1. Introduction

This thesis aims to advance our understanding of the most likely causes of recently observed changes in the climate system, primarily through a more detailed analysis of available upper air temperature records. It also advances from addressing simple “yes” or “no” questions of whether particular external forcing factors are important in explaining these changes to begin to consider the internal consistency of state-of-the-art climate models, at least within the troposphere. In section 1.1 the basic concepts of detection and attribution are presented, whilst sections 1.2 and 1.3 introduce the observed and modelled datasets used in this thesis. Section 1.4 provides a brief synopsis of the principal results from previous detection and attribution studies, methodological details being covered in chapter 4. In chapter 2 a quality control algorithm is applied to the HadRT upper air radiosonde temperature record (Parker et al., 1997). Before advancing to a formal quantitative detection study, an inter-comparison of modelled and observed datasets is undertaken in chapter 3. The sensitivity to potential sources of uncertainty of results of previous fixed-signal zonal-mean optimal detection studies is addressed in chapter 5. Chapter 6 advances previous detection studies to consider a number of tropospheric temperature variables under a common detection methodology and qualitatively assess whether the models are likely to be adequate (internally consistent) explanations of the observations. In chapter 7 a methodological framework is set out under which more formal, quantitative, statements regarding model adequacy could be achieved. Finally, in chapter 8 conclusions and avenues for future research are discussed.

1.1 The detection and attribution framework

Most scientists around the world believe that there is a discernible anthropogenic influence upon climate causing globally averaged near-surface and free troposphere temperatures to rise, among other changes to the climate system. If model projections of the evolution of this effect are correct, then global mean temperatures will increase further in coming decades, causing considerable net climate change – the anthropogenic greenhouse effect. For governments, society, and industry to seriously consider effective action to minimise and mitigate the effects of such a climate change, scientific evidence of a demonstrable anthropogenic cause is required. Climate change detection and attribution aims to provide this evidence.

The definitions of detection and attribution used here are derived from those discussed in chapter 8 of the 1995 IPCC (Intergovernmental Panel on Climate Change) Second Assessment Report (SAR) (Santer et al. 1996). These are the standard definitions used (albeit in various guises) in most subsequent climate change detection and attribution studies.

The detection problem is essentially one of finding a given model-calculated climate change signal in the observations. Such a signal will be embedded in noise due to natural internal climatic fluctuations, so the detection problem is far from simple. An estimation of the unforced internal climate variability (noise) is therefore essential to be able to claim significance. As instrumental records are relatively short and potentially influenced by anthropogenic (and other external forcing) effects (Jones and Hegerl, 1998), this must be derived from long control runs of GCM's (General Circulation Models), or simpler models.

A signal, or combination of signals, is detected if, and only if, the signal amplitude(s) in the observations is (are) greater than that expected by chance due to natural internal climate variability alone.

This statement indicates that the detection problem is a statistical one whereby one must reject the null hypothesis of natural climate variability explaining observed trends. The implications of a statistical Type I error (non-rejection) vis-à-vis that of a Type II error (rejection of a valid null) implies that detection studies should be looking for a high critical value acceptance of greater than 90% (Zwiers, 1999). This ensures that any claims of detection are likely to be conservative, and that we will not ambiguously claim the presence of an anthropogenic, or other external forcing influence, when none actually exists. If the model estimate of natural internal variability is grossly inadequate then problems are likely to arise in making unambiguous statements of detection.

Even once a signal has been detected, it does not directly imply cause and effect; this is the domain of attribution. Attribution is harder to attempt rigorously as, in the strictest sense, an infinite series of possible climate forcings and their combinations must be considered and rejected until only one remains. In practice, attribution is limited to a sensible number of physically plausible explanations of the observed climate change.

The observed climate change is attributed to a given signal (or combination of signals) if it is consistent with this (these), and only this (these), signal(s) from the entire population of plausible signals.

This is a test equivalent to non-rejection of the null hypothesis of consistency between the signal(s) and the observations. It should be noted that, in the frequentist statistical sphere of classical detection literature non-rejection of the null hypothesis at level P% does not imply acceptance of the alternative hypothesis at level (100 - P)%, due to the asymmetry of the statistical system (Levine and Berliner, 1999). This means that any claims of statistical significance levels in attribution studies are likely to be overly optimistic, the uncertainty limits being underestimated. Levine and Berliner (1999) propose an alternative approach, from which realistic attribution confidence statements can be made, whereby the statistical system is symmetrical. Allen et al. (2001) take a different approach to the attribution problem, where attribution is claimed for the most parsimonious combination of model signals that are found to be statistically consistent with the observations. However, regardless of methodological considerations, attribution does not rely solely upon significance levels, but rather on the rejection of all other plausible causes until only one remains. So as long as the uncertainty limits are sufficiently separated between competing signal combinations, attribution can be claimed. As for detection, statements of attribution are critically dependent upon the adequacy of the model-derived estimate of natural internal variability.

1.2 Observational datasets

The basic requirement for a successful outcome of any climate change detection and attribution study must be a set of suitable climate observations. Without such a resource there would be little chance of stringently confirming and, possibly, constraining model predictions of climate change. Observational climate datasets can in theory be observations of any parameter, or set of parameters, which measure the time varying behaviour of the atmosphere. This yields a potentially large set of observational datasets that may be useful in detection and attribution studies. Obviously objective selection criteria are required to narrow down this field.

The most important criterion is that the searched for signal in any parameter should be as near as can be achieved to orthogonal to the major axis of natural climate variation in the phase space of the treated dataset. This maximises the chances of any signal being detected, as the signal will be in a direction where it is dominant over natural internal climate variability. Santer et al. (1994) characterise this in terms of an analysis of Signal-to-Noise Ratios (SNRs) where 'noise' is derived from long control runs of the same GCM's as the signals. Ideally the noise estimate should include everything other than the signal(s) being considered. The results are likely to be critically model dependent. Three aspects of the signal are generally considered: the overall magnitude, orthogonality properties, and the trend significance. The full model field cannot be considered as current model control runs (with no changes in external forcings) are not run for a long enough time to gain sufficient statistical degrees of freedom to be able to consider the full spatial field on the multi-decadal timescales important in detection studies. This would require control runs of the order of at least several thousand years. Hence the statistics are realised in some form of truncated Empirical Orthoganal Function (EOF) space. Various criteria such as pattern correlation between signal and noise, number of EOF's required to explain P-critical (%) of the variance, or the time evolution of the signal to noise ratios could be used (Santer et al. 1994). In practice it is safest to use all such criteria. Previous studies (Santer et al. 1991, Santer et al. 1994, Santer et al. 1996) suggest that, of the commonly observed atmospheric variables, surface temperature and vertical temperature are both useful whereas precipitation and surface pressure are less so due to their lower SNRs.

Observations also need to exist for a continuous period, long enough for any underlying trend to be discernible from that due solely to natural internal climate variations. This critical trend length for successful detection studies considering near-surface temperatures has been found by various authors (see Santer et al. 1995,1996, Hegerl et al. 1996, Barnett et al., 1998, for example) to be at least 30 years. Applying a 30-year or longer trend length requirement dramatically reduces the set of suitable candidate variables. Furthermore, numerous authors (Santer et al. 1995, Hegerl et al. 1996,1997, Barnett et al. 1998) have illustrated that the longer the trend length under consideration, the greater the chance of a successful detection exercise. This is because, assuming natural internal variability is a stationary normal distribution, natural trends due to internal variability in any detection statistics will tend towards zero over long periods. It has also been noted that in certain detection algorithms using a short trend length, spurious claims of detection can be made by chance, whereby the statistic after a short further period of time is no longer significant (Barnett et al. 1998). This is to be expected to occur by chance due to the nature of the statistical analysis, whereby a given trend is said to be anomalous at the P% level when compared to natural variations. A more stringent criteria is suggested as a result by Barnett et al. (1998) whereby detection is claimed only when a statistical measure becomes significant and remains significant for a long period (order decades).

