
**Technical Report**
RAL-TR-97-040

# Checking for Model Consistency in Optimal Fingerprinting

**M R Allen and S F B Tett**

January 1998

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

The Central Laboratory of the Research Councils
Library and Information Services
Rutherford Appleton Laboratory
Chilton
Didcot
Oxfordshire
OX11 0QX
Tel: 01235 445384 Fax: 01235 446403
E-mail library@rl.ac.uk

**ISSN 1358-6254**

# Checking for model consistency in optimal fingerprinting

Myles R. Allen
Space Science Department, Rutherford Appleton Laboratory
Chilton, Didcot, OX11 0QX &
Department of Physics, University of Oxford
&
Simon F. B. Tett
Hadley Centre for Climate Prediction and Research
UK Meteorological Office
London Road, Bracknell, RG12 2SZ

Technical Report: RAL-TR-97040, August, 1997

**Abstract**

Current approaches to the detection of climate change and attribution of an anthropogenic influence on climate involve quantifying the level of agreement between one or more model-predicted patterns of anthropogenically-forced change and observed changes in the recent climate record. Analyses of uncertainty rely on simulated climate variability from a control run of a climate model. If the model gives an inaccurate representation of climate variability in the real world, all estimates of uncertainty, including statements of confidence in claims of attribution, are compromised. Any numerical representation of the climate system is likely to display too little variance on small spatial scales so there will always be aspects of model variability which are unrealistic, leading to a risk of spurious detection results. The risk is particularly severe if the detection strategy involves optimisation of signal-to-noise because unrealistic aspects of model variability may automatically be given high weight through the optimisation procedure. The solution is to confine attention to aspects of the model and of the real climate system in which the model simulation of internal climate variability is adequate – or, more accurately, cannot be shown to be inadequate. We propose a simple consistency check based on standard linear regression which can be applied to both the space-time and frequency-domain approaches to optimal detection and demonstrate the application of this check to the problem of detection and attribution of anthropogenic signals in the radiosonde-based record of recent trends in atmospheric vertical temperature structure. We also suggest results should be reported in terms of return-times in place of the usual confidence intervals, return-times being more transparent and less dependent on the distribution of climate noise.

# 1 Introduction

A common overall approach has emerged to the detection of anthropogenic climate change. A detection statistic is defined and evaluated in an observational dataset. This might be a global mean quantity (e.g. *Stouffer* et al., 1994); a model vs. observation pattern correlation (*Mitchell* et al., 1995a; *Tett* et al., 1996); the observed trend in pattern correlation (*Santer* et al., 1996); or some form of "optimised fingerprint" (*Hasselmann*, 1979; *Hannoschöck & Frankignoul*, 1985, *Bell*, 1986; *Hasselmann*, 1993, *Santer* et al., 1994a; *North* et al., 1995, *Hegerl* et al., 1996; *Stevens & North*, 1997). The same detection statistic is then evaluated treating sections of a control run of a climate model (in which there is no secular change in forcing) as "pseudo-observations" to provide an estimate of the distribution of that statistic under the null-hypothesis of no anthropogenic change. If the observed value of the chosen statistic lies in the uppermost $100P^{\text{th}}$ percentile of the distribution estimated from the control, then detection is claimed with a $100P\%$ risk of a type-1 error (so $P$ = probability of a false positive). Clearly, this approach to quantifying uncertainty, or the risk of error, in claims of detection and attribution requires complete confidence in the realism of the model simulation of internal climate variability.

*Hasselmann*, 1997, distinguishes between "detection" of anthropogenic climate change (ruling out, at a certain confidence level, the possibility that an observed change is due to internal variability alone) and "attribution" (demonstrating that the observed change is consistent with the predictions of a climate model subjected to a particular forcing scenario and inconsistent with all physically plausible alternative causal explanations). Formal attribution is clearly a much more demanding objective than detection. Indeed, as *Hasselmann*, 1997, observes, it is a logical impossibility unless we use physical arguments to confine attention *a priori* to a relatively small number of alternative explanations. The attribution framework proposed by *Hasselmann*, 1997, and implemented by *Hegerl* et al., 1997, also relies heavily on model-simulated climate variability, because "consistent" and "inconsistent" are formally defined as "within the bounds of variability as simulated by a particular climate model".

Following standard practice, we will distinguish between "internal" (unforced) climate variability and the climate system's response to time-varying natural forcings such as changes in the solar constant. If the temporal history of these natural forcings is known, and the response mechanism can be accurately modelled, these can be treated exactly like an anthropogenic forcing (e.g. *Hegerl* et al., 1997). If the forcing histories are unknown, they must be treated as sources of internal variability.

We have a number of *a priori* reasons to distrust model simulations of internal climate variability. On the simplest level, there are known sources of variability in the observational record (the simplest example being observation error) which are not represented in current models. Even if these additional sources are included in the model, it will always be the case that variability on small spatio-temporal scales is likely to be under-represented in any finite representation of a continuous turbulent system. Fortunately, we do not require a model simulation of internal variability to be accurate in every respect for the model to be used for uncertainty analysis in climate change detection and attribution. In principle, only those aspects of model behaviour which are relevant to the detection and attribution problem need to be realistic. For example, if

our chosen detection statistic is the global mean temperature, then all we require is an estimate of the variability of this quantity on the relevant timescales. The problem is determining which aspects of model variability are crucial to a particular detection or attribution problem and developing quantitative measures of model adequacy.

Simple checks, such as the comparison of global mean power spectra, can identify gross deficiencies in model variability, but the problem of how to remove the (presumed, but unknown) anthropogenic signal from the historical record prior to computing a power spectrum remains. Proxy and incomplete observations of the pre-industrial period (e.g. *Bradley & Jones*, 1993) can help here, but separating low-frequency climate variability from slow changes in the relationship between proxy observations and the climatic variables which they are supposed to represent remains a problem. There is also the intrinsic difficulty that paleo-climate observations are sparse, so a paleo-climate reconstruction of any climate index must be contaminated with the high-spatial-wavenumber components of variability which models are known to simulate poorly (*Stott & Tett*, 1997) and which, it is hoped, are irrelevant to climate change detection. This may be an issue for recent pioneering studies comparing model-simulated variability with the paleo-climate record (e.g. *Barnett* et al., 1996).

The other problem with global mean power spectra is that a deficiency in the model's internal variability may fail to show up in the global mean while having a significant impact on the chosen detection statistic (this is necessarily true if a "centred" statistic is used, which is defined to be independent of the global mean – *Santer* et al., 1993). Recognising this, *Hegerl* et al., 1996, use a linear response model to estimate and remove the anthropogenic signal from the historical record and then use the residual as an estimate of natural variability. While clearly an advance on simple power spectra, this approach relies uncomfortably on the adequacy of a very simple linear model for both the form and amplitude of the anthropogenic signal. They note that it would tend to give a very conservative estimate of uncertainty, because errors in the model compound genuine natural variability in the observations. This may be unimportant if all that is being tested is the null-hypothesis of zero climate sensitivity (or, more precisely, no response to the candidate forcing – the crudest form of "detection") but when these techniques are extended to the attribution problem, or to provide error estimates on forecasts of $21^{st}$ century climate change, an excessively conservative estimate of uncertainty is as misleading as an excessively optimistic one.

The crucial question is this: is the model simulation of internal climate variability adequate to quantify uncertainty in global change detection? Or to rephrase the question in a testable form: do we have reason to distrust the results of this particular application of the model? The notion of adequacy for a particular task is crucial. It will always be possible to identify deficiencies in some aspect of model climatology or simulated climate variability, and therefore misleading to insist that the model be absolutely realistic on all spatio-temporal scales before it can be trusted for climate applications. In the following section, we attempt to address this question in the context of the "optimal fingerprint" approach to climate change detection and attribution.

# 2 Fingerprinting as generalised linear regression

Although it has appeared in various guises (*Hasselmann*, 1979; *Bell*, 1986; *Santer* et al., 1994b; *North* et al., 1995; *Thacker*, 1996), the basic principle of "optimal" detection is the classical technique of generalised linear regression (see *Mardia* et al., 1979, for a helpful introduction). In order to stress this link, we use the standard notation of the linear regression literature. A set of $m$ "guess patterns", each consisting of a rank-$n$ vector representing the pattern of the climate system's response to a particular external forcing scenario, provide the independent variables of the regression model. We denote these guess patterns as the columns of the $n \times m$ matrix $\mathbf{X}$. Typical examples include the pattern of surface or vertical temperature change which is expected to result from increasing concentrations of greenhouse gases, anthropogenic sulphate aerosols, declining stratospheric ozone, aerosols from volcanic eruptions or some combination of these. The individual elements of $\mathbf{X}$ correspond to the local trend at a particular latitude-longitude or (in the "vertical detection" problem discussed here) latitude-height location. In our discussion here, we shall assume that $\mathbf{X}$ is real, although in the frequency-domain representation of *North* et al., 1995, elements may correspond to complex coefficients after the data have been Fourier transformed in time. The same basic principles apply in both cases (*Hegerl & North*, 1997).

Guess patterns may be defined *a priori*, or using simple physical arguments based on the pattern of the forcing (as in *Santer* et al., 1996) or by averaging the response to that forcing scenario from an ensemble of runs of a climate model (as in *Tett* et al., 1996). For consistency with *Hasselmann*, 1997, we shall base our optimisation procedure on the assumption that the guess patterns may be treated as noise free. With only a four-member ensemble in the example in section 5, this is clearly incorrect, so our procedure remains sub-optimal in this respect. We do, however, take residual noise in $\mathbf{X}$ into account in our analysis of uncertainty, as detailed below.

All current approaches to optimal detection are based on the assumption that the recent climate record may be represented as a linear superposition of these model-predicted guess-patterns plus an additive noise term. Thus the detection problem simply involves estimating the amplitude of these patterns in a rank-$n$ vector of observations, $\mathbf{y}$, or estimating the parameters $\boldsymbol{\beta}$ in the basic linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1}$$

where $\mathbf{u}$ is the "climate noise" term whose covariance is given by the $n \times n$ matrix $\mathbf{C}_N$:

$$\mathbf{C}_N \equiv \mathcal{E}(\mathbf{u}\mathbf{u}^T) \tag{2}$$

Under the assumption that $\mathbf{u}$ is multivariate normal (which we will return to below), the best (lowest variance) linear unbiased (BLUE) estimator of $\boldsymbol{\beta}$ in (1) may be found by introducing a "pre-whitening" coordinate transformation $\mathbf{P}$ such that

$$\mathcal{E}(\mathbf{P}\mathbf{u}\mathbf{u}^T\mathbf{P}^T) = \mathbf{P}\mathbf{C}_N\mathbf{P}^T = \mathbf{I}. \tag{3}$$

The term pre-whitening refers to the fact that the transformed noise, $\mathbf{Pu}$, appears to be "white" (uncorrelated and Gaussian distributed) in these transformed coordinates.

