Statistical Principles for Climate Change Studies

RICHARD A. LEVINE

Division of Statistics, University of California, Davis, Davis, California

L. MARK BERLINER

The Ohio State University and National Institute of Statistical Sciences, Columbus, Ohio

(Manuscript received 10 Sepetember 1997, in final form 10 April 1998)

ABSTRACT

Statistical principles underlying ''fingerprint'' methods for detecting a climate change signal above natural climate variations and attributing the potential signal to specific anthropogenic forcings are discussed. The climate change problem is introduced through an exposition of statistical issues in modeling the climate signal and natural climate variability. The fingerprint approach is shown to be analogous to optimal hypothesis testing procedures from the classical statistics literature. The statistical formulation of the fingerprint scheme suggests new insights into the implementation of the techniques for climate change studies. In particular, the statistical testing ideas are exploited to introduce alternative procedures within the fingerprint model for attribution of climate change and to shed light on practical issues in applying the fingerprint detection strategies.

1. Introduction

Predictions of climate change due to human-induced increases in greenhouse gas and aerosol concentrations have been an ongoing arena for debate and discussion. A major difficulty in early detection of changes resulting from anthropogenic forcing of the climate system is that the natural climate variability overwhelms the climate change signal in observed data. A number of schemes based on fingerprint methodologies have been developed to overcome this inherent problem [Bell (1982), (1986); Hasselmann (1979), (1993), (1997); Madden and Ramanathan (1980); and North et al. (1995); see Hegerl and North (1997) for a review and comparison of these approaches and Solow (1991) for discussion of some related statistical issues].

The basic idea behind fingerprint procedures is to express the climate data in terms of low-dimensional signal patterns. The intent is that these patterns describe and summarize defining characteristics of the signal. The fingerprints are constructed to maximize the signal-to-noise ratio in the observed climate data. Fingerprints thus provide an optimal dimension reduction of the full climate system enhancing climate change detection strategies (Hasselmann 1993).

The fingerprint approach, as introduced by Hasselmann (1979), entails a mathematical procedure for optimally detecting a climate change signal above the background natural climate variability noise. We will describe how such schemes are representable as optimal hypothesis testing procedures from the classical statistics literature. The formal statistical hypothesis testing framework provides insights into the implementations of the fingerprint detection strategies. We also discuss the statistical development of procedures used in attribution analyses. In particular, we can draw on the statistical literature to suggest alternative attribution methodologies.

Section 2 motivates the discussion of the statistical principles for climate change studies through a development of statistical models for climate change analysis. An understanding of the statistical issues in modeling the climate system is necessary for the statistical exposition of fingerprint models and methods. Sections 3a and 3b review statistical hypothesis testing and comparative experiments for detecting a climate change signal and assessing casual relationships (i.e., attribution). Section 4a reviews the fingerprint method as presented by Hasselmann (1993) and formulates the problem as a traditional most powerful statistical hypothesis test. Sections 4b and 4c compare and contrast current attribution methods with statistical approaches and offer some new suggestions. Section 5 considers practical issues of detecting climate change signals via fingerprints and discusses ideas for incorporating alternative statistical procedures in the detection schemes.

Corresponding author address: Dr. Richard A. Levine, Division of Statistics, Statistical Laboratory, University of California, Davis, Davis, CA 95616.
E-mail: levine@wald.ucdavis.edu

We recognize that the literature and current consensus regarding climate change are based on extensive and sophisticated scientific reasoning, of which only a portion is attributable to the results of statistical testing (IPCC 1996). However, since statistical testing does play a crucial role in these studies, delineation of the choice of appropriate methods and the interpretation of results is warranted.

2. Statistical formulations

The detection of a climate change signal and attribution of the potential signal to specific forcings requires understanding of the variation in the observed data, namely the climate change signal and natural climate variability. In this section, we discuss issues in modeling these two components of the climate change problem.

Let $\boldsymbol{\Psi} = \{\Psi_a\}$ denote an observed climate state where the index $a = (\mathbf{v}, \mathbf{x}, t)$ represents the dependence of each component on climate variables $\mathbf{v}$ (e.g., pressure, humidity, etc.), spatial coordinates $\mathbf{x}$, and time t. Dependence on location, time, and selected variables is suppressed. Define f_A to be a quantitative measure of anthropogenic forcings; define f_E to be other *external* forcings, such as solar input, volcanoes, etc.

a. Models

The discussion of modeling relevant to the climate change issue is motivated by two concerns. 1) Statistical tools are best derived and applied depending on the stochastic models believed to underlie the problem; 2) sources of our uncertainty, both regarding ''nature'' and our analyses must be identified to assess results and suggest actions. In particular, we seek formal answers to the question ''What is 'climate variability'?'' In the following sections we describe the modeling of the climate state $\boldsymbol{\Psi}$ from several viewpoints of the physical structure of the climate system and the associated uncertainties. We make no claim that these models are novel; for example, see Hasselmann (1976).

1) DETERMINISTIC VIEW

We may consider the climate state

$$\boldsymbol{\Psi} = \mu(i, f_A, f_E), \tag{1}$$

where i represents a quantification of Earth's *ensemble* membership (''initial condition''), and μ is some deterministic function, namely, the trajectory of an integration of a physical model.

2) ENSEMBLE-STOCHASTIC VIEW

Even if one accepts the determinism in Eq. (1), uncertainty arises due to lack of complete knowledge of i. Hence, we consider an *ensemble-stochastic* view in which i is viewed as a random quantity, say with density function Q. Specifically, consider a ''signal model''

$$\boldsymbol{\Psi} = \mu(\cdot, f_A, f_E) + \boldsymbol{\Psi}_e(i, f_A, f_E), \tag{2}$$

where an ensemble average $\mu(\cdot, f_A, f_E)$ is defined by

$$\mu(\cdot, f_A, f_E) = \langle\mu(i, f_A, f_E)\rangle = \int \mu(j, f_A, f_E)Q(j)\,dj.$$

Note that this argument corresponds to the usual ensemble averaging technique common in fluid dynamics. Also, in general, ensemble averages depend on Q.

3) OPERATIONAL-STOCHASTIC VIEW

Aspects in Eq. (1), beyond i, are unknown. In general, μ, f_A, and f_E are all unknown. The most expedient approach is to simply replace $\mu(\cdot, f_A, f_E)$ by $\hat{\mu}(\cdot, \hat{f}_A, \hat{f}_E)$, where ''hats'' denote estimates. However, this approach does not readily enable us to account for our uncertainties in the estimated quantities. To do so, we may also employ probabilistic representations of these uncertainties. An *operational-stochastic* approach endows all unknown quantities with probability distributions. (This view coincides with the Bayesian approach to statistics.)

Another source of error in this context accrues from the need to approximate μ via numerical computation. Even with no other uncertainties, nonlinearities in the physical models suggest numerical problems. These uncertainties can also be managed via probabilistic modeling.

To summarize these uncertainties, we consider a ''signal model''

$$\boldsymbol{\Psi} = \hat{\mu}(i, \hat{f}_A, \hat{f}_E) + \mathbf{O}(i, f_A, f_E), \tag{3}$$

where $\mathbf{O}(i, f_A, f_E)$ represents ''error.'' The specification of a probability distribution for $\mathbf{O}$ is based on the above sources of uncertainty.

