{v}^1, \tag{3.25}$$

that is, a rotation of the vector Palmer et al. study. Note that if $\mathbf{B}_1^{-1}$ and $\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0$ have the similar eigenstructures or commute, $\mathbf{B}_1\mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0 = \mathbf{G}_0^{\mathrm{T}}\mathbf{G}_0\mathbf{B}_1$ (e.g., if $\mathbf{B}_1$ is a scalar multiple of the identity matrix), then $\mathbf{K}_*^{\mathrm{T}} \approx \mathbf{v}^1$. However, in general, the A-optimal $\mathbf{K}_*$ does *not* necessarily have large (absolute) elements at the same locations where $\mathbf{v}^1$ does. Further, even if $\mathbf{v}_1$ is highly localized, $\mathbf{K}_*$ need not be. At the extreme, if $\mathbf{v}^1$ is composed of all zeros save a single element equal to 1, then $\mathbf{K}_*$ is the corresponding column of $\mathbf{B}_1^{-1}$.

To further understand the meaning of (3.8), note that by definition of the solutions to (3.22) we may write the random forecast error $\mathbf{e}_{01} = \mathbf{X}_1 - \boldsymbol{\mu}_1$ [see (2.8)] as

$$\mathbf{e}_{01} = \sum_{i=1}^{n}\alpha_i\mathbf{v}^i, \tag{3.26}$$

where the α_i are random variables. By the $\mathbf{B}_1^{-1}$ orthogonality of the $\mathbf{v}^i$'s [i.e., $(\mathbf{v}^i)^{\mathrm{T}}\mathbf{B}_1^{-1}\mathbf{v}^j = \delta_{ij}$], we have that

$$\mathbf{K}_*\mathbf{e}_{01} = (\mathbf{v}^1)^{\mathrm{T}}\mathbf{B}_1^{-1}\mathbf{e}_{01} \tag{3.27}$$

$$= \sum_{i=1}^{n}\alpha_i(\mathbf{v}^1)^{\mathrm{T}}\mathbf{B}_1^{-1}\mathbf{v}^i \tag{3.28}$$

$$= \alpha_1. \tag{3.29}$$

Hence, our optimal choice for $\mathbf{K}$ implies that we are to measure the $\mathbf{B}_1^{-1}$ projection of $\mathbf{e}_{01}$ onto $\mathbf{v}_1$. This suggests that adaptive observation locations should be biased *away* from those suggested by Palmer et al.

3) LORENZ AND EMANUEL (1998)

Their suggestion is most directly related to our sections 2b(1) and 3b. To achieve good estimation of $\mathbf{X}_1$, they suggest taking adaptive observations of its coordinates having the largest variances in $\mathbf{B}_1$. Note that this

does not coincide with our A optimality suggestion in section 3b, unless $\mathbf{B}_1$ is a diagonal matrix. However, in the single-observation context, their approach can be related to statistical D optimality. This result does not hinge on a small observation error approximation.

Recall from section 2b(2) that the plug-in-approximated D-optimal designs for prediction of $\mathbf{X}_2$ are those that minimize $\det(\mathbf{A}_1)$. That is, under the approximation, dynamics beyond t_1 play no role; see (2.23). Next suppose $\mathbf{K}$ is a row vector whose elements are all zero, except for one element equal to 1. One can show that

$$\det(\mathbf{A}_1) = \frac{\sigma^2}{\sigma^2 + \mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}} \det\mathbf{B}_1. \tag{3.30}$$

Thus, the D-optimal design corresponds to that site that makes $\mathbf{K}\mathbf{B}_1\mathbf{K}^{\mathrm{T}}$ as large as possible, namely, the location with largest variance.

A fourth approach is suggested by Langland and Rohaly (1996). In order to identify locations where changes in the analysis at $t = t_1$ would produce large changes in the forecast of $\mathbf{X}_2$, they propose calculating the gradient w.r.t. conditions at time t_1 of some scalar function, say h, of the forecast variables. Formally, they calculate

$$\frac{\partial h}{\partial \mathbf{X}_1} = \mathbf{G}(\boldsymbol{\nu}_0)^{\mathrm{T}} \frac{\partial h}{\partial \mathbf{X}_2}.$$

They suggest locating observations at sites where $|\partial h/\partial \mathbf{X}_1|$ is large. Palmer et al. discuss the relationship between this strategy and those based on the eigenvectors of (3.22).

Finally, the calculations proposed by both Bishop and Toth (1996) and Palmer et al. (1998) require simplification in practice because of computational constraints. These simplifications center on modeling the various required covariance matrices and, in the case of Bishop and Toth, on estimating how those matrices change as K varies. We refer readers to the above papers for further details.

4. Example

Lorenz and Emanuel (1998) considered a low-order model that exhibits chaotic behavior that in some respects resembles that of the atmosphere. The model consists of variables $\mathbf{X}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^{\mathrm{T}}$ defined at n points and evolving according to

$$\frac{dx_i}{dt} = -x_{i-2}x_{i-1} + x_{i-1}x_{i+1} - x_i + F, \tag{4.1}$$

where F is a forcing and the variable x_i is taken to be periodic:

$$x_{i+n} = x_i. \tag{4.2}$$

One may visualize that the n dimensionless variables represent the values of some atmospheric quantity at n sites that are equally spaced about a latitude circle, indeed, a ''toy'' equator. The forcing F appears to control the complexity of the system. Here $F = 8$ gives rise to a fairly complex system. For values of F tending to 8, the system undergoes a bifurcation from periodic to chaotic behavior. For additional discussion of the model and its motivation, see Lorenz and Emanuel (1998).

a. Description of an experiment

A reference or true state of the system, which will be subsequently sampled in our experiment, is computed as follows. First, in all calculations we set $n = 40$ and $F = 8$. Next, we choose initial values $x_i = i/10.0$, for $1 \le i \le 39$ and $x_{40} = x_1$. The first 6000 time steps are discarded as transients; the next 1000 steps are saved as the reference states.

Following the dimensionalization of Lorenz and Emanuel, 1 day corresponds to nondimensional time intervals in (4.1) of duration 0.2; hereafter, all times will be dimensional. We also assume, as did Lorenz and Emanuel, that sites 21–40 represent land stations and sites 1–20 represent ocean stations. The land stations receive observations every 12 h; observations over the ocean are made intermittently as described below.

The Jacobian matrices needed at various stages of design are computed using the tangent linear model related to (4.1), which was derived by differentiating (4.1).