Equation (3) is satisfied if $\mathbf{P}^T\mathbf{P} = \mathbf{C}_N^{-1}$, provided this inverse exists. Because $\mathbf{Pu}$ is indistinguishable from white noise, we may invoke the Gauss-Markov theorem to prove that the following estimator for $\boldsymbol{\beta}$ is BLUE:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{P}^T\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{P}^T\mathbf{P}\mathbf{y} = (\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{y}. \tag{4}$$

This is simply the ordinary least squares solution applied to the transformed (rotated and weighted) variables. The link to standard regression is most transparent in the case of a single-pattern with uncorrelated noise (i.e. when $\mathbf{X}$ has only a single column and $\mathbf{C}_N$ is diagonal):

$$\tilde{\beta} = \frac{\sum_i \frac{x_i y_i}{\lambda_i^2}}{\sum_i \frac{x_i^2}{\lambda_i^2}}, \tag{5}$$

where $\lambda_i^2$ is the expected noise variance in the $i^{\text{th}}$ component of $\mathbf{y}$.

For reference, the $\nu^{\text{th}}$ row of $\mathbf{X}^T\mathbf{C}_N^{-1}$ in (4) corresponds to the $\nu^{\text{th}}$ fingerprint $f_\nu^i$ in equation (30) of *Hasselmann*, 1997, while the matrix $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ corresponds to the metric $D_{\nu\mu}$ in his equation (31) and $\tilde{\boldsymbol{\beta}}$ corresponds to the detection coefficients, $d^\nu$ in his equation (33).

An estimate of the variance of $\tilde{\boldsymbol{\beta}}$ is given by

$$\tilde{V}(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X})^{-1} \tag{6}$$

which, provided $\mathbf{u}$ is multivariate normal, can be translated into a confidence ellipsoid. That is, the quantity:

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_m^2, \tag{7}$$

meaning that the LHS of (7) is distributed like $\chi^2$ with $m$ degrees of freedom. To bound the region corresponding to a given $P$-value (where $P$ is the probability that the true value of $\boldsymbol{\beta}$ lies *outside* this region), we find the critical value of $\chi^2$ for which $P(\chi^2 > \chi^2_{\text{crit}}) = P$ and plot the values of $\boldsymbol{\beta}$ for which the LHS of (7) is equal to this value. Again, in the single-pattern, uncorrelated noise, example, equation (6) becomes

$$\mathcal{E}[(\tilde{\beta} - \beta)^2] = \frac{1}{\sum_i \frac{x_i^2}{\lambda_i^2}}. \tag{8}$$

If we wish to compute the joint distribution of a subset of the parameters in the multi-pattern case, we simply extract the relevant rows and colums from $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ and evaluate (7) with this reduced number of degrees of freedom (see *Press* et al., 1992 for a clear discussion of this point). The confidence intervals thus obtained represent an estimate of our uncertainty in the factors by which we have to scale the model response to the various forcings to match what is taking place in the real world.

This estimate of the variance of $\tilde{\boldsymbol{\beta}}$ also provides an estimate of the implied uncertainty in any scalar linear diagnostic, $\phi$. With trivial exceptions, $\phi$ can always be represented as a projection of the observations onto a vector of weights, or $\phi = \mathbf{w}^T\mathbf{y}$. If the elements of $\mathbf{w}$ are all equal to $1/n$, for example, then $\phi$ is simply the global mean. If $\mathbf{w}$ is a unit vector, then $\phi$ is the value of the observation-vector at a particular location and so on.

Neglecting uncertainty in $\mathbf{X}$ as before, the variance of $\phi$ attributable to the uncertainty in $\tilde{\boldsymbol{\beta}}$ is:

$$\tilde{V}(\phi) = \mathbf{w}^T \mathbf{X} \tilde{V}(\tilde{\boldsymbol{\beta}}) \mathbf{X}^T \mathbf{w}. \tag{9}$$

By assessing the extent to which trends at individual locations or in global-mean quantities are consistent with optimal detection results in this way, we can move on from the simple yes/no question of whether the observations are globally consistent with the predictions of a climate model, to investigate which aspects of the observational record disagree most strongly with the model predictions, identifying likely model errors.

The fact that we are using a linear model in (1) does not mean that we cannot examine problems in which non-linearity is important. For example, suppose a model forced with the combined effects of changing sulphate-aerosol and greenhouse-gas levels gave a pattern of change which was significantly different to the sum of the patterns obtained in runs forced with each of these factors alone (significance might prove very difficult to establish without very large ensembles of runs, but suppose the non-linearity is strong enough that it is possible). We can then use the difference between the combined pattern and the sum of the two individual patterns to define a "fingerprint" of this non-linearity. This, too, can then be searched for in the observations to establish whether such non-linearity is detectable in the real world.

The key advantage of this regression-based approach over detection schemes based on pattern correlation (e.g. *Mitchell* et al., 1995a; *Santer* et al., 1996; *Tett* et al., 1996) is that it provides information on relative pattern amplitudes in model and observations: correlations convey no amplitude information. If the guess patterns are based on an ensemble-average of model simulations with forcing changes matched to the period of the observations, *and* the model has the timing and amplitude of the response to these forcing changes exactly right, then the expected value of the estimated pattern-amplitude coefficients, $\mathcal{E}(\tilde{\boldsymbol{\beta}})$, will be approximately unity. As noted above, this expectation is only approximate because the assumption that $\mathbf{X}$ is noise free is only strictly correct in the limit of an infinite ensemble. In general, noise in $\mathbf{X}$ will tend to bias $\tilde{\boldsymbol{\beta}}$ towards zero (*Mardia* et al., 1979) and increase the true variance in the estimator by a factor of approximately $1+1/M$, where $M$ is the ensemble size. We overcome this second problem by simply inflating $\tilde{V}(\tilde{\boldsymbol{\beta}})$ by this factor, but the bias in $\tilde{\boldsymbol{\beta}}$ remains, making the overall algorithm slightly over-conservative. The derivation of alternative unbiased estimators in the presence of noise in both $\mathbf{X}$ and $\mathbf{y}$ is straightforward (e.g. *Ripley & Thompson*, 1987), but we will examine these in detail elsewhere.

# 3 Estimating the climate noise covariance

The key difficulty in optimal fingerprinting is that $\mathbf{C}_N$ is unknown and is estimated from a control integration of the climate model thus:

$$\hat{\mathbf{C}}_N = \mathbf{Y}_C \mathbf{Y}_C^T \tag{10}$$

where the columns of $\mathbf{Y}_C$ represent a succession of $\mathbf{y}$-like vectors of "pseudo-observations" extracted from the control. As far as possible, these pseudo-observations must be

calculated in such a way as to mimic the observation vectors, including in particular applying the same observation mask to account for the effects of missing data.

Since $\mathbf{y}$ typically represents trends over a 30–50-year period, and control integrations are necessarily limited to 1,000-2,000 years duration, the number of independent vectors of "pseudo-observations" in a typical control run (the rank of $\mathbf{Y}$) is orders of magnitude less than the number of degrees of freedom, $n$. The estimated covariance matrix obtained from the control, $\hat{\mathbf{C}}_N$, is therefore non-invertible.

One solution to this problem is obtained by noting that we do not actually require $\mathbf{C}_N^{-1}$ for $\tilde{\boldsymbol{\beta}}$ to be BLUE. We only require that the transformation $\mathbf{P}$ is such that (3) is satisfied and the unit matrix on the RHS of (3) need not be of rank $n$. If we assume that $\hat{\mathbf{C}}_N$ provides a reliable estimate of the noise covariance only in the subspace spanned by the $\kappa$ highest-variance EOFs of the control (eigenvectors of $\hat{\mathbf{C}}_N$), then a natural transformation to use is $\mathbf{P} = \mathbf{F}^T$ where the columns of $\mathbf{F}$ are the $\kappa$ highest-variance EOFs of the control weighted by their inverse singular values (square root of the corresponding eigenvalues of $\hat{\mathbf{C}}_N$).

This is equivalent to using the Moore-Penrose pseudo-inverse, $\mathbf{F}\mathbf{F}^T$ in place of $\mathbf{C}_N^{-1}$. The pseudo-inverse based on the EOFs of the control seems the most natural one to use, but others are also possible: for example, *Hegerl* et al., 1996, use the EOFs of one of their forced runs. This seems reasonable when only a single forcing is under consideration, but introduces a bias towards one scenario over another when $m > 1$, which may be an important consideration in attribution studies. We are also concerned about the impact on algorithm stability of including basis-vectors which are known to be poorly sampled in the control integration: all things considered, using the EOFs of the control may compromise the power of the detection algorithm, we believe it is the approach least likely to give misleading results.

The problem is that key results depend critically and predictably on the choice of $\kappa$: in general, the estimated uncertainty envelope around $\tilde{\boldsymbol{\beta}}$ shrinks close to monotonically with increasing $\kappa$, so (in a detection problem) the confidence level at which the null-hypothesis of zero climate sensitivity can be rejected increases predictably with $\kappa$ even when this null-hypothesis is valid. The reason is that increasing $\kappa$ introduces EOFs in which the variance in the control is unrealistically low. These will automatically be given high weight by the optimisation procedure.