Depending on the goal of an analysis, we might need to combine the operational- and ensemble-stochastic views. Consider the model

$$\boldsymbol{\Psi} = \hat{\mu}(\cdot, \hat{f}_A, \hat{f}_E) + \boldsymbol{\Psi}_e(i, f_A, f_E) + \mathbf{O}(i, f_A, f_E), \tag{4}$$

where $\boldsymbol{\Psi}_e$ is intended to account for ensemble variation and $\mathbf{O}$ is intended to account for sources of operational uncertainties. To handle this model we must not only produce distributions for the two error vectors, but also their joint distribution (i.e., the dependence structure between errors).

The issue of the meaning of probability warrants a comment here. We are liberal in our use of probability in this discussion. In particular, probability statements need not only hinge on the ensemble or frequency interpretation. Rather, we are prepared to model uncertain quantities as being amenable to probability. This view is in concert with the Bayesian approach to statistics (Berger 1985; Bernardo and Smith 1994).

4) Intrinsic-stochastic view

By an *intrinsically stochastic* view we mean that $\boldsymbol{\Psi}$ is thought of as random with some probability distribution, say P. The key is that we would make this assumption even conditional on i, f_A, and f_E. A standard notation is to write

$$\boldsymbol{\Psi} \sim P(\cdot\,|i, f_A, f_E). \tag{5}$$

The notion is that no implementable form of the laws of physics permit us to envision a usable function μ that could completely determine $\boldsymbol{\Psi}$. Nevertheless, we can entertain a ''signal plus noise'' model of the form

$$\boldsymbol{\Psi} = \mu(i, f_A, f_E) + \mathbf{N}(i, f_A, f_E), \tag{6}$$

where μ is based on an appropriate physical model and $\mathbf{N}$ represents unexplained noise. Note that $\mathbf{N}$ is not interpreted in the same fashion as the $\boldsymbol{\Psi}_e(i, f_A, f_E)$ indicated in Eq. (2). The question of whether or not the $\mathbf{N}$ and the $\mathbf{O}$ actually represent different quantities may be mainly philosophical to some. In a sense we agree, though we also believe there is value to raising the two views.

For the model to be well-defined, distributional assumptions for the N-process, conditional on i, f_A, and f_E are needed. Typically, it is assumed that the expected value of N is zero. Note that this is an *assumption* in general, rather than a consequence of definition, as for anomalies from an ensemble average such as the $\boldsymbol{\Psi}_e$'s defined in Eq. (2).

As before, we might consider combining these various interpretations as

$$\boldsymbol{\Psi} = \hat{\mu}(\cdot, \hat{f}_A, \hat{f}_E) + \boldsymbol{\Psi}_e(i, f_A, f_E) + \mathbf{O}(i, f_A, f_E) + \mathbf{N}(i, f_A, f_E). \tag{7}$$

b. Statements of the climate change problem

Perhaps the most immediately relevant version of the climate change problem is the question, ''Is $\boldsymbol{\Psi}$ significantly different than it would have been had f_A been zero?'' Two clarifications merit emphasis. First, the meaning of ''significantly different'' is intended to involve the practical, physical importance of variations in $\boldsymbol{\Psi}$, as opposed to the notion of ''statistical significance.'' Second, in this query, we are concerned with *this* ensemble member, as opposed to the anticipated ensemble average impacts. In the notation of the previous section, we would study the quantity

$$\Delta_i = \mu(i, f_A, f_E) - \mu(i, 0, f_E), \tag{8}$$

if available, though practically, we would hope to study

$$\hat{\Delta}_i = \hat{\mu}(i, \hat{f}_A, \hat{f}_E) - \hat{\mu}(i, 0, \hat{f}_E). \tag{9}$$

These statements of the issue are problematic for most classical statistical analyses as we have but one realization under this Earth's initial conditions. Thus, we cannot produce the distributions for the errors $\mathbf{O}$ and $\mathbf{N}$. Assuming adoption of the model (2), more accessible statements of the climate change problem involve statements regarding ensemble means. ''Is $\mu(\cdot, f_A, f_E)$ significantly different from $\mu(\cdot, 0, f_E)$?'' or is

$$\langle\Delta_i\rangle = \langle\mu(i, f_A, f_E)\rangle - \langle\mu(i, 0, f_E)\rangle \tag{10}$$

significantly different from zero? Though more tractable than testing $\boldsymbol{\Psi}$ [or equivalently $\boldsymbol{\Psi}_e(i, f_A, f_E)$] directly, the issues of (1) practical versus statistical significance and (2) uncertainties in μ, f_A, and f_E arise.

In an intrinsically stochastic formulation, a natural statement of the problem is to ask whether or not $P(\cdot\,|i, f_A, f_E)$ and $P(\cdot\,|i, 0, f_E)$ [see (5)] tend to produce significantly different climates. Oddly, in this statement the actual data is not relevant, if both candidate probability distributions are known. That is, we are not necessarily interested in classifying $\boldsymbol{\Psi}$ as coming from one of these two distributions, unless we first can argue that the distributions really are practically different. Of course, in our setting, neither distribution is known.

Most climate change studies focus on changing signals, though climatologists also recognize that changes in *variability* are also possible and important. Though amenable to statistical analysis, this issue will not be treated here.

c. Other points

In viewing $\boldsymbol{\Psi}$ as the ''true'' values of climate variables, we should recognize that our observational data is not $\boldsymbol{\Psi}$ itself, but rather some version with *measurement error*. This additional source of error is often ignored in climate change studies or implicitly absorbed into error terms such as $\boldsymbol{\Psi}_e$, $\mathbf{O}$, and $\mathbf{N}$ defined earlier. However, measurement errors can have quite different structures than these other errors. As one example, we might expect nonstationarity in measurement errors over time, as technologies changed.

The usual formulations of the climate change problem found in the literature falls into the ensemble-stochastic view of section 2a(2). Specifically, the climate change problem is viewed from a signal-plus-noise standpoint with signal $\mu(\cdot, f_A, f_E)$ and noise $\boldsymbol{\Psi}_e(i, f_A, f_E)$. The problem is then divided into parts. 1) *Detection* of signal and 2) subsequent *attribution* of a detected signal to anthropogenic sources. In the remainder of this paper we focus on the statistical formulation of the usual techniques for detecting and attributing climate change. Thus, though we will briefly mention alternative statistical strategies that exploit the other models presented in section 2a, we leave discussion and development of the operationally and intrinsically stochastic views to future work.

3. Statistical analyses

a. Detection: Hypothesis testing

The detection of a climate change signal may be viewed as the statistical testing of a hypothesis. To explore this issue, we review the formal hypothesis testing framework as viewed by statisticians. Both the conceptual interpretation of statistical tests and some of the formal theory leading to criteria for selecting test procedures are discussed. These ideas and formalities have their origins in the work of R. A. Fisher, J. Neyman, E. S. Pearson, etc. See Fisher (1990) and Lehmann (1993) for in-depth discussions and additional references, and Lehmann (1986) for an extended presentation of the formal theory. Casella and Berger (1990) provides an introductory, textbook presentation of statistical testing.

Fisher's main view begins with the translation of a scientific theory into a statement concerning the probability distribution of some observable data. Let X represent some random variable or observable. Suppose that the probability distribution of X, say f, is known up to some parameter, θ. Common notation is that $X \sim f(x \mid \theta)$, where θ is unknown, but lies in some set Θ ($\theta \in \Theta$). Fisher assumed that the scientific hypothesis to be tested corresponds to a specific value for the unknown parameter, say θ_0. This value generates the *null hypothesis*, typically denoted by H_0: $\theta = \theta_0$. To test for ''significance'' in the data, Fisher devised the p-value. The p-value is the probability, assuming the null hypothesis is true, of observing data at least as unfavorable to H_0 as the actual data observed. The logic here is indirect; the suggestion is that small p-values are evidence against the null. It is important to note that the p-value is *not* the probability of the null, given the observed data. That quantity plays no role in classical testing.

