Assessing the robustness of zonal mean climate change detection

Peter W. Thorne, Phil D. Jones, Tim J. Osborn, and Trevor D. Davies

Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

Simon F. B. Tett, David E. Parker, Peter A. Stott, and Gareth S. Jones

Hadley Centre for Climate Prediction and Research, Met. Office, Bracknell, UK

Myles R. Allen

Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, University of Oxford, Oxford, UK

[1] We assess the robustness of previous optimal detection and attribution studies considering zonal-mean temperatures. Principal results, which have consistently pointed towards a demonstrable anthropogenic influence on recently observed upper air temperatures, are confirmed. Importantly our detection results are not critically dependent on the inclusion of stratospheric as well as tropospheric temperatures. We find that detection is dependent on input field pre-processing choices, and on the choice of detection algorithm. There are a number of cases where either no signals are detected, or results fail a consistency test. *INDEX TERMS:* 0325 Atmospheric Composition and Structure: Evolution of the atmosphere; 0365 Atmospheric Composition and Structure: Troposphere—composition and chemistry; 0350 Atmospheric Composition and Structure: Pressure, density, and temperature. **Citation:** Thorne, P. W., P. D. Jones, T. J. Osborn, T. D. Davies, S. F. B. Tett, D. E. Parker, G. S. Jones, and M. R. Allen, Assessing the robustness of zonal mean climate change detection, *Geophys. Res. Lett.*, *29*(19), 1920, doi:10.1029/2002GL015717, 2002.

1. Introduction

[2] Previous optimal detection and attribution studies have primarily considered near-surface or zonally averaged upper air temperatures [e.g. *Allen et al.*, 2001; *Barnett et al.*, 1999; *Hegerl et al.*, 1997; *Stott et al.*, 2001a, 2001b; *Tett et al.*, 1996, 1999, 2002 (hereafter T02)]. When results are based upon such a limited set of observed parameters, it is important that we rigorously assess their robustness. Although near-surface temperature detection studies have had numerous sensitivity studies applied [e.g. *Tett et al.*, 1999, T02; *Stott et al.*, 2001a, 2001b], little attention has been paid to the sensitivity of zonal-mean upper-air detection results. Here we test what we consider to be the likely major sources of uncertainty in previous studies [*Allen and Tett*, 1999 (hereafter AT99); T02]. If the results are robust, our confidence in the most likely causes of recently observed changes will be increased.

[3] We employ the optimal regression methodology, as described in detail elsewhere [AT99; *Allen and Stott*, 2001; T02]. The basic premise is that the observations can be recreated from a linear combination of the input model signal response fields and an additive noise term due to natural internal climate variability. The fields undergo a "rotation" to maximise the signal to noise ratio. This rotation is performed within the phase space spanned by the leading EOF's of an estimate of natural internal climate variability (the number included being termed the truncation). Model fields are assumed to be of the correct pattern and, therefore, the approach is one of signal fitting–finding the scaling required on the model signals to recreate the observations. Associated uncertainty remains in the individual signal amplitude estimates due to natural internal climate variability, observational uncertainty, model error, and methodological limitations, although only the first of these is accounted for in assessing the uncertainty range [AT99]. If the uncertainty range is entirely positive, then the signal is detected. In the present study, for consistency with previous detection analyses [e.g., AT99; T02], we use the 5–95% confidence interval, with a risk of false detection of ∼5%. If the uncertainty range encompasses unity, then we conclude that the unscaled model signal is a potentially consistent explanation of the observations (attribution). We include a consistency check on the residuals from the regression, under the assumption that they should be indistinguishable from an independent estimate of natural internal climate variability.

2. Experimental Design

[4] We repeat the experimental set-up of previous studies [AT99; T02]. Observed values are derived from the globally gridded HadRT upper air radiosonde temperature dataset [*Parker et al.*, 1997] at a monthly resolution. We use 8 standard WMO reporting levels from the lower troposphere (850 hPa) to the lower stratosphere (50 hPa). Data are firstly annually averaged (Dec–Nov years), with the criterion that at least 2 months of data must exist in 3 of 4 seasons of any year for an annual mean to be calculated. Analysis [*Thorne*, 2001] indicates that results of zonal-mean detection studies are, to first order, insensitive over a reasonable range of variations. Values are subsequently zonally averaged.

