Climate Dynamics (1999) 15:419–434

M. R. Allen · S. F. B. Tett

Checking for model consistency in optimal fingerprinting

Received: 9 December 1997 / Accepted: 24 December 1998

Abstract Current approaches to the detection and attribution of an anthropogenic influence on climate involve quantifying the level of agreement between model-predicted patterns of externally forced change and observed changes in the recent climate record. Analyses of uncertainty rely on simulated variability from a climate model. Any numerical representation of the climate is likely to display too little variance on small spatial scales, 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. 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 deficient. 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. The influence of anthropogenic greenhouse gases can be detected at a high confidence level in this diagnostic, while the combined influence of anthropogenic sulphates and stratospheric ozone depletion is less clearly evident. Assuming the time-scales of the model response are correct, and neglecting the possibility of non-linear feedbacks, the amplitude of the observed signal suggests a climate sensitivity range of 1.2–3.4 K, although the upper end of this range may be underestimated by up to 25% due to uncertainty in model-predicted response patterns.

M.R. Allen (✉)
Space Science Department, Rutherford Appleton Laboratory, Chilton, Didcot, OX11 0QX
Also at: Department of Physics, University of Oxford, UK
E-mail: m.r.allen@rl.ac.uk

S. F. B. Tett
Hadley Centre for Climate Prediction and Research, UK Meteorological Office, London Road, Bracknell, RG12 2SZ, UK

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 versus 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 and Frankignoul 1985; Bell 1986; Hasselmann 1993; Santer et al. 1994a; North et al. 1995; Hegerl et al. 1996; North and Stevens, 1998). 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 $100 \times P^{\text{th}}$ percentile of the distribution estimated from the control, then detection is claimed with a $100 \times P\%$ risk of a type-1 error (so P = probability of a false positive). Clearly, this approach to quantifying the risk of error 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 variability similar to internal variability: in this situation, we would simply increase the variance attributable to internal variability to take into account the additional variance due to these unknown external forcings.

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 time-scales. 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 incompletely known) anthropogenic signal from the historical record prior to computing a power spectrum remains, see Jones and Hegerl (1998) for a discussion of this point. Proxy and incomplete observations of the pre-industrial period (e.g. Bradley and 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 (e.g. Briffa et al. 1998). 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-wave-number components of variability which models are known to simulate poorly (Stott and Tett 1998) 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 (i.e. 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 21st 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 "response patterns", $\mathbf{x}_{k;k=1,m}$, each consisting of a rank-ℓ 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 as the columns of the $\ell \times m$ 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 the $\mathbf{x}_k$ might correspond to the local trend expected due to the k^{th} forcing scenario at a particular latitude–longitude or (in the "vertical detection" problem discussed here) latitude-height location. Alternatively, in the full "space-time" variant of the algorithm, they correspond to the expected response at a given point in both space and time. 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 and North 1997). 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 response patterns plus an additive noise term.

Response patterns may be specified *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 a particular 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 response patterns may be treated as noise free. If these patterns are derived (1) from simulations of the 21st century, as in Hegerl et al. (1996) or (2) from energy balance models, as in North and Stevens (1998) then this assumption is well justified, since in both cases the sampling uncertainty in the response patterns is essentially zero (repeating the experiment would yield an identical pattern). Both of these approaches have disadvantages, however: (1) assumes that response-patterns do not change over time, and is inapplicable to responses to natural forcing such as solar and volcanic activity, while (2) requires that the full response of the non-linear climate system is correctly represented by an energy balance model.

Following Mitchell et al. (1995a) and Tett et al. (1996) we prefer to compare like with like, basing our response-patterns on the mean of an ensemble of simulations of a fully non-linear GCM spanning the same period which is covered by the observations. The disadvantage of this approach is that the response patterns themselves, as well as the observations, are subject to sampling uncertainty: a second ensemble would yield a somewhat different ensemble mean response. If the model simulation of internal variability is correct, the variance in the response patterns from an M-member ensemble is approximately $1/M$ times the variance in the observations (exactly so if all distributions are Gaussian). With only a four-member ensemble in the example in Sect. 5, this introduces a bias towards zero in estimated pattern-amplitudes. Resolving this bias requires the introduction of non-linear estimators, which we will consider elsewhere. We do, however, make a first-order correction for noise in the response patterns in our analysis of uncertainty, as detailed later.

Once the response-patterns have been specified, the detection problem simply involves estimating the amplitude of these patterns in a rank-ℓ 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 $\ell \times \ell$ matrix $\mathbf{C}_N$:

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

$\mathscr{E}$ being the expectation operator. Under the assumption that $\mathbf{u}$ is multivariate normal (which we will return to), 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

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

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

The notion of a "low-variance estimator" may require some clarification for non-specialists: strictly interpreted, it means that *if* we were able to repeat the whole experiment (not just the model simulations, but the actual forcing of the real climate system over the 20th century) a large number of times, then the BLUE estimator for $\boldsymbol{\beta}$ would vary less between experimental realisations than any linear unbiased alternative. Given that we only have a single realisation of the real climate, this interpretation may seem rather artificial: Hasselmann (1993) and North et al. (1995) prefer to discuss these estimators in terms of signal-to-noise, while Hasselmann (1998) and Leroy (1998) recommend an explicit Bayesian treatment. We stress that, at the level of complexity we are dealing with here, the differences are a matter of interpretation, and that all these approaches should give essentially the same result: different interpretations may, however, suggest different avenues for further refinement of the technique. The Bayesian treatment, for example, is particularly well suited to the incorporation of prior information into the analysis, while couching everything in terms of classical linear regression suggests how other standard regression tools can be brought to bear on the problem.

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 (Mardia et al. 1979) to prove that the following estimator for $\boldsymbol{\beta}$ is BLUE:

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

introducing $\mathbf{F}^T \equiv (\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{C}_N^{-1}$ as the operator which extracts $\tilde{\boldsymbol{\beta}}$ from $\mathbf{y}$. Notice that this is simply the ordinary least squares solution applied to the transformed variables, $\mathbf{PX}$ and $\mathbf{Py}$. 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), in which case:

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

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

For reference, the v^{th} row of $\mathbf{X}^T\mathbf{C}_N^{-1}$ in Eq. (4) corresponds to the v^{th} fingerprint f_v^i in Eq. (29) of Hasselmann (1997) while the matrix $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ corresponds to the metric $D_{v\mu}$ in his Eq. (30) and $\tilde{\boldsymbol{\beta}}$ corresponds to the detection coefficients, d^v in his Eq. (32). The disadvantage of Hasselmann's (1997) definition of a fingerprint is that, in a multi-pattern problem, pattern-amplitudes are not estimated by operating directly on the observations with $\mathbf{X}^T\mathbf{C}_N^{-1}$ unless $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ is diagonal, which it generally is not. This led Hegerl et al. (1997) to introduce an "orthogonalised fingerprint" for display purposes, to make it clear which aspects of the observations play a dominant role in estimating the different elements of $\tilde{\boldsymbol{\beta}}$. Hegerl et al.'s (1997) idea, in a two-pattern problem, is to modify ("rotate") one row of $\mathbf{X}^T\mathbf{C}_N^{-1}$ to give $\mathbf{X}'^T\mathbf{C}_N^{-1}$ such that the rows of $\mathbf{X}'\mathbf{C}_N^{-1}$ are orthogonal under the metric defined by $\mathbf{C}_N$, meaning that $(\mathbf{C}_N^{-1}\mathbf{X}')^T\mathbf{C}_N(\mathbf{C}_N^{-1}\mathbf{X}') = \mathbf{X}'^T\mathbf{C}_N^{-1}\mathbf{X}'$ is diagonal.

Although it is only used for display, we have found this orthogonalisation procedure to be potentially confusing since it implies that different patterns are treated differently in the estimation of $\boldsymbol{\beta}$, which they are not: if $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ is non-diagonal, then none of the rows of $\mathbf{F}^T$ will be aligned exactly with Hasselmann's (1997) fingerprints. Some way of referring to $\mathbf{F}^T$ is required, since this is the actual operator used to extract $\tilde{\boldsymbol{\beta}}$ from the observations. Citing a much older literature, we could simply refer to it as the BLUE estimator for $\boldsymbol{\beta}$, but for consistency with climate change detection terminology,

we suggest that the m rows of $\mathbf{F}^T$-should be referred to as the "distinguishing fingerprints" in the multi-pattern detection problem. In physical terms, Hasselmann's (1997) fingerprints discriminate against noise with covariance $\mathbf{C}_N$, while the k^{th} distinguishing fingerprint distinguishes the response to forcing scenario k from alternative responses $\mathbf{x}_i$, $\mathbf{x}_j$, … etc. in the presence of this noise. We stress that this is an issue of terminology and presentation, not a substantive difference in approach.

