8. Conclusions

8.1 Summary of principal findings

This thesis set out to advance previous climate change detection and attribution studies by considering recently observed upper air temperature records, for the period 1958-1998, in a more rigorous manner than had previously been achieved (Santer et al., 1996a, Tett et al., 1996, 2001, AT99, Hill et al., 2001, G. S. Jones et al., 2001, for example). The scope was broadened to progress from detection statements, to use the results from quantitative detection studies to assess, in a simple manner, the likely tropospheric consistency of two state-of-the-art climate models. To this end observed near-surface temperatures for the same period were also considered.

If any model can be shown to be a consistent explanation of recently observed tropospheric temperature changes, then confidence in its predictions of future climate change will be increased. This is important if such models are going to be realistically used for impacts, adaptation, and mitigation purposes, amongst other applications. It should be stressed that the optimal detection methodologies used here have only considered changes at the largest time and space scales. Any model may be grossly inadequate at smaller time and space scales, even if the largest scale model features can be shown to be perfectly adequate. Therefore, the results of this thesis do not provide justification to use small-scale model output without discussing the many caveats. Furthermore, an implicit assumption is that the observations are a true manifestation solely of recent climate changes. There is evidence (Santer et al., 1999 for example) to suggest that there exist considerable uncertainties in the observed changes between available upper air temperature datasets. Further work aimed at resolving these uncertainties would increase confidence in the results of this thesis.

In chapter 1, the requirements for successful climate change detection studies were introduced, with particular emphasis upon those datasets used in this thesis. Principal results of a number of previous independent detection approaches considering a range of modelled and observational datasets were outlined. There is a consistent indication in these of a significant anthropogenic influence on climate, particularly in the latter half of the 20th Century. There is also more limited evidence, in some studies, for the influence of natural external forcings on climate, especially solar influences in the early part of the 20th Century. All previous results rely almost exclusively upon either near-surface or zonally averaged upper air temperature diagnostics. Reliance upon such a limited set of input atmospheric variables reduces confidence in the results. Unfortunately, there is a dearth of other suitably constrained observational datasets (such as water vapour, cloud cover, rainfall, or sea level pressure changes for example) that could be used in detection studies. Even if these did exist, then low SNRs may mean that they provide little extra useful information for detection purposes (Santer et al., 1994). Hence in this thesis, an attempt was made to ameliorate the situation by considering full spatial field near-surface and upper air temperatures under a common detection approach.

The HadCRUTv dataset sampling technique has been relatively stable, at least over the last 50 years, and rigorous quality control tests (Jones et al., 1999, 2001) have been applied. Both long-term homogeneity and error assessments for potential biases have been made (Jones et al., 2001, Folland et al., 2001). Hence confidence in the homogeneity of this series is high. In contrast, the HadRT upper air temperature record is based upon radiosonde ascents from 1958 to date. There have been numerous changes in instrumentation, calculation methods, ascent times, and launch location, amongst others, over time for the majority of stations within the radiosonde network. All such changes could result in inhomogeneities within the HadRT gridded upper air temperature product (Parker and Cox, 1995, Eskridge et al., 1995, Gaffen et al, 2000a for example). Two versions of the HadRT dataset were used in this thesis: HadRT2.1 and HadRT2.1s. Both versions have been corrected globally for known inhomogeneities post-1979 with reference to the MSUc satellite temperature record

(Christy et al., 1998). These corrections solely consider the temporal consistency of individual grid-box records, with no attempt to maintain any measure of spatial consistency (Parker et al., 1997). The difference between the datasets is that HadRT2.1s has not had any corrections applied within the troposphere.

Given that the HadRT record is likely to contain residual errors, chapter 2 investigated the spatial consistency of both versions. Statistical checks were instigated, using near-neighbour comparisons, to determine those grid boxes within well-sampled regions in each version of the HadRT dataset that are potentially dubious. These grid-box records were compared to available station metadata (Gaffen, 1996), and where a physically plausible reason existed for the suspected inhomogeneities, the grid-box series were deleted. The resulting datasets contain a few percent less datapoints, but this should be balanced against the fact that they are more spatially consistent. A simple qualitative analysis of the spatially quality controlled HadRT datasets showed that they should be of suitable quality for use in detection studies. Results of this analysis were found to be potentially sensitive to inclusion criteria in the calculation of annual and seasonal averages. The use of three versions of each HadRT dataset each with increasingly strict temporal grid-box value inclusion criteria for the calculation of annual mean values was advocated. The potential for further non-negligible residual errors in any of these HadRT dataset series is not discounted. So long as residual errors are primarily random and not systematic in nature it should not have a first-order effect upon subsequent analyses. A consideration of both HadRT2.1s and 2.1 should provide at least a weak indication as to the stability of results to at least some of the likely uncertainties in the observed record.