1) STEP 1

To start up the experiment, a first guess ($\boldsymbol{\nu}_0$) and associated covariance matrix for analysis errors ($\mathbf{A}_0$) at the present time t_0 are needed. To mimic a real analysis, we ran the following M iterations of data assimilations based on simulated data. At time step $t_{-1} = t_0 - 10M$ (M/2 days) prior to t_0, the true state is perturbed by adding independent Gaussian noises $N(0, \delta^2)$ at land sites and $N[0, (2\delta)^2]$ at ocean sites. This perturbed state is used as the best guess at that time. A hypothesized analysis error covariance matrix associated with this information is defined as follows: a diagonal matrix with variance $(2\delta)^2$ at ocean sites and δ^2 at land sites. At the next step $t_0 - 10(M - 1)$, this information is combined with simulated observations at all land stations as well as a single ocean site (described below). The observational errors are assumed to be independent and follow $N(0, \sigma_1^2)$ for land stations and $N(0, \sigma_2^2)$ for ocean. Data assimilation is computed at this stage using formulas in (2.9), (2.15), and (2.13). This procedure is repeated ($M - 1$) times up to time t_0.

The ocean site chosen in each step was that site that had the largest predictive variance (i.e., following the Lorenz and Emanuel approach). This feature of the spinup was introduced to reduce the possibility of producing dominant analysis variances at some ocean sites. Such sites could exert undue influence on subsequent adaptive designs. (As we will describe, despite our spinup, such dominant sites did occur.)

The point of this spinup strategy is to mimic the an-

TABLE 2. Various times for four cases. Note that the times are dimensional and measured in days. Day 1 is iterated 500 from the actual long run.

Case	Start-up time t_{-1}	Design time t_0	Observation time t_1	Forecast time t_2
1	1	11	11.5	12, 12.5, 13.5
2	2	12	12.5	13, 13.5, 14.5
3	6	16	16.5	17, 17.5, 18.5
4	7	17	17.5	18, 18.5, 19.5

ticipated structure of the adaptive observation setting, in which reasonably accurate, land-based data are available at the design stage. Specifically, the spinup introduces an unbalanced structure in the matrix A_0, in that sites over land should have much smaller variances than ocean sites.

2) STEP 2

At the present time t_0, we suppose there is an opportunity to observe one of the ocean sites, labeled 1–20, with error $N(0, \sigma_2^2)$, at time t_1. The goal here is to improve forecast at future time t_2.

We set $\delta = 0.3$, $M = 20$, $\sigma_1 = 0.5$, and $\sigma_2 = 0.5$. Four experiments were run at different times, within the same long run of the system. Table 2 provides design time, adaptive observation time, and three choices of prediction times for each of the four cases.

There are two additional clarifications. First, the adaptive designs followed the formulation of section 2a. In particular, while we used routine observations during spinup, we did not adjust for routine observations at the observation times such as the analyses outlined in section 2c. Second, all four spinups involved independent data. For example, the spinup for case 2 made no use of data from the spinup of case 1.

b. Results

To provide some intuition regarding results, Fig. 1 presents information summarizing the expected state of the system at time t_1 based on information available at time t_0 for case 2. (Figure 2 contains the same information for case 4.) Specifically, Fig. 1a shows the true value of $\mathbf{X}_1$ along with its prediction $\boldsymbol{\mu}_1$. Figure 1b presents the diagonal elements of the corresponding $\mathbf{B}_1$. Also based on $\mathbf{B}_1$, the third panel (Fig. 1c) provides information about the covariance structure for neighboring sites, that is, covariances of forecast errors between site x_i and $x_{i+\tau}$ where $\tau = 1$, 2, or 3 is plotted. Results for case 4 appear in Fig. 2.

Figure 3 shows the values of the average of the variances in case 1 as a function of adaptive observation site, corresponding to the data assimilation as well as three forecast times. The same information for cases 2–4 is reported in Figs. 4–6, respectively. In each of these four problems, our *A*-optimal solution corresponds to the site yielding the smallest average variance. Resulting optimal designs are summarized in Table 3.

The first observation in these results is that for this small experiment, A optimality and the strategies of Lorenz and Emanuel (1998) and Palmer et al. (1998) tend to agree if one seeks optimal estimation of $\mathbf{X}_1$. (An exception occurs in case 3, though the improvements of A optimality in expected average mean-squared estimation error appear modest.) Differences in results are more evident for forecasting. This is certainly plausible in comparison to Lorenz and Emanuel, since their suggestion does not ''look ahead'' via $\mathbf{G}_0$. Note that in constructing a 2-day forecast for case 1, the A-optimality solution suggests a relative savings of 19.4% [$(0.618 - 0.498)/0.618$] in predictive mean-squared error over the best estimation at t_1 approach. In case 3, for 2-day forecasting, we see a relative savings of 31.4% in using the A-optimal design versus that suggested by Palmer et al., while the savings over Lorenz and Palmer is only 8.4%. In case 4, A optimality showed a 24.5% relative savings over the other approaches for a 2-day forecast.

It is interesting to inquire about impacts of designing for a particular forecast time. That is, are designs for forecasting one day ahead at least reasonable for forecasting for less than one day ahead? To examine this question, first consider case 1. The A-optimal solution for 2-day forecasting is site 7. This site is also A optimal for a 1-day forecast, but we observe a relative savings loss of 3.6% in terms of average prediction variance compared to the best (site 9) for a half-day forecast. In case 2, the A-optimal (and Palmer et al.) procedure for 2-day forecasting is site 7, for which we expect an average prediction variance of 0.148 for a half-day lead. (Site 7 is also optimal for a 1-day lead.) The corresponding optimal value of 0.145 (for site 12) leads to a relative loss of only 2%. Case 3 also suggests little regret in designing for 2-day forecasts. However, in case 4, using site 17 gives relative savings loss of 30.7% for a half-day forecast. (Site 17 is optimal for 1-day forecasting.) Case 4 leaves the issue unsettled. If the tangent linear approximations driving the A-optimal 2-day solution are poor approximations to the true dynamics, then obviously the solution is not robust. On the other hand, if the tangent linearization is reasonable for 2 days, then designing only for the very short term is inefficient.