The estimate $\tilde{\boldsymbol{\beta}}$ is unbiased, so $\mathscr{E}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and its $m \times m$ covariance, $\mathscr{E}[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\,(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T]$, is given by

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

Provided $\mathbf{u}$ is multivariate normal, this can be translated into a confidence ellipsoid as follows. Equation (6) implies that

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

meaning that the left-hand side (LHS) of Eq. (7) is distributed as the sum of squares of m normally-distributed unit-variance random numbers, or χ^2_m. 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 χ^2 for which $P(\chi^2 > \chi^2_{\text{crit}}) = P$ and plot the values of $\boldsymbol{\beta}$ for which the LHS of Eq. (7) is equal to this value. Again, in the single-pattern, uncorrelated noise case, Eq. (6) becomes

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

It 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 columns from $\mathbf{X}^T\mathbf{C}_N^{-1}\mathbf{X}$ and evaluate Eq. (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, ϕ. With trivial exceptions, ϕ 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 $1/\ell$, for example, then ϕ is simply the global mean. If $\mathbf{w}$ is a unit vector, then ϕ is the value of the observation-vector at a particular location and so on. Neglecting uncertainty in $\mathbf{X}$ as before, the variance of ϕ attributable to the uncertainty in $\tilde{\boldsymbol{\beta}}$ is:

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

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 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. (Of course, if we use detection results to identify *and correct* model errors, there is a danger of circularity: a positive detection result with the corrected model no longer carries the same weight as one obtained with an independently-specified model. There are techniques for establishing whether an apparent improvement in model-data agreement simply due to overfitting, but to discuss these issues in detail is beyond our scope here. Validation against independent observations, preferably of different variables, would also help: exclusive reliance on temperature data is clearly unsustainable.)

The fact that we are using a linear model in Eq. (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-aeorsol 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. If it is detected, then a full non-linear treatment would be necessary to analyse it explicity: we simply note that the linear model can, in principle, be applied to the initial step of testing the null-hypothesis of complete linearity.

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 amplitudes of response-patterns in model and observations: correlations convey no amplitude information. If the response 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, $\mathscr{E}(\tilde{\boldsymbol{\beta}})$, will be approximately unity.

As noted, $\mathscr{E}(\tilde{\boldsymbol{\beta}})$ is only approximately unity because the assumption that $\mathbf{X}$ is noise free is only 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. This factor is obtained by assuming that the shapes of the columns of $\mathbf{X}$ are approximately constant, while their amplitude varies due to the small size of the ensemble. Suppose $\tilde{\boldsymbol{\beta}}_{obs}$ is the factor by which we would have to scale the "true" (infinite ensemble) model response pattern to reproduce the observations, and $\tilde{\boldsymbol{\beta}}_{ens}$ is the corresponding pattern amplitude in this particular M-member ensemble-mean. The value of $\tilde{\beta}$ we obtain from regressing the observations onto this ensemble mean is the ratio

$$\tilde{\beta} = \frac{\tilde{\beta}_{obs}}{\tilde{\beta}_{ens}} \sim \frac{N[\mathscr{E}(\tilde{\beta}_{obs}),\, V(\tilde{\beta})]}{N[\mathscr{E}(\tilde{\beta}_{ens}),\, \frac{1}{M}\sqrt{(\tilde{\beta})}]}. \tag{10}$$

Provided $O(\tilde{V}(\tilde{\beta})/M) \ll 1$, then $\mathscr{E}(\tilde{\beta}_{ens}) \simeq 1$ and this ratio may be approximated by

$$\tilde{\beta} \sim N[\mathscr{E}(\tilde{\beta}_{obs}),\, (1 + 1/M)\,\tilde{V}(\tilde{\beta})]. \tag{11}$$

We therefore simply inflate $\tilde{V}(\tilde{\beta})$ by this factor, but the bias in $\tilde{\beta}$ 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:

$$\hat{\mathbf{C}}_N = \frac{1}{n}\mathbf{Y}_N\mathbf{Y}_N^T \tag{12}$$

where the columns of $\mathbf{Y}_N$ represent a succession of n vectors of "pseudo-observations", $\mathbf{y}_N$, 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 30–50 y period, and control integrations are necessarily limited to 1000–2000 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 ℓ, the number of elements in $\mathbf{y}$. The estimated covariance matrix, $\hat{\mathbf{C}}_N$, 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 Eq. (3) is satisfied, and the unit matrix on the right-hand side (RHS) of Eq. (3) need not be $\ell \times \ell$. If we assume that $\hat{\mathbf{C}}_N$ provides a reliable estimate of the noise covariance only in the subspace spanned by the κ highest-variance "EOFs of the control" (eigenvectors of $\hat{\mathbf{C}}_N$), then a natural transformation to use is $\mathbf{P}^{(\kappa)}$ where the rows of $\mathbf{P}^{(\kappa)}$ are the κ highest-variance EOFs of the control weighted by their inverse singular values (square root of the corresponding eigenvalue of $\hat{\mathbf{C}}_N$). $\mathbf{P}^{(\kappa)}\hat{\mathbf{C}}_N\mathbf{P}^{(\kappa)T}$ is equal to the $\kappa \times \kappa$ unit matrix by construction.

This is equivalent to using the Moore-Penrose pseudo-inverse, $\mathbf{P}^{(\kappa)T}\mathbf{P}^{(\kappa)}$ 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, although 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 κ: in general, the estimated uncertainty envelope around $\tilde{\boldsymbol{\beta}}$ shrinks close to monotonically with increasing κ, so (in a detection problem) the confidence level at which the null-hypothesis of zero climate sensitivity can be rejected increases predictably with κ even when this null-hypothesis is valid. The reason is that increasing κ 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 $\mathbf{P}^{(\kappa)}\mathbf{C}_N\mathbf{P}^{(\kappa)T} = \mathbf{I}$ by construction, $\hat{\mathbf{C}}_N \neq \mathbf{C}_N$ because of the finite length of the control, so Eq. (3) is only approximately satisfied. Worse, because the EOFs of the control have been chosen to maximise variance in a particular segment, $\mathbf{Y}_{N_1}$, the transformation $\mathbf{P}^{(\kappa)T}$ is biased with respect to that segment. Applied to another, arbitrarily selected, segment of the control with estimated covariance matrix $\hat{\mathbf{C}}_{N_2} = (1/n_2)\mathbf{Y}_{N_2}\mathbf{Y}_{N_2}^T$, the diagonal elements of $\mathbf{P}^{(\kappa)}\mathbf{C}_{N_2}\mathbf{P}^{(\kappa)T}$ will, on average, tend to be less than unity (see North et al. 1982; von Storch and Hannoschöck 1986). 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.

To take this into account, we replace Eq. (6) with the estimate

$$\begin{aligned}\tilde{V}(\tilde{\beta}) &= \hat{V}(\mathbf{F}_1^T\mathbf{y}_{N_2}) \\ &= \frac{1}{n_2}\mathbf{F}_1^T\mathbf{Y}_{N_2}\mathbf{Y}_{N_2}^T\mathbf{F}_1 \\ &= \mathbf{F}_1^T\hat{\mathbf{C}}_{N_2}\mathbf{F}_1 \end{aligned} \tag{13}$$

where $\mathbf{F}_1 = (\mathbf{X}^T\hat{\mathbf{C}}_{N_1}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{C}}_{N_1}^{-1}$. Because $\mathscr{E}(\mathbf{F}_1^T\mathbf{y}_{N_2}) = \mathbf{0}$, the RHS of Eq. (13) is simply the standard estimate of the variance of $\tilde{\boldsymbol{\beta}}$ obtained by summing squares over n_2 realisations of $\mathbf{F}_1^T\mathbf{y}_{N_2}$. A scatter plot of these individual estimates provides a simple way of visualising the distribution. Note that Eq. (13) collapses to Eq. (6) in the limit of a long control integration, as $\hat{\mathbf{C}}_{N_2} \to \hat{\mathbf{C}}_{N_1} \to \mathbf{C}_N$.

Suppose the estimate $\hat{\mathbf{C}}_{N_2}$ has v degrees of freedom (if all the $\mathbf{y}_{N_2}$ are independent, then v would equal n_2 this is virtually never the case in practice) : Eq. (7) for the errors $\tilde{\boldsymbol{\beta}}$ is then replaced by

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[\tilde{V}(\tilde{\boldsymbol{\beta}})]^{-1}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \equiv \varepsilon^2(\boldsymbol{\beta}) \sim mF_{m,v}, \tag{14}$$

the standard F distribution with m and v degrees of freedom in the numerator and denominator, respectively taking into account sampling uncertainty in $\tilde{\boldsymbol{\beta}}$ and $\tilde{V}(\tilde{\boldsymbol{\beta}})$. A confidence ellipsoid around our "best-guess" value, $\tilde{\boldsymbol{\beta}}$, can be found by plotting the locus of points $\boldsymbol{\beta}$ for which $\varepsilon^2(\boldsymbol{\beta})$ is equal to the corresponding critical value of the $F_{m,v}$ distribution. (When only a single response pattern is under consideration, $m = 1$, the Student's t-distribution may be used instead, but this is trivially related to the $F_{1,v}$-distribution so we will not discuss it here for the sake of brevity.)