Fisher's notion of testing is that it can be used to provide quantified evidence against hypotheses. The intention is that if the p-value is very tiny (0.05 has become a common cutoff, though some suggest 0.01 as more appropriate), the scientific hypothesis corresponding to H_0 is untenable and warrants replacement. On the other hand, one cannot produce quantifiable evidence in favor of the null. However, Fisher noted that if one repeatedly tests a particular null and never achieves strong evidence against it, then one would continue to entertain that null hypothesis as a plausible theory.

A concept of data being ''unfavorable to H_0'' is required in the definition of the p-value. That is, we must have some notion of an alternative hypothesis in mind to implement the procedure. Alternatively, formal statistical tests do not indicate evidence about the null in a vacuum; rather, the quantifications involve comparisons against alternatives. To clarify, suppose that our statistical model is that $X \sim N(\theta, 1)$, which is shorthand for ''X has a normal (Gaussian) distribution with mean θ and variance 1.'' Some theory yields H_0: $\theta = 0$. We then observe $X = x_d$. (Assume that $x_d > 0$.) The p-value calculation says we should find the probability that a $N(0, 1)$ random variable is as unfavorable to the null as ''x_d.'' If alternatives of interest are simply that $\theta \neq 0$, the usual p-value is the probability that $|X| \geq x_d$, where $X \sim N(0, 1)$. However, if the plausible alternatives are that $\theta > 0$, the p-value is the probability that $X \geq x_d$, where $X \sim N(0, 1)$. In general this is no small matter; in this case the two p-values differ by a factor of 2.

J. Neyman and E. S. Pearson sought a more formal approach to testing. The notion is that a scientific theory partitions the set Θ into two sets, say Θ_0 and Θ_a. We seek a test of the competing null and alternative hypotheses: H_0: $\theta \in \Theta_0$ versus H_a: $\theta \in \Theta_a$. We can either 1) reject H_0, in favor of H_a, or 2) fail to reject H_0. (Failing to reject H_0 does not necessarily mean we accept H_0. Statisticians and textbooks vary on this point.) Since there are two possible actions, there are two kinds of errors. 1) Type I error: rejection of H_0 when it is in fact true, and 2) Type II error: failing to reject (or accepting) H_0 when it is false.

A *test* is a rule for picking an action, based on the data. We are to choose a subset known as the *rejection region*, say R, of the set of all possible values of the data X. Should the observed data $x \in R$, we reject H_0; else, we fail to reject H_0. The criteria for choosing R are based on probabilities of error. First, define the *power function* of the test R to be the probability, $\Pr_\theta(X \in R)$, of rejection for a fixed value of θ. Intuitively we wish to make the power small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_a$. However, there is clearly no unequivocal best test. (We could simply always reject, ensuring no error if $\theta \in \Theta_a$, or never reject, ensuring no error whenever $\theta \in \Theta_0$.) To make progress, the suggestion is to essentially fix one of the error probabilities. Namely, consider all tests that have the property that the probability of Type I error is less than some preassigned, small number α. [That is, $\Pr_\theta(X \in R) \leq \alpha$ for all $\theta \in \Theta_0$.] Among all such tests, we then try to find R to make the power large for those $\theta \in \Theta_a$. Should there exist an R that makes the power as large as possible for every $\theta \in \Theta_a$, we say it is a uniformly most powerful, level α test (UMP).

In many cases, UMP tests fail to exist. In such cases it is common to further restrict the class of tests. One such restriction is to consider those level α tests for which the power outside the null hypothesis is greater than α; that is, $\Pr_\theta(X \in R) \geq \alpha$ for all $\theta \in \Theta_a$. Such tests are said to be unbiased. Within this class one then searches for tests that maximize the power for all $\theta \in \Theta_a$. Such tests are uniformly most powerful unbiased, level α tests.

A second restriction often suggested is to consider those α level tests that display certain invariance properties. That is, tests that behave symmetrically under certain classes of transformations of the statistical formulation. For example, important classes of transformations involve ''change of units'' for physical measurements. Once an invariance theory is formulated, one can then search for tests maximizing power within the class of invariant α level tests. Further detail is beyond the scope of this paper.

Note that in both the Fisherian and Neyman–Pearson formulations, the two hypotheses are treated asymmetrically. The null hypothesis plays a special role in each case. (This issue is particularly important in developing techniques for attribution versus detection.) A fully decision theoretic approach, formulated early in the work of A. Wald, generally permits asymmetric treatment of actions. The key is that probabilities of errors are not necessarily the appropriate quantities for control and optimization. Rather, the (expected) losses associated with particular actions are minimized. Such procedures are particularly important in the study of remediation.

Finally, we mention the Bayesian approach. In this approach to statistics, all unknown quantities, including parameters, are viewed as random variables. In particular, the unknown, now ''random,'' θ is endowed with a prior probability distribution. We may compute the prior probability of H_0 based on this distribution. The theory of probability provides a mechanism (Bayes' theorem) for updating the distribution of θ conditional on the observed data, yielding the posterior distribution for θ. The Bayesian testing solution is then based on the posterior probabilities of the null and alternative hypotheses. We will pursue this avenue for climate change detection elsewhere; also, see Leroy (1998). For general discussion of both decision theory and Bayesian analyses, see Berger (1985).

b. Attribution: Statistical issues

Intuitively, the keys to attribution of a phenomenon is the explanation of the observed results in terms of *causes* coupled with the elimination of other potential causes. For the classical statistician, attribution is argued, though never unequivocally proven, from data obtained in *randomized* and *controlled* experiments. ''Control'' refers to the fact that experimental units are treated fairly under competing causes and differences are compared. ''Randomization'' is used to determine which units receive which treatments to (hopefully) guard against unforeseen potential causes and biases (Fisher 1990). Such approaches are not practical for climate change. First, the earth cannot serve as its own control. Historical records are subject to errors. Also, we believe a variety of climate changes, not anthropogenically induced, have occurred. Second, we cannot run small-scale, surrogate Earths under various conditions in a laboratory; the earth system is believed to be far too complex for relevant features to be captured by toys (Trenberth 1997). Hence, climatologists rely on large numerical models to provide controls, at the best, or at least some useful information about anticipated behavior of the twentieth century climate without anthropogenic forcing.

From a foundational point of view, genuine attribution of climate change is impossible. Hence, statistical arguments are needed to provide scientific explanation of our understanding of climate change issues. The role of statistics in attempting to establish causation is a rich and fascinating issue. For brevity we only offer the following references: Holland (1986) and Good (1983).

4. Statistics and fingerprinting

a. Detection as a hypothesis test

In order to formulate the fingerprint detection scheme as a statistical hypothesis test, we need to review and describe the model from which fingerprints are derived. We will utilize the notation and ideas of Hasselmann (1993), denoted *H* from here onward. The climate vector $\boldsymbol{\Psi}$ is assumed to be a linear combination of the signal $\boldsymbol{\Psi}^S$ and internal climate noise $\tilde{\boldsymbol{\Psi}}$ so that

$$\boldsymbol{\Psi} = \boldsymbol{\Psi}^S + \tilde{\boldsymbol{\Psi}}. \tag{11}$$

The natural variation $\tilde{\boldsymbol{\Psi}}$ is assumed to be a random *n*-vector with mean zero and dispersion (covariance matrix) $\mathbf{C}(\mathbf{x}, t)$. The space–time lagged covariances modeled through the matrix $\mathbf{C} = \mathbf{C}(\mathbf{x}, t)$ are presumed known for the moment. It may seem that we have changed notation from that in section 2. However, this notation emphasizes the fact that one can choose which of μ or $\hat{\mu}$ to assign as $\boldsymbol{\Psi}^S$ as well as which combination of $\boldsymbol{\Psi}_e$, $\mathbf{O}$, and $\mathbf{N}$ to assign to $\tilde{\boldsymbol{\Psi}}$. Of course, how one prescribes $\mathbf{C}$ depends critically on these choices.