The most obvious source of this problem, which is also the simplest to deal with, is that low-ranked EOFs of the control will generally contain unrealistically low variance due to sampling deficiencies: these correspond to state-space directions which were not visited during this relatively short control integration. Although $\mathbf{F}^T\hat{\mathbf{C}}_N\mathbf{F} = \mathbf{I}$ by construction, $\hat{\mathbf{C}}_N \neq \mathbf{C}_N$ because of the finite length of the control, so equation (3) is only approximately satisfied. Worse, because the EOFs of the control have been chosen to maximise variance in the particular segment, $\mathbf{Y}_C$, the transformation $\mathbf{F}$ is biased with respect to that particular segment. Applied to another, arbitrarily selected, segment of the control with covariance matrix $\hat{\mathbf{C}}_{N_2}$, the diagonal elements of $\mathbf{F}^T\hat{\mathbf{C}}_{N_2}\mathbf{F}$ will, on average, tend to be less than unity. This is important because it introduces a bias in the estimate of the covariance of $\tilde{\boldsymbol{\beta}}$ (*Bell*, 1986). Recognising this, *Hegerl* et al., 1996, stipulate that different control runs, possibly from different models, are used for optimisation and hypothesis

testing. Thus we replace (6) with the estimate

$$\tilde{V}'(\tilde{\boldsymbol{\beta}}) = (\mathbf{X}^T\hat{\mathbf{C}}_{N_1}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{C}}_{N_1}^{-1}\hat{\mathbf{C}}_{N_2}\hat{\mathbf{C}}_{N_1}^{-1}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{C}}_{N_1}^{-1}\mathbf{X})^{-1} \tag{11}$$

where $\hat{\mathbf{C}}_{N_1}$ and $\hat{\mathbf{C}}_{N_2}$ are estimated from different control integrations or independent segments of the same integration, $\mathbf{Y}_{C_1}$ and $\mathbf{Y}_{C_2}$. Notice that equation (11) collapses to equation (6) in the limit of a long control integration, as $\hat{\mathbf{C}}_{N_2} \to \hat{\mathbf{C}}_{N_1} \to \mathbf{C}_N$.

If the covariance estimate $\hat{\mathbf{C}}_{N_2}$ has $\nu$ degrees of freedom (for example, if it was computed by averaging over $\nu$ independent realisations of pseudo-observations extracted from the control), then equation (7) for the errors in $\tilde{\boldsymbol{\beta}}$ is replaced by

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[\tilde{V}'(\tilde{\boldsymbol{\beta}})]^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \epsilon^2(\boldsymbol{\beta}) \sim mF_{m,\nu}, \tag{12}$$

the standard $F$ distribution with $m$ and $\nu$ degrees of freedom in the numerator and denominator respectively. Thus a confidence ellipsoid around our "best-guess" value, $\tilde{\boldsymbol{\beta}}$, can be found by plotting the locus of points $\boldsymbol{\beta}$ for which $\epsilon^2(\boldsymbol{\beta})$ is equal to the corresponding critical value of the $F_{m,\nu}$ distribution.[1]

The RHS of equation (12) only converges to $\chi^2_m = mF_{m,\infty}$ (corresponding to an infinitely long control, in which case equations (12) and (7) become equivalent) for $\nu > 100$. In a 50-year diagnostic, this would require control runs of several thousand years, which are not generally available. Much attention has therefore been devoted to the estimation of $\nu$, the "true" number of degrees of freedom of a relatively short control integration – see, for example, *Zweirs & von Storch*, 1995, and references therein. This is important because an over-estimate of $\nu$, due to the neglect of serial correlation in $\mathbf{Y}_{C_2}$, can lead to spuriously high estimates of significance. *Zweirs & von Storch*, 1995, propose a correction for $\nu$ based on the assumption that the temporal evolution of all these scalar diagnostics in the control run can be represented by first-order autoregressive processes, or "AR(1) noise." The problem (noted by *Zweirs & von Storch*, 1995, themselves) is that the control model is not in fact a linear stochastic process at all, even though it may be indistinguishable from one, so there is no rigorous answer to the question of what is the "correct" value of $\nu$, and results can depend disconcertingly heavily on the method used to estimate it. For example, temporal correlations will generally depend on spatial scale: the projection of the control onto a highly structured spatial pattern may be much less autocorrelated in time than the projection onto a very smooth, large-scale pattern. In a multi-pattern analysis, which autocorrelation coefficient is appropriate? In the analysis presented here, we use the largest one, giving the most conservative estimate of $\nu$, but can see no rigorous justification for this choice.

A much more transparent approach, which does not rely on any degrees-of-freedom estimates at all, is to focus the reporting of results onto return-times rather than confidence intervals. We evaluate $\epsilon^2(\boldsymbol{\beta})$ over all vectors of pseudo-observations in $\mathbf{Y}_{C_2}$ (recalling that $\boldsymbol{\beta} = \mathcal{E}(\tilde{\boldsymbol{\beta}}) = \mathbf{0}$ in the control), and simply take the maximum value, $\epsilon^2_{\max}$ as indicating a "300-year error" (that is, the maximum Mahalanobis-weighted error in

[1] When only a single guess-pattern is under consideration, $m = 1$, the Student's $t$-distribution may be used instead, but this is trivially related to the $F_{1,\nu}$-distribution so we will not discuss it here for the sake of brevity.

$\boldsymbol{\beta}$ likely to be observed in a 300-year segment of the control, assuming that is what was available for hypothesis-testing). We then plot $\boldsymbol{\beta}$ for which the LHS of equation (12) equals $\epsilon^2_{\max}$. It may readily be shown that this corresponds approximately to a $P$-value of $\ln(2)/\nu$, that is, $\epsilon^2_{\max}$ will exceed this critical value of the underlying distribution in approximately 50% of cases, irrespective of the shape of that distribution.

Reporting return-times and approximate $P$-values based on the estimated degrees-of-freedom $\nu$ has four clear advantages over confidence intervals based on an assumption of multivariate normality:

1. they are conceptually simpler for presentation of results to non-specialists: "the control did not move outside this region in 300 years";
2. they rely much less on distributional assumptions – we require only that the distribution of $\tilde{\boldsymbol{\beta}}$ is radially symmetric under the norm defined by $\tilde{V}'(\tilde{\boldsymbol{\beta}})$, not that this distribution is Gaussian;
3. the role of the estimated degrees-of-freedom of $\mathbf{Y}_{C_2}$ is completely transparent;
4. and most importantly, they explicitly discourage claims of significance which involve extrapolation beyond the region explored by the control.

For example, if $\nu = 15$, so $\ln(2)/\nu = 0.05$, this the smallest P-value which can be quantified legitimately. If the observations lie well outside the region defined by $\epsilon^2_{\max}$, then all that can be said is that we have detected a model-data discrepancy at $P_{<0.05}$. Using an $F$-test to claim significance at the $P_{0.001}$ level, for example, implies we can extrapolate from observations of the central body of the distribution right out into the tails, which is clearly unsafe.

Whichever approach is adopted to define uncertainty intervals, we still rely on the assumption that $\hat{\mathbf{C}}_{N_1}$ and $\hat{\mathbf{C}}_{N_2}$ are individually realistic, or at least that errors in the representation of climate variability in the two control runs are unrelated. Even if separate models are used, any such independence assumption for different climate models is suspect, because these models have so much (often, entire components) in common. If, as is likely, both models display too little variance on small spatial scales, both $\hat{\mathbf{C}}_{N_1}$ and $\hat{\mathbf{C}}_{N_2}$ will be subject to a similar bias, compromising analysis of uncertainty.

Indeed, these biases due to systematic deficiencies in the simulation of climate variability in the control overwhelm, at least in the vertical detection problem, any bias due to random sampling in the estimation of $\mathbf{C}_N$. In the language of dynamical systems, even with these relatively short control integrations, the model attractor is already sufficiently well-sampled that the differences between its underlying shape and that of the attractor of real climate variability have already become more important than differences between the true shape and the sampled shape of the model attractor.

# 4 Consistency checks to detect model inadequacy

Having framed the optimal fingerprinting algorithm as a linear regression problem, a variety of simple checks for model adequacy immediately present themselves, drawn from

the standard statistical literature. For simplicity, following *Hasselmann*, 1997, we will focus on parametric tests based on the assumption of multivariate normality. To judge from the analyses we have performed to date, the assumption of normality is likely to be reasonably close to valid for temperature data on large spatio-temporal scales. Assuming normality for other data types (such as precipitation) would be more problematic. The problem is that formal tests for multivariate normality are not particularly powerful when applied to the relatively small sample sizes we encounter here (i.e. they are unlikely to identify weak departures from normality), so even if the data depart from normality, it would be difficult to identify and characterise the departure.

Our null-hypothesis, $\mathcal{H}_0$, is that the control simulation of climate variability is an adequate representation of variability in the real world in the truncated state-space which we are using for the analysis – i.e. the subspace defined by the first $\kappa$ EOFs of the control run does not include patterns which contain unrealistically low (or high) variance in the control simulation of climate variability. Because the effects of errors in observations are not represented in the climate model, $\mathcal{H}_0$ also encompasses the statement that observational error is negligible in the truncated state-space (on the spatio-temporal scales) that is used for detection. A test of $\mathcal{H}_0$, therefore, is also a test of the validity of this assumption.

If we are unable to reject $\mathcal{H}_0$, then we have no explicit reason to distrust uncertainty estimates based on our analysis. This does not, of course, mean that these uncertainty estimates are necessarily correct. It may mean only that the tests we have devised are not powerful enough to identify some crucial deficiency in model simulated variability. But it is important to recognise that the demonstration of internal consistency is all that can ever be expected from a formal attribution study. Proof that the model is "correct", meaning that every alternative has been taken into account and rejected, is a logical impossibility.

We formulate a simple test of this null-hypothesis as follows: if $\mathcal{H}_0$ is true then the residuals of regression,

$$\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}, \tag{13}$$

should behave like mutually independent, normally distributed random noise in the coordinate system (under the norm) defined by $\hat{\mathbf{C}}_N^{-1}$, so

$$\tilde{\mathbf{u}}^T \hat{\mathbf{C}}_N^{-1} \tilde{\mathbf{u}} \sim \chi^2_{\kappa-m}, \tag{14}$$

meaning that the LHS of equation (14) is distributed like the sum of the squares of $\kappa - m$ normally-distributed random variables. If an increase in $\kappa$ introduces EOFs of the control which contain unrealistically low variance, then the LHS of (14) will move to an improbably high percentile of the $\chi^2_{\kappa-m}$ distribution, and $\mathcal{H}_0$ will be rejected, giving us some warning that estimates of uncertainty are then likely to be unreliable.

Considering again the case of a single-pattern with uncorrelated noise, equations (13) and (14) become

$$\sum_{i=1}^{\kappa} \frac{(y_i - \tilde{\beta} x_i)^2}{\lambda_i^2} \sim \chi^2_{\kappa-1}. \tag{15}$$

Terms in which the control variance is unrealistically low correspond to small values of $\lambda_i^2$ which inflate the LHS of (15) into a high percentile of the $\chi^2$ distribution.