The RHS of Eq. (14) only converges to $\chi_m^2 = mF_{m,\infty}$ (corresponding to an infinitely long control, in which case Eqs. (14) and (7) become equivalent) for $v > 100$. 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 v, the "true" number of degrees of freedom of a relatively short control integration – see, for example, Zwiers and von Storch (1995) and references therein. This is important because an over-estimate of v, due to the neglect of serial correlation in $\mathbf{Y}_{N_2}$, can lead to spuriously high estimates of significance. Zwiers and von Storch (1995) propose a correction for v 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 Zwiers and 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 v, and results can depend disconcertingly 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 v, but can see no rigorous justification for this choice.

An alternative approach, which makes the role of the estimate of the degrees-of-freedom more transparent, is to focus the reporting of results onto return-times rather than confidence intervals. Assuming we have 300 years of control available for hypothesis-testing, we evaluate $\varepsilon^2(\boldsymbol{\beta})$ over all vectors of pseudo-observations in $\mathbf{Y}_{N_2}$ (recalling that $\mathscr{E}(\tilde{\boldsymbol{\beta}}) = \mathbf{0}$ in the control), and simply take the maximum value,

$$\varepsilon_{\max}^2 = \max(\mathbf{y}_{N_2}^T\mathbf{F}_1[\tilde{V}(\tilde{\beta})]^{-1}\mathbf{F}_1^T\mathbf{y}_{N_2}) \tag{15}$$

as indicating a "300-y error" (that is, an estimate of the maximum error in $\tilde{\boldsymbol{\beta}}$ we expect to observe in a 300-y segment of the control, where the "size" of the error is defined in terms of the estimated error covariance, $[\tilde{V}(\tilde{\beta})]^{-1}$). We then plot $\boldsymbol{\beta}$ for which the LHS of Eq. (14) equals $\varepsilon_{\max}^2$.

In the single-pattern case ($\mathbf{F}_1$ of rank one), equating $\varepsilon_{\max}^2$ with the LHS of Eq. (14) is equivalent to stating that the uncertainty range in $\tilde{\boldsymbol{\beta}}$ is given by $\pm\max(|\mathbf{F}_1^T\mathbf{y}_{N_2}|)$, or plus/minus the largest value of $\tilde{\boldsymbol{\beta}}$ estimated from the pseudo-observation vectors in $\mathbf{Y}_{N_2}$. The role of the estimated error covariance, $\tilde{V}(\tilde{\boldsymbol{\beta}})$ in Eq. (15) is simply to provide the shape of the confidence interval in the multi-pattern case.

If an estimate of degrees-of-freedom v is available, then $\varepsilon_{\max}^2$ represents a median estimator of the $100 \times (1 - P)^{\text{th}}$ percentile of the underlying distribution, where $P \approx \ln(2)/v$; that is, $\varepsilon_{\max}^2$ will exceed this critical value of the underlying distribution in approximately 50% of cases, even for non-Gaussian control distributions. If we want to avoid mentioning degrees of freedom at all, we can simply state that there is a 50% probability of the maximum error in $\tilde{\boldsymbol{\beta}}$ exceeding $\varepsilon_{\max}^2$ in any randomly-selected 300-y segment of the control or in any time-series with equivalent variability (trivially true, since there is no reason for $\varepsilon_{\max}^2$ form $\mathbf{Y}_{N_2}$ to be larger or smaller than that obtained from another realisation of equivalent variability).

More generally, the $k + 1^{\text{th}}$ largest value ($k = 0$ corresponding to the largest) of a diagnostic obtained from a control integration with v degrees-of-freedom is a median estimator of the $100 \times (1 - P)^{\text{th}}$

percentile of the underlying distribution if

$$\sum_{j=k+1}^{v} \binom{v}{j} (1-P)^j P^{v-j} = 0.5, \tag{16}$$

the 50th percentile of the cumulative binomial probability of $\geq (k+1)$ occurrences of an event of probability $(1-P)$ in a series of v trials.

Reporting return-times explicitly, accompanied by approximate P-values based on the estimated degrees-of-freedom v, 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 y";
2. They rely much less on distributional assumptions: we require only that the distribution of $\tilde{\beta}$ is radially symmetric under the norm defined by $\tilde{V}(\tilde{\beta})$, not that this distribution is Gaussian;
3. The role of the estimated degrees-of-freedom of $\mathbf{Y}_{N_2}$ is more transparent (we plot what the control actually did, and state what approximate P-value this corresponds to, rather than plotting P-values directly and relying on the validity of poorly understood distributional assumptions);
4. And most importantly, they explicity discourage claims of significance which involve extrapolation beyond the region explored by the control.

For example, if $v = 15$, so $\ln(2)/v \simeq 0.05$, this is the smallest P-value which can be quantified legitimately. If the observations lie well outside the region defined by $\varepsilon^2_{\max}$, 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.

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.

Our null-hypothesis, $\mathscr{H}_0$, 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 κ 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, $\mathscr{H}_0$ also encompasses the statement that observational error is negligible in the truncated state-space (on the spatio-temporal scales) used for detection. A test of $\mathscr{H}_0$, therefore, is also a test of the validity of this assumption.

If we are unable to reject $\mathscr{H}_0$, 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 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 $\mathscr{H}_0$ is true then the residuals of regression (see Eq. (1),

$$\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}, \tag{17}$$

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

$$r^2 = \tilde{\mathbf{u}}^T \hat{\mathbf{C}}_N^{-1} \tilde{\mathbf{u}} \sim \chi^2_{\kappa - m}, \tag{18}$$

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

The principle of the test may be clarified if we again consider the case of single-pattern with uncorrelated noise, in which case Eq. (17) and (18) become

$$\sum_{i=1}^{\kappa} \frac{(y_i - \tilde{\beta} x_i)^2}{\lambda_i^2} \sim \chi^2_{\kappa-1}. \tag{19}$$

Terms in which the control variance is unrealistically low correspond to small values of λ_i^2 which inflate the LHS of Eq. (19) into a high percentile of the χ^2 distribution.

In geometric terms, the χ^2 test involves summing residuals over all directions in the state-space defined by EOFs 1 to κ of the control which are orthogonal to the hyperplane defined by the response patterns, $\mathbf{X}$, where orthogonality is defined in terms of the metric given by $\hat{\mathbf{C}}_N^{-1}$ (i.e. $\mathbf{a}$ and $\mathbf{b}$ are orthogonal if $\mathbf{a}^T \hat{\mathbf{C}}_N^{-1} \mathbf{b} = 0$). If, by increasing κ, we introduce an EOF 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 r^2 faster than we would expect it to rise if the control variability is adequate, giving us some warning that uncertainty estimates are becoming unreliable. (A residual check based on the χ^2 statistics has been proposed independently by Leroy 1998, we are grateful to G. Hegerl for drawing our attention to this work.)

The basis of the χ^2 test is to estimate and remove all externally-forced signals and then to examine the residuals for consistency with the climate noise model based on the control. If, therefore there is 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 χ^2 test will fail to identify any inconsistency. It should be 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-years segments of the observational record as in Hegerl et al. (1996) then the χ^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-y trends in a 35-y dataset, so we only have a single $\mathbf{y}$ to work with.

If independent control runs are used for optimization and testing then, strictly speaking, an F-test should be used in place of the χ^2-test to take into account the effects of uncertainty in the projection of $\hat{\mathbf{C}}_{N_2}$ onto the EOFs of $\hat{\mathbf{C}}_{N_1}$:

$$\tilde{\mathbf{u}}^T \hat{\mathbf{C}}_{N_2}^{-1} \tilde{\mathbf{u}} \sim (\kappa - m) F_{\kappa - m, v} \tag{20}$$

In practice, we find it makes very little difference which test is applied, provided they are used to place an upper limit on the truncation level. Moreover, the F-test requires an estimate of v, the degrees of freedom of the estimate of $\hat{\mathbf{C}}_{N_2}$, problems with which have been noted. The χ^2-test corresponds to $v \to \infty$, and so will always be more likely than the F-test to indicate that the null-hypothesis of residual consistency should be rejected. Since, in this application, we are using the test to guard against including EOFs in which control variance is unrealistically low, this represents the more cautious option.