In practice, observations and signals are typically defined as anomalies from some baseline state, $\boldsymbol{\Psi}_0^S$. Here, $\boldsymbol{\Psi}_0^S$ may represent the climatology of the system, the climate state at a specific time, or a time series of the climatic variables under given ''control'' conditions. For example, in studying global trends in temperature over time, Santer et al. (1996) defined the baseline state as the average temperatures observed in the period 1880–1920. Alternatively, the vector $\boldsymbol{\Psi}_0^S$ may be defined as the expected temperature trend in a system where greenhouse gas and aerosol concentrations are not increasing. Of course, this control state can only be estimated, typically using climate model output. In each of these approaches to formulating a baseline, a level of uncertainty [corresponding to the use of estimated (hatted) quantities in section 2] is introduced though seldom quantified.

Interest lies in studying the observed climate state and expected signal as they differ from the baseline. To this end, redefine the ''data'' $\boldsymbol{\Psi}$ and signal $\boldsymbol{\Psi}^S$ to be an anomaly from $\boldsymbol{\Psi}_0^S$. If a climate signal above that present in the natural climate variability does not exist, we would expect the state of the system $\boldsymbol{\Psi}$, on average, to be no different from the baseline state $\boldsymbol{\Psi}_0^S$; that is $\boldsymbol{\Psi}^S$ will be zero. Hence, $\boldsymbol{\Psi}^S$ is the climate signal above and beyond the natural variability $\tilde{\boldsymbol{\Psi}}$.

Assume we have a set of *p* expected signal patterns represented by the *n*-vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$. These patterns may define climate change induced by increases in greenhouse gas concentrations, aerosol concentrations, regional climate changes, or other human activities expected to impact the climate system. These vectors are assumed known as generated by climate models or expert opinion. Assume that the signal vector $\boldsymbol{\Psi}^S$ lies in a space spanned by the vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$; that is,

$$\boldsymbol{\Psi}^S = \sum_{i=1}^{p} a_i \mathbf{g}_i \tag{12}$$

for some unknown set of coefficients $a_1, \ldots, a_p$. In matrix notation we write

$$\boldsymbol{\Psi}^S = \mathbf{Ga}, \tag{13}$$

where $\mathbf{G}$ is an $n \times p$ matrix with columns corresponding to the vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$ and $\mathbf{a}$ is a p-vector containing the coefficients of the linear combination in Eq. (12).

The fingerprint methodology of H reduces $\boldsymbol{\Psi}$ to a sequence of p low-dimensional detectors denoted d_i, through an application of linear filters or fingerprints, denoted $\mathbf{f}_i$; that is $d_i = \mathbf{f}_i^{\mathrm{T}}\boldsymbol{\Psi}$, $i = 1, \ldots, p$. Let $\mathbf{d}^S = (d_1^S, \ldots, d_p^S) = (\mathbf{f}_1^{\mathrm{T}}\boldsymbol{\Psi}^S, \ldots, \mathbf{f}_p^{\mathrm{T}}\boldsymbol{\Psi}^S)$ where superscripts T represent vector transpose. The goal, then, is to determine the statistical significance of the optimally detected signal. This optimally detected signal is constructed by maximizing the distance between the signal and natural climate variability or signal-to-noise ratio over all possible fingerprints. More specifically, the fingerprints $\mathbf{f}_i$ are chosen to maximize the quantity

$$\rho^2(\mathbf{d}^S) = (\mathbf{d}^S)^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{d}^S,$$

where the (i, j)th element of $\mathbf{D}$ is

$$\mathrm{cov}(\tilde{d}_i, \tilde{d}_j) = \mathrm{cov}(\mathbf{f}_i^{\mathrm{T}}\tilde{\boldsymbol{\Psi}}, \mathbf{f}_j^{T}\tilde{\boldsymbol{\Psi}}) = \mathbf{f}_i^{\mathrm{T}}\mathbf{C}\mathbf{f}_j.$$

Here, H shows that if the forced signal $\boldsymbol{\Psi}^S$ lies in a space spanned by known model prediction pattern vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$, then the solution to this maximization problem is $\mathbf{f}_i^* = \mathbf{C}^{-1}\mathbf{g}_i$. The statistical significance of the optimally detected signal is then determined through the statistic $\rho^2(\mathbf{d}^*)$ where the vector $\mathbf{d}^*$ is given by

$$\mathbf{d}^* = \mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Psi}. \tag{14}$$

There is a direct relationship between the optimal fingerprint method of H and most powerful tests of coefficients in a corresponding regression problem. To provide a perspective to this view, first note that the cornerstone of the approach is to make inferences about the coefficients $a_1, \ldots, a_p$. In settings in which the patterns are estimated from model runs, it would seem that these a_i's are also determined. However, we believe the view is that the models are capable of capturing the structure and patterns of climate change, reflected by the $\mathbf{g}_i$, but not the precise magnitudes as would be implied by both the a_i's and the $\mathbf{g}_i$'s taken as a group. Rather, the key is to treat the patterns as essentially correct, and then decide whether or not the patterns appear to be present in observational data by regressing the data on the patterns, that is, estimating the a_i's from the data.

To formally relate the signal detection problem as a statistical multiple regression problem, combining Eqs. (11) and (13) implies that,

$$\boldsymbol{\Psi} = \mathbf{Ga} + \tilde{\boldsymbol{\Psi}}. \tag{15}$$

In the statistical regression literature, $\mathbf{G}$ is termed the design matrix and $\mathbf{a}$ regression coefficients. Next, optimal estimates of the signal $\boldsymbol{\Psi}^S$ are obtained from the generalized least squares estimates (Weisberg 1985, section 4a) of the regression coefficients, denoted $\hat{\mathbf{a}}$. Namely,

$$\hat{\mathbf{a}} = (\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Psi}$$

and

$$\hat{\boldsymbol{\Psi}}^S = \mathbf{G}\hat{\mathbf{a}}.$$

The estimate $\hat{\mathbf{a}}$ is rightly viewed as the Gauss–Markov optimal procedure; note too that it is simply a rotation of the projected fingerprints $\mathbf{d}^*$ given in Eq. (14).

Statistical inference can be based on the fact that, under Gaussian assumptions, $\hat{\mathbf{a}}$ has a multivariate normal distribution:

$$\hat{\mathbf{a}} \sim N\,[\mathbf{a}, (\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1}]. \tag{16}$$

For example, consider testing the hypothesis of no signal in H that may be equivalently written as

$$H_0\colon \mathbf{a} = \mathbf{0} \quad \text{versus} \quad H_A\colon \mathbf{a} \neq \mathbf{0}. \tag{17}$$

If the dispersion matrix $\mathbf{C}$ is known, the uniformly most powerful invariant test of Eq. (17) rejects H_0 if

$$\begin{aligned} T &= \boldsymbol{\Psi}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G}(\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Psi} \\ &> \chi_p^2(1 - \alpha), \end{aligned} \tag{18}$$

where $\chi_p^2(1 - \alpha)$ is the critical value at level α based on the chi-square distribution with p degrees of freedom (see Lehmann 1986, section 8.7 or Seber 1984). The restriction to invariant tests arises here because no UMP test exists.

