Models versus radiosondes in the free atmosphere: A new detection and attribution analysis of temperature

F. C. Lott,[1] P. A. Stott,[1] D. M. Mitchell,[2] N. Christidis,[1] N. P. Gillett,[3] L. Haimberger,[4] J. Perlwitz,[5] and P. W. Thorne[6]

[1]   This analysis revisits detection and attribution of free atmosphere temperatures from radiosondes, almost a decade after previous studies. Since that time, data sets have not only become longer, but understanding of observational uncertainty has vastly improved. In addition, a coordinated set of experiments exploring the effects of human and natural forcings on past climate change has been made with a new generation of climate models. These advances allow a much more thorough analysis of the effects of modeling and observational uncertainty on attribution results than previously possible. Observational uncertainty is explored using multiple radiosonde reconstructions including those with ensembles of realizations exploring the effects of processing choices. Modeling uncertainty is explored by calculating multiple fingerprints of natural influence (from changes in solar irradiance and volcanic aerosols) and of human influence (due to greenhouse gases and due to the effects of combined anthropogenic forcings including stratospheric ozone depletion). With increased confidence over previous studies, human influences (both greenhouse gas and other anthropogenic forcings) have been detected in spatiotemporal changes in free atmosphere temperature from 1961 to 2010, irrespective of whether the full atmospheric column (30–850 hPa) is examined or purely the troposphere, with stratospheric ozone depletion dominating the cooling that has been observed in the lower stratosphere. Thus the advances of the last decade yield increased confidence that anthropogenic influences have made a substantial contribution to the evolution of free atmosphere temperatures.

Citation: Lott, F. C., P. A. Stott, D. M. Mitchell, N. Christidis, N. P. Gillett, L. Haimberger, J. Perlwitz, and P. W. Thorne (2013), Models versus radiosondes in the free atmosphere: A new detection and attribution analysis of temperature, *J. Geophys. Res. Atmos.*, *118*, 2609–2619, doi:10.1002/jgrd.50255.

1.   Introduction

[2] In the field of detection and attribution, fingerprint analysis [*Hasselmann*, 1979] is the standard technique to help determine the causes of observed climate change. A fingerprint is a pattern of changes in the climate, simulated in response to a given forcing. In the free atmosphere, the fingerprint from anthropogenic forcing is a pattern of a warming troposphere and cooling stratosphere [*Karoly et al.*, 1994]. This fingerprint is unlike that which would result from solar forcing alone, where an increase would be expected to warm the stratosphere and vary regionally in the troposphere [*Cubasch et al.*, 1997; *Santer et al.*, 2003; *Gray et al.*, 2010; *Hegerl et al.*, 2007]. It also differs from that due to stratospheric ozone depletion alone, which cools the stratosphere, and from explosive volcanic eruptions that inject aerosol into the stratosphere and which cause a periodic short-term warming of the stratosphere and cooling of the troposphere. The aim of this paper is to apply standard fingerprinting techniques to determine if any of these patterns can be detected in the observed free atmosphere temperatures since systematic quasi-global observations by weather balloons began in the late 1950s.

[3] While the resemblance between the observed pattern in the free atmosphere and the anthropogenic fingerprint has long suggested that this is the main cause of the climate change observed, detection and attribution studies seek to determine whether such observed changes could not have happened by chance, i.e., they are outside the range (at some

---

[1]Met Office Hadley Centre, Exeter, UK.
[2]Department of Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK.
[3]Canadian Centre for Climate Modelling and Analysis, Environment Canada, Victoria, British Columbia, Canada.
[4]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria.
[5]Cooperative Institute for Research in Environmental Sciences, University of Colorado, and NOAA Earth System Research Laboratory Physical Sciences Division, Boulder, Colorado, USA.
[6]Cooperative Institute for Climate and Satellites, NCSU and NOAA's National Climatic Data Center, Asheville, North Carolina, USA.

Corresponding author: F. C. Lott, Met Office Hadley Centre, Fitzroy Road, Exeter EX1 3BP, UK. (fraser.lott@metoffice.gov.uk)

significance level) that could be explained by natural internal variability alone (detection). They then seek to determine to what extent observed changes can be explained by natural and anthropogenic factors (attribution). *Santer et al.* [1995] is an early example of such a study, wherein the output of a climate model with a simple slab ocean represented the equilibrium climate response. It was found that observed changes up to 1993 could not be explained by natural factors alone. Subsequent detection and attribution studies of free atmosphere temperatures looked at the space time pattern of changes with coupled climate models [*Tett et al.*, 1996; *Allen and Tett*, 1999; *Thorne et al.*, 2002, 2003; *Jones et al.*, 2003]. The results of these studies supported the earlier conclusion in finding a significant human influence, in particular from greenhouse gases, on observed free atmosphere temperature changes.