A limitation of the simple χ^2 check, which would be shared by any univariate summary statistic, is that if the model displays too much variance on large spatial scales, this may mask the introduction of EOFs in which model variance is too low (since the mean over all scales may still look reasonable). A solution to this problem would be to examine a running χ^2-statistic, based on residuals $\kappa - \eta$ to κ: more importantly, the general evolution of the statistic must be examined, rather than reliance on an automated check.

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 response patterns (columns of $\mathbf{X}$) and rotated fingerprints (rows of $(\mathbf{X}^T \mathbf{C}_N^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C}_N^{-1}$). As soon as this correlation begins to drop rapidly with truncation, 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 response pattern and fingerprint is unaffected. Hegerl 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 variability 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 response 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 κ 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 relatively *ad hoc* criterion, we compare r^2 from Eq. (18) with the standard χ^2 distribution to establish the maximum value of κ 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 κ smaller than this limits. 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 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: this is the example considered by Tett et al. (1996). The observation vector, **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 50 hPa). Annual averages were computed to 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 Chichón 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 Fig. 1 shows the resulting pattern of vertically resolved temperature changes.

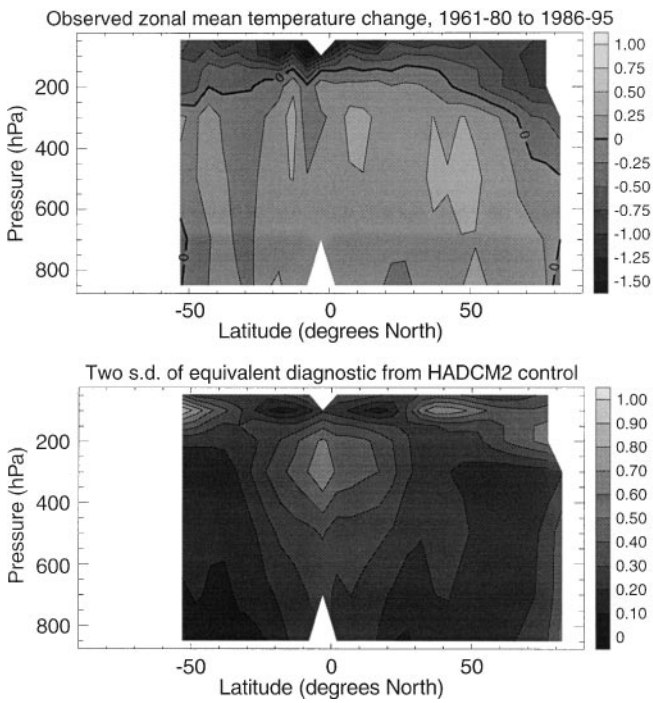

Fig. 1 *Upper panel*, vertical pattern of zonal mean temperature difference between the period 1961–80 and 1986–95, excluding years contaminated by volcanic eruptions. Note the overall pattern of stratospheric cooling and tropospheric warming. *Lower panel*, two standard deviations of the same diagnostic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

We also extract precisely the same diagnostic (applying the same missing data mask to the zonal mean temperatures, giving rise to the discrepancies between the figures displayed here and those in Tett et al. 1996, see, Appendix) 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 Fig. 1 shows two standard deviations of our chosen diagnostic estimated from 40 35-y long segments extracted at 10-y intervals from a 426-year control integration and masked using the pattern of missing data in the observations: the columns of $\mathbf{Y}_{N_1}$ (a separate 310-y segment is used to provide $\mathbf{Y}_{N_2}$ 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 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 model. In the first ensemble of four integrations (intialised 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 cholrofluorocarbons (expressed as equivalent-CO_2) for the period 1860 to 1996. The upper panel of Fig. 2 shows the ensemble mean of an identical diagnostic to that shown in the upper panel of Fig. 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 parametrisation of the effects of sulphate aerosols (Mitchell et al. 1995b) and an estimate for 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 $\mathbf{Y}_{N_1}$. Using mass weighting means that high-ranked EOFs have

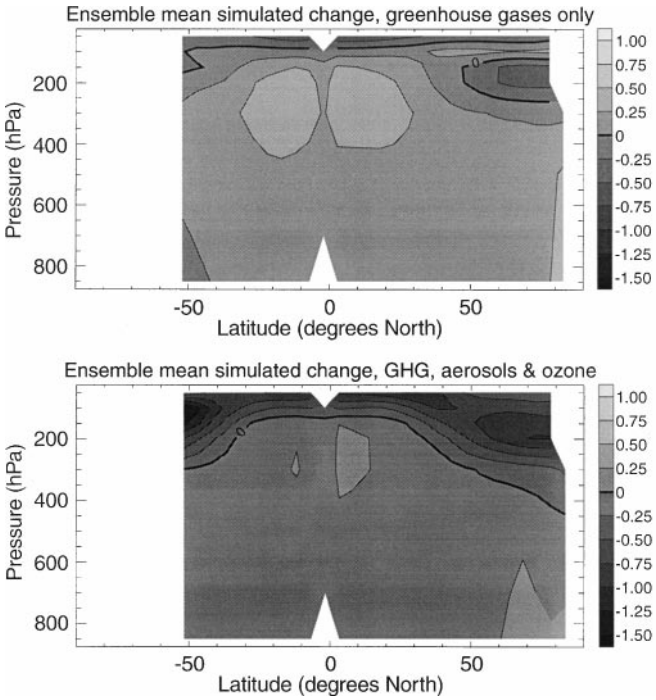

Fig. 2 Model-predicted changes over the period 1961–95 based on the ensemble mean of four integrations of the HadCM2 climate model forced with the effects of changing greenhouse gases (*upper panel*) and including the effects of sulphate aerosols and declining stratospheric ozone (*lower panel*)

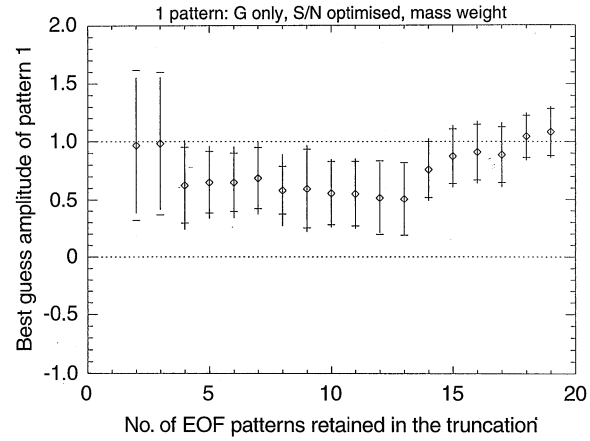

Fig. 3 Estimated amplitude of G (greenhouse gas pattern) versus rank of the detection space (number of EOFs retained of the control). *Diamonds*, "best guess"; *vertical bars*, $P_{0.05}$ confidence interval based on an assumed Gaussian distribution; *Dashes*, "310-y error": $\pm\sqrt{\varepsilon^2_{\max}}$ observed in a 310-y control integration. Note how error-bars decline as we include more EOFs: what is the "correct" truncation/error-bar?

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 stropospheric variability (Gillett et al. 1998), 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 a scaled version of G (greenhouse gas pattern) with additive climate noise. The diamonds in Fig. 3 show β_G, the estimated amplitude of the G pattern, as a function of the rank of the detection space (κ = number of EOFs retained of the control). Vertical bars show the $P_{0.05}$ (two-tailed) confidence interval based on an assumed Gaussian distribution with $\nu = 12$ degrees of freedom in $\mathbf{Y}_{N_2}$. Estimates of ν taking into account lag-1 autocorrelation in $\mathbf{Y}_{N_2}$ range from 11 to 28, depending on the precise diagnostic considered, but the higher values are clearly unrealistic since there are only 8–9 non-overlapping 35-y segments in this control segment. The value of $\nu = 12$ implies that a $\sim 50\%$ increase in degrees of freedom has been gained by overlapping vectors of pseudo-observations, which is consistent with standard spectral estimation theory (Allen and Smith 1996, Appendix).

The horizontal dashes in Fig. 3 show the "310-y error" range, that is, $\pm$ the largest absolute pattern-amplitude ($\sqrt{\varepsilon^2_{\max}}$) observed in a 310-y control integration. These ranges approximately match the parametric ranges, indicating that the Gaussian assumption and our estimate of ν are, in this instance, reasonably accurate. We stress that ν is a very uncertain quantity, so these 310-y error ranges represent a more robust diagnostic than the standard confidence intervals.

Figure 3 indicates that $\mathscr{H}(\beta_G = 0)$, 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). For truncations $\kappa \leq 13$, however, $\mathscr{H}(\beta_G = 1)$, that the model-predicted amplitude is correct, can also be rejected at $P_{<0.05}$, except at the lowest truncations, where the detection space appears to be inadequate to resolve the signal. The key point to note is that error bars consistently decline as we increase the truncation level. Before drawing any further conclusions, therefore, we need to establish the maximum truncation at which the model is reliable.