[5] Model data at annual resolution on the three-dimensional model grid are used from the HadCM2 [*Johns et al.*, 1997], and HadCM3 [*Pope et al.*, 2000] coupled climate models, and are bilinearly interpolated to the coarser resolution observed grid before being sub-sampled to the observed coverage. We use these subsampled model fields in subsequent comparisons. Anomalies of both modeled and

observed fields are calculated using a common reference period of 1960–1995. The difference fields between the 1960–1980 and 1985–1995 period averages are used as input to the regression. The signal responses considered for each model are those for well-mixed greenhouse-gases plus sulfate aerosols (GS henceforth), stratospheric ozone depletion (O henceforth), and the combined response to solar [*Lean et al.*, 1995] and volcanic [*Sato et al.*, 1993] forcings (NATURAL henceforth). For HadCM2, which does not have a NATURAL ensemble we create an estimate of NATURAL by linearly combining the individual solar and volcanic forcing response estimates and inflating our uncertainty estimates to account for the increased uncertainty. Differences exist between the two models for similar forcings, particularly in the manner in which the anthropogenic forcings are parameterized [See *Tett et al.*, 1999, T02; *Stott et al.*, 2001a]. This, and other differences [*Pope et al.*, 2000; *Gordon et al.*, 2000], provide for a degree of independence.

3. Sources of Analytical Uncertainty

[6] The results of any detection study are likely to depend on pre-processing choices, as well as on details of the detection algorithm applied.

[7] In previous studies, data have been mass weighted [AT99; T02]. This emphasizes tropospheric values, whilst giving less weight to those in the stratosphere. Other weighting schemes are possible [*Gillett et al.*, 2000]. We also consider volume weighting, which places much higher emphasis on stratospheric values. If we could estimate the true covariance of the climate system for optimization purposes, the choice of weighting, being solely a linear transformation of the input fields, would have no effect upon the results. However, for both models, the control runs are not sufficiently long, so the full covariance matrix used in the optimisation cannot be well-estimated.

[8] We consider four cases for the number of grid-box values required in any zonal band for a zonal-mean value to be calculated (1, 3, 5, and 7 (AT99 and T02 use 3)). As the criterion becomes increasingly strict, the standard error of the resulting dataset is likely to decrease as the estimated zonal-mean becomes closer to the true zonal-mean. Further, because spatial quality control on the HadRT dataset [*Thorne*, 2001] has only been possible in well sampled regions, we expect the stricter zonal means to be better constrained. There will also be increasing emphasis on well-sampled Northern Hemisphere mid-latitudes and tropospheric values [*Thorne*, 2001] which may affect the results.

[9] There is uncertainty in the observed HadRT fields [*Parker et al.*, 1997]. To partly address its effects we use two versions of the dataset: HadRT2.1 and HadRT2.1s. Both have been modified globally for known heterogeneities with reference to co-located MSUc data [*Christy et al.*, 1998]. HadRT2.1s has had modifications applied solely within the stratosphere (at 200hPa and above), whereas HadRT2.1 is adjusted throughout.

[10] Previous studies have considered results from an Ordinary Least Squares (OLS henceforth) approach [AT99; T02], under which the model signals are assumed to be correct. A scaling factor is applied to the result to account for the known uncertainty in the model signals due to finite ensemble size [AT99]. There is also a Total Least

Squares (TLS henceforth) detection methodology, which explicitly accounts for the presence of noise within the signals as well as the observations [*Allen and Stott*, 2001]. This has been shown to have little effect upon detection, but potentially large effects on the scalings for near-surface temperature detection studies [*Stott et al.*, 2001b]. We compare OLS with TLS results. We caution that the consistency tests are not identical between the approaches, especially for small truncations at which TLS tends to be more liberal, and that they should only be used for guidance in combination with our estimators.

[11] Zonal-mean detection studies have been criticized for reflecting the recently observed pattern of stratospheric cooling and tropospheric warming, rather than details of the pattern [*Legates and Davies*, 1997]. Although this divergence is predicted in the models in response to anthropogenic forcing of climate, we additionally consider a troposphere-only diagnostic defined only between the 850 and 300 hPa levels.