Before advancing to a formal quantitative detection exercise, in chapter 3 the HadRT dataset fields were compared to HadCM2 and HadCM3. There exist sufficient differences between both the models, and the forcings applied, to treat them as being independent. The use of truly independent model results, insofar as any GCMs can be, would increase confidence in the results of such simple inter-comparisons, and

more formal quantitative detection studies. Considering increasingly complex input fields, qualitative evidence from both HadCM2 and HadCM3 indicated the necessity for anthropogenic influences to be included to adequately explain recently observed upper air temperature changes. Simple statistical indicators of grid-box and full spatial field similarity were also employed to gain a more quantitative measure of model 'skill'. The significance of any discrepancies was assessed by using a section of HadCM3 control run to estimate the likely natural internal climate variability, with results treated conservatively. It was shown that even using such simple statistical indicators, some evidence exists for a demonstrable anthropogenic influence in both models. There also exists much weaker evidence for both solar and volcanic influences. No claims of unambiguous detection or attribution were made based upon this analysis due to the simplicity of the statistical indicators considered. Demonstrable differences existed in results between the models for equivalent forcings. Therefore, it was stressed, in agreement with previous analyses (Hegerl et al., 2000, Barnett et al., 2000), that reliance upon output from a single model will likely yield ambiguous conclusions in detection studies.

The optimal regression detection methodologies used were formally introduced in chapter 4 and, where possible, related to other climate change detection approaches. The basic premise of the optimal regression approach is that the observations are made up of a linear combination of the forcing response signals and an additional noise term (AT99, AS01). Both observational and model fields are optimised by rotating the fields such that they focus on those regions of phase space where the signal is dominant over noise due to natural internal climate variability. Output from the regression can most easily be visualised as a cloud of plausible solution points. These can be explicitly mapped as a Probability Density Function (PDF), of the same dimensionality as the number of input model fields. If the uncertainty range is entirely positive in any signal direction at a given confidence limit (typically 90%) then the signal is detected. Further, if the range contains unity, then it is also a consistent explanation of the observations. Two versions of the detection regression algorithm approach have been used here: Ordinary Least Squares (OLS) (AT99), and

Total Least Squares (TLS) (AS01). The underlying difference between being that TLS explicitly accounts for uncertainty in the input model signals, whereas OLS solely makes an *ad hoc* scaling factor correction to account for this uncertainty. Under both approaches, the residuals are compared to an independent section of control to allow for a consistency test.

Results of two previous zonal-mean difference field (1985-1995 minus 1960-1980) OLS detection studies, considering HadCM2 and HadCM3 model fields (AT99, Tett et al., 2001), were revisited in chapter 5. The sensitivity of these previously published results was considered to numerous input field pre-processing choices, and whether noise in the signals was implicitly accounted for. Exact results were seen to be especially sensitive to how the input fields were weighted prior to input, with those diagnostics giving higher weighting to the troposphere (the method used previously) most likely to yield consistent residuals. This systematic behaviour most likely reflects gross model inadequacies within the stratosphere (Gillett et al., 2000, Collins et al., 2001), along with increasing observational error at these heights (Gaffen et al., 2000a). Other pre-processing choices led to less distinct effects on the results, although some systematic differences remained. Of particular interest was a systematic increase in residual test failure when HadRT2.1 was used as the observed dataset. The effects of all pre-processing choices were almost entirely in whether the residuals were consistent and, if so, whether any signals were detected. Therefore, previous zonal-mean detection studies can only ever have been conservative. For both HadCM2 and HadCM3, the most likely cause of recently observed zonal-mean upper air temperature changes was found to be a well-mixed greenhouse-gases and sulphate aerosols (GS) signal. There was also evidence, albeit weaker, of combined natural (solar and volcanic) external, and stratospheric ozone depletion signals. Under a TLS approach, natural forcings were systematically more likely to be detected than under an OLS approach. Importantly, the presence of the temperature trend differential across the tropopause is not required for the successful detection of an anthropogenic (GS) influence in either model, although a stratospheric ozone depletion signal becomes much less detectable. This refutes previous criticisms that

the detection of an anthropogenic influence in previous zonal-mean detection studies is critically dependent upon this divergence in trends (Legates and Davies, 1997).

