

## **7. A framework for quantitatively advancing detection studies to begin to rigorously consider the internal consistency of state-of-the-art climate models**

*“Stage 4 studies differ from the earlier categories in their use of a number of climate variables simultaneously. Thus the searched-for anthropogenic signal in a Stage 4 investigation might consist of patterns of change in surface air temperature, precipitable water, diurnal temperature range, etc. Each variable would be defined in at least two (and possibly three or four) dimensions. The individual components of this multi-variable “fingerprint” might be decided upon by preliminary analysis of the signal-to-noise properties of a wide range of climate variables in model data. At Stage 4, “successful detection” would mean that the multi-variate, space-time varying climate-change signal from a model experiment forced with transient increases in anthropogenic emissions was in good accord with available observations”*

*Santer et al., 1996, IPCC Second Assessment Report (SAR)*

In the IPCC SAR, the most advanced detection study conceived of was a Stage 4 study. Implementation of such a study has not been achieved to date, primarily due to the lack of a range of suitably constrained observational datasets of sufficient length. The availability of both HadCRUTv (Jones et al., 2001) near-surface, and HadRT (Parker et al., 1997) upper air temperature records, for the common period of 1958 to date, means that such studies should now be possible considering tropospheric temperatures. In chapter 6, a number of tropospheric temperature variables based upon these data were considered independently in detection studies under a common OLS detection framework (AT99) for HadCM2 and HadCM3, and the results were then qualitatively intercompared. This, it could be (liberally) argued, was a first attempt at a Stage 4 study. It would be highly desirable, however, to elicit a more formal, quantitative measure as to whether the climate models being considered are demonstrably adequate explanations of recent climate change.

The definition of Stage 4 studies provided in the SAR is at best ambiguous. There are two obvious routes that could be pursued to gain a quantitative detection statistic. Firstly, a single application of any detection algorithm could be undertaken with all the input fields combined together, yielding a single result. Alternatively, an application of the same detection algorithm could be applied to each field, and the results compared in a quantitative manner to determine whether the model is, at least in some limited sense, a consistent explanation of recently observed climate changes.

Undertaking an approach whereby the atmospheric variables being considered are combined to yield a single input field to any detection algorithm has some immediately obvious limitations. The most immediate problem when considering optimal detection techniques is that combining the fields will yield a larger input to the detection algorithm. This makes it much harder to accurately estimate the covariance matrix (due to natural internal climate variability), upon which the optimisation is based. Hence such an approach could lead to both poorly constrained, and highly biased, estimators under optimal approaches to detection and attribution such as those employed in this thesis (AT99, AS01). This is especially pertinent given the relatively short model control runs currently available to estimate the covariance matrix for both optimisation and hypothesis testing purposes.

Furthermore, although it is a potentially powerful statistical approach, as more variables are added it would become increasingly difficult to find out where the information yielding the result effectively arises from within such a (relatively) complicated input field. This would mean the loss of potentially useful information as to whether, and if so where or in what variables, the model is overestimating / underestimating or otherwise grossly mis-representing the atmospheric variability and / or signal response. Taking such an argument to its logical conclusion, the presence of a single variable that a model grossly fails to predict in any meaningful manner may present a null detection result when all other variables are, at least in a statistical sense, in agreement. This is, of course, an important result as the model is an inadequate explanation of all the variables being considered, being only as good as its worst variable. However, it is also an inadequate explanation of only a single variable amongst the many being considered, and expectations must be that any model will not be adequate in all regards, as it is solely a numerical approximation to the true climate system. Furthermore, the lack of information on what in the model is

effectively causing such a null result means that the modelling community cannot easily address it.

Given such obvious potential problems in the combined input approach, a conceptual framework is proposed and justified here, to apply the alternative approach of inter-comparing the results of the detection algorithm when it is applied to each variable in turn. A method is outlined, using the optimal regression approach of AT99 to yield a number of PDFs as to where the true solution lies in signal phase space. These individual estimates could subsequently be employed to yield a quantitative assessment as to whether the results are internally consistent, and if they are then what the true solution is most likely to be. Section 7.1 briefly recaps the OLS regression technique of AT99. In section 7.2, conditions are relaxed from a perfect world situation towards the real world, to identify those factors which are likely to be important in any consideration of model internal consistency. Finally, section 7.3 proposes a methodological approach to moving forwards to gain the desired quantitative measures of model adequacy. Development and application of the precise methodology are left for future work by others.