Notice the test statistic T in Eq. (18) is equivalent to the statistic utilized to assess the significance of the optimally detected signal in H. If $\mathbf{f}^* = \mathbf{C}^{-1}\mathbf{G}$ denotes the optimal fingerprint,

$$\begin{aligned} T &= \boldsymbol{\Psi}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G}(\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\Psi}^{-1} \\ &= \boldsymbol{\Psi}^{\mathrm{T}}\,\mathbf{f}^*(\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1}(\mathbf{f}^*)^{\mathrm{T}}\boldsymbol{\Psi} = \rho^2(\mathbf{d}^*). \end{aligned} \tag{19}$$

In the single pattern problem ($p = 1$),

$$\boldsymbol{\Psi} = a\mathbf{g} + \tilde{\boldsymbol{\Psi}},$$

where a is an unknown constant. It can be shown that the test in Eq. (18) is also a UMP-unbiased test (Lehmann 1986, sections 5.8 and 7.7 or Seber 1984).

1) Unknown C

If the dispersion $\mathbf{C}$ is unknown, H suggests estimating the correlation structure from historical data independent of the observations $\boldsymbol{\Psi}$. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ denote the historical data and $\overline{\mathbf{Y}}$ denote the mean of the historical sample. If each $\mathbf{Y}_i$ is identically distributed $N(\boldsymbol{\tau}, \mathbf{C})$ with unknown mean $\boldsymbol{\tau}$, then the maximum likelihood estimator

$$\hat{\mathbf{C}} = \frac{1}{m}\sum_{i=1}^{m} (\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^T$$

may be utilized as an estimate for the dispersion $\mathbf{C}$. Substituting this estimate into the test statistic in Eq. (18) suggests that a test of Eq. (17) would reject the null hypothesis when

$$T_Y = \boldsymbol{\Psi}^{\mathrm{T}}\hat{\mathbf{C}}^{-1}\mathbf{G}(\mathbf{G}^{\mathrm{T}}\hat{\mathbf{C}}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\hat{\mathbf{C}}^{-1}\boldsymbol{\Psi}$$

is large. Unfortunately, the determination of a ‘‘large’’ T_Y is not as easily accomplished as in the known dispersion case. However, if m is large, use of the χ^2 critical value as in Eq. (18) may be reasonable.

2) Fingerprinting as regression

The regression formulation as described above allows for an implementation of a sequential fingerprint detection scheme when the signal patterns $\mathbf{g}_i$, $i = 1, \ldots, p$ (which may represent changes in climate due to carbon dioxide, aerosols, regional variables, etc.), form a hierarchical set. Namely, when the climatologist wishes to assess the significance of patterns in a prescribed order of importance, for example, carbon dioxide first, followed by aerosols, etc. More specifically, multipattern signal detection is typically performed as follows. Initially test for the appearance of the pattern $\mathbf{g}_1$ in the data. If this pattern is detected, add the second pattern $\mathbf{g}_2$, and so on until a pattern is no longer detected in the observations. The multipattern detection is thus reduced to a sequence of tests on univariate detectors.

This procedure is analogous to a forward stepwise routine for building a regression model (Weisberg 1985, chapter 8). Let a_i denote the regression coefficient corresponding to pattern $\mathbf{g}_i$. We sequentially test if each a_i is zero given all other signal patterns $\mathbf{g}_1, \ldots, \mathbf{g}_{i-1}$ are in the regression model. Continue adding patterns to the regression equation until the first acceptance of a null hypothesis. Hence, rather than testing the hypothesis in Eq. (17) that all regression coefficients $a_1, \ldots, a_p$ are zero, test each regression coefficient separately and sequentially.

Providing these analogies between fingerprinting and statistical regression analysis is not merely an academic exercise. Rather it enables transfer of technology as well as cautions. First, the interpretation of parameters in multiple regression can be problematic. While the notation is casual, the meaning of a_1 in the univariate model $a_1\mathbf{g}_1$ is not the same as that in the model $a_1\mathbf{g}_1 + a_2\mathbf{g}_2$. In the second case, a_1 is roughly the rate of change in the signal in varying the first pattern while holding the second pattern constant. In some not entirely pathological examples, a significant, say positive a_1 based on the first model may become insignificant or even change sign when $\mathbf{g}_2$ is added to the model. More generally, relationships among the patterns can create difficulties in interpretation and analysis. The issue is known as multicollinearity in statistics (Weisberg 1985, chapter 8). Some concerns can be mitigated by regressing on appropriately defined empirical orthogonal functions of the patterns. (This is known as principal components regression in statistics.) However, this creates difficulties in interpretation of the derived ‘‘patterns.’’ Related problems arise in defining a ‘‘carbon dioxide plus aerosols’’ pattern. Interactions among these forcings generally suggest that this pattern is not simply obtained from the pattern for carbon dioxide and the pattern for aerosols.

A second issue involves the above sequential, forward stepwise analysis. In general, the results strongly depend upon the order in which patterns are added. Further, the advertised α levels for the sequence of tests are typically not valid. Indeed, these procedures are notorious for overstating the overall statistical significance of results.

b. Attribution

Suppose the detection test of the previous section does indeed reject the null hypothesis of no climate change signal above baseline. We cannot immediately attribute the perceived signal to any of the patterns $\mathbf{g}_1, \ldots, \mathbf{g}_p$. Though we cannot obtain pure causal relationships in this context, one can test if the forcing patterns $\mathbf{G}$ are consistent with the observations.

Hasselmann (1997) pursues this line of inquiry, roughly as follows. Climate model data is used to construct an estimate of the expected signal amplitude $\hat{\mathbf{a}}_M$ under a selected forcing specification; the subscript M indicates a model-based derivation. An estimate $\hat{\mathbf{a}}_{\mathrm{obs}}$ can be constructed from the climate observations. A comparison of $\hat{\mathbf{a}}_M$ and $\hat{\mathbf{a}}_{\mathrm{obs}}$ will determine if the observed climate signal is consistent with the expected, model-based signal.

The uncertainty associated with estimating the climate change signal via $\hat{\mathbf{a}}_M$ and $\hat{\mathbf{a}}_{\mathrm{obs}}$ is derived from the model-based and natural climate variability, respectively. To this end, Hasselmann (1997) assumes

$$\hat{\mathbf{a}}_M \sim N(\mathbf{a}_M, \mathbf{C}_M)$$

and

$$\hat{\mathbf{a}}_{\mathrm{obs}} \sim N(\mathbf{a}, \mathbf{C}_{\mathrm{obs}}),$$

where $\mathbf{a}_M$ and $\mathbf{a}$ are the true coefficients and $\mathbf{C}_M$ and $\mathbf{C}_{\mathrm{obs}}$ denote the variability in the estimates. Note that

$$\mathbf{C}_{\mathrm{obs}} = (\mathbf{G}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{G})^{-1},$$

where $\mathbf{C}$ is the dispersion matrix of the natural climate variability as discussed earlier.