In geometric terms, the $\chi^2$ test involves summing residuals over all state-space directions in the subspace defined by EOFs 1 to $\kappa$ of the control which are orthogonal to the hyperplane defined by the guess-patterns, $\mathbf{X}$, where orthogonality is defined in terms of the metric given by $\hat{\mathbf{C}}_N^{-1}$ (i.e. two vectors, $\mathbf{a}$ and $\mathbf{b}$ are orthogonal if $\mathbf{a}^T\hat{\mathbf{C}}_N^{-1}\mathbf{b} = 0$). If, by increasing $\kappa$, we introduce an EOFs in which control variance is unrealistically low then the component of that EOF which lies in the plane defined by $\mathbf{X}$ will tend to distort uncertainty analysis in the regression but, at the same time, the component orthogonal to $\mathbf{X}$ will tend to inflate the LHS of (14) faster than we would expect it to rise if the control variability is adequate, giving us some warning that uncertainty estimates are becoming unreliable.[2]

We stress that if there is a component of natural variability that is incorrectly simulated by the control and is associated with a pattern *identical* to the predicted pattern of anthopogenic change, the $\chi^2$ test will fail to identify any inconsistency. It should be intuitively clear that, with only a single vector of observations, $\mathbf{y}$, an error in simulated variability whose properties are statistically identical to the predicted anthropogenic change cannot, by definition, be identified through statistical analysis. If, on the other hand, a series of detection experiments are performed, for example on successive 50-year segments of the observational record as in *Hegerl* et al., 1996, then the $\chi^2$ test can readily be generalised to check directions lying in the plane defined by $\mathbf{X}$, provided that some sort of smoothness assumption could be made concerning the temporal evolution of the anthropogenic signal. For the sake of simplicity, we postpone discussion of this generalisation to a future publication. In the "vertical detection" problem we use as the example here, this option is not available because we are investigating 35-year trends in a 35-year dataset, so we only have a single $\mathbf{y}$ to work with.

If independent control runs are used for optimisation and testing then, strictly speaking, an $F$-test should be used in place of $\chi^2$ to take into account the effects of uncertainty in the projection of $\hat{\mathbf{C}}_{N_2}$ onto the EOFs of $\hat{\mathbf{C}}_{N_1}$. In practice, we have found this makes very little difference, because the tests are being used simply to place a crude upper limit on the truncation level.

Aware that truncating at too high a level raises problems in optimal fingerprinting, *Hegerl* et al., 1996, use a simple criterion to determine the truncation level based on the correlation between the unrotated guess patterns (columns of $\mathbf{X}$) and rotated fingerprints (rows of $(\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_N^{-1}$). If this correlation drops below some cutoff value, they conclude that the optimisation is "introducing noise" and reduce the truncation. In advocating something slightly more complicated, we feel obliged to detail what we see as the potential problems with the *Hegerl* et al., 1996, approach while stressing that there is no reason why their approach and ours should not give similar results in a particular application. The key problem with the *Hegerl* et al., 1996, correlation criterion is that it is insensitive to the global variance in the control. If the model consistently underestimates variability on all spatio-temporal scales then the rotation at a given truncation and therefore the correlation between guess-pattern and fingerprint is unaffected. *Hegerl*

[2]A residual check based on the $\chi^2$ statistics has been proposed independently by *Leroy*, 1998 – we are grateful to G. Hegerl for drawing our attention to this work.

et al., 1996, use other indicators like the power spectra of global mean quantities to check that global variance in the control is not inconsistent with the observations, but because these indicators are not specific to the truncated state-space used for detection, their use might lead to the model being rejected even when model variability is realistic in that truncated space. Perhaps worse, a problem in model variablity which did not happen to project onto the global mean might pass unnoticed.

A second problem with the *Hegerl* et al., 1996, correlation criterion is that it may render optimisation useless in precisely the situation where it is most needed. When the unrotated guess patterns are completely dominated by regions or spatio-temporal scales in which the climate noise is also very high, the correlation criterion may indicate truncating at a value of $\kappa$ which excludes all EOFs containing a reasonable level of signal-to-noise even when there is a genuinely detectable signal and the control simulation of natural variability is perfectly adequate.

Instead of selecting the truncation level *a priori* or using some more-or-less arbitrary criterion, we evaluate (14) against the standard $\chi^2$ distribution to establish the maximum value of $\kappa$ for which the control still gives a believable estimate of climate noise. Detection can then only be claimed if the null-hypothesis of zero climate sensitivity can be rejected for values of $\kappa$ smaller than this limit. An example of the application of this test to the "vertical detection" problem is given in the following section.

# 5 An example: the "vertical detection problem"

We examine results from the application of the algorithm described above to the comparison of the observational record of atmospheric vertical temperature structure over the period 1961–1995 with a series of simulations from the HadCM2 (*Johns* et al., 1997) coupled climate model: the example considered by *Tett* et al., 1996. The observation vector, $\mathbf{y}$, is based on operationally received radiosonde data expressed as anomalies about the 1971-90 period. These were monthly averaged on a 10° longitude by 5° latitude grid on standard pressure levels (850, 700, 500, 300, 200, 150, 100 and 50hPa). Annual averages were computed at each latitude/pressure point in which there were more than 8 months with data.

Following *Tett* et al., 1996, we compute vertical profiles of the zonal mean differences between the period 1961–80 and 1986–95. To minimise the impact of volcanos, data for 1963–4 (Mt. Agung) and 1992 (Mt. Pinatubo) are omitted (the eruption of El Chichon in 1981 should not affect this particular diagnostic, being outside either period). Latitude/pressure points with fewer than 20% (50%) of the years with data in the 1961–80 (1986–95) periods are also set to missing. The upper panel in figure 1 shows the resulting pattern of vertically resolved temperature changes.

We also extract precisely the same diagnostic (applying the same missing data mask to the zonal mean temperatures) from a series of experiments performed with the HadCM2 coupled general circulation model. The resolution of both atmosphere and ocean components of the model is 3.75° longitude by 2.5° latitude with 19 vertical levels in the atmosphere and 20 in the ocean. This model has been extensively investigated for global change detection and prediction purposes (e.g. *Mitchell* et al., 1995a; *Johns* et al.,

1997), and generates internal variability when integrated in a "control" configuration (no change in forcing) which compares reasonably well with that observed in the real world (*Tett* et al., 1997). The lower panel in figure 1 shows two standard deviations of our chosen diagnostic estimated from 40 35-year-long segments extracted at 10-year intervals from a 426-year control integration and masked using the pattern of missing data in the observations: the columns of $\mathbf{Y}_{C_1}$ (a separate 310-year segment is used to provide $\mathbf{Y}_{C_2}$ for hypothesis-testing). The trends in the observations are evidently significant relative to internal climate variability as simulated by HadCM2. The question we address here is whether they can they be attributed to anthropogenic influences.

We compare these observed zonal mean temperature changes with changes simulated in two sets of experiments performed with the HadCM2 climate model. In the first ensemble of four integrations (initialised from points in the control integration separated by 150 years), denoted G, the model was forced with the effects of observed changes in $CO_2$, methane and chlorofluorocarbons (expressed as equivalent-$CO_2$) for the period 1860 to 1996. The upper panel shows the ensemble mean of an identical diagnostic to that shown in figure 1 extracted from the model years 1961–95. A second ensemble of four integrations, denoted GSO and shown in the lower panel, included a simple parameterisation of the effects of sulphate aerosols (*Mitchell* et al., 1995b) and an estimate of the effect of declining stratospheric ozone after 1974 based on extrapolating trends observed by the Total Ozone Mapping Spectrometer for the period 1979 to 1989.

The contribution of changing aerosols to the vertical pattern of temperature change, modelled in a third ensemble (GS) in which ozone levels were held constant, is relatively minor. For the sake of brevity we do not discuss GS results here, but for the vertical detection problem, they are generally similar to results from G.

In all the results reported here, we use a mass-based weighting on all patterns. This has no direct impact on the estimation step once the truncation space has been defined (because the climate noise covariance provides its own, physically-based, weighting function), but it does impact the EOF-decomposition of $\mathbf{Y}_{C_1}$. Using mass weighting means that high-ranked EOFs have substantial loading in the troposphere, whereas high-ranked EOFs based on a volume weighting, for example, are completely dominated by the stratosphere. This turns out to be important because the model simulation of stratospheric variability is less realistic than its simulation of tropospheric variability, so the use of a volume-based (log-pressure) weighting function leads to the model being rejected by our internal consistency checks before we find we can detect anything.

We begin by testing a simple univariate model: assuming that the observations consist only of a scaled version of G (greenhouse gas pattern) with additive climate noise. The diamonds in figure 3 show $\beta_G$, the estimated amplitude of the G pattern, as a function of the rank of the detection space ($\kappa$ =no. of EOFs retained of the control). Vertical bars show the $P_{0.05}$ (two-tailed) confidence interval based on an assumed Gaussian distribution, while the horizontal dashes show the "310-year error" range, ±the largest absolute estimated pattern-amplitude ($\sqrt{\epsilon^2_{\max}}$) observed in a 310-year control integration. These ranges approximately match, as we would expect if there are 10-15 degrees of freedom in $\mathbf{Y}_{C_2}$ (since there are fewer than 10 non-overlapping 35-year segments in this control integration, this indicates a modest increase in degrees of freedom has been gained by