A final important point concerns anticipated differences in results for various approaches. We suggest that in the presence of sites with relatively large variances (diagonal elements) in $\mathbf{B}_1$, any sensible approach to adaptive observation selection for either estimation of $\mathbf{X}_1$ or short-term forecasting will suggest observing those sites. The situation is clearest in case 4. In that example, site 12 has an extraordinarily large ($\mathbf{B}_1$) variance (see Fig. 2). This appears to dominate Palmer et al. in this case, despite the use of the same dynamics

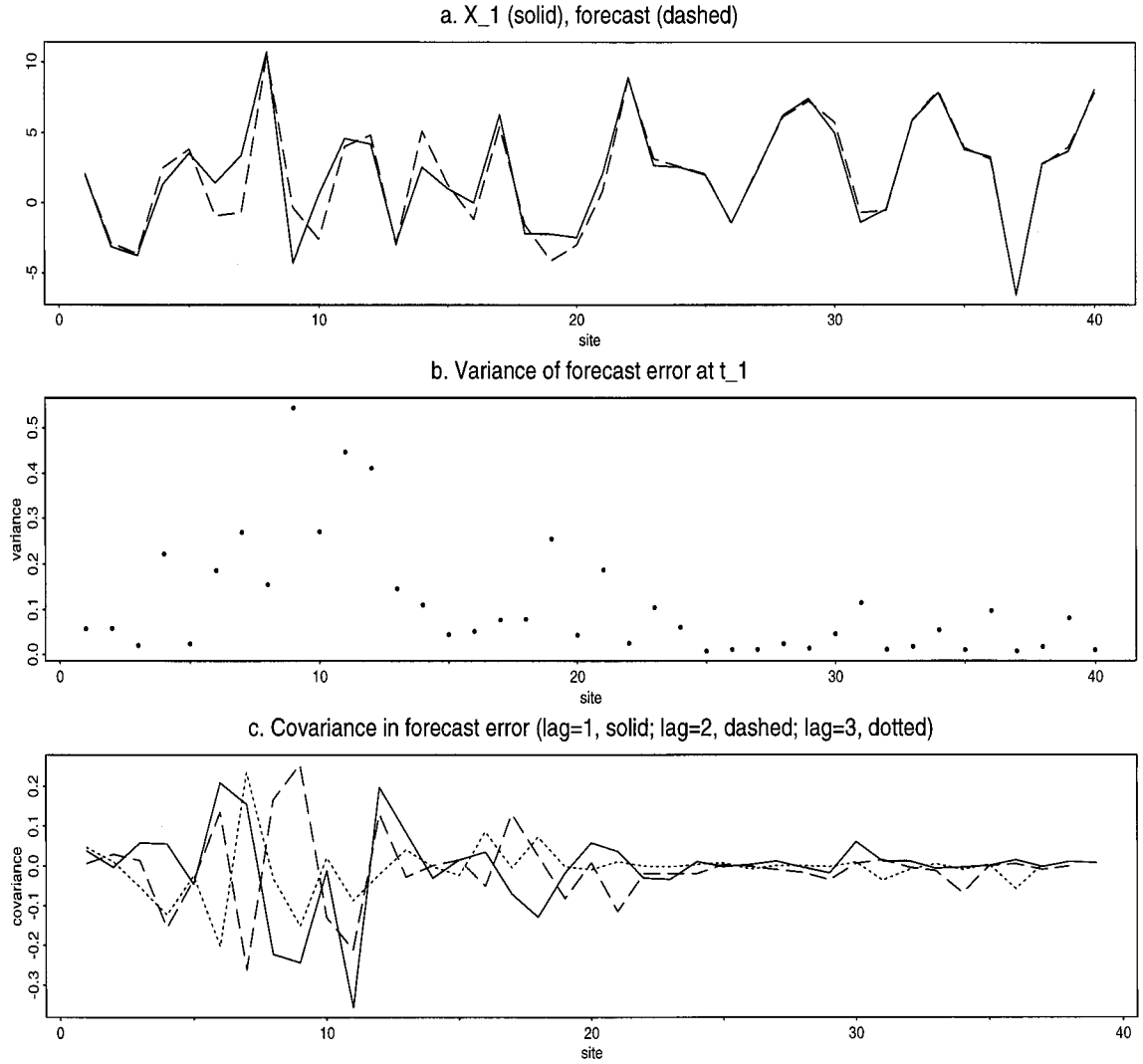

FIG. 1. Predicted field ($\boldsymbol{\mu}_1$) and covariance ($\mathbf{B}_1$) for case 2.

for long-term forecasting as in the A-optimal approach. Of course, $\mathbf{B}_1$ variance determines the Lorenz and Emanuel suggestion, but also appears to dominate A optimality for short-term forecasting. This may seem at odds with our section 3b, where we suggested that both variance and correlation drive the A-optimal solution. That claim is correct; this system, however, produces highly oscillatory sample paths and spatial correlation patterns in $\mathbf{B}_1$, leading to a diminished role for spatial correlation.

5. Discussion

a. Summary

While there has been much recent interest in the problem of adaptive observations, all investigations to date begin from the intuitive notion that adaptive observations seek to improve the quality of the analysis or the skill of subsequent forecasts. Our fundamental contribution is a rigorous statistical formulation of the adaptive design problem. This formulation provides a precise mathematical framework for further understanding of the importance of various components of the problem, such as the uncertainty of the current analysis, the dynamics of error evolution, the form and errors of observations, data assimilation, and the choice of criterion to be optimized.

We discuss properties of optimal statistical solutions in a variety of contexts. To enable calculation of optimal solutions, several approximations were used. First, criteria to be optimized may depend on (expected) forecasts from future times. At the time of design such forecasts are unknown (since they depend on data not yet observed). Such forecasts were replaced by longer-lead forecasts from the time of adaptive design. Second, analysis and forecast errors are assumed to be Gaussian and to evolve linearly in time. (As discussed below, neither