A test of consistency between the model-predicted and observed climate signals analyzes the hypothesis

$$H_0\colon \mathbf{a} = \mathbf{a}_M \quad \text{versus} \quad H_A\colon \mathbf{a} \neq \mathbf{a}_M. \qquad (20)$$

The normality assumptions and an independence supposition yield the distribution

$$\hat{\mathbf{a}}_{\mathrm{obs}} - \hat{\mathbf{a}}_{M} \sim N(\mathbf{0}, \mathbf{C}_{\mathrm{obs}} + \mathbf{C}_{M}),$$

under the null hypothesis. The null hypothesis is rejected if

$$(\hat{\mathbf{a}}_{\mathrm{obs}} - \hat{\mathbf{a}}_{M})^{\mathrm{T}}(\mathbf{C}_{\mathrm{obs}} + \mathbf{C}_{M})^{-1}(\hat{\mathbf{a}}_{\mathrm{obs}} - \hat{\mathbf{a}}_{M}) > \chi^2_{1-\alpha},$$

where $\chi^2_{1-\alpha}$ is the critical value at level α based on the chi-square distribution with p degrees of freedom. If this test fails to reject, evidence for consistency is claimed.

If $\mathbf{C}$ and $\mathbf{C}_M$ are unknown, one can construct estimates $\hat{\mathbf{C}}$ and $\hat{\mathbf{C}}_M$ via climate model control runs. In particular, $\mathbf{C}$ can be estimated from climate model output in which anthropogenic influences on the climate system are negligible. (One could also estimate $\mathbf{C}$ based on data as in the earlier detection analysis.) Here, $\mathbf{C}_M$ can be estimated from repeated runs of the climate model utilized to construct $\hat{\mathbf{a}}_M$. The consistency test is then based on Hotelling's T-distribution or a variant thereof (analogous to the detection test derived in section 4a). See Hegerl et al. (1997) for an application. Note that direct substitution of estimated covariances into tests derived under known covariance assumptions are best viewed as approximations. See section 5 for more discussion.

These consistency tests have several statistical deficiencies. First, as mentioned in section 3, some statisticians do not accept the notion of demonstrating evidence for a null hypothesis. This is, the consistency hypothesis, $\mathbf{a} = \mathbf{a}_M$, cannot be ''accepted'' in a Fisherian convention. Even if one relaxes this convention, it is difficult to ascribe an error rate or measure of confidence if one accepts the null. While we know what it means to reject a null hypothesis at some α level, this does not provide a valuable error rate if we accept. The error of interest (Type II, as described in section 3) is that of incorrectly accepting. The probability of this error (one minus the ''power'') depends on the true value of $\mathbf{a}$, which is unknown.

Second, the consistency test is performed upon rejecting the null hypothesis of the detection test in Eq. (17). This two-step procedure complicates a correct calculation of error probabilities in the consistency test. Specifically, such calculation must account for the probability of incorrectly detecting a climate change when one does not exist and the probability of correctly detecting a climate change when one does exist as well as the probability of incorrectly rejecting the null hypothesis in Eq. (20).

c. Alternative procedures for attribution

The search for consistency described in section 4b is analogous to demonstrations of *equivalence* in the statistics literature. For example, biostatisticians are often confronted with the problem of determining if generic drugs are equivalent to a brand name in that they become available at the drug action site at approximately the same rate and concentration. Government regulations require the proof of bioequivalence before a generic drug may be marketed. (See Berger and Hsu 1996.)

The test for consistency is essentially one of geoequivalence: testing if the observed and model-predicted climate change signals are equivalent. We can thus utilize the established bioequivalence testing procedures for a test of geoequivalence. These procedures are derived from a test of the hypothesis

$$H_0\colon \mathbf{a} \neq \mathbf{a}_M \quad \text{versus} \quad H_A\colon \mathbf{a} = \mathbf{a}_M. \tag{21}$$

Notice the null hypothesis from the consistency test (20) appears in the alternative here.

An easily applicable test of (21) was developed by Brown et al. (1995). Let $\boldsymbol{\theta} = \mathbf{a} - \mathbf{a}_M$, $\hat{\boldsymbol{\theta}} = \hat{\mathbf{a}}_{\mathrm{obs}} - \hat{\mathbf{a}}_M$, and $\boldsymbol{\Sigma} = \mathbf{C}_{\mathrm{obs}} + \mathbf{C}_M$. Under the assumption that

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}),$$

we have evidence against the null hypothesis and in favor of geoequivalence at level α if the (confidence) set

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta})^{1/2} \leq z_\alpha + (\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta})^{-1/2}(\hat{\boldsymbol{\theta}}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta})\} \tag{22}$$

does not contain zero. Here, z_α is the α critical level from a standard normal distribution. These confidence sets are optimal in the sense that they minimize the expected volume of the set among all confidence sets with coverage probability $1 - \alpha$ (Brown, et al. 1995).

If $\boldsymbol{\Sigma}$ is unknown, approximations can again be based on independent estimates of the required dispersion matrices.

Alternatively, the so-called simple versus simple hypothesis testing formulation poses the detection–attribution problem as a single hypothesis test. Recall we have estimates of both $\mathbf{a}$ and $\mathbf{a}_M$ as derived from observed and climate model data respectively. The test of the hypothesis

$$H_0\colon \mathbf{a} = \mathbf{0} \quad \text{versus} \quad H_A\colon \mathbf{a} = \mathbf{a}_M, \tag{23}$$

arguably considers detection and consistency in one analysis. If the null hypothesis is rejected, one might claim evidence that a signal exists and is consistent with the climate model predicted signal, analogous to the conclusions sought by Hasselmann (1997).

For $\mathbf{C}$ assumed known, the most powerful α level test is known to be ''reject H_0'' if

$$T_M = \frac{\mathbf{a}_M^{\mathrm{T}}\mathbf{C}_{\mathrm{obs}}^{-1}\hat{\mathbf{a}}_{\mathrm{obs}}}{(\mathbf{a}_M^{\mathrm{T}}\mathbf{C}_{\mathrm{obs}}^{-1}\mathbf{a}_M)^{1/2}} > z_\alpha, \tag{24}$$

where z_a is the appropriate critical value based on a standard normal distribution (e.g., $z_{.05} = 1.645$).

This simple versus simple view admits an enhancement. The test in Eq. (24) is actually UMP for testing the hypotheses

$$H_0 : \mathbf{a} = \mathbf{0} \quad \text{versus} \quad H_A : \mathbf{a} = d\mathbf{a}_M, \qquad d > 0, \tag{25}$$

where d is unspecified. (Note that any specific value assigned to $d > 0$ would cancel in the calculation of

T_M.) That is, the magnitude of $\mathbf{a}_M$ is not important; rather, the test looks in the direction indicated by $\mathbf{a}_M$. See Hegerl et al. (1997) for discussion and implementations.

5. Practical considerations and discussion

a. Covariance estimation

It is well-recognized in the climate change literature that the estimation of dispersion matrices (e.g., $\mathbf{C}$) is a difficulty regardless of one's view of the detection-attribution problems. Misrepresentation of the correlation structure may lead to incorrect conclusions (Hegerl and North 1997). Hasselmann (1997) suggests that dispersion matrices ''can be estimated from observations or model simulations with sufficient accuracy for a meaningful signal-to-noise analysis.'' However, we believe that translation of ''sufficient accuracy'' may be difficult.

Often, a long control run from a climate model with no anthropogenic forcing is utilized to estimate $\mathbf{C}$. Under assumptions of separability between the temporal and spatial correlations, the estimation procedure is tractable (see, for example, Hasselmann 1993; Hegerl et al. 1997; and Santer et al. 1995a for applications). However, improved climate model representations of the natural climate variability as well as reasonable statistical techniques for estimating $\mathbf{C}$ are necessary to realize the full potential of the fingerprint detection scheme (Hegerl et al. 1997).