A further constraint is that the observations must be homogeneous through time. If the observation time, method, location, or adjustments applied to observed series for real-time operational purposes have changed through time then this will result in potential inhomogeneities within the series (Parker and Cox, 1995, Gaffen, 1994, Eskridge et al., 1995, Jones et al., 1999 and references therein). Homogeneity can be assessed using available station meta-data, and by various automated statistical techniques to check for outliers and break points in a series, and then adjusted to create a homogeneous series (see Jones et al., 1999, and references therein for a more complete discussion on this point with regards to the surface temperature record, or Eskridge et al., 1995 for upper air temperatures). However, the potential for further residual inhomogeneities (e.g. urbanisation influences for surface data) remaining within any homogenised series should not be discounted. There are also known problems with removing suspected inhomogeneities in that they may in fact be real rather than 'apparent' changes as climate change could naturally occur in a step-like fashion (Gaffen et al., 2000a). Another complication in using fully automated checks is that the resulting series will be critically dependent upon the criteria used in identifying and correcting for break points (Gaffen et al., 2000a).

The observed variable being considered must be measured at discrete points across the globe. A parameter such as global mean temperature is of limited use in any study of detection, and even more so in attribution because similar changes in such simple (univariate) indicators could be caused by a variety of different combinations of forcing factors and internal variability. Studies using such univariate diagnostics have been referred to as Type I studies in the IPCC SAR (Santer et al. 1996). They can almost invariably be fitted to any given forcing by fine tuning and introducing terms into the statistical transfer function which transforms the forcing (e.g a history of solar activity) to the observed diagnostic (see for example Tol and De Vos, 1998 who use global mean temperature in a Bayesian approach). This is obviously of dubious scientific value and, even if physically plausible terms derived from state-of-the-art climate models are used in the transfer function, it is unlikely to bring us closer to either unambiguous detection or attribution. Therefore the variable needs to be measured at many points to enable a spatial or spatio-temporal (multivariate) approach to detection to be employed. It is convenient to grid the variable in a manner which avoids having to employ complex comparison techniques (such as neural networks) with the gridded model output. The gridding can be a simple box average (e.g. Jones et al., 1999, 2001 for surface data), or a more complex approach whereby weighting of individual stations is attempted based upon grid-box representativeness. Parker et al. (1997), for upper air data, use an inverse distance weighting from the centre of the grid-box.

It is useful if the reporting network is of high density as this reduces the standard error due to sampling (Jones et al. 1997a, 1999, 2001). Expectations are that a single station reading will have a higher variance than an average of many regional station readings. The standard error will be critically dependent upon the station variance and inter-record correlation within a gridbox (Jones et al. 1997a). This sampling error is an extra source of observational uncertainty in any detection study, although Jones et al. (1999, 2001) show that it is not the major source of uncertainty, at least in the near-surface temperature record. The effects of sampling error can also in theory be corrected for, at least in the near-surface temperature record to yield a variance corrected version (Jones et al., 2001). This does not necessarily imply that such corrections will be possible in all other observed climate series.

Most recent detection studies look for some form of diagnostic of increasing pattern similarity between the observed variable and a modelled response(s) to external forcing, either natural or anthropogenic (Tett et al., 1999, 2001, Stott et al., 2001, Hegerl et al., 1996, 1997, 2000 Santer et al., 1995, for example). Therefore it is reasonable to expect that the spatial coverage coherency of the observational field will affect the chances of a successful outcome. A spatially coherent coverage field of observations will enable more accurate pattern estimation in those areas which are well-observed. However, this is potentially at the cost of representing patterns or values elsewhere in the global field, which a more scattered observational field may better pick out. To date little attention has been paid to this criterion, although model data (for HadCM2) have been shown to be generally representative only at the largest scales for near-surface temperatures (Stott and Tett, 1998).

It is obvious from the above that a spatially coherent variable with a large correlation decay length would be preferable. Osborn (1997) shows that for precipitation, which has a small correlation decay length, both point and areal values can be very different and respond in divergent manners to the same forcing, both in terms of averages and extremes. There is no reason to suspect that this is not also true, if less markedly so, for other variables. Therefore care must be taken in ensuring that one is comparing truly grid-box (rather than point) representative observations with the gridded GCM output.

Taken together these criteria lead to an almost empty set of observational datasets which can be used in detection studies. Realistically one is limited to using near-surface temperatures (Jones et al. 2001) and the radiosonde temperature record (Parker et al. 1997). Only these records are of the necessary length (greater than 30 years), are well enough sampled and constrained, and exhibit relatively high SNRs. However, there are still residual uncertainties in both datasets, and specifically in recent temperature trend differences between the lower troposphere and the near-surface (NRC, 2000). In section 1.2.1 the construction of the near-surface temperature record is described, section 1.2.2 summarises the radiosonde record, and section 1.2.3 discusses recent work towards reconciling the most recent 20-year trend discrepancy between the two datasets.

1.2.1. The HadCRUTv gridded surface temperature record

The HadCRUTv surface temperature record is well documented (see for example Jones, 1994, 1995, Jones et al., 1997a, 1999, 2001 and references therein). The values are composed of station temperature anomalies over land and sea surface temperature (SST) anomalies over the ocean. All monthly station data have been quality controlled, and corrected for known / suspected inhomogeneities before being added to the database. Grid box values are simply an average of all station anomaly values within the box, with no attempt being made to account for distance from the centre of the grid-box (Jones, 1994). This is the major temperature dataset used in those detection studies reported in both the IPCC SAR (Santer et al. 1996) and TAR (Third Assessment Report) (Mitchell and Karoly et al., 2001), and many other detection studies to date (Allen et al., 2000, 2001, Stott et al., 2001, Tett et al., 1999, 2001, Hegerl et al., 1996, 1997, Santer et al. 1995, for example). The HadCRUTv dataset in its latest version (Jones et al., 2001) exists as a 5° longitude x 5° latitude globally gridded, variance corrected, set of grid-box monthly anomalies, from the 1961-90 average, from the middle of the 19th Century to the present day. The global average temperature record is shown in Figure 1.1, and exhibits a long term (although by no means linear) warming trend in the series over this period.

The areal coverage of HadCRUTv decreases before the 1950's and again before the early 1900's (Santer et al., 1995), and slightly in recent times due to delays in data availability in near real time (Jones et al., 1999). Jones et al. (1997a) show through a frozen-grid analysis that the higher variability in global-mean near-surface temperature during the 19th century is primarily due to the sparser data coverage at that time, at least over land. Data variability is also greater when fewer stations contribute to a grid-box value due to an increase in the sampling standard error (see Jones et al., 1999 and earlier discussion), as is the case in some regions before 1951. A variance-corrected version of these data (Jones et al., 2001) is used in this thesis. HadCRUTv accounts for the effect of the changing density of reporting stations both over land and in ocean regions (Jones et al., 2001). The adjustments have no effect on data coverage, and the hemispheric average series are not significantly changed, although there are slight changes in the time series in some cases at continental and regional scales. Further, Jones et al. (2001) comment upon how variance corrections will yield more stable EOF's, an important consideration in detection studies using the optimal regression methodology of Allen and Tett (1999) and Allen and Stott (2001) (see chapter 4).

Some criticisms have been made of the near-surface temperature record over time (see NRC, 2000 for a summary), casting doubt upon whether the observed century scale global-mean warming has actually occurred. Jones et al. (1999), Peterson et al. (1999), and NRC (2000) dismiss the commonly-argued urban heat island effect as a plausible explanation, estimating its maximum potential contribution to be, at most, an order of magnitude smaller than the observed hemispheric-scale surface warming in the last century. Residual inhomogeneities are harder to constrain, but gross errors (both systematic and random) are routinely corrected for or removed by the quality control criteria and statistical tests. Folland et al. (2001) address a number of previously identified potential sources of uncertainty, including previously implemented corrections to the dataset, and find that trends remain highly significant since 1861. More subtle inhomogeneities could explain some residual trends but would have to be similarly positively biased over large regions to account for the observed changes. In all likelihood, individual erroneous grid-box trends will exhibit essentially random biases. Physical mechanisms for a (global) systematic positive bias to explain the observed trend are hard to imagine and, therefore, it is concluded that the observed trend is likely to be real and not a sampling artefact. Further credence is given to the validity of the temperature record of Jones et al. (2001) as two other independently produced surface temperature datasets broadly agree with the trends on global and regional scales (Quayle et al., 1999, Hansen et al., 1999, see also the discussion in NRC, 2000). However, care should be taken not to read too much into this as much of the data are common to all three near-surface temperature datasets, with the differences being primarily in treatment and assimilation of the raw data, rather than the raw data themselves. Totally independent reconstructions from borehole temperature measurements confirm the warming trend (Huang et al., 2000), although uncertainties remain over the effects of land-use changes during this period on these borehole temperature series (Mann, 2001).