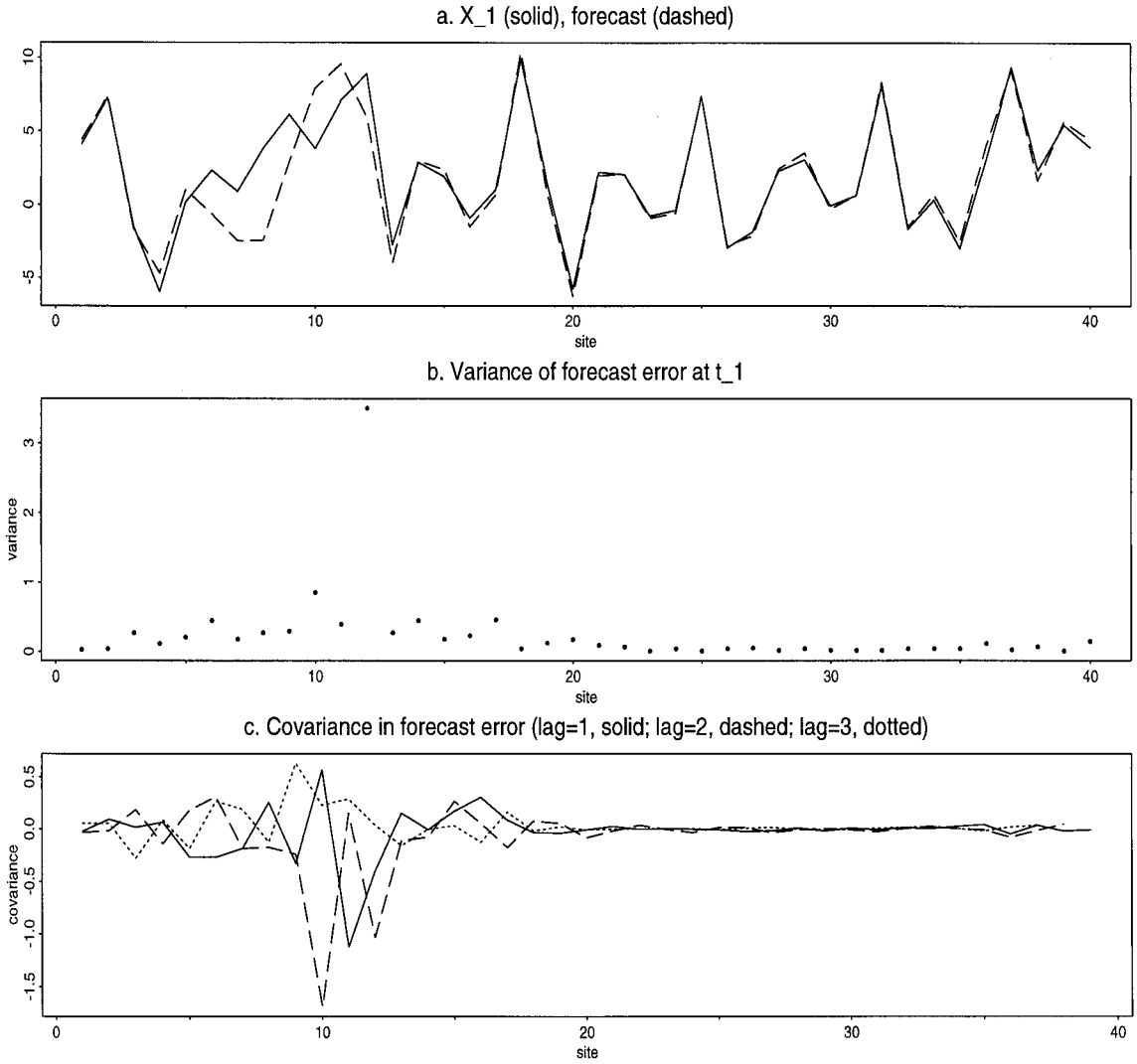

FIG. 2. Predicted field ($\boldsymbol{\mu}_1$) and covariance ($\mathbf{B}_1$) for case 4.

of the Gaussian nor linear assumptions are required in the formulation of the problem.) For a single, accurate observation, the optimal observation is that which ''observes'' the projection onto errors at the current time (in either an analysis or a short-term forecast) of the structure that will evolve subsequently into the leading eigenvector of the forecast error covariance matrix at the desired forecast time.

The relationships of other proposed approaches to optimal statistical solutions are discussed in general and in the context of a numerical example involving the 40-variable model of Lorenz and Emanuel (1998). This example suggests that other proposed approaches may often be close to a statistically optimal solution, at least for this highly idealized model. However, this may not be representative of results in more general situations.

Further improvement in adaptive observing strategies over the heuristic techniques employed to date (such as in FASTEX) will require progress in several areas. First, it is clear that much of the subtlety of the problem arises when the covariance matrices for the analysis and forecast have nontrivial structure. Understanding and estimating this structure are difficult and are the subject of much current research in data assimilation. For adaptive observations, there is the additional difficulty that, in practice, data assimilation schemes only approximate the optimal estimates [(2.12)–(2.15)] assumed here and the covariances in question will depend on the specific scheme employed to assimilate data. Even given good estimates of the covariance matrices, work will remain to be done in computing the optimal design, especially incorporating the constraints and form of existing observational platforms.

Practical implementation of adaptive strategies also awaits the advent of novel observing techniques and technologies to replace the use of manned aircraft. In-

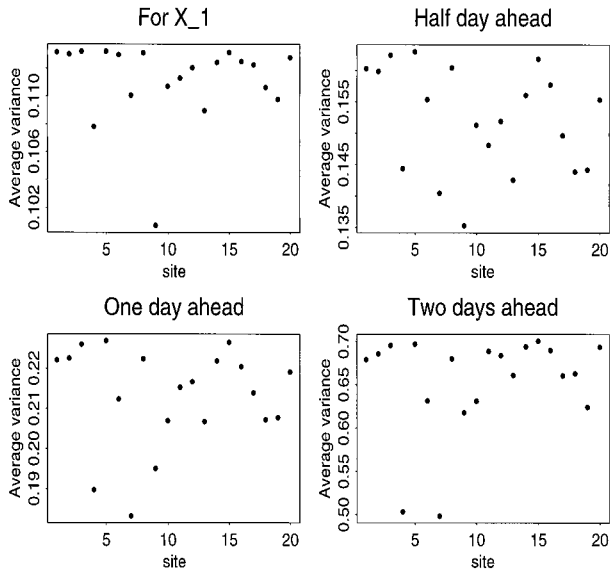

FIG. 3. Average predictive variances in case 1.

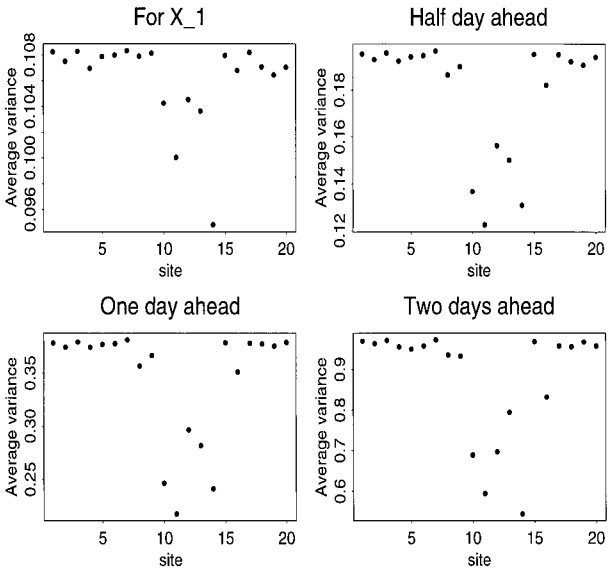

FIG. 5. Average predictive variances in case 3.

deed, the framework presented here may have its soonest practical application in the adaptation of existing systems such as the radio-sounding network.

b. Assumptions

Our statistical formulations rely heavily on selection of the mean-squared error criterion to gauge predictive accuracy, tangent linear approximations, and knowledge of analysis error covariance matrices (this aspect was discussed in section 5a), and appear to also rely on Gaussian distributional assumptions. Extremely important issues involve the impacts of departures from these assumptions. A complete analysis is beyond the scope of this article, though the following perspectives merit discussion.