Use of observations, as in section 4a(1) rather than model output may well appear to be more readily defensible than relying on models only. However, observations potentially contain variations and spatial dependences not entirely attributable to true natural climate variability. First, they contain measurement errors and biases. Further, data such as $\boldsymbol{\Psi}$ and the $\mathbf{Y}_i$'s in section 4a(1) are typically produced using objective analysis type procedures. Such procedures can induce spatial structure due to human averaging that need not reflect nature. A second potential issue involves the choice of baselines. Recall that $\boldsymbol{\Psi}$ is actually an anomaly from a particular baseline (denoted by $\boldsymbol{\Psi}_0^S$ in section 4). Clearly, if we use the procedure in section 4a(1) it is important that the resulting estimate of $\mathbf{C}$ is not biased by differences between $\boldsymbol{\Psi}_0^S$ and $\overline{\mathbf{Y}}$.

Even if all of the potential bias issues raised above are judged to have negligible effects, results based on the use of estimates as if they are ''true'' parameter values must be viewed as approximations. The problem of course is gauging the quality of approximation. The problem is not easy. For example, recall the key development in section 4 involved the estimate $\hat{\mathbf{a}} = (\mathbf{G}^T\mathbf{C}^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{C}^{-1}\boldsymbol{\Psi}$ and its distribution given in Eq. (16). If we simply replace C by an unbiased estimate [as in section 4a(1)], the resulting estimate of a is no longer unbiased nor normally distributed. While this is not severe as $m \rightarrow \infty$, gauging the impact for a particular m is not trivial. As noted in *H,* simulation can be a valuable tool in this process. Of course, simulation is merely an alternative approximation, useful primarily when the cost in performing massive simulation experiments is small.

b. Correlation based detection

A common implementation of the fingerprint notion is to detect a climate change signal by studying some measure of dependence, for example, correlation between the expected, climate model constructed anthropogenic signal and climate observations. [Barnett (1986) provides the first application of this technique. For further applications, see Karoly et al. (1994), Santer et al. (1996), Santer et al. (1995b), Tett et al. (1996), and Wigley and Santer (1990).] The fingerprint in these applications is a time series of the correlation-like statistics measuring the temporal change in similarity between the expected (under some specification of forcing) climate signal and the climate observations.

One motivation for such studies is a relationship between these techniques and the methodology presented in section 4a. More generally, regression and correlation analyses are related. Suppose that for each fixed time epoch, we let an s-vector $\mathbf{g}$ be a model-derived estimate of the climate signal change under some forcing specification. (s denotes the number of spatial grid points considered.) That is, $\mathbf{g}$ represents spatial departures from a mean model behavior. Assume that for each time period t, the s-vector of climate observations $\boldsymbol{\Psi}_t$ (actually, the departures from the spatial mean of the observations, which are in turn anomalies from some baseline) follows the model

$$\boldsymbol{\Psi}_t = a\mathbf{g} + \tilde{\boldsymbol{\Psi}}_t. \qquad (26)$$

Paralleling section 4a, assume that for each t, $\tilde{\boldsymbol{\Psi}}_t \sim N(\mathbf{0}, \mathbf{C}_s)$, where $\mathbf{C}_s$ is an $s \times s$ covariance matrix, and a is an unknown regression coefficient. Hence, $\boldsymbol{\Psi}_t$ is a snapshot of a particular climate variable anomaly in space at one point in time.

Consider (dependence on time is suppressed) the generalized least squares estimate of a at time t,

$$\hat{a}_t = (\mathbf{g}^T\mathbf{C}_s^{-1}\mathbf{g})^{-1}\mathbf{g}^T\mathbf{C}_s^{-1}\boldsymbol{\Psi}_t.$$

Under the above assumptions, we have that each $Z_t = (\mathbf{g}^T\mathbf{C}_s^{-1}\mathbf{g})^{1/2}\hat{a}_t$ is normally distributed with mean a and variance one, and can be used to test the hypothesis

$$H_0\colon a = 0 \quad \text{versus} \quad H_A\colon a \neq 0,$$

at that time.

We remark that in some cases authors have not adjusted for $\mathbf{C}_s$, and consider the correlation statistic

$$r_t = (\mathbf{g}^T\mathbf{g}\boldsymbol{\Psi}^T\boldsymbol{\Psi})^{-1/2}\mathbf{g}^T\boldsymbol{\Psi}_t.$$

If $\mathbf{C}_s$ is an identity matrix it can be shown that testing the null H_0: $a = 0$ is equivalent to testing the null that the ''true'' correlation between the observations and the model predictions is zero, based on r_t. However, two critical points arise. First, the suggestion that $\mathbf{C}_s$ is the identity seems untenable, leading us to recommend testing based on statistic Z_t rather than r_t.

Second, a more foundational concern arises. Formally, genuine correlation is measured between two random quantities by definition. That is, a true correlation being tested actually means that one is assuming that **g** is a realization of some stochastic process. But here's the rub: if we endow the model with a stochastic background, then the spatial covariance matrix of **g** should also enter into the analysis. (This object also arose in the geoequivalence test.) See Smith (1997) for discussion and suggestions for handling both covariance structures.

Current use of correlation-like schemes measure similarity between climate observations and expected climate signal by studying the trend of r_t over time. The idea is analogous to allowing the regression coefficient a in Eq. (26) to vary in time, denoted say $a(t)$. Again, such trend analysis should take account of the spatial correlations (both of the data and model results) as well as potential temporal covariance structures.

c. Interpretation of testing results

Given the scientific and societal importance of climate change studies, it is worthwhile to place the value of classical statistical hypothesis testing results in perspective. Hegerl and North (1997), Hegerl et al. (1997), and Hegerl et al. (1996) suggest that the hypothesis test in Eq. (17) and its cousins are conservative in the sense that a bad ''guess'' for **G** ''will not wrongly detect change where change does not exist'' (Hegerl and North 1997). A statistical defense for this claim requires that an error rate be appended. That is, if the null hypothesis is true, incorrectly rejecting it and claiming evidence for change is an event of probability α. This claim is indeed correct no matter how well or poorly the alternatives have been constructed.

Of course, this claim hinges on the assumptions under which the probability calculation is done (e.g., normality, etc.). However, more subtle concerns may arise. The claim associates no climate change with the null hypothesis holding exactly. It need not carry over if one thinks of the test as an approximation to a test of a fuzzy null hypothesis that $\mathbf{a} \approx \mathbf{0}$. This is particularly problematic in that we may have little confidence in the meaning of precise, nonzero values of the a_i's. Hence, quantifying the level of approximation desired in the fuzzy null is difficult.

To clarify potential sources of concern, we consider a very special example: a test of the hypotheses given in Eq. (23) in the case of a single pattern ($p = 1$) where $a_M > 0$ and assuming that $\mathbf{C} = \mathbf{I}$. It can be shown that the rejection rule reduces to

$$T_M = \frac{\mathbf{g}^{\mathrm{T}}\boldsymbol{\Psi}}{(\mathbf{g}^{\mathrm{T}}\mathbf{g})^{1/2}} > z_\alpha. \tag{27}$$

The sign of a_M determines the test, but the value of a_M is irrelevant. The condition $T_M = z_\alpha$ generates a hyperplane in the data space perpendicular to the vector **g**.