The singular value spectrum of the control, shown in Fig. 4, gives little indication of the 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 and Smith 1996), no such break is evident, so we require other truncation criteria.

The solid line in Fig. 5 shows the evolution of $(\kappa - m)/r^2$ as defined in Eq. (18) as a function of truncation. Although we test r^2 directly using a Eq. (18), we plot this quantity because it may be interpreted physically as the cumulative ratio of model/observed

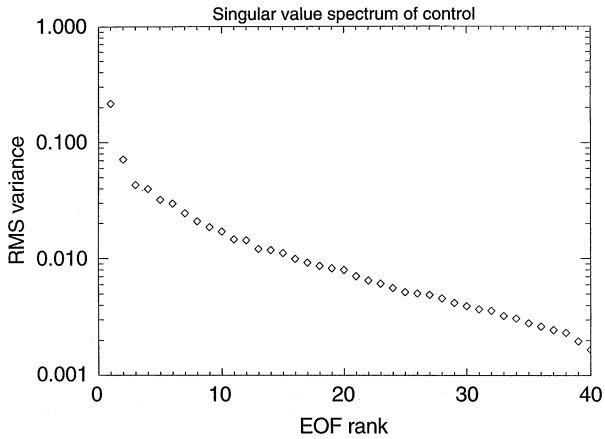

Fig. 4 The spectrum of singular values of the control. There is no sharp break in the spectrum, giving no indication of an appropriate truncation

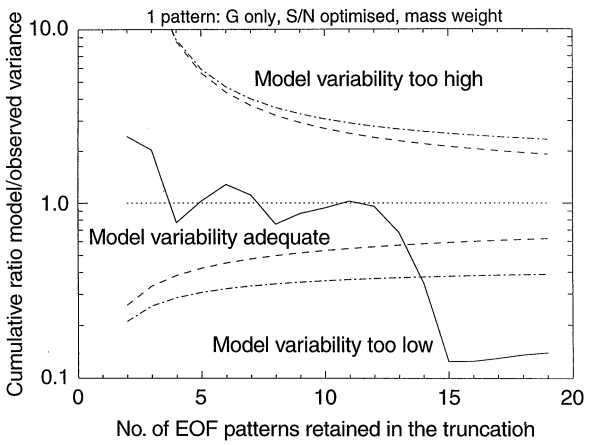

Fig. 5 *Solid line*, evolution of $(\kappa - m)/r^2$ (cumulative model/observed residual variance ratio) with truncation. Model-simulated variability appears to be approximately correct for low truncations, and consistently low for higher truncations. *Dashed* (*dash-dot*) *line*, 5–95% range of the $(\kappa - m)/\chi^2_{\kappa-m}(1/F_{\kappa-m,\nu})$ distribution. If the *solid line* moves outside this range, uncertainty estimates will be unreliable

residual variance. For truncations ≤ 12 it varies around unity, indicating the model variability is consistent with observed, while for higher truncations it drops rapidly outside the range indicated by the $\chi^2_{\kappa-m}$ and $F_{\kappa-m,\nu}$ distributions. The range based on $\chi^2_{\kappa-m}$ is consistently narrower than that based on $F_{\kappa-m,\nu}$, indicating that the χ^2 test should be used in preference to the F-test to provide a conservative truncation point given the difficulties in estimating ν. A similar truncation point is indicated if we reverse the control segments $\mathbf{Y}_{N_1}$ and $\mathbf{Y}_{N_2}$, suggesting that variability is consistently underestimated in these low-ranked EOFs, rather than just in this particular segment.

Levels of internal variability can vary by more than a factor of two in variance between different models (Kim et al. 1996). If the r^2 values in Fig. 5 are reduced by this amount, the χ^2 test indicates that uncertainty estimates are unreliable for truncations as low as 7. For small truncations, $\kappa = 4$–6, the test is simply not powerful enough to identify this model-data discrepancy.

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, $\mathscr{H}(\beta_G = 1)$ is rejected. A univariate model based on the GSO pattern appears to do rather better, shown in Fig. 6. Again, 12 is the maximum allowable truncation, at which point $\mathscr{H}(\beta_{GSO} = 0)$ can be rejected, while $\mathscr{H}(\beta_{GSO} = 1)$ cannot. Note how the estimate of β_{GSO} varies with truncation, emphasising the need for objective truncation criteria: at the lowest truncations, we are failing to represent the signal adequately, so results are essentially arbitrary. Estimates then stabilise up to $\kappa = 12$, upon which they begin to vary again, presumably because we are introducing EOFs containing unrealistically low variance which are being given high weight in the optimisation.

It would be incorrect to conclude on the basis of this improvement alone 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 the associated 310-y return

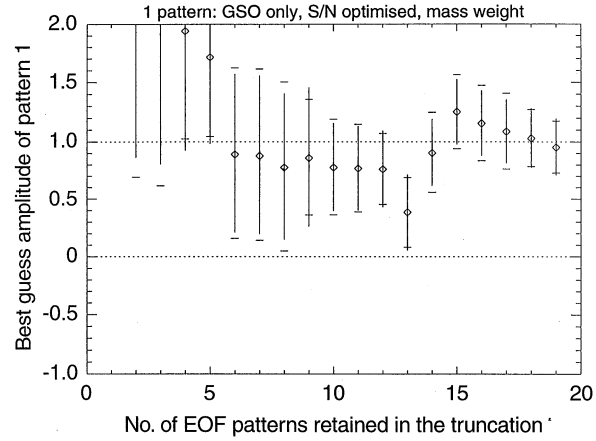

Fig. 6 Estimated amplitude of *GSO* (greenhouse gases, aerosols and ozone pattern) versus rank of the detection space. Note how, unlike in the case of the G pattern, $\mathscr{H}(\beta_{GSO} = 1)$ cannot be rejected at 12-EOF truncation

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 Fig. 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 parametrization 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 $P_{0.1\text{–}0.05}$ 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 of 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-y return envelope is illustrated in Fig. 8, which shows the joint distribution of *G* and *GSO* pattern amplitudes, with S/N optimisation, computed from the columns of $\mathbf{Y}_{N_2}$. The ellipse, by construction, passes through the largest excursion from the origin as defined in Eq. (15). For comparison, the dashed and dotted lines show the $P_{0.1}$ and $P_{0.05}$ confidence intervals respectively computed using Eq. (14) with $\nu = 12$, the estimated degrees of freedom taking into account lag-1 auto-correlation in the control. As we would expect for this ν, the 310-y return envelope lies between the two. 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, we expect results to be critically dependent on the choice of κ. 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.

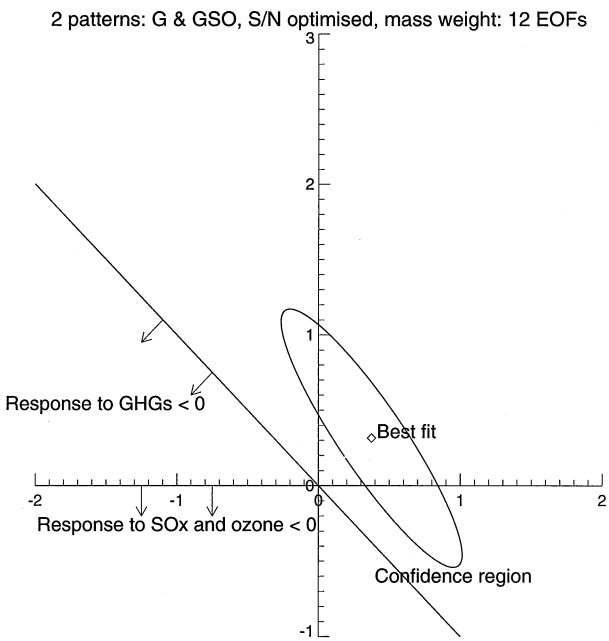

Fig. 7 Best-fit amplitudes $\tilde{\beta}$ and associated uncertainty ranges for the model-predicted patterns of change due to greenhouse gases alone (*horizontal axis*) and the combined effects of greenhouse gases, sulphates and ozone (*vertical axis*). Estimates based on $\kappa = 12$ highest-ranked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