First, the reader should not equate ''statistical design'' with linearization plus Gaussian assumptions. The latter are simply a particular framework that is familiar and in which we can readily construct criteria for optimization. If these assumptions are deemed to be untenable, one would construct alternative procedures for constructing approximate predictive distributions (e.g., ensemble forecasting–based ideas; see Bishop and Toth 1996). Similarly, our criteria were motivated by mean-squared prediction error minimization. If this criterion

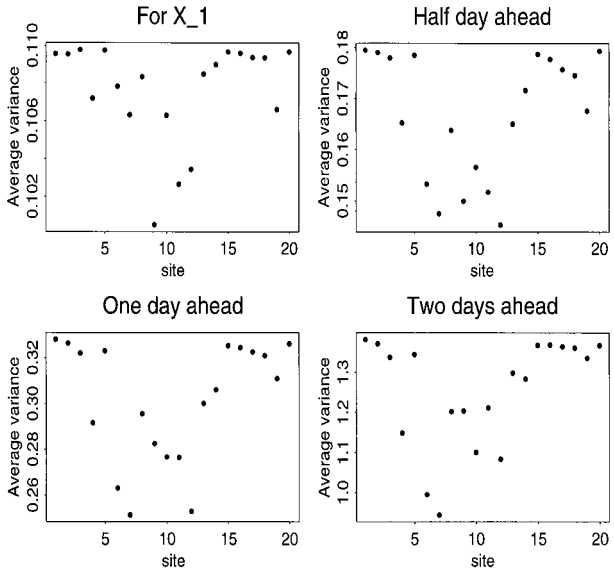

FIG. 4. Average predictive variances in case 2.

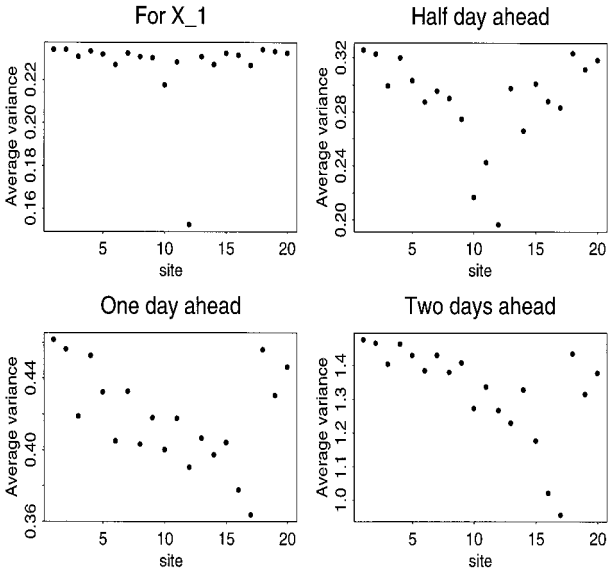

FIG. 6. Average predictive variances in case 4.

TABLE 3. Optimal design example results. For optimal estimation at time t_1 and the three forecast times, the optimal sites are indicated for three methods: A optimality, Palmer et al., and Lorenz and Emanuel. The numbers in parentheses are the average prediction variance achieved for the corresponding design.

Case	Forecast time (days)	A optimal	Palmer et al.	Lorenz and Emanuel
1	0	9 (0.101)	9 (0.101)	9 (0.101)
	0.5	9 (0.135)	9 (0.135)	9 (0.135)
	1	7 (0.183)	9 (0.195)	9 (0.195)
	2	7 (0.498)	4 (0.503)	9 (0.618)
2	0	9 (0.101)	9 (0.101)	9 (0.101)
	0.5	12 (0.145)	7 (0.148)	9 (0.150)
	1	7 (0.251)	12 (0.253)	9 (0.283)
	2	7 (0.944)	7 (0.944)	9 (1.203)
3	0	14 (0.095)	13 (0.104)	11 (0.100)
	0.5	11 (0.123)	13 (0.150)	11 (0.123)
	1	11 (0.218)	13 (0.282)	11 (0.218)
	2	14 (0.544)	13 (0.794)	11 (0.594)
4	0	12 (0.152)	12 (0.152)	12 (0.152)
	0.5	12 (0.196)	12 (0.196)	12 (0.196)
	1	17 (0.364)	12 (0.390)	12 (0.390)
	2	17 (0.956)	12 (1.267)	12 (1.267)

is not considered primary, the statistical method could be applied to whatever is viewed more appropriate (at least in principle).

Second, suppose one is interested in relaxing the Gaussian assumptions, but maintaining linearization. Then analyses hinge on what choices are made for the predictive covariance matrices, not the assumptions made to derive them. This article compared results based on linearization-based covariances. If one wishes to ignore all Gaussian assumptions and rather view these derivations as an extended Kalman filter, so be it. The comparisons are still valid. However, if the Gaussian approximation breaks down due to multimodality or nonnormal tails in the predictive distributions, the proper issue for examination is the potential meaninglessness of any analysis based on means and covariances. These objects can be poor summaries of such nonnormal distributions.

Acknowledgments. We are grateful to R. Gall, K. Emanuel, R. Errico, E. Lorenz, D. Nychka, T. Palmer, and three referees for their comments and suggestions. NCAR is sponsored by the National Science Foundation. This research was partially supported by the NCAR/NSF Geophysical Statistics Project.

APPENDIX A

Statistical Design of Experiments

The statistical design of experiments is a fundamental subdiscipline of statistics. Owing to seminal work of R. A. Fisher early in the twentieth century and extended by G. E. P. Box and others, statistical design has become a stalwart in many scientific endeavors. The idea is to design reasonable, informative experiments that can be of value when one wishes to ''learn'' about the phenomenon at hand. In combination with mathematical propositions of design problems, a tradition of optimal experimental design developed from early contributions of J. Kiefer and V. V. Fedorov. (These two notions are not mutually exclusive, of course.)

In the optimal design mode, the development of a design criterion is coupled with choice of an estimation or prediction rule. Indeed, designs are usually developed assuming one will use optimal estimators or predictors. (Recall how the use of conditional expectations as predictors arose in the MSE example of section 2.) A statistician's tenet is that to design an optimal experiment, one must know what will be done with the data obtained.

In the Bayesian statistics tradition (Bernardo and Smith 1994), all quantities are treated as if they are random variables. Hence, Bayesian experimental design is virtually identical to design for prediction. Our use of the modifier ''Bayesian'' in this paper could easily be replaced by ''stochastic'' or ''probabilistic.'' In reading the literature the choice of what to call the type of analysis varies considerably, suggesting that there are many approaches to the problem. However, nearly all formal approaches involve the principle indicated above, in concert with the statistical view described above.

For general introductions to optimal statistical design of experiments, see Fedorov (1972), Pukelsheim (1993), and Silvey (1980). See Chaloner and Verdinelli (1995) for a review of Bayesian experimental design. See Aitchison and Dunsmore (1975) for background discussion of statistical prediction. Also, see Ford et al. (1989) for review of special issues in experimental design in nonlinear contexts.