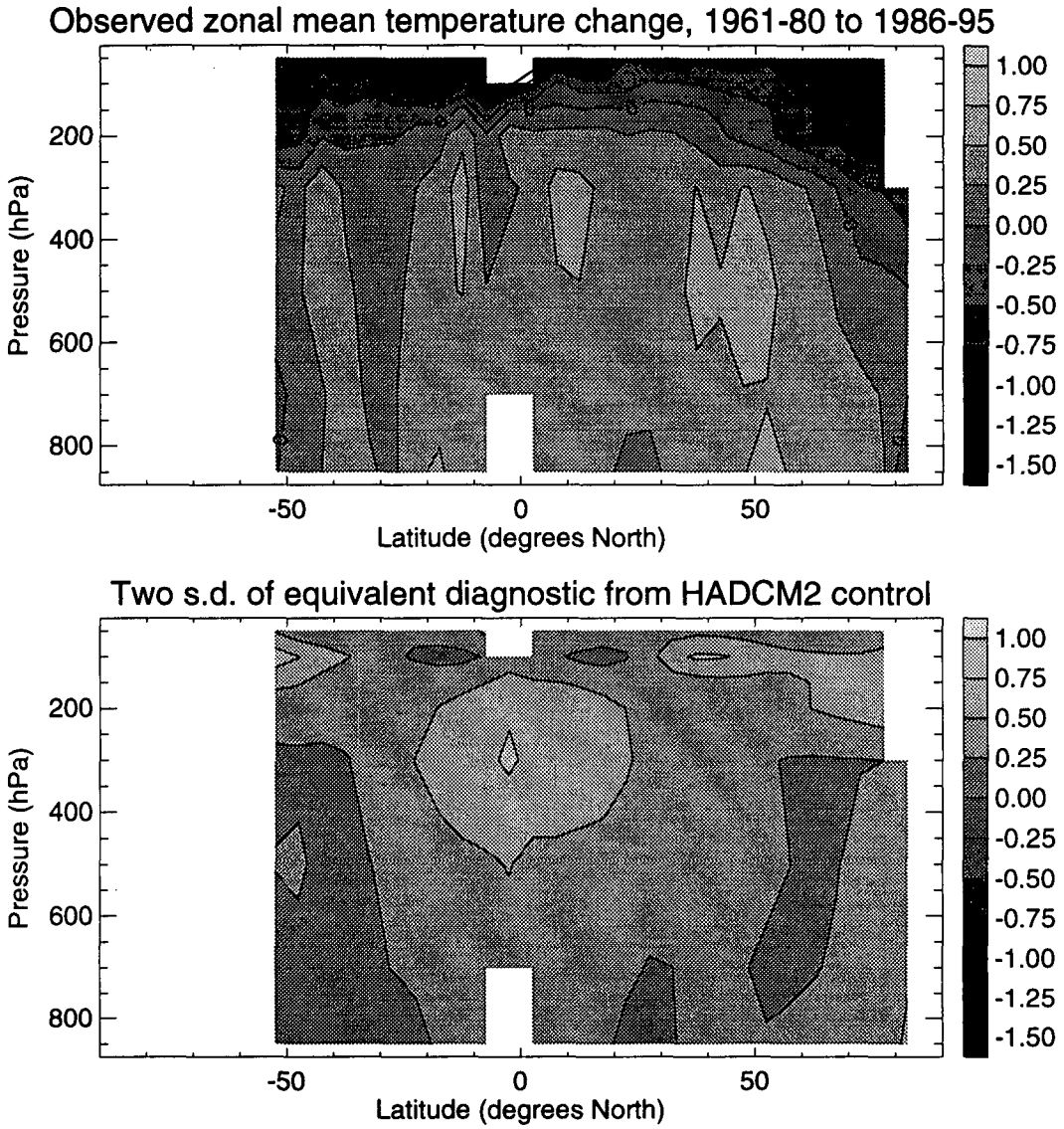


Figure 1: Upper panel: Vertical pattern of zonal mean temperature difference between the period 1961–80 and 1986–95, excluding years contaminated by volcanic eruptions. Note the overall pattern of stratospheric cooling and tropospheric warming. Lower panel: Two standard deviations of the same diagnostic estimated from 40 35-year-long segments extracted at 10-year intervals from a 426-year control integration of the HadCM2 climate model.

overlapping vectors of pseudo-observations).

Figure 3 indicates that $\mathcal{H}(\beta_G = 0)$ – the hypothesis that the amplitude of the greenhouse gas pattern is zero in the observations – can be consistently rejected at $P_{<0.05}$ (0.05 being approximately the smallest $P$-value we can quantify with a control integration of this length) except for the lowest truncations, where the detection space is clearly inadequate to resolve the signal. For truncations $\kappa \leq 13$, however, $\mathcal{H}(\beta_G = 1)$ – that the model-predicted amplitude is correct – can also be rejected at $P_{<0.05}$. The key

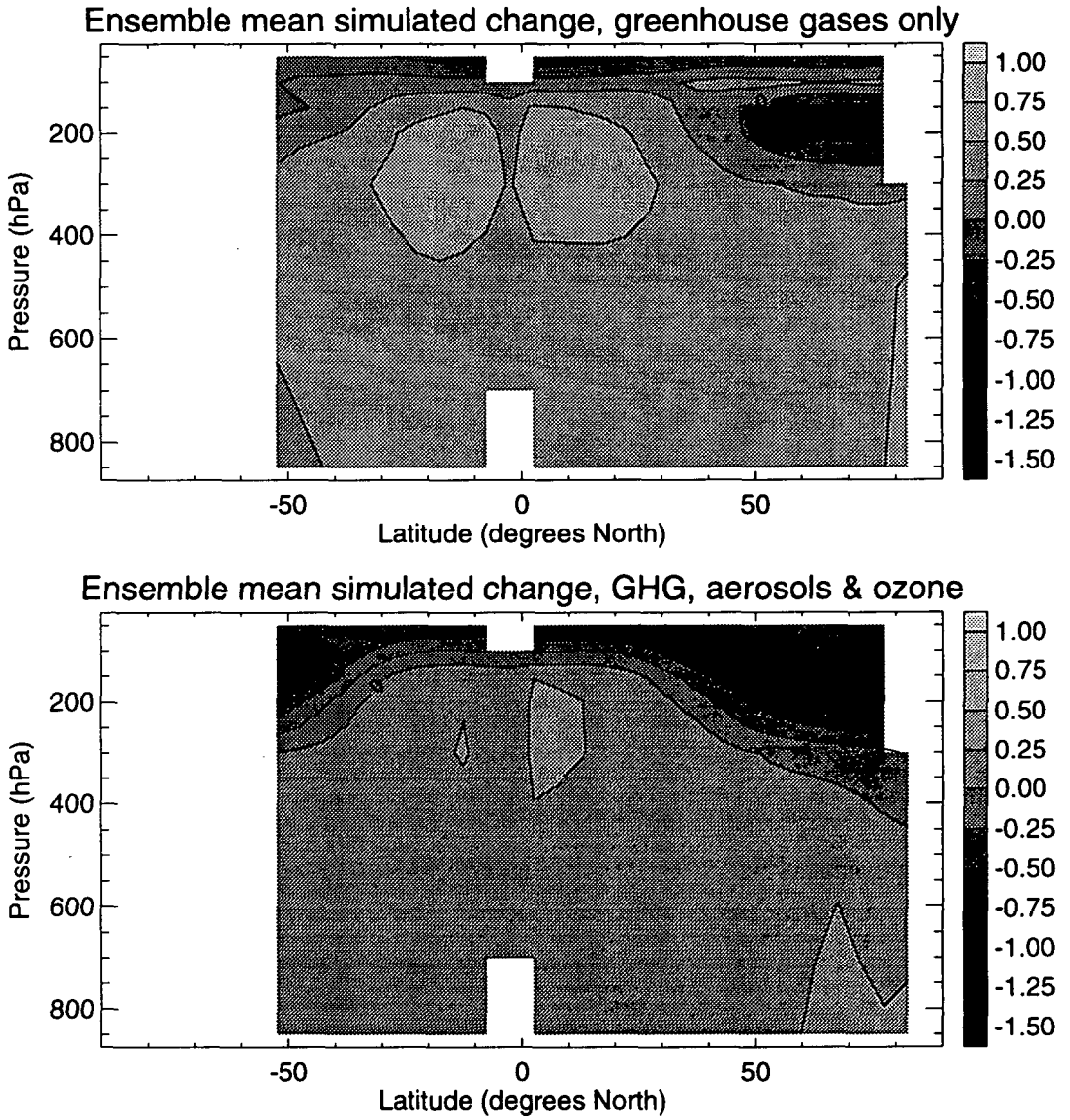


Figure 2: Model-predicted changes over the period 1961–95 based on the ensemble mean of four integrations of the HadCM2 climate model forced with the effects of changing greenhouse gases (upper panel) and including the effects of sulphate aerosols and declining stratospheric ozone (lower panel).

point to note is that error bars consistently decline as we increase the truncation level, including more EOFs of the control in the detection space. Before drawing any further conclusions, therefore, we need to establish the maximum truncation at which the model is reliable (or, to be precise, cannot be shown to be unreliable).

The singular value spectrum of the control gives us little help in choosing an appropriate truncation. Were this to consist of a small number of large singular values followed by a sharp cutoff, we would truncate after the cutoff. As is generally the case in geophysical systems (*Allen & Smith*, 1996), no such "noise floor" is evident in the singular value spectrum shown in figure 4, so we require other truncation criteria.

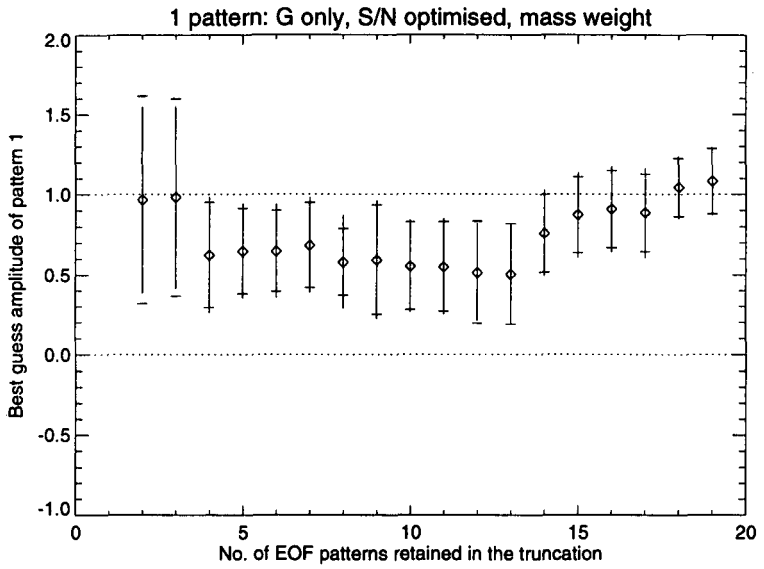


Figure 3: Estimated amplitude of G (greenhouse gas pattern) versus rank of the detection space (no. of EOFs retained of the control). Diamonds: "best guess". Vertical bars: $P_{0.05}$ confidence interval based on an assumed Gaussian distribution. Dashes: "310-year error" - $\pm\sqrt{\epsilon^2_{\max}}$ observed in a 310-year control integration. Note how error-bars decline as we include more EOFs: what is the "correct" truncation/error-bar?

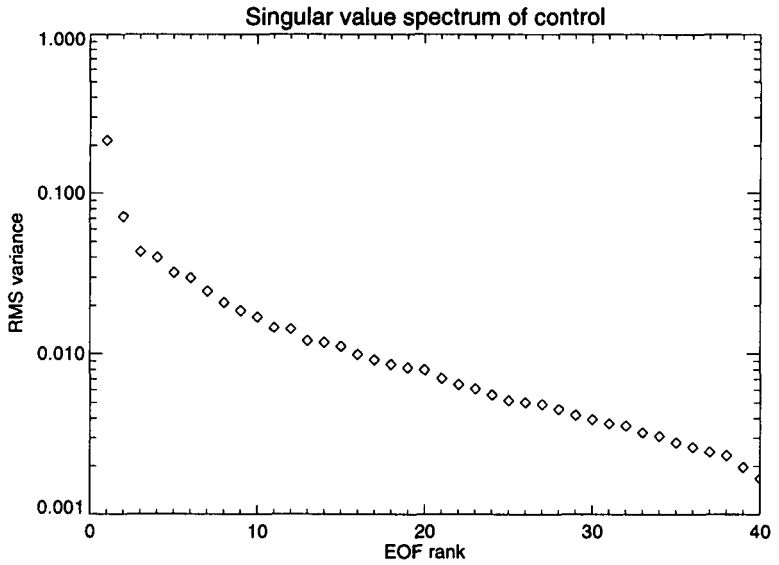


Figure 4: The spectrum of singular values of the control. There is no sharp break in the spectrum, giving no indication of an appropriate truncation.