Consider a two-dimensional data example. Suppose $\mathbf{g} = (1\ 3)^{\mathrm{T}}$ and $\alpha = .05$ ($z_{.05} = 1.645$). The rejection hyperplane is the line

$$\Psi_2 = -\frac{1}{3}(\Psi_1 + z_\alpha\sqrt{10}).$$

This line intersects the line generated by $\mathbf{g} = (1\ 3)^{\mathrm{T}}$ at the point (0.52, 1.56). Hence, for observed data (0.525, 1.575), one would claim detection at an $\alpha = 0.05$ error rate. However, suppose our computer model had suggested the value of $a_M = 1.05$. We may not be comfortable with a claim of detection since the actual data is neither compatible with the null hypothesis nor with the suggested alternative. Intuitively, this data leaves us ''50–50'' on which hypothesis we believe, or makes us question the validity of either hypothesis or our other assumptions. The situation is particularly disturbing if a_M is much larger than 1.05, since the data is more compatible with the null (that is being rejected), than with the entire model output (that is being accepted). This concern is amplified in cases for which the data fall in the rejection, yet are far away from the anticipated alternative **g**. Such concerns may well be a motivation for the consistency test of H discussed early.

Parallel comments apply to the detection test in Eq. (17). The rejection region defined in Eq. (18) is the complement of an ellipsoid centered at **0**. The orientation of this ellipsoid is determined by the $\mathbf{g}_i$. For highly eccentric ellipsoids, there may be substantial probability of rejection at a values that might well be viewed as part of a fuzzy null.

In part the general concern we are raising is one of practical versus statistical significance. The classical tests described in this article are not designed to be useful in assessing practical significance of results. That is, a ''detected'' signal need not be an important signal. Similarly, meaningful changes may not be readily detectable (i.e., the climate signals exist in regions where our tests have low power). It bears repeating that the significance discovered in testing refers to data, not to the quantities tested. For example, rejection of the null H_0: $\mathbf{a} = \mathbf{0}$ in Eq. (17) does not mean that we have evidence that **a** is significantly different from zero in terms of climatic behavior.

Another viewpoint is that the sort of tests reviewed here tend to overstate the evidence against the null hypothesis. For further discussion, see Berger and Delampady (1987) and Berger and Sellke (1987).

Acknowledgments. This work was supported by NCAR's Geophysical Statistics Project, funded by NSF Grant #DMS-9312686. We are grateful to J. Berger, A.

Dempster, G. Hegerl, R. Madden, and T. Wigley for valuable discussions regarding climate change studies. Two referees provided extremely valuable comments, criticisms, and suggestions leading to substantial improvements in this presentation.

REFERENCES

Barnett, T. P., 1986: Detection of changes in global tropospheric temperature field induced greenhouse gases. *J. Geophys. Res.,* **91,** 6659–6667.

Bell, T. L., 1982: Optimal weighting of data to detect climatic change: Application to the carbon dioxide problem. *J. Geophys. Res.,* **87,** 11 161–11 170.

——, 1986: Theory of optimal weighting of data to detect climatic change. *J. Atmos. Sci.,* **43,** 1694–1710.

Berger, J. O., 1985: *Statistical Decision Theory and Bayesian Analysis.* 2d ed. Springer-Verlag, 617 pp.

——, and M. Delampady, 1987: Testing precise hypotheses. *Stat. Sci.,* **2,** 317–352.

——, and T. Sellke, 1987: Testing of a point null hypothesis: The irreconcilability of significance levels and evidence. *J. Amer. Stat. Assoc.,* **82,** 112–139.

——, and J. C. Hsu, 1996: Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Stat. Sci.,* **11,** 283–319.

Bernardo, J. M., and A. F. M. Smith, 1994: *Bayesian Theory.* John Wiley and Sons, 586 pp.

Brown, L., G. Casella, and J. T. G. Hwang 1995: Optimal confidence sets, bioequivalence, and the limacon of Pascal. *J. Amer. Stat. Assoc.,* **90,** 880–889.

Casella, G., and R. L. Berger, 1990: *Statistical Inference.* Duxbury, 650 pp.

Fisher, R. A., 1990: *Statistical Methods, Experimental Design, and Scientific Inference.* University Press.

Good, I. J., 1983: *Good Thinking: The Foundations of Probability and its Applications.* University of Minnesota Press, 332 pp.

Hasselmann, K., 1976: Stochastic climate models. Part I: Theory. *Tellus,* **28,** 473–485.

——, 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology of Tropical Oceans,* D. B. Shaw, Ed., Royal Meteorology Society, 251–259.

——, 1993: Optimal fingerprints for the detection of time-dependent climate change. *J. Climate,* **6,** 1957–1971.

——, 1997: Multi-pattern fingerprint method for detection and attribution of climate change. *Climate Dyn.,* **13,** 601–611.

Hegerl, G. C., and G. R. North, 1997: Statistically optimal approaches to detecting anthropogenic climate change. *J. Climate,* **10,** 1125–1133.

——, H. von Storch, K. Hasselmann, B. D. Santer, C. Cubasch, and P. D. Jones, 1996: Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J. Climate,* **9,** 2281–2305.

——, K. Hasselmann, C. Cubasch, J. F. B. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, 1997: On multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dyn.,* **13,** 613–634.

Holland, P. W., 1986: Statistics and causal inference. *J. Amer. Stat. Assoc.,* **81,** 945–970.

Intergovernmental Panel on Climate Change, 1996: *Climate Change 1995: The Science of Climate Change.* Cambridge University Press, 878 pp.

Karoly, D. J., J. A. Cohen, G. A. Meehl, J. F. B. Mitchell, A. H. Oort, R. J. Stouffer, and R. T. Wetherald, 1994: An example of fingerprint detection of greenhouse climate change. *Climate Dyn.,* **10,** 97–105.

Lehmann, E. L., 1986: *Testing Statistical Hypotheses.* 2d ed. Wiley 600 pp.

——, 1993: The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *J. Amer. Stat. Assoc.,* **88,** 1242–1249.

Leroy, S. S., 1998: Detecting climate signals: Some Bayesian aspects. *J. Climate,* **11,** 640–651.

Madden, R. A., and V. Ramanathan, 1980: Detecting climate change due to increasing carbon dioxide. *Science,* **209,** 763–768.

North, G. R., and K.-Y. Kim, 1995: Detection of forced climate signals. Part II: Simulation results. *J. Climate,* **8,** 409–417.

——, ——, S. S. P. Shen, and J. W. Hardin, 1995: Detection of forced climate signals. Part I: Filter theory. *J. Climate,* **8,** 401–408.

Santer, B. D., U. Mikolajewicz, W. Brüggemann, U. Cubasch, K. Hasselmann, H. Höck, E. Maier-Reimer, and T. M. L. Wigley, 1995a: Ocean variability and its influence on the detectability of greenhouse warming signals. *J. Geophys. Res. Oceans,* **100,** 10 693–10 725.

——, K. E. Taylor, T. M. L. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, 1995b: Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dyn.,* **12,** 77–100.

——, and Coauthors, 1996: A search for human influences on the thermal structure of the atmosphere. *Nature,* **382,** 39–46.

Seber, G. A. F., 1984: *Multivariate Observations.* Wiley, 686 pp.

Smith, R. L., 1997: Detecting signals in climatological data. *Bull. Int. Stat. Inst.,* **LVII,** 211–214.

Solow, A. R., 1991: On the statistical comparison of climate model output and data. *Greenhouse-Gas-Induced Climate Change: A Critical Appraisal of Simulations and Observations,* M. E. Schlesinger, Ed., Elsevier, 505–510.

Tett, S. F. B., J. F. B. Mitchell, D. E. Parker, and M. R. Allen, 1996: Human influence on the atmospheric vertical temperature structure: Detection and observations. *Science,* **274,** 1170–1173.

Trenberth, K. E., 1997: The use and abuse of climate models. *Nature,* **386,** 131–133.

Weisberg, S., 1985: *Applied Linear Regression.* Wiley.

Wigley, T. M. L., and B. D. Santer 1990: Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.,* **95,** 851–865.