Design in the context of prediction of stochastic processes also has a history. A related topic is the *sequential* design of experiments, referring to designing a sequence of experiments. General background and further references can be found in Chernoff (1972) and Titterington (1980). Beyond the statistics and probability literature, extensive use and development of these ideas can be found in various disciplines. For example, as a referee of this paper pointed out, techniques of statistical design analogous to those presented here are discussed in some engineering literatures; see El-Jai and Pritchard (1988), Fedorov and Müller (1989), and Omatu and Seinfeld (1989) for pertinent discussion and review.

In practice, design of experiments is based on optimization with respect to a specified class of possible experiments. While our focus here is on limitations dictated by adaptive observation procedures, the principles could be applied to the design of all data collection procedures, including monitoring systems. Further, analyses balancing economic costs of data collection combined with computational overhead versus economic value of improved weather forecasts can be pursued in

principle, though such suggestions are beyond the scope of this paper.

Criteria for optimal designs

Recall that best MSE-based prediction let to optimization of (expected) trace of the conditional covariance matrix of $\mathbf{X}$, say $\mathbf{B}$. More generally, a variety of criteria aimed at making $\mathbf{B}$ ''small'' have been studied. Of course, $\mathbf{B}$ is a matrix so the notion of the best (expected) $\mathbf{B}$ cannot be defined. To make the problem mathematically meaningful, some criterion function, generically denoted by $\mathcal{F}$, taking $\mathbf{B}$ into a scalar must be specified. We present a brief overview here. [See Chaloner and Verdinelli (1995) and Silvey (1980) for in depth reviews; the following presentation relied heavily on that of Silvey.] As we develop these we suspend constantly writing that we optimize expected values (with respect to the data $\mathbf{Y}$) of these criteria.

1) D OPTIMALITY

Suppose rather than simply predicting $\mathbf{X}$, we wish to provide a prediction region: that is, we are to calculate a set such that, conditional on $\mathbf{y}$, the probability that $\mathbf{X}$ is in that region is some specified value, say P. (Note the parallel to ''confidence intervals'' in estimation.) Typically, we wish to choose the region as small as possible while satisfying the probability condition. Under the Gaussian assumptions used here, the desired prediction region takes the form of an ellipsoid in n dimensions, given by

$$\{\mathbf{x} : (\mathbf{x} - E(\mathbf{X}|\mathbf{y}))^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - E(\mathbf{X}|\mathbf{y})) \leq c_P\}, \quad \text{(A.1)}$$

where c_P is a constant. Geometry tells us that the volume of this ellipsoid is proportional to the square root of the *determinant* of $\mathbf{B}$. Designs that minimize the determinant of $\mathbf{B}$ are D optimal.

2) G AND E OPTIMALITY

Next, suppose we actually plan to predict a particular linear combination of $\mathbf{X}$, say $\mathbf{L}^{\mathrm{T}}\mathbf{X}$, for some vector $\mathbf{L}$. The predictive variance is readily found to be $\mathbf{L}^{\mathrm{T}}\mathbf{B}\mathbf{L}$. We would find designs that minimize this quantity. (This is sometimes called C optimality.) A generalization is to consider a collection of interesting $\mathbf{L}$ and minimize maximum predictive variance over that class. This is G optimality. If the class is chosen to be the set of all n vectors satisfying $\mathbf{L}^{\mathrm{T}}\mathbf{L} = 1$, we refer to the approach as E optimality. Note that an E-optimal design minimizes the largest eigenvalue of $\mathbf{B}$.

3) A OPTIMALITY

Rather than the ''minimax'' notion of G optimality, we may wish to minimize an average prediction variance. Formally, we impose a probability distribution on $\mathbf{L}$. We then would find designs to minimize $E(\mathbf{L}^{\mathrm{T}}\mathbf{B}\mathbf{L})$, where this expectation is taken with respect to the distribution on $\mathbf{L}$. A mathematical fact implies that for any distribution on $\mathbf{L}$,

$$E(\mathbf{L}^{\mathrm{T}}\mathbf{B}\mathbf{L}) = \mathrm{tr}[\mathbf{B}E(\mathbf{L}\mathbf{L}^{\mathrm{T}})], \quad \text{(A.2)}$$

where tr indicates the trace operation. Designs minimizing the right-hand side of (A.2) are A optimal. Note that the minimization of tr($\mathbf{B}$), without the weights indicated in (A.2.), can be motivated as minimizing the simple average of the predictive variances of the individual components of $\mathbf{X}$ (i.e., the diagonal elements of $\mathbf{B}$ are these variances).

TABLE B1. Guide to key definitions for case 1.

	Expected value (analysis or forecast)	Covariance
At time t_0		
To estimate $\mathbf{X}_0$	$\boldsymbol{\nu}_0$ (input)	A_0 (input)
To forecast $\mathbf{X}_\alpha$	$\boldsymbol{\mu}_\alpha$ (B.2)	B_α (B.3)
To forecast $\mathbf{X}_1$	$\boldsymbol{\mu}_1$ (2.8)	B_1 (2.9)
At time t_α, after $\mathbf{Y}_\alpha$		
To estimate $\mathbf{X}_\alpha$	$\boldsymbol{\nu}_\alpha$ (B.5)	A_α (B.4)
To forecast $\mathbf{X}_1$	$\boldsymbol{\mu}_1^\alpha$ (B.6)	B_1^α (B.7)
At time t_1, after observation		
To estimate $\mathbf{X}_1$	$\boldsymbol{\nu}_1^\alpha$ (B.9)	A_1^α (B.8)
To forecast $\mathbf{X}_2$	$\boldsymbol{\mu}_2^\alpha$ (B.11)	B_2^α (B.12)

APPENDIX B

Technical Results for Alternative Data Collection Procedures

a. Case 1

At time t_α, $t_0 < t_\alpha < t_1$, we will observe a dataset, $\mathbf{Y}_\alpha$. Analogous to (2.10), assume that

$$\mathbf{Y}_\alpha = \mathsf{M}\mathbf{X}_\alpha + \boldsymbol{\epsilon}_\alpha, \quad \text{(B.1)}$$

where M is a $r \times n$ matrix. Assume that the measurement error vector $\boldsymbol{\epsilon}_\alpha \sim N(0, \boldsymbol{\Upsilon})$ and that the error vectors $\boldsymbol{\epsilon}$ [see (2.10)] and $\boldsymbol{\epsilon}_\alpha$ are independent of each other.