4. Results for Full Zonal-Mean Fields

[12] We consider the results at two truncations, 21, the estimated degrees of freedom of the shortest control segment available to perform optimization, and 11, half this value. This provides an indication of sensitivity to changing truncation. We note only whether the residuals pass the consistency test and, if so, then whether any signals are detected. Where signals are detected, we note whether they are potentially consistent explanations of the observed changes. In all cases we consider results for a multiple regression with all three signals as input for consistency with T02, which we treat as our baseline. Analysis in T02 failed the consistency test on the residuals above truncation 7, but detected a HadCM3 GS signal at this truncation.

[13] We first consider sensitivity to the choice of weighting. Volume weighting yields inconsistent residuals in the large majority of cases at both truncations, whereas mass weighting has fewer failures (results not shown). The HadRT radiosonde data is likely to exhibit increasing errors with altitude [*Parker and Cox*, 1995; *Parker et al.*, 1997], whilst both HadCM2 and HadCM3 grossly underestimate natural internal variability in the stratosphere and potentially contain errors in the signals [*Gillett et al.*, 2000; *Collins et al.*, 2001]. Volume weighting will give more weight to the stratosphere and, therefore, the most likely explanation for the residual test failures is an increase in observational and particularly model error. Using TLS rather than OLS reduces the number of residual test failures under both weighting schemes. However mass weighting continues to exhibit fewer failures. Under neither OLS or TLS are the signals detected systematically biased between the weighting schemes. We therefore concentrate hereafter upon results for mass-weighted input fields.

[14] Table 1 illustrates results for the full (850–50 hPa) zonal-mean field. When OLS and HadRT2.1s are used GS is detected at truncation 11, regardless of the number of grid-box values per zone required. There is also a NATURAL influence detected, although this is dependent both upon the model and number of grid-box values required and, therefore, our confidence in this result is reduced. For HadRT2.1, residuals are inconsistent in many cases, and

Regression type		OLS				TLS			
Truncation / No. of values required in zonal band		1	3	5	7	1	3	5	7
11	Model and HadRT versions								
	HadCM2 + HadRT2.1s	GS	GS	GS	GS	NAT			
	HadCM3 + HadRT2.1s	GS,NAT	GS	GS	GS				
	HadCM2 + HadRT2.1					NAT*			
	HadCM3 + HadRT2.1	GS,NAT							
21									
	HadCM2 + HadRT2.1s	GS	GS*,O,NAT	GS*,O,NAT	GS*,O,NAT	GS,NAT	GS*,NAT*	GS*,NAT*	GS*,NAT*
	HadCM3 + HadRT2.1s								GS*,O*,NAT*
	HadCM2 + HadRT2.1						GS,NAT*	GS*,NAT*	GS*,NAT*
	HadCM3 + HadRT2.1							O*,NAT*	GS,O*,NAT*

Residuals inconsistent
* Signal significantly underestimated
$ Signal significantly overestimated

Table 1. Detection results considering full-field zonal-mean input and a range of potential uncertainties. In all cases the results are for a three-way regression of GS, O, and NATURAL. Our baseline case is that of T02 (HadRT2.1, HadCM3, 3 values required, and OLS regression) and is highlighted by a bold box outline and darker shading. In line with their results we find that the residuals are inconsistent at the truncations being considered.

signals are detected only in one case. Corrections to HadRT2.1 have been implemented in time alone, with no spatial consistency requirement [*Parker et al.*, 1997]. These corrections may, in at least some cases, therefore move the dataset away from the leading spatio-temporal modes of natural climate variability, inflating the residuals and hence the likelihood of consistency test failure. Corrections also have a systematic effect upon the signal scaling estimates, reducing them slightly such that signals are systematically less likely to be detected.

[15] At truncation 21, OLS results only yield consistent residuals when HadRT2.1s is used with HadCM2 model data. In this case GS, O, and NATURAL are all detected with some confidence. The GS signal tends to be significantly underestimated in HadCM2. Further analysis yields a tendency for all three signal scaling estimates to increase with truncation. This is not solely the result of known biases in the OLS estimators as it is repeated, although to a lesser extent, for TLS estimators [*Thorne*, 2001]. We therefore caution against applications of a regression approach to constrain future climate change predictions [*Allen et al.*, 2000, 2001] without consideration of the stability to changing truncation over a reasonable range.