Measurement errors (other than gross errors), which are by nature assumed to be random both in space and time, have not been implicitly taken into account by Jones et al. (2001). Hegerl et al. (2001) attempt to quantify the likely effect on detection studies using a Monte-Carlo approach to randomly seed errors in realisations of the gridded near-surface temperature record. It is assumed that the error is WHITE in nature (independent identically distributed noise). Hegerl et al. (2001) consider both spatially correlated and uncorrelated errors. Spatially correlated errors between grid boxes is an ultra-conservative treatment, as this would require spatially consistent errors, which given their suspected random nature is highly unlikely. A spatially systematic bias could occur due to coincident instrument or observing practice changes over large regions (Gaffen et al., 2000a), but at the surface the procedures have remained relatively stable over time (at least compared to upper air temperatures (1.2.2)) making this highly unlikely. Even taking this conservative approach the maximum possible trend in global mean near-surface temperatures gained by Hegerl et al. (2001) is only half that of the observed trend. This does not imply necessarily that such errors should be ignored, especially at smaller scales, but for the purposes of detection and attribution studies, which tend to concentrate on the largest temporal and spatial scales, these errors are not first order in HadCRUTv, and can therefore be discounted.

1.2.2. The HadRT radiosonde temperature record

The HadRT radiosonde temperature dataset consists of a 10° longitude x 5° latitude gridded temperature record on 9 standard WMO reporting levels throughout the troposphere and lower stratosphere (850, 700, 500, 300, 200, 150, 100, 50, and 30 hPa levels) from 1958 to date (Parker et al., 1997). Values are quoted as temperature anomalies in degrees Celsius from the 1971-90 average. The versions used in this thesis are HadRT 2.1 and 2.1s. These records have been corrected globally for known post 1979 inhomogeneities by reference to collocated data from the MSUc record (Christy et al., 1998, 2000 and references therein) in the same manner as version HadRT1.1 was over Oceania (Parker et al., 1997). Climatologies are recalculated after corrections have been applied. In version 2.1s corrections are made solely in the stratosphere, whereas in version 2.1 they are made throughout the depth of the atmosphere.

Parker et al. (1997) note that raw radiosonde temperature data used in the construction of the HadRT dataset have been widely affected by both random and systematic errors. The radiosonde recording network has been designed and implemented with operational meteorology rather than consistent climatological recording in mind. Gaffen (1994, 1996) has attempted to collate a history of metadata for individual stations, including changes in instrumentation, observation times, corrections applied, station location, and balloon type, amongst others. This work has been augmented by a number of other investigators (see for example Eskridge et al., 1995, Parker and Cox, 1995, Luers and Eskridge, 1998). It is only with such a complete station history, as well as an inter-comparison of instrumentation, that any small amplitude systematic temporal and spatial errors can be corrected. Parker et al. (1997) correct for known inhomogeneities based upon available station metadata (Gaffen, 1996) where a significant difference can be found in the individual station series using a student's t-test statistic. This reduces systematic errors in individual grid-box series, but is by no means ideal. Corrections applied to the HadRT temperature record are seasonally invariant. A more complex database is currently under development, which should reduce further any residual systematic biases present in the HadRT radiosonde record (Eskridge et al. 1995, CARDS webpage, GUAN webpage).

In chapter 2 of this thesis, the previous analysis of Parker et al. (1997) is augmented by extending quality control criteria to an analysis based upon a comparison with neighbouring grid boxes on both annual and seasonal timescales. Two simple statistical measures are used and are treated conservatively to minimise the chances of disposing of valid data. Such an exercise is important if, as is the case here, one is searching for increasing pattern similarity between model scenarios and the observations. The HadRT temperature record also exhibits much sparser spatial coverage than that of the HadCRUTv near-surface temperature record and is highly biased towards Northern Hemisphere land areas. A HadRT dataset coverage map for the 850 hPa level, providing an illustration of tropospheric coverage characteristics, is given in Figure 2.11. It should be noted that coverage also decreases rapidly within the lower stratosphere, especially during the early part of the record, due primarily to radiosonde balloon burst. This is discussed in further detail in Chapter 2.

Santer et al. (1999) explicitly recommend that more than one observed dataset should be used in any detection exercise in the free troposphere. Available records arise from radiosondes, MSU satellite records (Christy, 1995, Christy et al., 1998, 2000), and operational reanalyses. However, Barnett et al. (1999), Santer et al. (1999) and Stendel et al. (2000), for example, state that the current generation of reanalysis products (NCEP, Kalnay et al., 1996; and ECMWF, Gibson et al., 1997) are not of sufficient quality for such purposes. Further, the MSU satellite record (Christy et al., 2000) is not likely to be of a sufficient length (Barnett et al., 1998, 1999) and has remaining attendant uncertainties and problems due to being a depth-integrated quantity rather than a point measurement (NRC, 2000). There is no alternative gridded radiosonde temperature dataset available at this time. Therefore in this thesis attention is limited solely to the HadRT record, although the use of additional observed products in future work would be highly desirable.

1.2.3 Reconciling recent trend differences between near-surface and radiosonde temperature records

It has been shown that over approximately the last two decades (1979-2001) near-surface and lower tropospheric temperature trends (principally MSU2LT, centred around 740 hPa, Christy et al., 2000) have not been exactly coincident on a global average basis. The near-surface record has exhibited net warming, but little, if any, warming has been observed in the lower troposphere (Santer et al. 2000, 2000a; NRC, 2000 amongst others). This discrepancy leads some to question whether recently observed near-surface temperature trends are in error. The use of short-term trends can be wholly misleading. Analyses using longer periods (1958-99 and 1964-99) show agreement between the near-surface and (radiosonde based) lower tropospheric temperature trends (Angell, 2000, Brown et al., 2000, Gaffen et al., 2000, and Jones et al., 1997). If the most recent disparity is important then it must be right to consider not only the last two decades but also the longer 40 year record, as the differences must have been of opposite sign in these first two decades of the longer period. The recent NRC report "Reconciling observations of global temperature change" (2000) debates the most recent trend discrepancies in terms of the three major temperature datasets available and their uncertainties. In the previous two sub-sections residual uncertainties in each of the datasets used in this thesis have been discussed. In this sub-section potential reasons for the recently observed discrepancies are discussed.

The MSU series of Christy et al. (2000) is used in inter-comparisons between the near-surface and lower tropospheric temperature records. The MSU record, which exists from 1979 to date, uses satellite records of up-welling microwave radiation from oxygen molecules (the wavelength of which is temperature dependent) to make layer average temperature estimates. Three layer average series are derived (MSU4, MSU2, and MSU2LT), each of which gains its peak information from a different portion of the atmosphere (stratosphere, upper troposphere, and lower troposphere respectively). For the purposes of the NRC (2000) study, MSU2LT is considered which has the peak signal from ~740mb, but contains information from the surface up to ~300mb. Residual uncertainties remain relating to orbital effects (Wentz and Schabbel, 1998), instrumentation drift, and continuity between satellite platforms (NRC, 2000). There is also only the one dataset (albeit in numerous versions) and therefore confidence in the trends is reduced as they may be susceptible to the algorithms used in converting the radiation measurements to layer-average temperatures (NRC, 2000).