## **7.1 The OLS regression algorithm revisited**

Here, an application of the OLS optimal regression detection methodology of AT99 is proposed based upon the ability of the regression algorithm to explicitly calculate the Probability Density Function (PDF) of the  $\beta$  field, the range of plausible amplitudes of individual input model signals in the observations. Only the OLS approach is considered in this chapter, as it has the advantage of always yielding normally distributed solutions, although there is no likely fundamental reason why TLS regression results (AS01) could not also be used.

The basic premise of the OLS regression approach is that the observations can be expressed as a linear combination of physically plausible forcing responses, and an additional noise term due to natural internal variability. Optimisation of both the signals, and the observations, is undertaken with respect to an estimate of the covariance due to natural internal climate variability, to yield the best linear unbiased

estimator of the model signal amplitudes in the observations. The covariance is unknown and must be estimated from a segment of model control. There will be uncertainty in the signal amplitude estimators, which can be considered as a cloud of plausible solution values. To avoid bias in the estimated covariance (uncertainty) of the signal amplitude estimators, the calculation of the solution PDFs is carried out with respect to a covariance estimate based upon an independent section of control. The output will be approximately F-distributed in the signal phase space in the limit of a finite model control run. Isopleths of a PDF can be produced by specifying critical values for the F-distribution, and evaluating the loci of values that they satisfy. A consistency test is undertaken upon the residuals, to ensure that they are similar to an independent section of control. A complete derivation of the OLS solution is given in sub-section 4.2.1.

## **7.2 Relaxing the conditions from an ideal situation to the real world situation**

In this section, a highly simplified climate system is considered. This system consists of just six tropospheric variables. It also responds to only two time-varying external forcings. In the real world, the climate system both contains a much larger variable set, and responds to a much larger number of external forcings.

In an ideal world scenario, climate scientists would have at their disposal a perfect model, which had been run for both an infinitely long control run and an infinite number of ensemble members for each forcing (each started under an independent set of initial conditions from the control). These ensembles would perfectly capture the transient response to the changing forcings, to yield a pure realisation of the signal response. There would also exist an infinite set of observations that perfectly captured the true spatio-temporal history of each and every atmospheric variable over a continuum of real worlds. In such a situation, detection and attribution would not be required, but if any form of detection statistic were implemented then the final result would intuitively consist of a number of points which exactly overlaid yielding the true amplitude of the model signal responses in the observations,  $\beta_{\text{true}}$ . In the simple example used here, for the OLS approach, results would yield six values at

the point (1,1), as the model is perfect (Figure 7.1a). There would be no uncertainty in the estimates, as only this single point could ever be satisfied by every one of the infinite number of possible solutions.

Relaxing the requirements slightly, it is highly unreasonable to assume a continuum of real worlds that can be observed. One is, therefore, limited to a single realisation of the observations. Each of the realisations of the observations (for each of the six variables) is assumed to consist of a linear combination of the true climate system response, and an additive noise term due to internal climate variability. Uncertainty has been introduced into the detection algorithm and, therefore, the result will now be dependent upon the noise present in the observations. It is likely, on physical grounds, that the noise term will be correlated, but not identically, between the six observed variables, at least for tropospheric variables as the troposphere is well mixed. The presence of noise in the observations will affect the calculated OLS estimators  $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_6$ , of the true solution  $\beta_{\text{true}}$  (which is assumed to be the same for all six). Each variable will yield a  $\chi^2$  distributed (infinite control run leads to the F-distribution collapsing to a  $\chi^2$  distribution) PDF estimate as to the value  $\beta_{\text{true}}$ . Solution PDFs will be different shapes, depending upon the true covariance of the atmospheric variables being considered in the bivariate signal phase space, which is highly unlikely to be identical for different variables. Therefore, although the noise in each realisation of the observations includes a common component, this change is likely to project differently onto the bivariate signal phase space, leading to a separation of the solutions. The additional uncorrelated noise component in the six sets of observations will tend to move the individual estimators  $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_6$ , of  $\beta_{\text{true}}$  further apart. Intuitively, all six PDFs would be expected to contain the point (1,1) at the 90% confidence interval (Figure 7.1b).