[16] When TLS is used, the residuals always pass the consistency test. At truncation 11 there are far fewer cases of positive signal detection, seemingly reducing our confidence in the OLS results. Further analysis shows this to be almost entirely due to an increase in the lower- and

especially upper-bounds of the uncertainty in the estimates at this truncation rather than any systematic negative bias in the estimates themselves. In contrast, there is an increase in the frequency of occurrence of detection at truncation 21, at least partly due to the residuals being consistent in all cases. Under a TLS approach, a NATURAL signal is far more frequently detected, although GS remains detectable. In most cases both GS and NATURAL forcing responses are significantly underestimated in amplitude within both models. This is at odds with results from T02 for HadCM3 who found under an OLS approach that the signal responses were overestimated for GS. We stress that our results are in good agreement with T02 for the same combination of choices, and therefore this discrepancy is due to a combination of the choice of regression approach and the truncation being considered. The detection of an O influence is seen to be model-dependent under a TLS approach, with detection only occurring for HadCM3, and then only in cases where 5 or more gridboxes are required.

5. Results for Tropospheric Only Fields

[17] Table 2 repeats the analysis of Table 1 for the troposphere-only (850–300 hPa) data. Under an OLS approach, results at truncation 11 are similar, with the systematic effect of choice of HadRT version remaining. Therefore detection of a GS influence is not dependent upon the divergence in recent temperature trends across the

Regression type		OLS				TLS			
Truncation / No. of values required in zonal band		1	3	5	7	1	3	5	7
11	Model and HadRT versions								
	HadCM2 + HadRT2.1s	GS	GS		GS				
	HadCM3 + HadRT2.1s		GS	GS	GS				
	HadCM2 + HadRT2.1								
	HadCM3 + HadRT2.1								
21									
	HadCM2 + HadRT2.1s		GS, NAT		GS*,NAT		GS*,O,NAT*		GS,NAT*
	HadCM3 + HadRT2.1s	GS$		GS,NAT	GS,NAT				
	HadCM2 + HadRT2.1								NAT*
	HadCM3 + HadRT2.1								GS,O,NAT

Residuals inconsistent
* Signal significantly underestimated
$ Signal significantly overestimated

Table 2. As Table 1 except considering troposphere-only zonal-mean fields.

tropopause being included. At truncation 21, both models in combination with HadRT2.1s sometimes yield consistent residuals, in contrast to the full-field results for HadCM3. This is most likely due to lower observational and particularly model errors within the troposphere than the stratosphere. In both models GS and NATURAL influences are detected. An O influence is no longer detected, most likely because the majority (but not all) of the signal response arises within the stratosphere. Whenever HadRT2.1 is used at truncation 21 the residuals fail the consistency test.

[18] Results from the TLS approach exhibit a larger sensitivity to whether a troposphere-only diagnostic is considered. However the residuals of the regression are always consistent. At truncation 11 there are no longer any signal detections, and at 21 the number of occurrences of signal detection is reduced. This latter result may be due to fewer data going into a troposphere-only analysis. The NATURAL response remains likely to be underestimated within both models, but evidence for the GS response being underestimated is more ambiguous than for the full-field analysis. Weak evidence exists for a detectable O influence, although this may be due to chance alone.

6. Conclusions

[19] We have assessed the robustness of previous optimal regression zonal-mean detection studies. The chances of successful detection vary greatly with a number of potential sources of uncertainty. However, we found no evidence for potential systematic biases in terms of which signals are detected. Without considering such sources of uncertainty, as has been the case previously, we can only ever have failed to detect a signal when it is in fact detectable, rather than falsely detected signals which do not exist. We conclude that previous zonal-mean temperature detection studies can only ever have been conservative. Our results point towards a combination of anthropogenic and natural external forcings as being the most likely explanation of recent upper air temperature changes. A combined greenhouse-gases and sulfate aerosols signal is robustly detected, whereas the detection of natural external (solar and volcanic) and especially stratospheric ozone depletion signals is more uncertain.

[20] Results are highly sensitive to the weighting applied to the input field, with only those weightings that emphasize tropospheric values regularly yielding consistent residuals. The choice of the number of grid-box values required in any zonal band for a zonal-mean to be calculated has some impact. The choice of corrected or uncorrected observational dataset version being considered has a systematic effect, particularly on results using an OLS approach. The version that has been adjusted within the troposphere is less likely to have consistent residuals or detect any signals. This likely relates to adjustments being applied in a spatially incoherent manner [*Parker et al.*, 1997]. When we account for noise in the model signal fields (TLS), there is both a reduction in the frequency of consistency test failure, particularly at higher truncations, and an increased confidence in the presence of a natural external forcings signal within the observations. We also find some sensitivity to the choice of climate model and caution that results should ideally be considered from multiple models. Finally, we refute previous suggestions that the zonal-mean detection studies results are primarily dependent upon the inclusion of stratospheric values.