In chapter 6 a number of full spatial field (longitude, latitude) annually averaged tropospheric temperature variables were considered under a common space-time optimal detection approach for the period 1960-1995. Input was Large Area Averages (LAAs) for non-overlapping five year periods. The troposphere is well mixed on the timescales considered and, hence, detection results should be consistent between these variables if the models are adequate explanations of recently observed tropospheric climate change. Numerous sensitivity studies were performed. Four LAA choices were considered, two of which were optimised with regard to the data coverage. Furthermore, results were considered over a range of truncations to check that they were not sensitive to this choice. As a further check, global-mean and more complex spatio-temporal reconstructions were considered, based upon the OLS regression results, to ensure that the results were not grossly inaccurate in representing the observed changes. Results for both models indicated a detectable anthropogenic greenhouse-gases influence with high confidence, and sulphate aerosols influence with medium confidence. Limited evidence existed for a detectable volcanic forcing influence. Solar and stratospheric ozone depletion signals were found not to be a detectable influence upon tropospheric temperatures. Results for each signal differed according to model, tropospheric temperature variable, input pre-formatting, and truncation. Such differences are expected to occur by chance. Whether any of the differences were significant was not explicitly tested in any quantitative manner. Any seemingly gross differences were noted, and potential reasons discussed. When HadRT2.1 was used as the observed dataset, there was in most cases an increase in frequency of failure of the consistency test on the residuals. Based upon this and results from chapter 5 it was concluded that this most likely related to at least some of the corrections applied to HadRT2.1 being unrealistic.

Further consideration was given solely to the preferred LAA input diagnostic. Free troposphere layer average temperature results from both models for anthropogenic

forcings were seen to be approximately correct over a range of truncations. Estimates for near-surface temperatures and lapse rate diagnostics, however, indicated that the modelled anthropogenic forcing response is overestimated in magnitude in both models. Analysis of TLS results showed that the discrepancy in the results was highly unlikely to be due to known biases in the OLS algorithm (AS01). Variations in the OLS estimators for different tropospheric temperature variables might reasonably be expected to occur by chance due to natural internal climate variability, and other uncertainties (finite model, observational errors etc.) alone. Simple qualitative analysis of the OLS results for both models indicated that this might be the sole cause of the observed discrepancy. In chapter 7, a potential methodological framework to gain a quantitative model consistency statistic was discussed. If it were possible to prove a significant discrepancy between the individual temperature variable estimators, then it might prove difficult to apportion the source to model or observational error, without having recourse to a greater number (order of magnitude) of observed and modelled datasets. If the discrepancy were to arise due to model error, then both models overestimate near-surface changes. Conversely if the source were observational error, then the observations underestimate near-surface temperature changes in response to anthropogenic forcings. This latter result would be in marked contrast to previous studies attempting to explain observed discrepancies between the lower troposphere and the near-surface (NRC, 2000, Santer et al., 2000, 2001).

8.1.1 Principal conclusions

- Detection results for both HadCM2 and HadCM3 indicate that the most likely causes of recently observed near-surface and upper air temperature trends are anthropogenic forcings. There is more limited evidence for detectable volcanic forcing influences over this period.
- Results are robust to the likely major potential sources of uncertainty under optimal detection approaches, for both zonal-mean and more complex spatio-temporal inputs.
- These principal results are in agreement with previous studies, providing increased confidence in the presence of a human influence on climate.
- Both HadCM2 and HadCM3 are found to be potentially adequate explanations of recently observed tropospheric temperature changes, although it was not possible to quantify the degree of adequacy.
- There is evidence that at least some of the corrections applied at the grid-box level to HadRT by Parker et al. (1997) are sub-optimal, most likely due to considering solely temporal rather than spatio-temporal consistency. Furthermore, there remain gross residual errors in this product, which were removed prior to the detection and attribution analyses detailed here. Future detection studies would benefit greatly from an improved radiosonde temperature dataset.