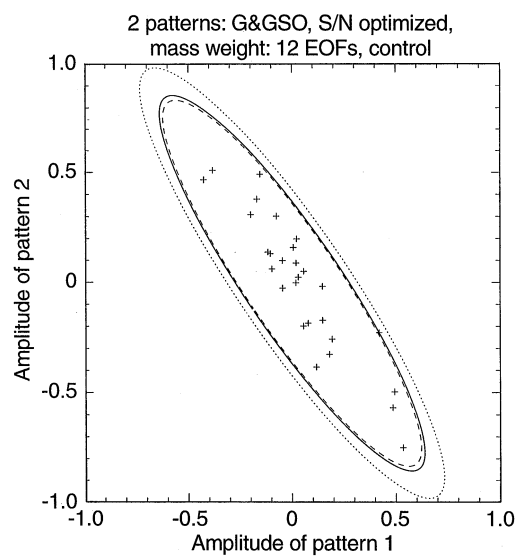

Fig. 8 The joint distribution of *G* and *GSO* pattern amplitudes, with S/N optimisation, in segments of a 310-y control integration. *Solid ellipse* shows largest noise-weighted excursion from the origin, *dotted/dashed lines* show $P_{0.05}/P_{0.1}$ confidence intervals based on an assumed Gaussian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

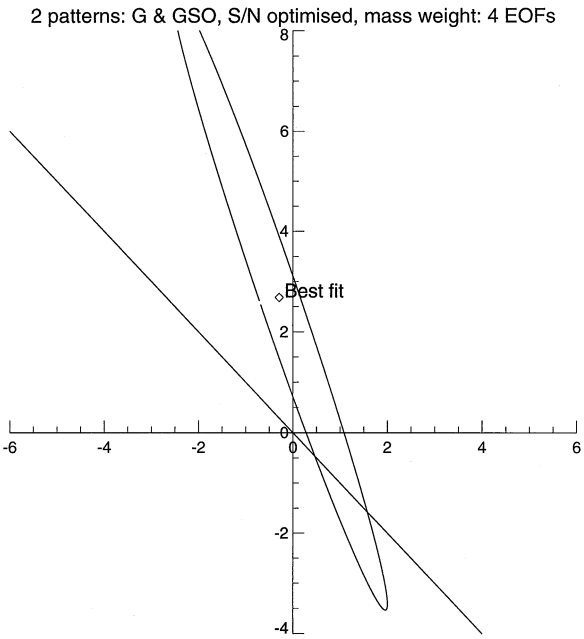

Fig. 9 Best-fit $\tilde{\beta}$ and associated uncertainty ranges with very low truncation: $\kappa = 4$. The detection space is unable to represent the signal, leading to very large uncertainties. Note *revised axes*

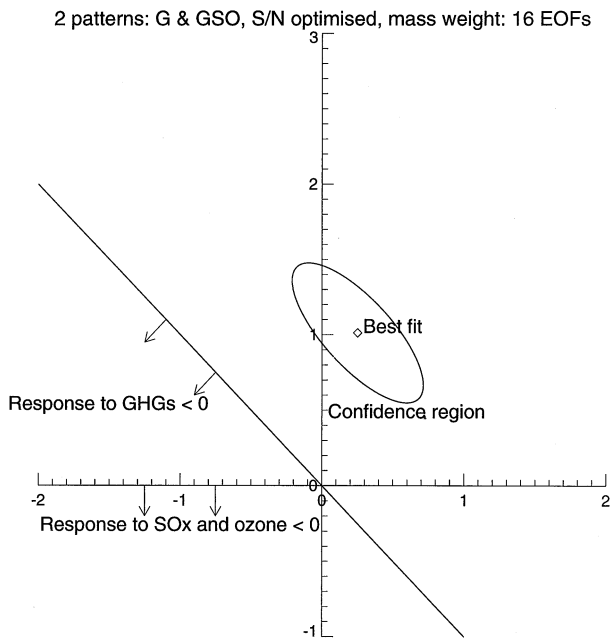

Fig. 10 Best-fit $\tilde{\beta}$ and associated uncertainty ranges with excessively high truncation: $\kappa = 16$. Inclusion of high-ranked EOFs containing unrealistically low variance leads to misleadingly small estimated uncertainties

Qualitatively different results emerge from the adoption of different truncations, graphically illustrating the need for objective criteria to determine the appropriate truncation level. The evolution of $P(\chi^2)$ with truncation in the bivariate model is very similar to the univariate case shown in Fig. 5. *P*-values for the χ^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, so results at $\kappa = 16$ are meaningless.

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

6 Implications for climate sensitivity

If an anthropogenic signal is indeed detectable in the recent climate record, then it should be possible to exploit this signal to quantify how the climate system responds to changing radiative forcing. We believe there has been insufficient emphasis to date on the physical implications of detection and attribution results, so we conclude with a preliminary illustration, subject to many caveats, of how such an exercise might proceed.

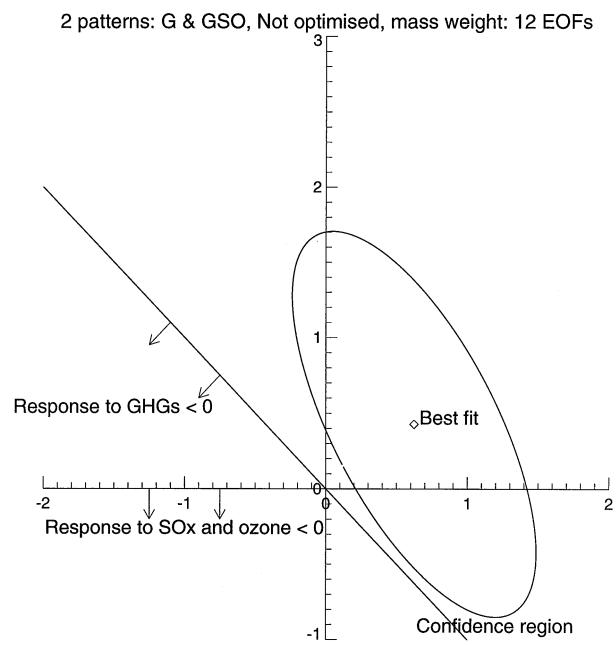

Fig. 11 An example of the benefits of optimisation: best-guess pattern amplitudes in the bivariate detection model with 12 EOF truncation but without S/N optimisation

Of particular interest is the response of the climate, on decadal time-scales, to rising greenhouse gases. If the various forcings had been prescribed separately in the original runs, we could infer this from the estimated amplitude of the greenhouse response, which we shall call $\tilde{\beta}_{GHG}$. Under the (very restrictive) assumption that the timing of the model response is correct, a $\tilde{\beta}_{GHG}$ range including unity would imply that the true climate sensitivity is consistent with the sensitivity of the model.

Since various forcings were prescribed simultaneously in the *GSO* run, interpretation is slightly more complicated, but an estimate of climate sensitivity can nevertheless be derived. Suppose $\mathbf{x}_{GHG}$ and $\mathbf{x}_{SO}$ represent the responses to greenhouse gases and the combined effects of sulphates and ozone respectively in the real world. Given the assumption of linearity, our best-fit regression model then becomes

$$\begin{aligned}\tilde{\mathbf{y}} &= \tilde{\beta}_G\mathbf{x}_{GHG} + \tilde{\beta}_{GSO}(\mathbf{x}_{GHG} + \mathbf{x}_{SO}) \\ &= (\tilde{\beta}_G + \tilde{\beta}_{GSO})\mathbf{x}_{GHG} + \tilde{\beta}_{GSO}\mathbf{x}_{SO} \\ &= \tilde{\beta}_{GHG}\mathbf{x}_{GHG} + \tilde{\beta}_{SO}\mathbf{x}_{SO}. \end{aligned} \quad (21)$$

Hence the sum $\tilde{\beta}_{GHG} = \tilde{\beta}_G + \tilde{\beta}_{GSO}$ gives an estimate of the scaling on the *total* greenhouse response required to match observations, and thus an estimate of the climate sensitivity, taking into account our uncertainty in the amplitude of the response to sulphates and ozone, while (in this two-pattern regression), the $\tilde{\beta}_{SO} = \tilde{\beta}_{GSO}$ estimate is only dependent on the sulphate/ozone signal. Because the regression algorithm is linear, precisely the same results would be obtained by subtracting $\mathbf{x}_G$ from $\mathbf{x}_{GSO}$ initially to give a separate $\mathbf{x}_{SO}$ pattern to input to the regression. This pattern would, however, be intrinsically noisier than $\mathbf{x}_{GSO}$, so we prefer to display results based on the simulations which were actually performed and derive the parameters of interest from the estimates afterwards. Since, at this stage, we are ignoring the effect of noise in the model-predicted patterns, it makes no practical difference which approach is taken.

At 12 EOF truncation, the scaling factor $\tilde{\beta}_{GHG}$ lies in the range 0.35–1.0 (310-y error, which is approximately the 5–95% interval) see Fig. 12. This implies, on the basis of this diagnostic, that the model is either overpredicting the response to greenhouse gases by almost a factor of 3 or (at the other end of the range) that the amplitude of the model response is approximately correct. We stress that these estimates are subject to a known bias towards zero due to noise in the estimated response-patterns, as discussed above. Preliminary results suggest that the application of an unbiased algorithm (Ripley and Thompson 1987) raises the upper end of the range by up to 25%, with little change to the lower end, bringing the central estimate ("best guess") closer to, but still less than, unity. The reason the model may be overpredicting the response to greenhouse gases needs further investigation, but it may be related to the strong warming of the tropical upper troposphere observed in the model, which Tett et al. (1996) argued is likely to be a model error.