Global average temperature series hide many differences in sampling distribution in both space and time as well as changes in instrumentation, corrections applied and other changes. Therefore, the use of a global mean diagnostic for comparison purposes may yield entirely spurious results. Spatially masking the MSU temperature record to that of the observational radiosonde record has an effect, both on a year-to-year basis and also on decadal trends, tending to increase the observed MSU 2LT warming (Christy et al. 2000, Santer et al. 1999). Masking also has an effect upon the correlation with the independent radiosonde record, degrading the agreement with the Angell (1988) record, which has been used previously to validate the MSU series (Santer et al. 1999, 2000). Similarly, masking the HadCRUT (an earlier version of the near-surface temperature record to that used in this thesis) (Jones et al., 1999) observations to that of the HadRT series increases the agreement of HadCRUT with both the HadRT and the MSU lower tropospheric temperature series (Santer et al. 2000). Furthermore, the choice of trend fitting method can greatly affect the calculated global mean trends, at least within the lower troposphere (Santer et al., 2000a).

The correlation between globally averaged values of the HadRT temperature series and the MSU record is susceptible to the method used in interpolating the level-specific radiosonde temperatures to layer-average MSU equivalent temperatures (Santer et al., 1999, Hurrell et al., 2000). The MSU 2LT series also has approximately 20% of its signal derived from surface microwave emissions which have no equivalent in the available HadRT radiosonde temperatures, and therefore will add 'noise' in any comparison if the surface radiance characteristics change over time due to natural or anthropogenic processes. Hurrell and Trenbeth (1997, 1998) argue that, for this reason, the MSU2 record (which does not incorporate a surface component, but gains its peak signal from higher in the troposphere (c. 400 hPa) and includes a stratospheric component) is more stable.

Even subsequent to masking datasets and taking into account uncertainties due to procedures used in the treatment of the data (Santer et al., 1999), discrepancies between the available observed near-surface and lower tropospheric temperature series remain. These discrepancies, or at least some of them, are almost certainly real (NRC, 2000, Chase et al., 2000). Expectations are that, over all time and space scales, near-surface and lower tropospheric temperatures need not perfectly co-vary. Sections of control runs from GCM's can be used to provide independent estimates of the natural variation, although with the obvious caveat that they are assumed to be a realistic representation of the real world. Santer et al. (2000), using such an approach, show that although natural internal variability may go some way towards explaining the observed discrepancy it cannot explain all of it, unless the natural variability is significantly underestimated in magnitude within the models. This approach neglects the possibility that the trend discrepancy may be a transient response to a combination of anthropogenic and natural external climate forcings as well as internal climate variability. Removing the effects of stratospheric ozone, ENSO and volcanic eruptions has been shown to bring the records into much closer agreement (Santer et al., 2001), as well as improve the model agreement with the observations (Bengtsson et al., 1999). To date the effects of anthropogenic forcings have not been rigorously assessed. Recent trends at the surface have tended to be dominated by increasing minimum temperatures, and at night, particularly in mid to high latitude wintertime, the lower troposphere tends to be more effectively de-coupled from the surface (Hurrell et al., 2000). Finally, as stated previously, the opposite trend occurs over the period 1958-78 with lower troposphere temperatures warming faster than surface temperatures (Gaffen et al., 2000, Jones et al., 1997, Brown et al., 2000, amongst others).

1.3 The Hadley Centre General Circulation Models

The second requirement for a successful detection and attribution study is a realistic simulation both of the expected transient climate change to a number of candidate forcings, and of natural climate variability. Unfortunately, we have not been studying the atmosphere for long enough, nor in enough detail, to have an adequate realisation of its natural variability characteristics on the decadal-to-centennial timescales which are important in climate change detection and attribution studies. Even if they existed, such observations would contain information pertaining to the effects of changing external (natural and anthropogenic) forcings, although some of these might be estimated and removed (Jones and Hegerl, 1998). However, following such a methodology it would be difficult to attach rigorous confidence estimates on the variability within any corrected series being due to natural internal variability alone. Nor are we able to perform repeat experimentation under numerous forcings upon the real world system. Therefore, we must rely upon numerical simulations of the climate using GCM's (General Circulation Models). This thesis uses two versions of the Hadley Centre GCM: HadCM2 (Mitchell et al. 1995, Johns et al. 1997) and HadCM3 (Pope et al., 2000, Gordon et al., 2000).

It is not the desired purpose of this thesis to go into extensive detail on the development and physics of climate models. However, a certain amount of information on models and their limitations is required to understand the implications of their use in detection and attribution studies. GCMs aim to produce physically based estimates of the effects of perturbations on the transient climate state (Johns et al. 1997). Due to computational constraints, models are run in some form of coarse resolution three-dimensional (latitude, longitude, height) grid co-ordinate system. The primitive equations are solved on this grid co-ordinate system by numerical finite difference approximations for both the oceanic and atmospheric components and the models then forward-stepped in time. Models generally consider land surface, hydrological and cryospheric processes in addition to atmospheric and oceanic components in an attempt to yield an approximation to the real world. They contain hugely simplified orography and land/sea data masks. Sub-gridbox scale processes are parameterised within the models. To date, little attention has been given to the sensitivity of models to uncertainty in these parameterisations. Given the likely uncertainties inherent in using such a modelling approach, detection and attribution studies effectively degenerate to being a model validation exercise where the question is: “Is the model an adequate representation of the recently-observed climate?”. It is likely that, at the grid-scale and time-step resolutions, variability will be under-estimated as models are using finite linear approximations to predict an infinite non-linear system (Allen and Tett, 1999). Stott and Tett (1998) have shown how only at the largest scales is HadCM2 near-surface temperature variability adequate, and this is likely to hold true for other models and variables, at least on theoretical grounds.

The Hadley Centre models are global, fully coupled GCM's with both atmosphere and ocean components having a resolution of 96 x 73 (3.75° x 2.5°) grid boxes in HadCM2, but a much finer (1.25° x 1.25°) ocean component in HadCM3. This is equivalent to a horizontal surface resolution of 3.75° longitude by 2.5° latitude in both models’ atmospheric components. They also each have a total of 19 layers in the atmosphere and 20 in the oceans. Seasonal flux corrections are applied at discrete spatial points to compensate for model drift into an unrealistic climatology for HadCM2, but no such corrections need to be applied to the HadCM3 model (Mitchell et al. 1995, Pope et al., 2000). Although such 'flux' corrections are undesirable they are felt to be preferable to model drift in HadCM2 (Johns et al., 1997). A long control run of over 1000 years with no changes in external forcings has been run for both models. Both these control integrations exhibit essentially no long-term global mean trends in almost all diagnostics (Johns et al., 1997, Pope et al., 2000, Gordon et al., 2000). However, there is a known cold bias in the troposphere and warm bias in the stratosphere (Johns et al. 1997, Pope et al., 2000) for both models. HadCM2 also has a winter warm bias in the troposphere north of 50°N (Johns et al., 1997). This may be related to the recently observed positive Arctic Oscillation (AO) phase (Thompson and Wallace, 2000) which is not captured in the model (Gillett et al., 2000a) and is within the period of observations used in the Johns et al. (1997) validation exercise.

Both HadCM2 and HadCM3 have been used to develop a baseline climatology through the long control run segment; and to study the likely impacts of changing external forcings, both natural and anthropogenic, over time. The evolution of any single model run response to an external forcing will be a superposition of the true signal response and 'noise' due to internal model climate variability. In an attempt to better quantify the likely climatic response to any given forcing each model is run a number of times to form an ensemble-mean response to each external forcing. Initial conditions for each ensemble member are taken randomly from the model control run, thus ensuring that the individual ensemble members are consistent with the model. Resulting differences between ensemble members are then considered to be due to internal climate variability alone. By averaging over a number of ensemble members, the SNR is increased (see Allen and Tett, 1999). The resulting ensemble mean provides a best guess of the likely climatic response to the forcing under consideration. It is desirable for the ensemble population size to be large in order to maximise the SNR. At the limit of an infinite ensemble population the pure model forcing-response signal will be gained by this process. Unfortunately, the ensembles used here are (generally) only four-member ensembles and therefore noise is likely to constitute a significant proportion of any ensemble average (Allen and Stott, 2001). More information on the ensembles described here can be found in Mitchell et al. (1995), Mitchell and Johns (1997), Tett et al. (1999, 2001), and Stott et al. (2001). The ensembles are also briefly summarised in Table 1.1.