The following calculations are summarized in Table B1. At time t_0 we compute a linearization [analogous to (2.3)] for the distribution of $\mathbf{X}_\alpha$. Assuming model dynamics $\mathbf{X}_\alpha = f_\alpha(\mathbf{X}_0)$ and defining $\mathsf{F}_\alpha(\boldsymbol{\nu}_0)$ to be the appropriate Jacobian, we have $\mathbf{X}_\alpha \sim N(\boldsymbol{\mu}_\alpha, \mathsf{B}_\alpha)$, where

$$\boldsymbol{\mu}_\alpha = f_\alpha(\boldsymbol{\nu}_0) \quad \text{(B.2)}$$

and

$$\mathsf{B}_\alpha = \mathsf{F}_\alpha(\boldsymbol{\nu}_0)\mathsf{A}_0\mathsf{F}_\alpha(\boldsymbol{\nu}_0)^{\mathrm{T}}. \quad \text{(B.3)}$$

[Compare these expressions to (2.8) and (2.9).] We then update based on the observed data $\mathbf{y}_\alpha$, that is, conditional on $\mathbf{Y}_\alpha = \mathbf{y}_\alpha$, $\mathbf{X}_\alpha|\mathbf{y}_\alpha \sim N(\boldsymbol{\nu}_\alpha, \mathsf{A}_\alpha)$, where

$$\mathsf{A}_\alpha = (\mathsf{B}_\alpha^{-1} + \mathsf{M}^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}\mathsf{M})^{-1} \quad \text{(B.4)}$$

and

$$\boldsymbol{\nu}_\alpha = \mathbf{A}_\alpha(\mathbf{B}_\alpha^{-1}\boldsymbol{\mu}_\alpha + \mathbf{M}^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}\mathbf{y}_\alpha). \quad \text{(B.5)}$$

Next, our predictive distribution, conditional on the data $\mathbf{y}_\alpha$, for $\mathbf{X}_1$ can be approximated by employing a linearization (from time t_α). Assuming that $\mathbf{X}_1 = f_{1-\alpha}(\mathbf{X}_\alpha)$ and defining $\mathbf{F}_{1-\alpha}(\boldsymbol{\nu}_\alpha)$ to be the appropriate Jacobian, we have that $\mathbf{X}_1|\mathbf{y}_\alpha \sim N(\boldsymbol{\mu}_1^\alpha, \mathbf{B}_1^\alpha)$, where

$$\boldsymbol{\mu}_1^\alpha = f_{1-\alpha}(\boldsymbol{\nu}_\alpha) \quad \text{(B.6)}$$

and

$$\mathbf{B}_1^\alpha = \mathbf{F}_{1-\alpha}(\boldsymbol{\nu}_\alpha)\mathbf{A}_\alpha\mathbf{F}_{1-\alpha}(\boldsymbol{\nu}_\alpha)^{\mathrm{T}}. \quad \text{(B.7)}$$

To set up the design step, compute A_1^α using (2.13) with all B_1 there replaced by B_1^α:

$$\mathbf{A}_1^\alpha = [(\mathbf{B}_1^\alpha)^{-1} + \mathbf{K}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{K}]^{-1}. \quad \text{(B.8)}$$

The analyzed field $\boldsymbol{\nu}_1^\alpha$ is computed based on (2.15):

$$\boldsymbol{\nu}_1^\alpha = \boldsymbol{\mu}_1^\alpha - \mathbf{B}_1^\alpha\mathbf{K}^{\mathrm{T}}(\boldsymbol{\Sigma} + \mathbf{K}\mathbf{B}_1^\alpha\mathbf{K}^{\mathrm{T}})^{-1}(\mathbf{K}\boldsymbol{\mu}_1^\alpha - \mathbf{y}). \quad \text{(B.9)}$$

To form the conditional distribution of $\mathbf{X}_2$ given $\mathbf{y}_\alpha$ and $\mathbf{y}$, follow the derivation of (2.18):

$$\mathbf{X}_2|\mathbf{y}_\alpha, \mathbf{y} \sim N(\boldsymbol{\mu}_2^\alpha, \mathbf{B}_2^\alpha), \quad \text{(B.10)}$$

where

$$\boldsymbol{\mu}_2^\alpha = g(\boldsymbol{\nu}_1^\alpha) \quad \text{(B.11)}$$

and

$$\mathbf{B}_2^\alpha = \mathbf{G}(\boldsymbol{\nu}_1^\alpha)\mathbf{A}_1^\alpha\mathbf{G}(\boldsymbol{\nu}_1^\alpha)^{\mathrm{T}}. \quad \text{(B.12)}$$

The matrix $\mathbf{G}$ is the collection of first partials of the transformation g; in (2.20) $\mathbf{G}$ is evaluated at $\boldsymbol{\nu}_1$.

Note that in selecting adaptive observations with the intent of achieving optimal estimation of $\mathbf{X}_1$, we encounter an extra difficulty. Because of the extra linearization step at time t_α, $\mathbf{A}_1^\alpha$ is a function of the routine, yet unobserved, observations $\mathbf{y}_\alpha$. [This is similar to the situation in section 2b(2).] In principle, design criteria would be based on expectations over both these routine observations. Approximations can be found by analog to the idea in section 2b(2). Further, in design for optimal prediction at time t_2, the relevant covariance matrix $\mathbf{B}_2^\alpha$ depends (in a complicated fashion) on both the routine and adaptive data.

b. Case 2

To incorporate both datasets, we find formulas corresponding to (2.13) and (2.15). This is a standard problem in Bayesian analysis. The easiest way of representing the answer is to 1) replace $\mathbf{B}_1$ and $\boldsymbol{\mu}_1$ everywhere in (2.13) and (2.15) by

$$\mathbf{B}^* = (\mathbf{B}_1^{-1} + \mathbf{M}^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}\mathbf{M})^{-1} \quad \text{(B.13)}$$

and

$$\boldsymbol{\mu}^*{}_1 = \mathbf{B}^*(\mathbf{B}_1^{-1}\boldsymbol{\mu}_1 + \mathbf{M}^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}\mathbf{y}_\alpha), \quad \text{(B.14)}$$

and then 2) proceed as in section 2a. The idea is that we formally can first assimilate the routine data, leading to (B.13) and (B.14), and then assimilate the adaptive observations as in section 2a. Justification of this is an argument in probability theory. This solution does hinge on the assumption that the measurement errors in these datasets are independent.

TABLE B2. Guide to key definitions for case 3.