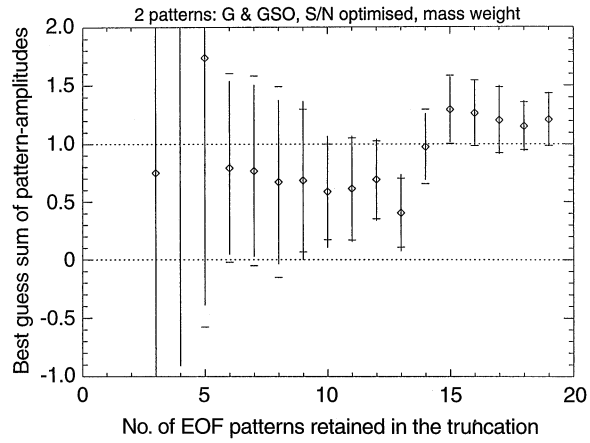

Fig. 12 Translating optimal detection results into estimates of climate parameters: the sum of *G* and *GSO* pattern amplitudes gives an estimate of the scaling required on the total greenhouse response to match observations, and thus an estimate of the climate sensitivity

Specifically, the model-predicted negative lapse-rate feedback (a decrease in lapse rate accompanying a tropospheric warming) appears to be unrealistically strong (Tett et al. 1997; Gillett et al. 1998), which would mean that the model could be overpredicting the mid- and upper-tropospheric temperature response to increasing greenhouse gases while getting the surface temperature response correct. If true, this would mean that our estimate of the climate sensitivity based on tropospheric temperature changes would be a systematic underestimate of the true (surface temperature) sensitivity. Work on more direct estimates of the sensitivity, based on surface temperature changes, is in progress.

The equilibrium sensitivity, *S*, of HadCM2 to doubling CO_2 is 3.3 K (based on an extended integration of the coupled model C. A. Senior, *personal communication*) so, to first order, we can translate our uncertainty range in $\tilde{\beta}_{GHG}$ into an "observed" range for the climate sensitivity of 1.2–3.4 K (rounding to 2 significant figures), reiterating that the upper end of this range may be underestimated by up to 25% due to our neglect of noise in the model-predicted patterns. We stress that there is a considerable element of extrapolation in this estimate: we are assuming that the timing of the model-predicted response is correct, that the response is sufficiently linear that we can make inferences from the observed response to intermediate-amplitude forcing changes to the full doubled-CO_2 sensitivity, and that the pattern of response is independent of its amplitude.

There are, in fact, strong physical arguments that the timing of the response itself depends on the sensitivity see, for example, Hansen et al. (1985), in which case the

"transfer function" relating the sensitivity to the observed response-pattern-amplitude, $S = S(\beta_{GHG})$, would necessarily be non-linear. Quantifying the strength of this non-linearity represents work in progress. If, however, we assume the relevant oceanic processes are correctly simulated, two points on the transfer function are known: $[\beta_{GHG} = 0, S = 0]$ (zero response implies zero sensitivity) and $[\beta_{GHG} = 1, S = 3.3]$ (correct response implies the model-predicted sensitivity). As long as we are interpolating between these two points (i.e. $\beta_{GHG} \leq 1$, scaling *down* from the model prediction), the impact of non-linearity in $\mathscr{S}(\beta_{GHG})$ is limited, but it would clearly be unsafe to extrapolate far outside the $0 \leq \beta_{GHG} \leq 1$ interval.

The assumption that the pattern of response is independent of its amplitude is also questionable. For example, on physical grounds, we would expect the rate of stratospheric cooling to be virtually independent of the tropospheric feedbacks which primarily determine climate sensitivity. To quantify accurately the physical implications of detection results, it will eventually become necessary to decompose detection diagnostics, such as fingerprint patterns, into components which depend on key climate parameters, like sensitivity, and components which do not. Again, this generalisation must be pursued in subsequent work. Bearing in mind all these caveats, however, we present this sensitivity estimate as an example of how, as the signal of anthropogenic climate change emerges from the noise, optimal detection results may be used to obtain information on physically-interpretable climate parameters.

Given the estimate $\tilde{\beta}$ and its associated uncertainty $\tilde{V}(\tilde{\beta})$, and bearing in mind these caveats, we can reconstruct the best-guess trend at each latitude/pressure point and the corresponding $P_{0.05}$ confidence interval using Eq. (9). Maximum and minimum reconstructed trends, taking into account internal variability illustrated in Fig. 1, are shown in Fig. 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 (Fig. 1) and model are clearly inconsistent. The χ^2 test described, being based on a global summary statistic, might well fail to identify local model-data discrepancies. For example, the observed cooling at ~ 50 hPa 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.

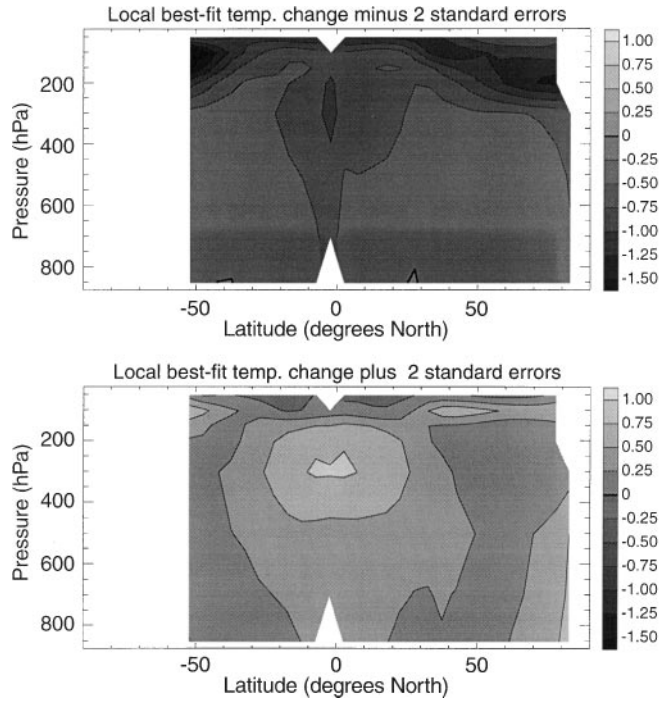

Fig. 13 Maximum and minimum ($P_{0.05}$ one-tailed limits) local trends indicated by the detection model. Locations where the observations (Fig. 1, *top panel*) lie outside this range indicate systematic model deficiencies

7 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 demonstrably inaccurate. We have presented a simple check (the χ^2-test or F-test on residuals) which should detect gross model inadequacies and demonstrated its application to the "vertical detection problem", examining decadal changes in atmospheric vertical temperature structure over the period 1961–1995. The HadCM2 control integration was found to be an inadequate model of internal variability in the particular diagnostic examined, provided a mass-weighting scheme was used to focus the analysis on the troposphere. Observed residual variability (after anthropogenic signals had been removed) was found to be inconsistent with the model when a volume-weighted diagnostic was used, highlighting known deficiencies in the model simulation of stratospheric variability.

Under the mass-weighted scheme, the influence of anthropogenic greenhouse gases on atmospheric vertical temperature structure was detected unambiguously at a high confidence level. Taking into account the model-predicted effects of sulphates and ozone improved the overall fit, but not enough for us to claim unambiguous detection of a sulphate/ozone signal. Because the greenhouse and sulphate/ozone signals are not

orthogonal to each other, assuming no prior knowledge of any of these signal amplitudes leaves us with an ambiguity: either the real climate response to greenhouse gas increase is weaker than that predicted by the model and the response to sulphate and ozone changes is negligibly small, or the model-predicted amplitude of both signals is approximately correct.

Even allowing for this uncertainty in the sulphate/ozone signal, we estimate (on the basis of the mass-weighted fingerprint, at the 95% confidence level) that the response to greenhouse gas increase is 0.35–1.0 times the model-predicted value. Assuming the timing of the model response is correct, this implies a 5–95% range in the climate sensitivity to doubling CO_2 of 1.2–3.4 K, although the upper end of this range is likely to be biased towards zero due to sampling uncertainty in the model-predicted patterns. This use of optimal detection results to obtain observationally-based estimates of climate parameters and to identify systematic model deficiencies represents, we believe, an exciting and novel research opportunity provided by the emergence of the anthropogenic signal.

Finally, 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. An alternative, non-parametric, approach is presented which avoids the most restrictive of these distributional requirements.