Natural external forcings considered in both models are variations in solar output and explosive volcanic activity. Solar radiation received at the top of the atmosphere has varied both due to sunspot activity and longer-term variations over the observed solar record. Two reconstructions of solar activity are used in HadCM2, those of Lean et al. (1995) (LBB), and Hoyt and Schatten (1993) and Willson (1997) (SOL). HadCM3 considers the forcing of Lean et al. (1995) in its ensemble labelled SOLAR. Both models parameterise this forcing by changing the value of the solar constant in the model (Tett et al., 2001). The major discrepancy between the two solar reconstructions used in HadCM2 over the 20th Century is that the middle 20th Century peak in activity, and thus forcing, is shifted by one solar sunspot cycle. In HadCM2 both solar activity ensembles start at the end of the 19th Century and run up until 1995. In addition each has a single ensemble member which extends further back (to 1700 for SOL and 1799 for LBB), from which the other three ensemble members are initiated in 1890 by perturbing the field with control data. Therefore, at least in the late 19th Century, these ensemble members may not be completely independent. HadCM3 SOLAR ensemble members start in the late 19th Century and continue until 1999.

When explosive volcanic eruptions occur they inject volcanic aerosols into the stratosphere, where they have a residence time typically of the order of a couple of years. Volcanic aerosols absorb and scatter incoming solar radiation, reducing radiation receipt at the surface, and thus leading to a net tropospheric cooling and stratospheric warming. In the models this volcanic forcing is derived from the dust veil index (Sato et al. 1993) and input, as zonal concentrations in four discrete bands, in an attempt to replicate the observed zonal volcanic aerosol concentration distribution, assuming a uniform mass mixing ratio above the model tropopause (Tett et al., 2001). The zonal distribution is critically dependent upon the latitude at which the eruption took place, with tropical eruptions having the greatest global coverage effect. For both HadCM2 and HadCM3 the volcanic ensemble members (VOL and

VOLCANIC respectively) are initialised in the late 19th Century and continue up until 1997 and 1999 respectively.

In addition to these individual natural external forcing ensembles, HadCM3 has an ensemble of runs which incorporates both solar (Lean et al. 1995) and volcanic (Sato et al. 1993) changes simultaneously. This ensemble is started in the late 19th Century continuing up until 1999 and is given the label NATURAL.

Both solar and volcanic forcing histories are only really well observed and constrained for the last 20 years. Prior to this they are based upon either evidence from proxy sources or very sparse direct observations. Other solar and volcanic forcing series estimates exist (Solanki and Fligge, 2000, Fligge and Solanki, 2000, Bard et al., 2000, Stothers, 1996 for example). However, each model ensemble takes up a large amount of computational time and therefore only a very few natural forcings ensembles can realistically be considered. Considering a number of different forcing history ensembles would help reduce uncertainties to the climatic response inherent in the natural forcing histories prior to the latest 20 years.

Anthropogenic forcings considered in increasing order of complexity are well-mixed anthropogenic greenhouse-gases (labelled as G for HadCM2 and GHG in HadCM3), well-mixed anthropogenic greenhouse gases and sulphate aerosols (GS in HadCM2, TROP-ANTHRO in HadCM3), and well-mixed anthropogenic greenhouse-gases, sulphate aerosols and stratospheric ozone (GSO in HadCM2, ANTHRO in HadCM3).

For HadCM2 the G forcing uses reconstructed emissions of all greenhouse gases (represented as CO_2 equivalent concentration) from 1860 until 1990. Forcing is then projected up until 2100 based upon a projection of compound 1% year-on-year increases (Mitchell et al., 1995, Mitchell and Johns, 1997). Mitchell and Johns (1997) note how this is a greater increase than that given under IPCC emissions scenario IS92A – Business As Usual (BAU). They estimate that by 2100 the forcing is over-estimated by 12% compared to this scenario. Individual ensemble members extend from 1860 until 2100 (Johns et al., 1997). In HadCM3 greenhouse gases are treated separately (rather than as a CO_2 equivalent concentration) and there are coupled chemical models which simulate the overall changes in individual greenhouse gas concentrations (Tett et al., 2001). The gases considered are: CO_2, CH_4, N_2O, and six (H-)CFCs and are assumed to have globally uniform mass mixing ratios. Again, members extend from 1860 to 1990 based upon observational estimates. Subsequently, the concentrations are calculated based upon emissions scenario B2 of the IPCC SRES report (Nakicenovic et al., 2000).

In HadCM2 the GS run series has the same greenhouse gas forcing but also incorporates the direct effect of sulphate aerosols, which scatter incoming solar radiation, parameterised by an increase in the clear-sky albedo (Johns et al., 1997). Therefore if the model grid-box is covered in cloud this forcing has no effect upon the model calculated values (Mitchell and Johns, 1997). The forcing applied is seasonally and geographically invariant, being a scaled version of current emissions patterns back in time from 1860-1990 (Mitchell and Johns, 1997). Future emissions were estimated for 2050 under the BAU scenario and sulphate loading values linearly interpolated between 1990 and 2050 to yield a forcing estimate (Mitchell and Johns, 1997). After 2050, values are just a scaled version of the 2050 pattern to agree with the BAU scenario. In HadCM3 TROP-ANTHRO runs include a more realistic pattern of forcings based upon an interactive sulphur cycle scheme to yield the distribution of sulphate aerosols at each timestep which is then passed to the model radiative scheme (Tett et al., 2001). The patterns of emissions scenarios are also based upon more realistic estimates of past and future emissions than was the case for HadCM2. The first order indirect effect of sulphate aerosols, due to changes caused in cloud droplet characteristics, is included in a rather crude manner (Tett et al., 2001 give details), although not any second order indirect effects on cloud lifetime. The TROP-ANTHRO ensemble also incorporates modelled changes in tropospheric ozone concentrations calculated in an off-line chemistry model and interpolated between the calculated periods (Tett et al., 2001). Both the evolving pattern and magnitude of the direct and indirect sulphate aerosol effects are less certain than those of well-mixed greenhouse-gases (Tett et al., 2001, Stott et al., 2001 amongst others).

For HadCM2 GSO is identical to the GS series except that post 1974 the effects of stratospheric ozone depletion due, at least primarily, to anthropogenic CFC and H-CFC emissions are taken into account (Tett et al., 1996). In HadCM3 the ANTHRO series is similarly identical to the TROP-ANTHRO series except post 1974 when, again, changes in stratospheric ozone concentrations are included (Tett et al., 2001).

1.4 Previous detection studies

The importance of detection and attribution studies has been recognised since the potential for a distinct large-scale anthropogenic effect on climate was first seriously considered in the 1970's. With increasing model resolution, realism, and power, more and more independent modelling efforts (albeit based on the same underlying modelling premise) over the past decade, and advances in observational datasets, detection and attribution approaches have also advanced and diversified. In chapter 4 the evolution and statistical methodology of some, but by no means all, of these approaches are outlined. Here, major results of **quantitative** detection and attribution studies to date are summarised showing the breadth of independent approaches and modelled and observed datasets which have **consistently demonstrated anthropogenic influences** in the recent climate records of a number of variables (principally near-surface and zonally-averaged temperatures). There also exist numerous examples of qualitative studies identifying both consistencies and inconsistencies between observations and simulations of climate change for a diverse range of indicators. Effectively, such studies degenerate into a model validation exercise and so yield little, if any, extra useful information in terms of either detection or attribution. These are summarised in the IPCC SAR (Santer et al., 1996) and TAR (Mitchell and Karoly et al., 2001).