The solid line in figure 5 shows the evolution of $P(\chi^2)$ (probability of obtaining a value of $\chi^2$ greater than or equal to that observed if the noise model is adequate) as a function of truncation. $P(\chi^2) \simeq 0.5$ for truncations lower than 12 EOFs, indicating the amplitude of model-simulated variability is approximately correct, and it diminishes

rapidly towards zero between 12 and 14 EOFs, indicating 12 is the maximum reliable truncation-level. To demonstrate what happens with an inadequate noise model, the dashed line shows the same diagnostic in a case where temperature anomalies in the control run have been divides by a factor of $\sqrt{2}$ (halving the variance). Discrepancies of this order between model and observed internal variability have often been noted in the literature (*Kim* et al., 1996). The $\chi^2$ test indicates that uncertainty estimates are unreliable for truncations as low as 7. For very small truncations, $\kappa$ =4–6, the test is simply not powerful enough to identify this model-data discrepancy.

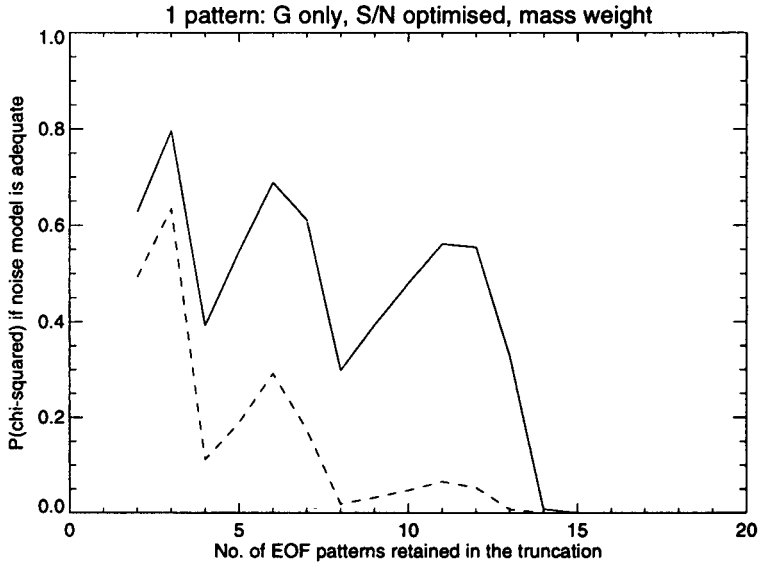


Figure 5: Solid line: evolution of $P(\chi^2)$ (probability of obtaining a value of $\chi^2 \geq$ that observed if the noise model is adequate) with truncation. $P(\chi^2) \rightarrow 0$ between 12 and 14 EOFs, indicating model is inadequate for truncations greater than these. For lower truncations, model-simulated variability appears to be approximately correct: $P(\chi^2) \simeq 0.5$. Dashed line: evolution of $P(\chi^2)$ when control variance is artificially reduced by a factor of 2.

We conclude that, over truncations at which the model can be relied upon, the G pattern significantly overestimates the response in the real world – that is, $\mathcal{H}(\beta_G = 1)$ is rejected. A univariate model based on the GSO pattern appears to do rather better, shown in figure 6. Again, 12 is the maximum allowable truncation, at which point $\mathcal{H}(\beta_{GSO} = 0)$ can be rejected, while $\mathcal{H}(\beta_{GSO} = 1)$ cannot. It would, however, be incorrect to conclude from this improvement that the combined influence of sulphates and ozone is detectable in the observations – it might be the case that the model sensitivity to greenhouse gas increase is too strong and the sulphates and ozone forcing is simply compensating for this error. To establish whether both effects are detectable, we need to investigate a bivariate detection model.

The bivariate model is that the observations consist of a linear superposition of the G and GSO patterns with an additive noise term. We apply the optimal fingerprinting algorithm (4) to estimate pattern-amplitudes and associated uncertainty ranges with G and GSO patterns providing the columns of $\mathbf{X}$. Best-guess pattern amplitudes, $\tilde{\boldsymbol{\beta}}$, and

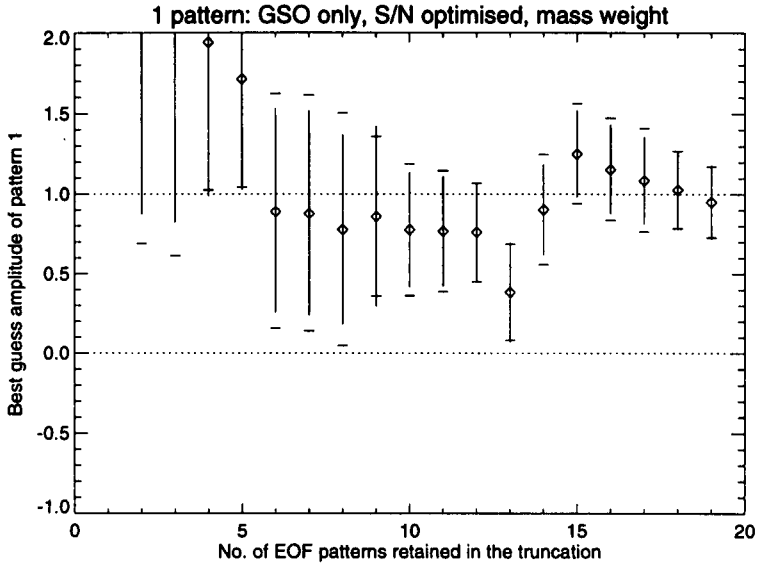


Figure 6: Estimated amplitude of GSO (greenhouse gases, aerosols and ozone pattern) versus rank of the detection space. Note how, unlike in the case of the G pattern, $\mathcal{H}(\beta_{GSO} = 1)$ cannot be rejected at 12-EOF truncation.

the associated 310-year return envelope (somewhere between the $P_{0.1}$ and $P_{0.05}$ confidence interval, depending on the unknown true degrees of freedom of the control) are shown in figure 7, with G pattern-amplitude on the horizontal axis, GSO on the vertical. Because the effects of greenhouse gases are present in both runs, patterns are highly correlated, so the ellipse is far from circular. The point [0,1], corresponding to exact agreement with the GSO prediction, lies within the confidence bound. The point [1,0], exact agreement with G, is excluded. The best-fit is obtained at the point [0.4,0.3], indicating the model overpredicts the response to greenhouse gases by $\sim 30\%$, and overpredicts the combined response to sulphates and ozone by a factor of three. This is consistent with the results of *Tett* et al., 1996, who found that a 50% reduction in the amplitude of the model-predicted response to ozone depletion improved the fit to observations. Both errors in the response and the crudeness of the parameterization used for ozone trends are likely to be responsible. The hypothesis of a zero or negative (meaning the model predicts the wrong sign) response to greenhouse gases can be excluded at the $P_{0.1-0.05}$ confidence level on the basis of these data, but if we assume no prior knowledge of the amplitude of the greenhouse gas response, the observations do not exclude the possibility a zero response to sulphates and ozone. We stress that this does not mean that the response to sulphates and ozone *is* zero, simply that the pattern of response predicted by the HadCM2 model (which could be incorrect) is not detectable using this algorithm in this particular diagnostic.

The origin of the 310-year return envelope is illustrated in figure 8, which shows the joint distribution of G and GSO pattern amplitudes, with S/N optimisation, computed from the columns of $\mathbf{Y}_{C_2}$. The ellipse, by construction, passes through the largest Mahalanobis-weighted excursion from the origin. For comparison, the dashed and dotted lines show the $P_{0.1}$ and $P_{0.05}$ confidence intervals respectively computed using equation

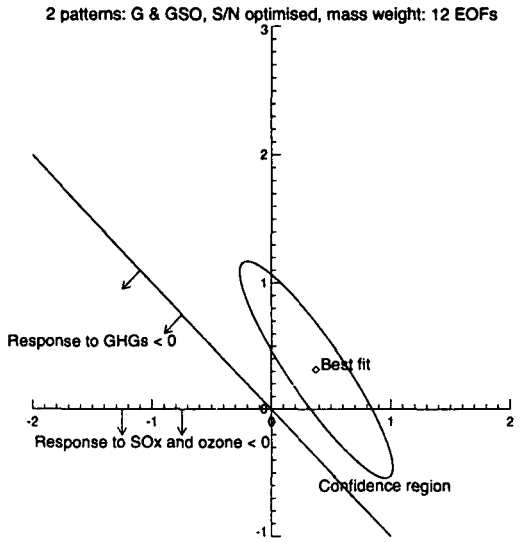


Figure 7: Best-fit amplitudes $\tilde{\beta}$ and associated uncertainty ranges for the model-predicted patterns of change due to greenhouse gases alone (horizontal axis) and the combined effects of greenhouse gases, sulphates and ozone (vertical axis). Estimates based on $\kappa = 12$ highest-ranked EOFs of the control.

(12) with $\nu \simeq 15$, the estimated degrees of freedom taking into account lag-1 autocorrelation in the control. As we would expect for this $\nu$, the 310-year return envelope lies between the two, closest to the $P_{0.1}$ contour. As discussed above, we conclude it would be unwise to attempt to quantify absolute (unsigned) $P$-values much less than 0.1 on the basis of this length of control. This is certainly intuitively plausible: the control segment used is approximately 10 times as long as the observational record, so $P_{0.1}$ is a natural lower limit on claims which can be made without extrapolation.

Figures 7 and 8 show results for $\kappa = 12$, confining the detection space to the 12 highest-ranked EOFs of the control. As argued above, we expect results to be critically dependent on the choice of $\kappa$. This is indeed the case. Figure 9 shows the corresponding result with $\kappa = 4$: in this case the truncation is too severe and the signals cannot be represented at all, resulting in large confidence intervals and complete loss of significance. Figure 10 shows the result of truncating at $\kappa = 16$: the confidence region is now much smaller, and we appear to be able to reject the hypothesis of zero response to sulphates and ozone. The evolution of $P(\chi^2)$ with truncation in the bivariate model is, however, very similar to its evolution in the univariate model, shown in figure 5, indicating that 12 is the maximum reliable truncation, so results at $\kappa = 16$ are meaningless.