Appendix: discrepancies with Tett et al. (1996)

When using segments of "pseudo-observations" from a model control integration as a basis for uncertainty analysis in a detection study, it is important that the same data gaps which occur in the observations are imposed on these segments of model data, since missing data increases the variance in any diagnostic. Owing to a coding error, this was not imposed on the control segments in Tett et al. (1996). The revised Table 1 from that study, using the full length of the control integration which was not available at that time, is shown here:

The revised Table 1 might be interpreted as giving some support to the hypothesis that declining stratospheric ozone had an influence on recent temperature trends, since only scenarios involving ozone indicate pattern correlation or congruence values with the observations which are significantly different from zero. The corrected version of Table 2 of that study, however, now indicates that none of these scenarios is significantly better than any other under these correlation-based diagnostics. Tett et al.'s (1996) detection of combined influence of greenhouse gases, sulphate aerosols and stratospheric ozone depletion, as specified in the GSO and SENS1 experiments, still stands, but on the basis of these R and g statistics, it is impossible to say which of these three forcings is primarily responsible for the agreement. Regression-based diagnostics, as used in the present study, are much easier to interpret when multiple forcing scenarios are involved. This illustrates the importance of careful treatment of missing data in detection studies, earlier noted by Santer et al. (1993). This is an area which has still not been completely resolved, since due to computational constraints, we still use zonal mean diagnostics without imposing the meridional sampling pattern.

Table 1 Revised version of Table 1 from Tett et al. (1996) study using full length of control integration and correcting error in application of missing data mask. R and g values show pattern correlation and congruence statistics between observed pattern of zonal mean temperature change between 1961–80 and 1986–95 and model predictions under greenhouse-gas-only (*G*) , greenhouse-plus-sulphate (*GS*), greenhouse-sulphate-and-ozone (*GSO*) and revised greenhouse-sulphate-and-ozone, halving the effect of ozone (SENS1). Numbers in brackets show the largest observed R or g value in 169 35-year segments from the HadCM2 control integration. Only the patterns involving ozone influence are detectably different from zero with this diagnostic

Signal	Mass weighting		Volume weighting	
	R	g	R	g
GSO	**0.78** (0.65)	**0.78** (0.55)	**0.79** (0.73)	**0.81** (0.57)
GS	0.73 (0.76)	0.72 (0.91)	0.81 (0.87)	0.60 (0.89)
G	0.70 (0.83)	0.69 (0.94)	0.81 (0.88)	0.51 (0.92)
SENS1	**0.78** (0.70)	**0.82** (0.87)	**0.82** (0.81)	**0.79** (0.76)

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. We are indebted to Catherine Senior for advice and unpublished results concerning the sensitivity of HadCM2 and to David Sexton for highlighting the problem in Tett et al. (1996) detailed in the Appendix. MRA was supported by the NERC/RAL Ocean Dynamics Service Level Agreement and by a NERC Advanced Research Fellowship, SFBT by the UK Department of the Environment, Transport and Regions under contract PECD 7/12/37.

References

Allen MR, Smith LA (1996) Monte Carlo SSA: detecting irregular oscillations in the presence of coloured noise. J Clim 9 : 3373–3404

Barnett TP, Santer BD, Jones PD, Bradley RS, Briffa KR (1996) Estimates of low frequency natural climate variability in near-surface air temperature. Holocene 6 : 255–263

Bell TL (1986) Theory of optimal weighting to detect climate change. J Atmos Sci 43 : 1694–1710

Bradley RS, Jones PD (1993) 'Little Ice Age' summer temperatures: their nature and relevance to recent global warming trends. The Holocene 3 : 367–376

Briffa KR, Schweingruber FH, Jones PD, Osbron TJ, Shiyatov, SG, Vaganov EA (1998) Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. Nature 391 : 678–682

Gillett NP, Allen MR, Tett SFB (1998) Modelled and observed variability in atmospheric vertical temperature structure. Clim Dyn, to appear

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

Hansen J, Russell G, Lacis A, Fung I, Rind D, Stone PA (1985) Climate response times: dependence on climate sensitivity and ocean mixing. Science 229 : 857–859

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

Hasselmann K (1993) Optimal fingerprints for the detection of time dependent climate change. J Clim 6 : 1957–1971

Hasselmann K (1997) On multifingerprint detection and attribution of anthropogenic climate change. Clim Dyn 13 : 601–611

Hasselmann K (1998) Conventional and Bayesian approach to climate change detection and attribution. Q J R Meteorel Soc (to appear)

Hegerl GC, North GR (1997) Comparison of statistically optimal approaches to detecting anthropogenic climate change. J Clim 10 : 1125–1133

Hegerl GC, von Storch H, Hasselmann K, Santer BD, Cubasch U, Jones PD (1996) Detecting greenhouse gas-induced climate change with an optimal fingerprint method. J Clim 9 : 2281–2306

Hegerl G, Hasselmann K, Cubasch U, Mitchell JFB, Roeckner E, Voss R, Waszkewitz J (1997) On multi-fingerprint detection and attribution of greenhouse gas and aerosol forced climate change. Clim Dyn 13 : 613–634

Johns TC, Carnell RE, Crossley JF, Gregory JM, Mitchell JFB, Senior CA, Tett SFB, Wood RA (1997) The Second Hadley Centre coupled ocean-atmosphere GCM: model description, spin-up and validation. Clim Dyn 13 : 103–134

Jones PD, Hegerl GC (1998) Comparisons of two methods of removing anthropogenically related variability from the near-surface observational temperature field. J Geophys Res 103 : 13 777–13 786

Kim KY, North GR, Hegerl GC (1996) Comparisons of the second-moment statistics of climate models. J Clim 9 : 2204–2221

Leroy S (1998) Detecting climate signals, some Bayesian aspects. J Clim 11 : 640–651

Mardia KV, Kent JT, Bibby JM (1979) *Multivariate analysis*. Academic Press, New York

Mitchell JFB, Johns TC, Gregory, JM, Tett SFB (1995a) Climate response to increasing levels of greenhouse gases and sulphate aerosols. Nature 376 : 501–504

Mitchell JFB, Davis RA, Ingram WJ, Senior CA (1995b) On surface-temperature, greenhouse gases, and aerosols – models and observations. J Clim 8 : 2364–2386

North GR, Stevens MJ (1998) Detecting climate signals in the surface temperature record. J Clim 11 : 563–577

North GR, Bell TL, Cahalan RF, Moeng FJ (1982) Sampling errors in the estimation of empirical orthogonal functions. Mon Weather Rev 110 : 699–706

North GR, Kim KY, Shen SSP, Hardin JW (1995) Detection of forced climate signals, 1: filter theory. J Clim 8 : 401–408

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical recipes in FORTRAN*: *the art of scientific computing*, 2 edn. Cambridge University Press, Cambridge, UK

Ripley BD, Thompson M (1987) Regression techniques for the detection of analytical bias. Analyst 112 : 377–383

Santer BD, Wigley TML, Jones PD (1993) Correlation methods in fingerprint detection studies. Clim Dyn 8 : 265–276

Santer BD, Mikolajewicz U, Bröggemann W, Cubasch U, Hasselmann K, Höck H, Maier-Reimer E, Wigley TML (1994a) Ocean variability and its influence on the detectability of greenhouse warming signals. J Geophys Res 100 : 10 693–10 725

Santer BD, Brüggemann W, Cubasch U, Hasselmann K, Höck H, Maier-Reimer E, Mikolajewicz U (1994b) Signal-to-noise analysis of time-dependent greenhouse warming experiments. Part 1: pattern analysis. Clim Dyn 9 : 267–285

Santer BD, Taylor KE, Wigley TML, Johns TC, Jones PD, Karoly DJ, Mitchell JFB, Oort AH, Penner JE, Ramaswamy V, Schwarzkopf MD, Stouffer RJ, Tett S (1996) A search for human influences on the thermal structure of the atmosphere. Nature 382 : 39–46

Stott PA, Tett SFB (1998) Scale-dependent detection of climate change. J Clim (to appear)

Stouffer RJ, Manabe S, Vinnikov KY (1994) Model assessment of the role of natural variability in recent global warming. Nature 367 : 634–636

Tett SFB, Mitchell JFB, Parker DE, Allen MR (1996) Human influence on the atmospheric vertical temperature structure: detection and observations. Science 247 : 1170–1173

Tett SFB, Johns TC, Mitchell J (1997) Global and regional variability in a coupled AOGCM. Clim Dyn 13 : 303–323

Thacker WC (1996) Climate fingerprints, patterns and indexes. J Clim 9 : 2259–2261

von Storch H, Hannoschöck G (1986) Statistical aspects of estimated principal vectors (EOFs) based on small sample sizes. J Clim 24 : 716–724

Zwiers FW, von Storch H (1995) Taking serial correlation into account in tests of the mean. J Clim 8 : 336–351