The simplest quantitative detection studies are two dimensional in nature, using global mean trends in atmospheric variables. These are termed Stage I detection studies, and as discussed previously could easily yield ambiguous results (see 1.2.3). Most studies of this type have used global mean near-surface temperatures, either from instrumental records or paleo-reconstructions, or both. Section 8.4.1 of the IPCC SAR (Santer et al., 1996) provides a good review of these Stage I studies and their limitations. Levitus et al. (2001) have recently used observed changes in global mean ocean heat content to attribute observed changes to a combination of anthropogenic and volcanic influences. Tol and De Vos (1998) have employed a

Bayesian approach to show a significant relationship with atmospheric CO_2 and indicate that natural variability is unlikely to explain the large-scale observed surface temperature trend. The Bayesian approach is potentially very powerful (Hasselmann, 1998, Berliner et al., 2000), removing at least some of the uncertainties and caveats attached to conventional studies. However, to date the full potential of the Bayesian approach has not been employed within the sphere of climate change detection and attribution, in anything but an idealised setting.

Stage II detection exercises (Santer et al., 1996) involve looking for a signal in three dimensions (either on a surface (longitude, latitude, time) or zonally averaged (latitude, height, time)). Stage III detection studies similarly search for a signal in this space, but they optimise the searched-for signal in some respect before carrying out detection. These two distinct approaches have been employed to elicit the most recent detection and attribution results, and hence they are concentrated upon in this section. The methodologies are summarised in chapter 4. To date only one Bayesian type study has been applied to either Stage II or III detection exercises in an idealised setting (Berliner et al., 2000), hence the studies summarised here result from the more classical frequentist statistical approaches. A further advanced stage was conceived in the IPCC SAR (Santer et al., 1996) labelled Stage IV whereby a number of climate variables are considered simultaneously in a detection exercise. This approach has yet to be implemented in any study.

The earliest studies (Stage II) used pattern correlations and are the main component of the detection chapter in the IPCC SAR (Santer et al., 1996). These studies have employed both centred correlation statistics (R(t)) which remove the global mean change and uncentred correlation statistics (C(t)). In both cases the signal is a fixed model derived signal, and only a single forcing scenario (or linear combination of signals) can be considered at any time. Uncentred (C(t)) statistics tend to be dominated by the global mean change and therefore are of limited use in discriminating between competing forcings, not really being any advance on Type I studies. Long term trends in the statistics (trends in the statistics when the fixed signal is correlated against the smoothed observational series) are used and compared to those from models run in control mode to assess trend significance. Using centred statistics and a single model, Santer et al. (1995) find a significant correlation for 50-

year near-surface temperature trends in JJA and SON if anthropogenic greenhouse gases and sulphate aerosol forcings are considered.

Several studies have considered zonally-averaged (latitude, height, time) temperature trends using a pattern correlation approach (Santer et al., 1996a, Karoly et al., 1994, Tett et al., 1996, Folland et al., 1998). Santer et al. (1996a) show that the observations are best explained by a linear combination of greenhouse gases, sulphate aerosols and ozone changes. Tett et al. (1996) augment this study by considering a more realistic sulphate forcing scenario. They find that when stratospheric ozone changes are taken into account the upper atmosphere temperature record is better captured but the lower atmosphere is too cool when compared to the observations. As a sensitivity study the stratospheric ozone depletion forcing is halved, this scenario provides the best results and removes some of the discrepancy in the lower atmosphere. Allen and Tett (1999) reassess the results of Tett et al. (1996) in the light of a masking deficiency in the model data whereby all data had been included rather than sub-sampling to the available observational field, and find that this does not significantly affect previous conclusions. Karoly et al (1994) also find a significant anthropogenic signal, although changes in stratospheric ozone and tropospheric sulphate aerosol concentrations are not considered in their study. Using an atmosphere-only model forced with observed SSTs, and a radiosonde field reconstructed from its leading eigenvectors (EOFs), Folland et al. (1998) find that observed trends cannot be explained by changing SSTs alone, and that anthropogenic forcings must be considered. Even though SSTs have increased, this in itself is not sufficient to explain the observed changes in atmospheric temperature structure.

Various criticisms have been made of pattern correlation approaches to detection and attribution, primarily by Legates and Davies (1997). Many of the criticisms have been shown by Wigley et al. (2000) to be either implicitly taken into account by the pattern correlation approach, or pertaining to technical details of the unrealistic hypothetical simple example presented by Legates and Davies (1997) and their treatment of it. Critically, Wigley et al. (1998) show that the trends found in Santer et al. (1995) for the surface temperature are comparable to those which would be found if the signal response were perfectly known for an emerging signal given the attendant noise due to internal variability (estimated from a GCM).

Optimal detection techniques (Stage III studies), which are heavily referenced within the IPCC TAR (Mitchell and Karoly et al., 2001), consider the signals and observations in some form of truncated phase space, and rotate the statistic such that it searches for each signal in a direction which maximises the SNR. The phase space in which the detection is performed is truncated as sufficient independent noise realisations do not exist for optimisation in the full field space (this would require a GCM control run of order a million years). A useful by-product of such approaches is that it has been shown that only the largest spatial scales are considered, and furthermore that models only adequately capture the variability at such large scales (Barnett et al., 2000, Stott and Tett, 1998). This is expected when models attempt to realise an infinite non-linear system by linear finite numerical approximations (Allen and Tett, 1999).

Numerous approaches have been taken to the optimisation and subsequent detection algorithms, all of which can be shown to be broadly equivalent (Hegerl and North, 1997) and can be couched in terms of linear regression (Allen and Tett, 1999, Levine and Berliner, 1999). A regression approach has advantages in terms of being able to estimate signal strengths directly and consider a number of signals simultaneously. The optimal regression methodology of Allen and Tett (1999) and Allen and Stott (2001) is discussed in detail in chapter 4. Three distinct approaches to the signal derivation have arisen: space-time, (Allen et al., 2000, 2001, Tett et al., 1999, 2001, Stott et al., 2001, 2001a, G.S.Jones et al., 2001); fixed signal (Hegerl et al., 1996, 1997, 2000, Barnett et al., 1999, 2000, Allen and Tett, 1999, Tett et al., 2001); and frequency-space (North and Stevens, 1998). This section concentrates on the first two of these approaches. Space-time approaches consider the transient nature of the signal, whereas a fixed signal approach assumes that the signal is of fixed pattern and it is solely the signal amplitude that varies. If the model correctly predicts any transient changes in the climate response, i.e. if they exist, then the former approach will be more powerful than the latter. However, an emerging signal will tend to be highly contaminated by noise and therefore a large ensemble must be used in any space-time approach to ensure against sub-optimal results, especially if noise in the signals is not implicitly factored into the analysis (Allen and Stott, 2001, Stott et al., 2001a).

A number of optimal detection studies have considered changes in near-surface temperatures using different models and accounting for various sources of potential error. Results for space-time signals have been found, for the detection of a well-mixed greenhouse-gases signal, to be generally insensitive to signal processing procedures, scales considered, and estimates of natural variability (Tett et al., 1999,2001 Stott et al., 2001). For HadCM2, results are also insensitive to whether noise in the signals is implicitly taken into account in the detection algorithm, although the range of uncertainty in signal strength is generally increased, especially in the upper-estimates (Stott et al., 2001a). Results have also been shown to be generally consistent between models forced with anthropogenic forcing histories for both fixed and space-time signals, as well as robust to various methodological differences (Allen et al., 2000, 2001, Barnett et al., 1999, 2000, Hegerl et al., 1996, 1997, 2000, Gillett et al., 2001). All these studies generally detect a distinct greenhouse gas influence with a more uncertain sulphate forcing influence.

Near-surface temperature optimal detection studies that consider natural forcings in some cases also yield a significant solar signal and, depending upon the time period under consideration, a volcanic signal. This is particularly the case in the early 20th Century climate (Tett et al., 1999, 2001, Stott et al., 2001, 2001a, Allen et al., 2001, Hegerl et al., 1997, 2000, Barnett et al., 1999). These results are less robust to changes in parameterisations than is the detection of anthropogenic influences. The use of decadal chunks of annually averaged temperatures in the space-time studies may be sub-optimal for detecting natural forcings; volcanic signals are short-lived (of the order of years) and the solar forcing cycle probably has a frequency peak at approximately 11 years (the sunspot cycle). Stott et al. (2001) show how, by reducing the temporal resolution in a space-time approach from decadal to annual, a clear volcanic signal from Pinatubo can be detected during the 1990s. Importantly, numerous studies have shown that natural forcings on their own are an inadequate representation of late 20th Century climate and that an anthropogenic influence is required (Hegerl et al., 1997, 2000, Tett et al., 1999, 2001, Stott et al., 2001, 2001a, Barnett et al., 1999).