Given current computing capabilities, even if a perfect model did exist, then it would only be able to be run for a definite finite length of control, and a small population of members for each ensemble response. This adds uncertainty to the solution in two ways. Firstly, the control sections used for the optimisation and hypothesis testing will be finite. The PDF distribution therefore becomes F rather than  $\chi^2$  distributed. An assumption is made in AT99 that the uncertainty arises solely in the magnitude

and not the shape of the principal modes of variability through sub-sampling of an infinite control. This is unlikely to be strictly true, but provides a good first order approximation, and is seen not to significantly affect detection results (Tett et al., 1999). Secondly, there will be uncertainty in the model-derived signal response estimates. This will affect the estimators  $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_6$ . AT99 note how this tends to bias the estimators towards zero, especially for weak or poorly constrained signals (see AS01 for an explanation), and increase the true variance in the estimator by a factor of approximately  $1+1/M$ , where  $M$  is the number of ensemble members used in the derivation of each signal. The effects of using a finite perfect model should, therefore, be primarily to increase the uncertainty in the solution PDFs, at least in the presence of strong signals. This will systematically reduce the signal detectability (whether the amplitude of a signal in the observations is positive definite at a given confidence interval), but will increase the chances of the estimates being consistent with both the observations (1,1), and each other, at any given confidence interval (Figure 7.1c).

The chances of any current model (or any future model) being a perfect model are vanishingly small. A model is, as the name suggests, only an approximation to the real world, it does not (and cannot) purport to be a true representation of the observed system in every minute detail. Therefore, the problem being considered here degenerates to one of the form: "is the model an adequate representation of the observations?". If it is adequate, then it should be impossible to prove that it is inadequate. Some form of test aimed at disproving model inadequacy is the only test that can realistically be applied. Any test must, therefore, determine whether, for the model being considered, there plausibly exists an estimator  $\tilde{\beta}_{true}$  of the true signal amplitudes.

There are additional uncertainties within the system being considered in the real world. The forcing histories used to derive the model forcing response patterns are increasingly poorly constrained back in time, such that there may be non-negligible errors in the signals being considered, which are not directly related to the model itself. It would also be naïve to assume that the observations are entirely free of residual sampling errors (chapter 2). These uncertainties will likely affect the

estimators  $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_6$  in the simple example discussed here by adding further random and systematic errors to the estimators (Figure 7.1d). It is difficult to conceive how these errors can be parameterised without access to an order of magnitude more models and observational dataset realisations than are currently available. However, any test for model adequacy should be able to identify cases when these sources of error are causing gross errors in the individual estimators  $\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_6$ .

Finally, Figure 7.1e gives actual results for the HadCM2 G+S signal combination for the six tropospheric temperature variables considered in Chapter 6. Purely qualitatively, this is indistinguishable from Figure 7.1d, the hypothetical situation. This does not imply that HadCM2 is demonstrably adequate. The hypothetical solution has been constructed by placing subjective noise estimates into the system, and making assumptions about the nature of this noise, which may or may not be correct. The Figure is solely used here for illustrative purposes, to enable an understanding of the likely nature of a consistent result.

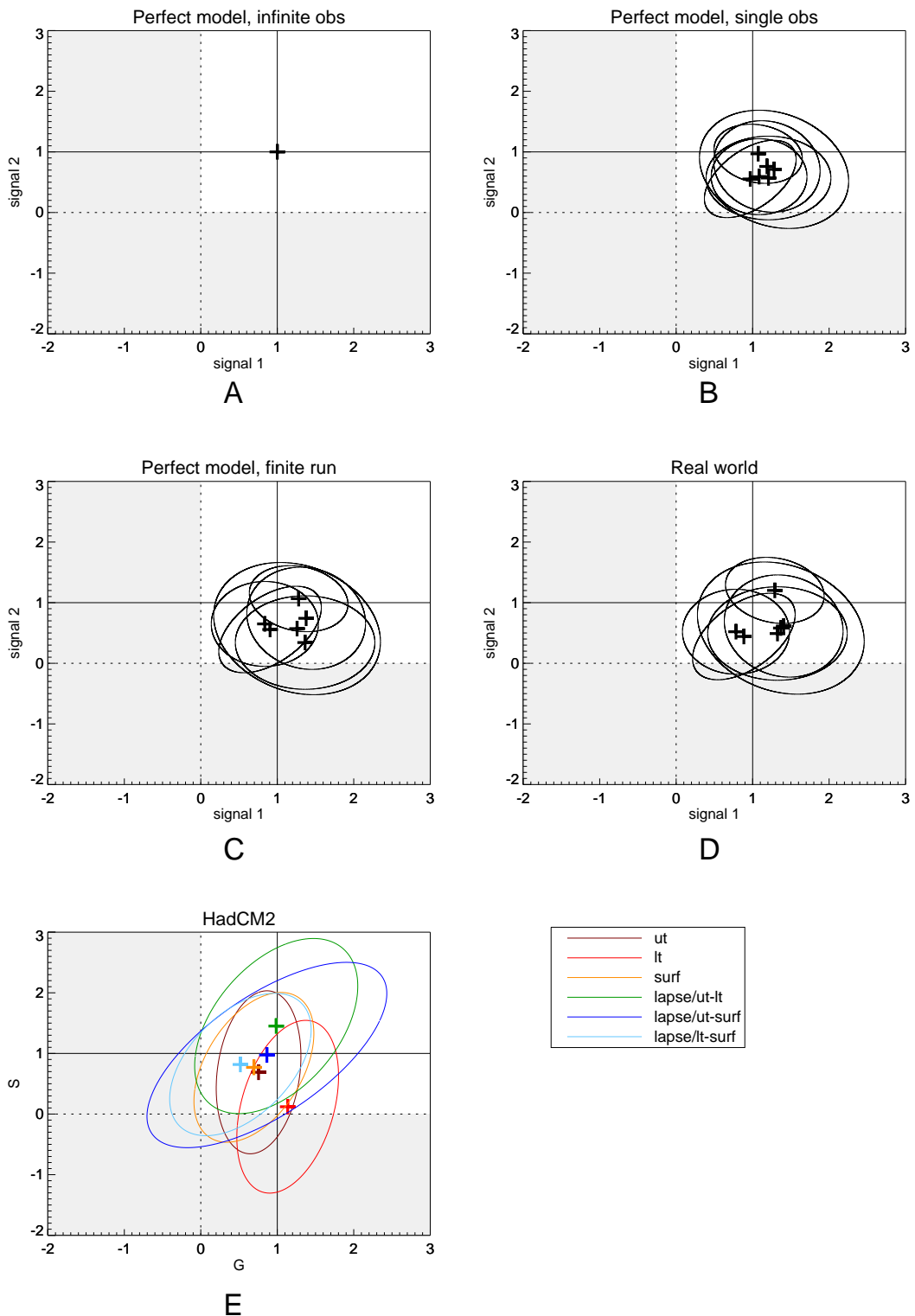
### 7.3 Checking for model consistency: A conceptual framework