	Expected value (analysis or forecast)	Covariance
At time t_0		
To estimate $\mathbf{X}_0$	$\boldsymbol{\nu}_0$ (input)	$\mathbf{A}_0$ (input)
To forecast $\mathbf{X}_1$	$\boldsymbol{\mu}_1$ (2.8)	$\mathbf{B}_1$ (2.9)
At time t_1, after observation		
To estimate $\mathbf{X}_1$	$\boldsymbol{\nu}_1$ (2.15)	$\mathbf{A}_1$ (2.13)
To forecast $\mathbf{X}_\beta$	$\boldsymbol{\mu}_\beta$ (B.16)	$\mathbf{B}_\beta$ (B.17)
At time t_β, after observation		
To estimate $\mathbf{X}_\beta$	$\boldsymbol{\nu}_\beta$ (B.19)	$\mathbf{A}_\beta$ (B.18)
To forecast $\mathbf{X}_2$	$\boldsymbol{\mu}_2^\beta$ (B.20)	$\mathbf{B}_2^\beta$ (B.21)

c. Case 3

Assume that the routine observations to be observed at t_β follows the model

$$\mathbf{Y}_\beta = \mathbf{M}\mathbf{X}_\beta + \boldsymbol{\varepsilon}_\beta, \quad \text{(B.15)}$$

where $\mathbf{M}$ is a $r \times n$ matrix and $\boldsymbol{\varepsilon}_\beta \sim N(0, \boldsymbol{\Upsilon})$. (There is no reason that $\mathbf{M}$, r, and $\boldsymbol{\Upsilon}$ need not be the same specifications as in case 1.)

The following calculations are summarized in Table B2. For optimal prediction of $\mathbf{X}_2$, first propagate (2.11) forward in time to t_β, yielding the approximate distribution of $\mathbf{X}_\beta$ given $\mathbf{y}$. To do this assume that $\mathbf{X}_\beta = g_\beta(\mathbf{X}_1)$ and define $\mathbf{G}_\beta(\boldsymbol{\mu}_1)$ to be the appropriate Jacobian. It follows that $\mathbf{X}_\beta|\mathbf{y} \sim N(\boldsymbol{\mu}_\beta, \mathbf{B}_\beta)$, where

$$\boldsymbol{\mu}_\beta = g_\beta(\boldsymbol{\nu}_1) \quad \text{(B.16)}$$

and

$$\mathbf{B}_\beta = \mathbf{G}_\beta(\boldsymbol{\nu}_1)\mathbf{A}_1\mathbf{G}_\beta(\boldsymbol{\nu}_1)^{\mathrm{T}}. \quad \text{(B.17)}$$

After observing $\mathbf{y}_\beta$, this distribution would be updated to yield

$$\mathbf{A}_\beta = [(\mathbf{B}_\beta)^{-1} + \mathbf{M}^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}\mathbf{M}]^{-1} \quad \text{(B.18)}$$

and analyzed field

$$\boldsymbol{\nu}_\beta = \boldsymbol{\mu}_\beta - \mathbf{B}_\beta\mathbf{M}^{\mathrm{T}}(\boldsymbol{\Upsilon} + \mathbf{M}\mathbf{B}_\beta\mathbf{M}^{\mathrm{T}})^{-1}(\mathbf{M}\boldsymbol{\mu}_\beta - \mathbf{y}_\beta). \quad \text{(B.19)}$$

Finally, assume $\mathbf{X}_2 = g_{2-\beta}(\mathbf{X}_\beta)$ and let $\mathbf{G}_{2-\beta}$ denote the Jacobian of $g_{2-\beta}$, evaluated at (B.19). Then $\mathbf{X}_2|\mathbf{y}, \mathbf{y}_\beta \sim N(\boldsymbol{\mu}_2^\beta, \mathbf{B}_2^\beta)$, where

$$\boldsymbol{\mu}_2^\beta = g_{2-\beta}(\boldsymbol{\nu}_\beta) \quad \text{(B.20)}$$

and

$$\mathbf{B}_2^\beta = \mathbf{G}_{2-\beta}\mathbf{A}_\beta\mathbf{G}_{2-\beta}^{\mathrm{T}}. \quad \text{(B.21)}$$

REFERENCES

Aitchison, J., and I. R. Dunsmore, 1975: *Statistical Prediction Analysis.* University Press, 273 pp.

Bernardo, J. M., and A. F. M. Smith, 1994: *Bayesian Theory*. Wiley, 604 pp.
Bishop, C., and Z. Toth, 1996: Using ensembles to identify observations likely to improve forecasts. Preprints, *11th Conf. on Numerical Weather Prediction*, Norfolk, VA, Amer. Meteor. Soc., 72–74.
Chaloner, K., and I. Verdinelli, 1995: Bayesian experimental design: A review. *Stat. Sci.,* **10,** 273–304.
Chernoff, H., 1972: *Sequential Analysis and Optimal Design.* Society for Industrial and Applied Mathematics, 119 pp.
Courtier, P., 1997: Dual formulation of four-dimensional variational assimilation. *Quart. J. Roy. Meteor. Soc.,* **123,** 2449–2461.
Ehrendorfer, M., and J. J. Tribbia, 1997: Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.*, **54,** 286–313.
El-Jai, A., and A. J. Pritchard, 1988: *Sensors and Controls in the Analysis of Distributed Systems.* John Wiley and Sons, 125 pp.
Fedorov, V. V., 1972: *Theory of Optimal Experiments.* Academic Press, 292 pp.
——, and W. Müller, 1989: Comparison of two approaches in the optimal design of an observation network. *Statistics,* **19,** 339–351.
Ford, I., D. M. Titterington, and C. P. Kitsos, 1989: Recent advances in nonlinear experimental design. *Technometrics,* **31,** 49–60.
Joly, A., and Coauthors, 1997: The Fronts and Atlantic Storm-Track Experiment (FASTEX): Scientific objectives and experimental design. *Bull. Amer. Meteor. Soc.,* **78,** 1917–1940.
Langland, R., and G. Rohaly, 1996: Adjoint-based targeting of observations for FASTEX cyclones. Preprints, *Seventh Conf. on Mesoscale Process*, Reading, United Kingdom, Amer. Meteor. Soc., 369–371.
Lorenc, A. C., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.,* **112,** 1177–1194.
Lorenz, E. N., and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulations with a small model. *J. Atmos. Sci.,* **55,** 399–414.
Omatu, S., and J. H. Seinfeld, 1989: *Distributed Parameter Systems: Theory and applications.* Oxford University Press, 430 pp.
Palmer, T. N., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics, and adaptive observations. *J. Atmos. Sci.,* **55,** 633–653.
Pukelsheim, F., 1993: *Optimal Design of Experiments.* Wiley, 480 pp.
Rao, C. R., 1973: *Linear Statistical Inference and Its Applications.* 2d ed. Wiley, 656 pp.
Silvey, S. D., 1980: *Optimal Design.* Chapman and Hall, 86 pp.
Snyder, C., 1996: Summary of an informal workshop on adaptive observations and FASTEX. *Bull. Amer. Meteor. Soc.,* **77,** 953–961.
Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, 613 pp.
Titterington, D. M., 1980: Aspects of optimal design in dynamic systems. *Technometrics,* **22,** 287–299.
West, M., and J. Harrison, 1989: *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, 704 pp.