References

Allen, M. R., and P. A. Stott, Estimating signal amplitudes in optimal fingerprinting, Part 1: Theory, Submitted to *Clim. Dyn.*, 2001.

Allen, M. R., and S. F. B. Tett, Checking for model consistency in optimal fingerprinting, *Clim. Dyn.*, 15, 419–434, 1999.

Allen, M. R., P. A. Stott, J. F. B. Mitchell, R. Schnur, and T. L. Delworth, Quantifying the uncertainty in forecasts of Anthropogenic climate change, *Nature*, 407, 617–620, 2000.

Allen, M. R., et al., Quantifying anthropogenic influence on recent near-surface temperature change, Accepted by *Surveys in Geophysics*, 2001.

Barnett, T. P., et al., Detection and attribution of recent climate change: A status report, *BAMS*, 80, 2631–2659, 1999.

Christy, J. R., R. W. Spencer, and E. S. Lobl, Analysis of the merging procedure for the MSU daily temperature time series, *J. Clim.*, 11, 2016–2041, 1998.

Collins, M., S. F. B. Tett, and C. Cooper, The internal climate variability of HadCM3, a version of the Hadley Centre coupled model without flux adjustments, *Clim. Dyn.*, 17, 61–81, 2001.

Gillett, N. P., M. R. Allen, and S. F. B. Tett, Modelled and observed variability in atmospheric vertical temperature structure, *Clim. Dyn.*, 16, 49–61, 2000.

Gordon, C., et al., The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments, *Clim. Dyn.*, 16, 147–168, 2000.

Hegerl, G. C., et al., Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change, *Clim. Dyn.*, 13, 613–634, 1997.

Johns, T. C., et al., The second Hadley Centre coupled ocean atmosphere GCM: Model description, spinup and validation, *Clim. Dyn.*, 13, 103–134, 1997.

Lean, J., J. Beer, and R. Bradley, Reconstruction of solar Irradiance since 1610–Implications for climate-change, *Geophys. Res. Lett.*, 22, 3195–3198, 1995.

Legates, D. R., and R. E. Davies, The continuing search for an Anthropogenic climate change signal: Limitations of correlation-based approaches, *Geophys. Res. Lett.*, 24, 2319 –2322, 1997.

Parker, D. E., and D. I. Cox, Towards a consistent global climatological rawinsonde database, *Int. J. of Clim.*, 15, 473–496, 1995.

Parker, D. E., et al., A new global gridded radiosonde temperature data base and recent temperature trends, *Geophys. Res. Lett.*, 24, 1499–1502, 1997.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3, *Clim. Dyn.*, 16, 123–146, 2000.

Sato, M., J. E. Hansen, M. P. McCormick, and J. B. Pollack, Stratospheric aerosol optical depths (1850–1990), *JGR*, 98, 22,987–22,994, 1993.

Stott, P. A., et al., Attribution of twentieth century temperature change to natural and anthropogenic causes, *Clim. Dyn.*, 17, 1–17, 2001a.

Stott, P. A., M. R. Allen, and G. S. Jones, Estimating signal amplitudes in optimal fingerprinting. Part II: Application to general circulation modelling, Submitted to *Clim. Dyn.*, 2001b.

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

Tett, S. F. B., et al., Causes of twentieth-century change near the Earth's surface, *Nature*, 399, 569–572, 1999.

Tett, S. F. B., et al., Estimation of natural and anthropogenic contributions to 20th Century temperature change, Accepted by *JGR*, 2002.

Thorne, P. W., Advancing climate change detection and attribution studies in the free atmosphere, Ph.D. thesis, University of East Anglia, Norwich, 2001.

———————————

P. W. Thorne, P. D. Jones, T. J. Osborn, and T. D. Davies, Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK.

S. F. B. Tett, D. E. Parker, P. A. Stott, and G. S. Jones, Hadley Centre for Climate Prediction and Research, Met. Office, Bracknell, UK.

M. R. Allen, Atmospheric, Oceanic and Planetary Physics, Clarendon Laboratory, University of Oxford, Oxford, UK.