Qualitatively different results emerge from the adoption of different truncations, graphically illustrating the need for objective criteria to determine the appropriate truncation level. Results from the $\chi^2$ diagnostic are very similar to the univariate case shown in figure 5. $P$-values for the $\chi^2$ statistics remain around the $50^{\text{th}}$ percentile until $\kappa = 12$–13, at which point they collapse towards zero. This is clearly the maximum truncation at which we should trust our analysis model.

The benefits of optimisation are illustrated in figure 11, which shows results from

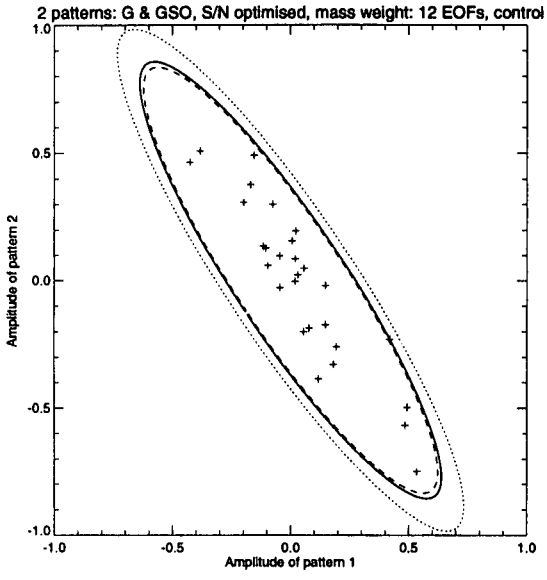


Figure 8: The joint distribution of G and GSO pattern amplitudes, with S/N optimisation, in segments of a 310-year control integration. Solid ellipse shows largest noise-weighted excursion from the origin, dotted/dashed lines show $P_{0.05}/P_{0.1}$ confidence intervals based on an assumed Gaussian distribution.

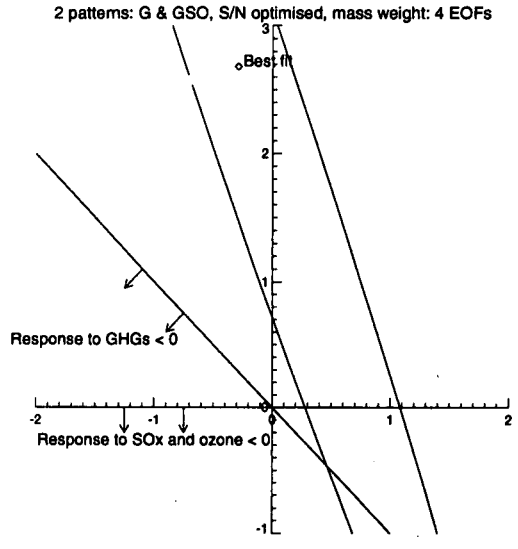


Figure 9: Best-fit $\tilde{\beta}$ and associated uncertainty ranges with very low truncation: $\kappa = 4$. The detection space is unable to represent the signal, leading to very large uncertainties.

precisely the same bivariate detection model based on a 12-EOF detection space but without weighting by the inverse noise variance (i.e. giving equal weight to errors in all 12 EOFs, corresponding to an ordinary least squares estimate). The best-guess pattern-amplitude is very similar to the optimised case, as would be expected because the ordinary least squares estimator is unbiased, but the uncertainty envelope is much larger.

Figure 12 illustrates how we can translate optimal detection results into estimates

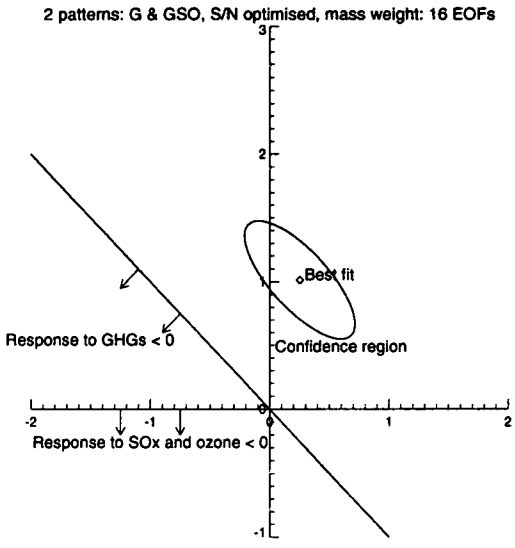


Figure 10: Best-fit $\tilde{\beta}$ and associated uncertainty ranges with excessively high truncation: $\kappa = 16$. Inclusion of high-ranked EOFs containing unrealistically low variance leads to misleadingly small estimated uncertainties.

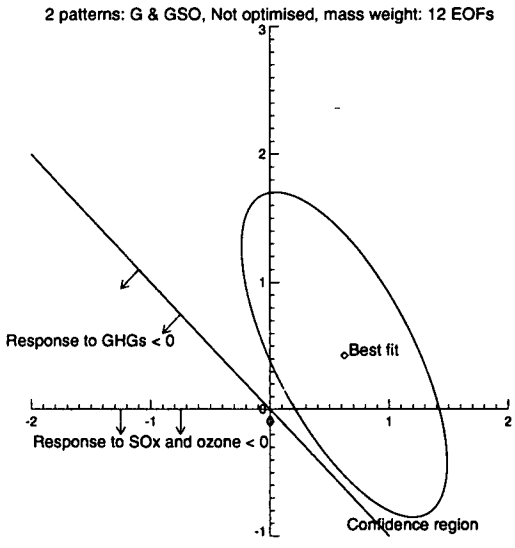


Figure 11: An example of the benefits of optimisation: best-guess pattern amplitudes in the bivariate detection model with 12 EOF truncation but without S/N optimisation.

of physically-interpretable climate parameters. The sum of the G and GSO pattern-amplitudes from the bivariate model gives an estimate of the scaling required on the total greenhouse response required to match observations, taking into account our uncertainty in the amplitude of the response to sulphates and ozone. At 12 EOF truncation, this scaling factor lies in the range 0.4–1. On these decadal timescales, HadCM2 behaves as if it has a sensitivity of ~2.8K (*Houghton* et al., 1996) implying an "observed" sensitivity

range of 1.1–2.8K, assuming the surface temperature response scales with that of the free troposphere. Considerable care must be taken in interpreting these numbers, because when this model is subjected to a much longer doubled-$CO_2$ integration, the sensitivity appears to rise to ~3.2K (Senior, *pers. comm.*) indicating the difficulty of interpreting diagnostics relating to different timescales in terms of a single summary statistic. Results quoted here, being based on an analysis of a 35-year record, clearly pertain directly only to what can be said about climate parameters and associated uncertainties on similar timescales.

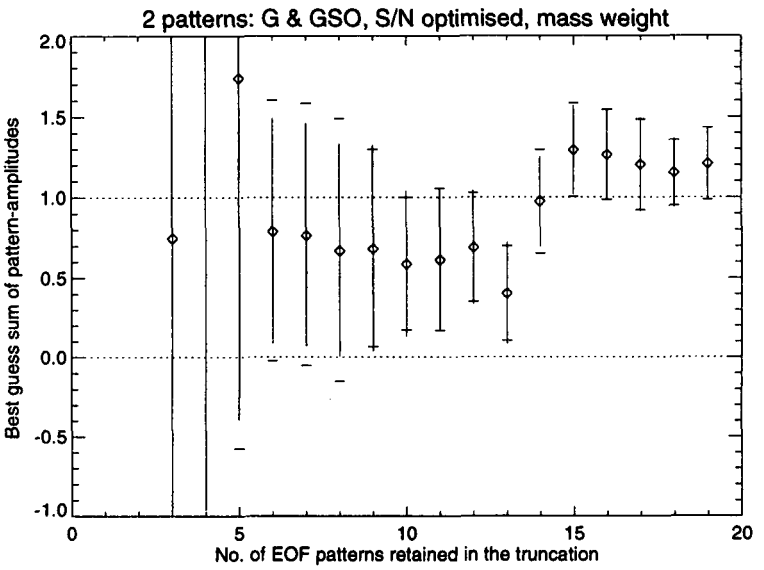


Figure 12: Translating optimal detection results into estimates of climate parameters: the sum of G and GSO pattern amplitudes gives an estimate of the scaling required on the total greenhouse response to match observations.

Given the estimate $\tilde{\boldsymbol{\beta}}$ and its associated uncertainty $\tilde{V}(\tilde{\boldsymbol{\beta}})$, and bearing in mind these caveats, we can reconstruct the best-guess trend a each latitude/pressure point and the corresponding $P_{0.05}$ confidence interval using equation (9). Maximum and minimum reconstructed trends, taking into account internal variability illustrated in figure 1, are shown in figure 13. Note that these are not themselves realisable patterns because uncertainties are correlated between locations (that is, a high positive trend in one region may be associated with a high negative trend in another and so forth). These maxima and minima provide, however, an indication of where the model-predicted trends may be consistent with the observations when subject to an appropriate scaling, and allow us to identify regions in which observations (figure 1) and model are clearly inconsistent. The $\chi^2$ test described above, being based on a global summary statistic, might well fail to identify local model-data discrepancies. For example, the observed cooling at ~50hPa in the extratropical stratosphere is considerably larger than the maximum model-predicted cooling, indicating an unambiguous model deficiency (it seems implausible that this error could be attributed to problems with the prescribed forcing). Over most of the troposphere, however, the observations lie within the range of possible model-predicted trends.