Incorporating seasonal information into the searched for signal reduces signal degeneracy (when a signal or combination of signals is similar to another signal) and increases the power of the regression algorithm to distinguish between forcings (Stott et al., 2001). The summer and autumn seasons exhibit highest SNRs (Stott et al., 2001). Barnett et al. (1999) and Hegerl et al. (1997,2000) consider solely summer season temperatures for this reason.

Changes in zonal mean temperatures have been considered by Allen and Tett (1999) and Tett et al. (2001) using a fixed signal optimal detection methodology. Allen and Tett (1999) detect a combined signal of greenhouse gases, sulphate aerosols, and ozone, using HadCM2 fields, consistent with the pattern correlation study of Tett et al. (1996). Tett et al. (2001) using the same input fields for HadCM3 data find that they can detect both anthropogenic and natural external forcing influences on the zonal mean temperature field. Both studies do not undertake any sensitivity studies; confidence in the result would be increased if they were found to be robust to uncertainties. In chapter 5 such an analysis is undertaken with a modified version of the observed upper air temperature dataset used in these studies. Hill et al. (2001), using a space-time approach, detect a solar signal as well as anthropogenic influences in HadCM2. G. S. Jones et al. (2001) augment these studies by undertaking a space-time study considering near-surface and zonally averaged upper air temperatures to yield what is effectively a crude 4-dimensional input pattern of late 20th Century temperature change. This study gives detectable anthropogenic and volcanic influences on the climate of the last 40 years, although not solar influences.

Barnett et al. (2001) have undertaken a study, using a single model, considering changes in oceanic heat content in a number of ocean basins. Regardless of whether optimisation is applied, an anthropogenic signal is detected and found to be consistent with the observations. However, natural external forcings are not considered and therefore rigorous attribution of the observed changes to an anthropogenic cause is not possible. The control run of the model used in the study is also relatively short and may exhibit drifts, although Barnett et al. (2001) state that sensitivity studies indicate that this should not be a first order effect. Further work considering more models and incorporating the effects of natural external forcings is required to confirm the principal findings of this study.

Optimal space-time studies have also been used in an attempt to attribute recent changes in observed near-surface temperatures to given causes. Using HadCM2 (Stott et al., 2001, Tett et al., 1999) and HadCM3 (Tett et al. 2001) late 20th Century near-surface temperature changes have been attributed to anthropogenic influences, although considering the whole 20th Century solar influences are also detected. Over the entire 20th Century the net change in solar forcing used in these studies has been almost zero. This raises an important point: a signal need not be important in explaining recent climate change, even if it is detected, since detection is solely a statistical parameter. The solar forcing on the timescales of a few decades under consideration here is essentially cyclical around a stationary mean and therefore yields little long-term near-surface temperature trend on multi-decadal timescales (relative to the short-term fluctuations). Allen et al. (2001) extend these studies by explicitly considering noise in the signals due to finite ensemble effects, and using a larger population of models, only some of which have ensembles based upon natural external forcings. Their study consistently yields an anthropogenic influence in the latter 20th Century, with a less certain solar and volcanic component. Hegerl et al. (1997, 2000) using a fixed signal approach, also tentatively attribute recently observed climate changes to anthropogenic and natural forcings.

Extending consideration beyond a simple yes-no detection exercise, Allen et al. (2000, 2001) illustrate how optimal detection studies which use a regression algorithm can be employed to consider certain aspects of the climate system. As an example, they attempt to constrain uncertainty in future predictions of global mean near-surface temperature by various models, by ascertaining the range of plausible signal strengths in the available observations and scaling future model predictions by this range. The argument is that this will hold under a transient forcing. It is believed to be unlikely that it would continue to hold after stabilisation of greenhouse gas levels in the atmosphere (Myles Allen, personal communication, 2000). Of course, changes in global mean temperature, whilst being a large component of the climate variance, is not very useful from impacts, adaptation, and mitigation perspectives, but the power exists within the algorithm to extend it to consider other variables which may have greater applicability.

Both the pattern and optimal studies summarised above have numerous caveats applied to them. Some of these have been addressed; others are more general in nature. All detection studies rely critically upon realistic estimates of natural variability, derived either from model runs with no external forcing, or from observations to claim significance. For this purpose the observations are potentially polluted by signals from natural external sources as well as anthropogenic influences, and although these can be estimated and removed (Jones and Hegerl, 1998), the result is model dependent, and residual errors may remain. There is also a risk of introducing a common factor effect in using such filtered observations (Santer et al., 1993), especially if optimisation is being carried out, whereby the fields are made to become artificially similar and will yield ambiguous results. Equally, model estimates of internal variability may be compromised by their inability to accurately represent some major modes of atmospheric variability (El Niño Southern Oscillation (ENSO), Quasi-Biennial Oscillation (QBO), North Atlantic Oscillation (NAO)/Arctic Oscillation (AO), stratosphere-troposphere interactions, amongst others), as well as oceanic variability on longer time scales. Gillett et al. (2000a) have shown that the detection of an anthropogenic signal in the near-surface temperature record for the latter 20th Century is insensitive to whether the recent positive trend and variability in the AO (Thompson and Wallace, 2000) is correctly simulated in HadCM2.

Some studies have tested the sensitivity to estimates of natural variability by inflating the model-derived variance (Tett et al., 1999, 2001 for example). Such an approach implicitly assumes that the pattern of natural variability is correctly estimated and only the amplitude is wrong; an assumption that almost certainly is not strictly true, especially in the limit of a finite control run (Allen and Tett, 1999). Errors in the estimate of natural variability are most important in optimal studies. The internal variability is employed in the optimisation procedure and therefore the optimal statistic will only consider areas of phase space sampled in the control. If the control underestimates a mode of variability then it will focus the optimised statistic on this region in the reduced phase space, reducing the chances of a successful detection exercise. Noise in the observations is considered highly unlikely to mimic the complex spatio-temporal fingerprints used in optimal studies and therefore the chances of yielding a false positive result are remote for this source of error. Equally, if the model control overestimates the natural variability in a given mode then the signal will be damped in this mode. Therefore an optimal approach can only ever yield a conservative result, being critically dependent upon an adequate model estimate of natural internal variability. Further, there is residual uncertainty in the observational fields. It is likely that all observed fields maintain at least some biases and random errors (Barnett et al., 1999). The question remains as to whether a signal can be detected if residual uncertainties remain both in the observations and in the model derived estimates of variability. In both cases errors can only lead to conservative conclusions whereby detection is rejected, unless the observational error has a similar pattern to an incorrect model signal, which is highly unlikely. Hegerl et al. (2001) take into account the potential effects of observational errors and show that, even making the most pessimistic assumptions, such errors will not yield a first order effect upon results of fixed signal detection studies.

A consistency test has been proposed for the optimal regression approach by Allen and Tett (1999), whereby the residuals of the regression are compared to an independent realisation of the natural internal variability (generally, an independent section of control) in an f-test. If the residuals are found to be inconsistent with the independent noise realisation then the result is rejected as a plausible explanation of the observed climate change. The test is relatively weak, as it does not consider the shape of the residuals but only their amplitude.