A test for internal model consistency must effectively be a test for whether there plausibly exists a solution or range of solutions,  $\beta_{true}$ , for any given model, which satisfies all the OLS estimators for the individual atmospheric variables being considered  $(\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_N)$ . The most intuitive approach to the problem is to **gain an unbiased estimate as to the distribution**  $\tilde{\beta}_{true}$  (where  $\sim$  indicates the best-guess estimate) from the individual estimators  $(\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_N)$ . This likely requires recourse to the Bayesian (probabilistic) family of statistical approaches, using the ability of the OLS regression algorithm to return PDF estimators as to the true solution. In terms of increasing complexity, the problem could be viewed in a conditional probability framework (Von Storch and Zwiers, 1999), as an explicitly Bayesian problem (Carlin and Louis, 2000), or in terms of relative entropy (Uffink, 1995 (and later web update)). The key problem in any of these approaches is that the individual component estimators are not independent of one another, certainly within the

troposphere. Therefore, to avoid bias in the estimator  $\tilde{\beta}_{true}$ , an orthogonality constraint may be required. Development of this estimator is beyond the scope of this thesis, and is left to future work by others.

Any resulting unbiased estimate  $\tilde{\beta}_{true}$  could then be compared pair-wise to the component estimators  $(\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_N)$ , to yield whether the model is plausibly an internally consistent explanation of the observations. The most obvious test is some form of distance statistic such as the Mahalanobis distance (Kotz et al., 1983, Von Storch and Zwiers, 1999) which is conditioned upon the two component distributions, although this assumes that the estimator  $\tilde{\beta}_{true}$  will be normally distributed. This is not a new idea to the atmospheric science community (Stephenson, 1997), although it has not to date been considered for such an application. There is likely to be a potential circularity argument in any such distance statistic test, especially for small populations of  $\beta$  estimates (where  $N$  is small), whereby the estimate  $\tilde{\beta}_{true}$  is not (ever) truly independent of the estimates to which it is being compared. This should still, however, provide for a relatively weak internal model consistency check. If the model cannot be shown to be a demonstrably inadequate explanation of the observations, then the PDF estimator of  $\tilde{\beta}_{true}$  could be treated as the best estimate of the model signal amplitudes in the observations, with associated uncertainty. This may have distinct advantages, as this estimator is likely, by construction, to have smaller uncertainty than any of the individual component estimators. The normal detection and attribution tests could then be applied to this estimator.

## Example of the likely behaviour of OLS regression results



**Figure 7.1** Illustration of the likely effects of relaxing from an ideal world to a real world situation, and results for HadCM2 G + S signal combination. In each case only the 90% confidence interval is plotted. Key value abbreviations are given in chapter 6.