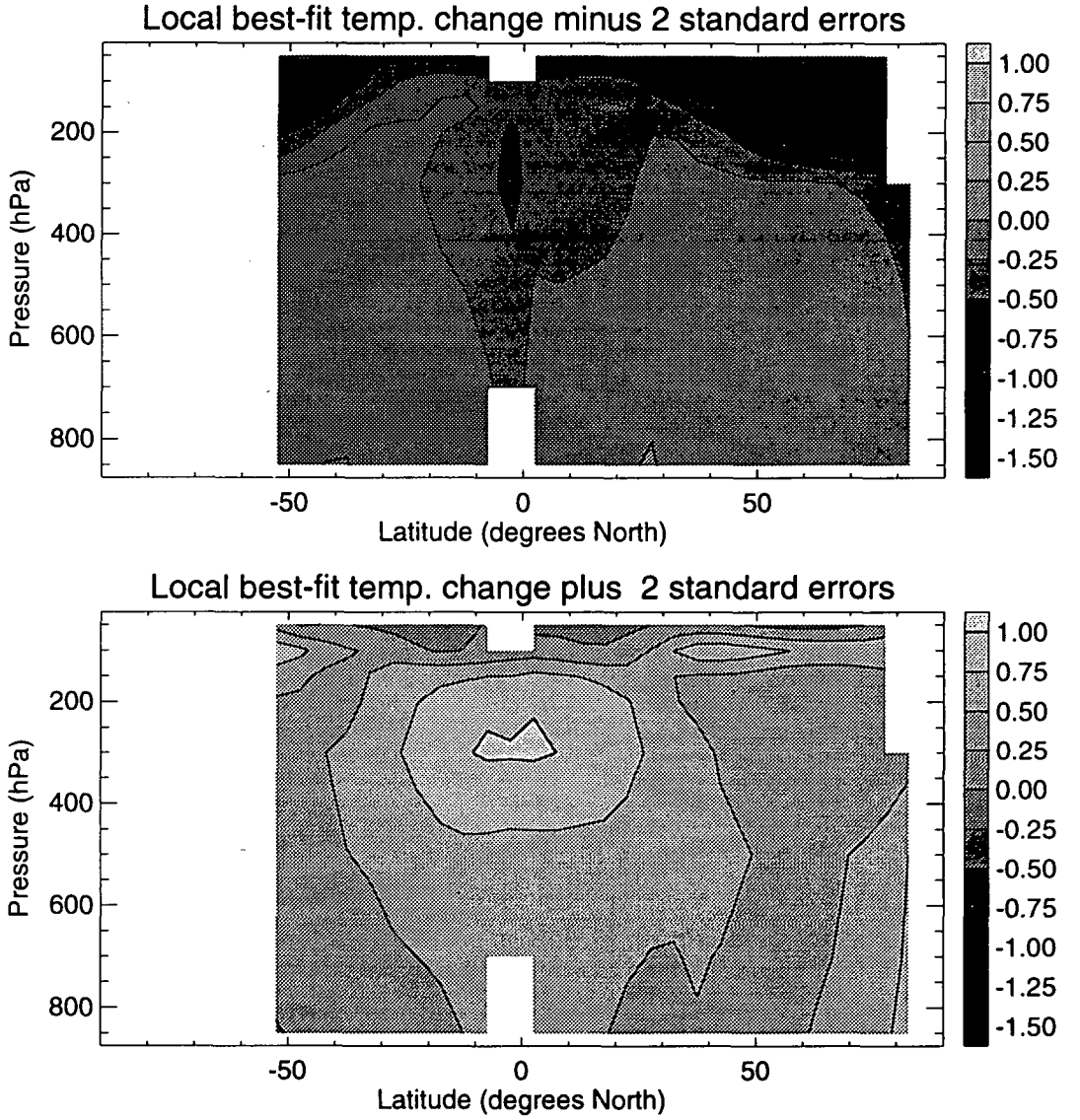


Figure 13: Maximum and minimum ($P_{0.05}$ one-tailed limits) local trends indicated by the detection model. Locations where the observations (figure 1, top panel) lie outside this range indicate systematic model deficiencies.

# 6 Summary

Formulating the optimal fingerprinting algorithm as a linear regression problem suggests some simple consistency checks for detection model adequacy whose primary purpose is to ensure that uncertainty estimates based on model-simulated variability are not completely inaccurate. We have presented a simple check (the $\chi^2$-test on residuals) which should detect gross model inadequacies and demonstrated its application to the "vertical detection problem", examining recent trends in atmospheric vertical temperature structure. We have also suggested that uncertainties should be presented in terms of return-times rather than confidence intervals based on an assumption of

multivariate normality. Conventional confidence intervals require an estimate of the degrees of freedom of the control, which is invariably uncertain, and may involve extrapolation from the body of the "climate noise" distribution into the distribution's tails. Without *a priori* reason to believe that climate noise is exactly Gaussian (and with good reason to believe it is not), such extrapolation is clearly unsafe. Finally, we illustrate how results from optimal detection may be used to obtain observationally-based estimates of climate parameters and identify systematic model deficiencies.

## Acknowledgements

The motivation for this work was primarily provided by *Hasselmann*, 1997, and *Hegerl* et al., 1997 – we would like to thank both Klaus Hasselmann and Gabriele Hegerl for fruitful discussions of their work. Thanks are also due to Art Dempster, Chris Forest, Geoff Jenkins, Gerry North, John Mitchell, Ben Santer and Peter Stott. MRA was supported by the NERC/RAL Ocean Dynamics Service Level Agreement and by a NERC Advanced Research Fellowship, SFBT by the U.K. Department of the Environment, Transport and Regions under contract number PECD 7/12/37.

## References

Allen, M. R., & Smith, L. A. Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. *J. Climate*, **9**, 3373–3404. 1996.

Barnett, T. P., Santer, B. D., Jones, P. D., Bradley, R. S., & Briffa, K. R. Estimates of low frequency natural climate variability in near-surface air temperature. *Holocene*, **6**, 255–263. 1996.

Bell, T. L. Theory of optimal weighting to detect climate change. *J. Atmos. Sci.*, **43**, 1694–1710. 1986.

Bradley, R. S., & Jones, P. D. 'Little Ice Age' summer temperatures: their nature and relevance to recent global warming trends. *The Holocene*, **3**, 367–376. 1993.

Hannoschöck, G., & Frankignoul, C. Multivariate statistical analysis of sea surface temperature anomaly experiments with the GISS general circulation model. *J. Atmos. Sci.*, **42**, 1430–1450. 1985.

Hasselmann, K. On the signal-to-noise problem in atmospheric response studies. *Pages 251–259 of:* Shawn (ed), *Meteorology of Tropical Oceans.* Royal Meteorological Society. 1979.

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

Hasselmann, K. On multifingerprint detection and attribution of anthropogenic climate change. *Climate Dynamics.* to appear. 1997.

Hegerl, G., Hasselmann, K., Cubasch, U., Mitchell, J. F. B., Roeckner, E., Voss, R., & Waszkewitz, J. On multi-fingerprint detection and attribution of greenhouse gas and aerosol forced climate change. *Climate Dynamics.* to appear. 1997.

Hegerl, G. C., & North, G. R. Statistically optimal methods for detecting anthropogenic climate change. *J. Climate*, **10**. in press. 1997.

Hegerl, G. C., von Storch, H., Hasselmann, K., Santer, B. D., Cubasch, U., & Jones,

P. D. Detecting greenhouse gas-induced climate change with an optimal fingerprint method. *J. Climate*, **9**. 1996.

Houghton, J. T., *et al.* (eds). *Climate Change 1995: The Science of Climate Change.* Cambridge Univ. Press. 1996.

Johns, T. C., Carnell, R. E., Crossley, J. F., Gregory, J. M., Mitchell, J. F. B., Senior, C. A., Tett, S. F. B., & Wood, R. A. The Second Hadley Centre coupled ocean-atmosphere GCM: model description, spin-up and validation. *Climate Dynamics*, **13**, 103–134. 1997.

Kim, K. Y., North, G. R., & Hegerl, G. C. Comparisons of the second-moment statistics of climate models. *J. Climate*, **9**, 2204–2221. 1996.

Leroy, S. Detecting climate signals, some Bayesian aspects. *J. Climate.* submitted. 1998.

Mardia, K. V., Kent, J. T., & Bibby, J. M. *Multivariate Analysis.* Academic Press. 1979.

Mitchell, J. F. B., Johns, T. C., Gregory, J. M., & Tett, S. F. B. Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature*, **376**, 501–504. 1995a.

Mitchell, J. F. B., Davis, R. A., Ingram, W. J., & Senior, C. A. On Surface-Temperature, Greenhouse Gases, And Aerosols - Models And Observations. *J. Climate*, **8**, 2364–2386. 1995b.

North, G. R., Kim, K. Y., Shen, S. S. P., & Hardin, J. W. Detection of forced climate signals, 1: Filter theory. *J. Climate*, **8**, 401–408. 1995.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. *Numerical Recipes in FORTRAN: the Art of Scientific Computing.* 2 edn. Cambridge Univ. Press. 1992.

Ripley, B. D., & Thompson, M. Regression techniques for the detection of analytical bias. *Analyst*, **112**, 377–383. 1987.

Santer, B. D., Wigley, T. M. L., & Jones, P. D. Correlation methods in fingerprint detection studies. *Climate Dynamics*, **8**, 265–276. 1993.

Santer, B. D., Mikolajewicz, U., Bröggemann, W., Cubasch, U., Hasselmann, K., Höck, H., Maier-Reimer, E., & Wigley, T. M. L. Ocean variability and its influence on the detectability of gr eenhouse warming signals. *J. Geophys. Res.*, **100**, 10693–10725. 1994a.

Santer, B. D., Brüggemann, W., Cubasch, U., Hasselmann, K., Höck, H., Maier-Reimer, E., & Mikolajewicz, U. Signal-to-noise analysis of time-dependent greenhouse warming experiments. Part 1: pattern analysis. *Climate Dynamics*, **9**, 267–285. 1994b.

Santer, B.D., Taylor, K.E., Wigley, T.M.L., Johns, T.C., Jones, P.D., Karoly, D.J., Mitchell, J.F.B., Oort, A.H., Penner, J.E., Ramaswamy, V., Schwarzkopf, M.D., Stouffer, R.J., & Tett, S. A Search for Human Influences on the Thermal Structure of the Atmosphere. *Nature*, **382**, 39–46. 1996.

Stevens, M. J., & North, G. R. Detection of the Climate Response to the Solar Cycle. *J. Atmos. Sci.* in press. 1997.

Stott, P. A., & Tett, S. F. B. Scale-dependent detection of climate change. *J. Climate.* in preparation. 1997.

Stouffer, R. J., Manabe, S., & Vinnikov, K. Y. Model assessment of the role of natural variability in recent global warming. *Nature*, **367**, 634–636. 1994.

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

**247**, 1170–1173. 1996.

Tett, S. F. B., Johns, T. C., & Mitchell, J. Global and regional variability in a coupled AOGCM. *Climate Dynamics*, **13**, 303–323. 1997.

Thacker, W. C. Climatic fingerprints, patterns and indexes. *J. Climate*, **9**, 2259–2261. 1996.

Zweirs, F. W., & von Storch, H. Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336–351. 1995.