Detection studies also assume that the model-derived signals are accurate representations of an evolving climate under the forcing in question. Both the correlation and optimal approaches assume that the pattern is correct but the amplitude not necessarily so, although correlation and some optimal approaches cannot effectively differentiate between pattern and amplitude effects. Numerous optimal approaches have considered different model signals and concluded that the results, at least for near-surface temperatures, are generally independent of the model used to derive the signal at the large scales considered (Hegerl et al., 2000, Barnett et al., 1999, 2000 Allen et al., 2001). These studies tend to yield a significant detection of anthropogenic influences in late 20th Century surface temperatures, although perhaps significantly, the exact details differ depending upon the pre-processing applied. Gillett et al. (2001) illustrate how differences in the signals detected by

Hegerl et al. (2000) (fixed signal, summer only), and Allen et al. (2001) (time-space signal, annual), for the same models and forcings can primarily be explained by differences in signal pre-processing algorithms, rather than any fundamental differences between the different optimal detection approaches employed. Further, the result of Gillett et al. (2001) suggests that there may not be a one size fits all "optimal" technique that is indeed optimal for all possible applications. Finally, a number of studies recognising the potential problems of using a single realisation of an emerging signal have taken to using ensemble realisations to better constrain the results (Stott et al., 2001, 2001a, Tett et al., 1999, 2001, Allen et al., 2000, 2001). This is consistent with an explicit recommendation of Barnett et al. (2000) who found that single member ensembles could yield highly erroneous results.

Residual uncertainties remain regarding the prescribed forcings the models are run with (Barnett et al., 1998, Stott et al., 2001). This is especially the case for natural forcings, which previous to about 20 years ago are based on best available proxy measures or sparse observations. Tett et al., (1999) and Stott et al., (2001) attempt to quantify the uncertainty in solar forcing history by considering signals based upon two solar reconstructions which had their mid-20th Century radiation peak separated by a complete solar cycle (Lean et al., 1995, Hoyt and Schatten., 1993). They find that detection, at least in the HadCM2 model, is potentially dependent upon the ensemble realisation, and therefore accurate forcing histories are essential to constrain uncertainties. However, these are likely to be difficult to obtain through proxy indicators.

Our confidence in the results of current detection studies will be increased if an anthropogenic signal can be detected in a broader range of both climate parameters and GCM's, considering a range of detection methodologies and numerous potential sources of uncertainty. Confidence would also be increased in model realism if the results of these detection studies were found to be, at least potentially, internally consistent for any given model(s).

1.5 Conclusions

In this chapter previous work in the sphere of climate change detection and attribution has been summarised to enable the remainder of this thesis to be placed in context. Detection and attribution were formally defined before advancing to consider the components necessary for a successful detection and attribution study. Particular focus was placed on the observed near-surface (HadCRUTv, Jones et al., 2001) and upper air (HadRT, Parker et al., 1997) globally gridded datasets employed in this thesis. The two versions of the Hadley Centre GCM considered were also briefly summarised. Subsequently, consideration was given to previously published detection results. Although there exists a range of results, these studies consistently suggest a demonstrable anthropogenic influence and, where considered, more tentative evidence for natural external influences, upon 20th Century climate. This result is seen to be at least to first order independent of both dataset and methodological considerations, although the range of both is limited. Previous detection studies have considered solely near-surface temperatures, ocean heat content, or zonally averaged upper air temperatures. In the remainder of this thesis it is aimed to ameliorate this situation by advancing to consider the full-field upper air temperature record in a rigorous manner.

Table summarising the modelled forcing histories for HadCM2 and HadCM3.

Forcing	**HadCM2 ensembles**	**HadCM3 ensembles**
Well-mixed greenhouse gases	G 1860-2100 (Considered as CO_2 equivalent)	GHG 1860-2100 (Individual constituent gases considered)
Greenhouse gases plus anthropogenic sulphate aerosols	GS 1860-2100 (Direct effect of sulphate aerosols only)	TROP-ANTHRO 1860-2100 (includes tropospheric ozone changes)
Greenhouse Gases plus Anthropogenic sulphate aerosols plus changes in Stratospheric Ozone	GSO 1860-1995	ANTHRO 1860-2100
Solar	SOL, LBB 1890-1995, 1890-1996 (First members start in 1700 and 1799)	SOLAR 1860-1999
Volcanic	VOL 1890-1997	VOLCANIC 1860-1999
All natural		NATURAL 1860-1999

Table 1.1 Summary of HadCM2 and HadCM3 ensembles giving the acronyms and the run dates. A detailed description is given in Section 1.3 and Stott et al., 2001 (HadCM2) and Tett et al., 2001 (HadCM3).

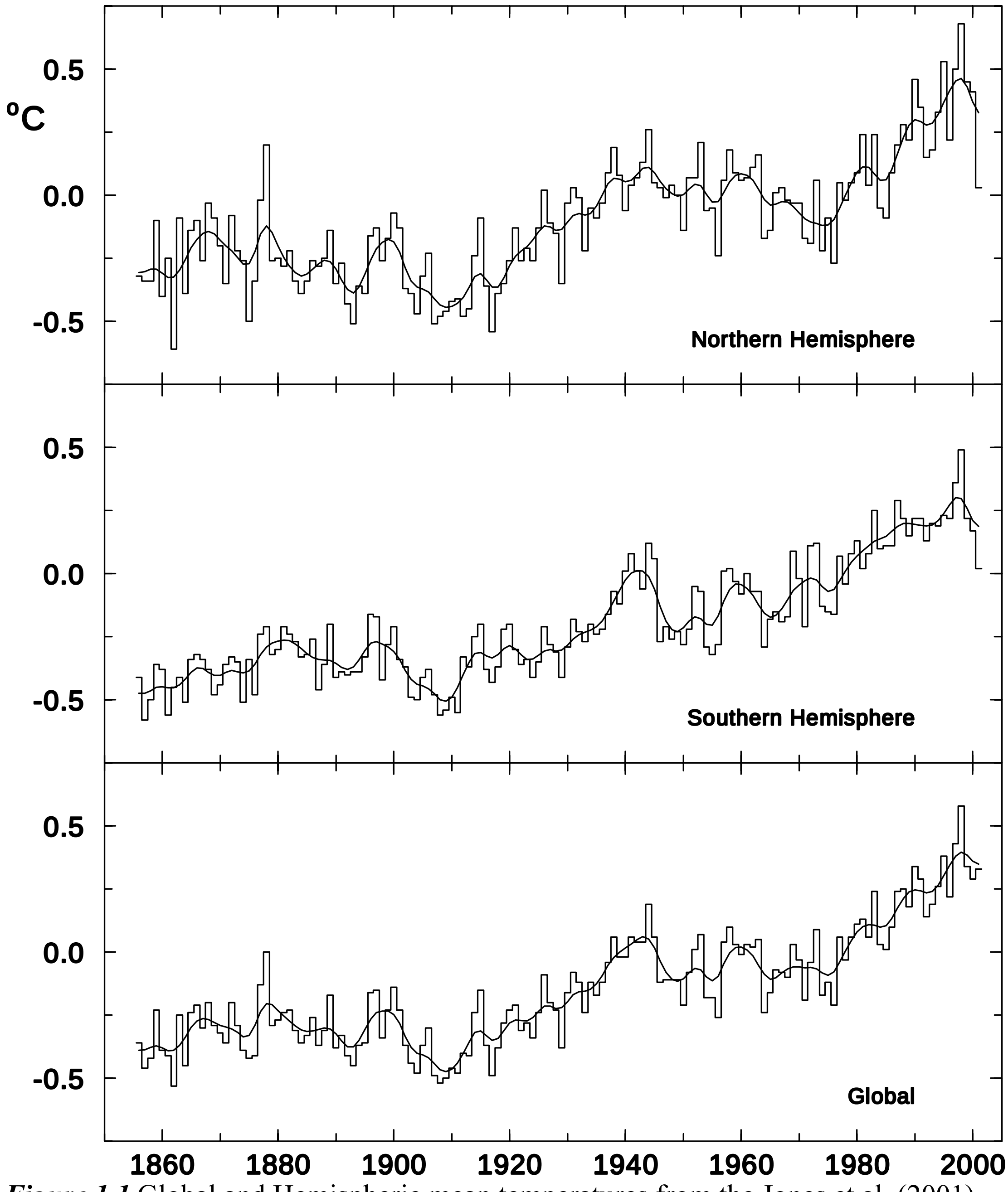

Figure 1.1 Global and Hemispheric mean temperatures from the Jones et al. (2001) dataset. Reproduced from http://www.cru.uea.ac.uk/cru/data/temperature/