8.2 Remaining uncertainties and avenues for further research

As with all previous detection studies, **the results detailed in this thesis are critically dependent upon the adequacy of both the models and the observations**. This dependency is ameliorated by a consistency test on the residuals from the regression, which should identify gross inadequacies, as well as the suite of further sensitivity studies considered. This consistency test is, however, a fairly weak test as it solely considers the amplitude and not the shape of the residuals. Scope remains for the development of a more optimal test on the residuals of the regression, which can at least begin to account for their shape as well as their magnitude. Numerous additional quantitative and qualitative studies to those carried out in this thesis have indicated that the models considered are unlikely, at least for the purposes of this thesis, to be grossly inadequate. Although residual errors cannot be entirely ruled out in either of the observed temperature datasets used, they are likely to be primarily random rather than systematic in nature and, therefore, lead to conservative detection results. It is imperative that modelling improvements, model validation studies, and iterative improvements made to the observed datasets be continued, in order to further constrain residual uncertainties.

The use of two versions of the HadRT dataset means that at least some (but only some) of the uncertainty in the observed upper air temperatures has been explicitly addressed. Analysis in chapter 2 indicated that the HadRT temperature record is highly unlikely to be free of residual errors. Further, detection results from chapters 5 and 6 indicate that, although corrections applied by Parker et al., (1997) within the troposphere may have reduced the vertical errors within the dataset, they are likely to have been spatially sub-optimal. That is, the corrections look nothing like the leading modes of spatio-temporal variability in either GCM considered in this thesis. There remains scope for improving the methodology used in correcting the HadRT dataset for suspected inhomogeneities, such that both spatial and temporal consistency of the data are maintained in a more meaningful manner. This is probably best achieved through consideration of the individual station series rather than the gridded product.

Following the application of a quality control algorithm the HadRT product could then be regridded. Future detection studies would be helped by the development of such an improved gridded upper air temperature product. Uncertainties in the observed near-surface temperature series were not explicitly addressed, although alternative gridded products do exist and should be additionally used in future work. The magnitude of these uncertainties is unlikely to be as great as for upper air temperatures, at least at the large scales considered in detection, so this should not affect any of the primary conclusions of this thesis in a systematic manner.

Results detailed in this thesis arose from a consideration solely of a single suite of observational datasets, and only two versions of the Hadley Centre's GCM. Confidence in the results would be increased if the results could be shown to be insensitive to the use of both alternative observed datasets, and models (particularly those from other modelling centres). Further, this thesis has considered only one tropospheric climate parameter (temperatures), and from this inferred that the models are at least potentially internally consistent explanations of the recently observed tropospheric climate changes. To verify this result it would be desirable to repeat the analysis incorporating other additional tropospheric parameters. There may be problems in detecting signals under such approach, as other parameters are likely to have lower SNRs (Santer et al., 1994). However, so long as the signals were not significantly noise polluted, questions of internal model consistency and, therefore, attribution, could still be robustly addressed, even if larger uncertainty limits meant that the results did not directly aid detection. At least for now this is unlikely to be possible as suitable observed datasets do not exist and satellite and reanalysis products are too short and / or are likely to contain distinct errors and inhomogeneities. Efforts to construct suitable observed datasets for other (global) tropospheric parameters than temperature that can be usefully employed in detection studies should be a matter of priority for the atmospheric science community.

Expanding further, this thesis has only considered the atmospheric component of the climate system, and the potential internal model consistency of only the tropospheric

temperature component. There remains scope for future work to consider additional components of the climate system simultaneously (stratospheric and ocean temperatures for example) to yield a more stringent evaluation of the likely adequacy of the models. To date ocean temperatures have only just begun to be considered (Barnett et al., 2001, Levitus et al., 2001) in formal detection studies, and stratospheric temperature records suffer from poor observational data coverage. Therefore extending the analyses in this thesis to consider further sub-units or components of the climate system is highly unlikely to be a trivial exercise, at least in the near future.

Finally, in chapter 7 the possibility of using detection results to consider questions of model internal consistency quantitatively rather than qualitatively was addressed. If it were possible to construct a methodology that yielded unbiased estimators as to the true solution from individual component solutions then a whole suite of further applications of detection studies pertaining to the adequacy of models could result. More work into the feasibility and development of a methodology to attain such an unbiased estimate would be highly desirable.