[4] It should be noted, however, that much of the work on free atmosphere changes from radiosonde observations is now a decade old. In more recent years, improved radiosonde data sets have been developed (detailed in section 2), understanding of observational uncertainty has improved [*Titchner et al.*, 2009], and many innovations have been made between the previous phase of the Coupled Model Intercomparison Project (CMIP3) [*Meehl et al.*, 2007] and the current phase (CMIP5) [*Taylor et al.*, 2011]. For most centers, these include new model versions, greater horizontal and vertical resolution, and more realistic external forcings (see supporting information). CMIP5 models all consider past stratospheric ozone depletion and its recovery in the future. The majority of CMIP5 models prescribe ozone changes by utilizing the time varying ozone database according to *Cionni et al.* [2011] or a modification of this database. The rest of the models include a stratospheric chemistry scheme and calculate ozone changes interactively [*Eyring et al.*, 2013].

[5] In addition to these improvements, this study is enhanced by another 10 years of observational data having been gathered. This should make it easier to discriminate between natural and anthropogenic factors (since detection and attribution requires multi-decadal trends to distinguish forced trends from natural internal variability [*Santer et al.*, 2011]). Thus revisiting this topic is timely. This paper focuses on radiosonde data sets, which now span over 50 years, 20 years more than satellite data. This increased period (from which this study focuses on 1961–2010) should prove useful for detecting changes in climate. This paper will use the approaches to detection and attribution in the free atmosphere taken by *Tett et al.* [1996], *Allen and Tett* [1999], *Thorne et al.* [2002, 2003], and particularly *Jones et al.* [2003] to assess this new data. For companion studies analyzing satellite data, see *Santer et al.* [2012].

2. Data Analysis Method

2.1. Data Sets

[6] A number of new radiosonde data sets have been developed since the studies of a decade ago. Following the review by *Thorne et al.* [2011] and having assessed which data sets and ensembles had coverage for the entire period of record, four data sets were selected for analysis.

[7] The first of these is HadAT2 [*Thorne et al.*, 2005]. Of the observational data sets, this has the least spatial coverage and thus is used as a common mask for all other data, both observations and models, to allow a like-for-like comparison. (Note that the change in zonal trends found due to coverage differences in the observational mask was typically less than 0.1 K/decade, in line with the findings of *Mears et al.* [2011]). HadAT employed a neighbor-composite-based breakpoint identification and adjustment approach with manual intervention. Adjustments for the post-2003 portion of the record are assessed manually every quarter and adjustments therefore have been applied to the whole of the record.

[8] The other three observational data sets are from the RICH/RAOBCORE family [*Haimberger et al.*, 2012]. The first of these sets used is RAOBCORE 1.5, which uses the ERA-40 [*Uppala et al.*, 2005] and ERA-Interim reanalyses [*Dee et al.*, 2011] to detect and adjust for non-climatic breakpoints. The other two are the ensembles of realizations known as RICH-obs 1.5 and RICH-τ 1.5. Both of these consist of 32-member ensembles where each member reflects different adjustment processing decisions (such as minimum number of data points or treatment of transitions), with breakpoint detection derived from RAOBCORE (i.e., the breakpoint locations are identical across RAOBCORE and all RICH members). RICH-obs makes adjustments by directly comparing neighbor station time series, while RICH-τ compares the differences between the time series and the ERA-Interim background for each neighbor and uses this returned series to calculate the target station adjustment.

[9] In order to detect climate forcings in the observational data, simulations with the relevant combinations of forcings over the full period are required. Therefore the selection of model data sets was limited by the need for each model to have runs with natural forcings (NAT) only, as well as runs with only greenhouse gas forcings (GHG) and finally with all historical (i.e., anthropogenic and natural) forcings (ALL), between 1961 and 2010 available on the CMIP5 [*Taylor et al.*, 2011] archive at the time the analysis was undertaken (Spring 2012). This led to the models shown in Table 1 being used.

[10] All data sets were recomputed as a common temperature anomaly relative to the 1961–1990 climatology, re-gridded by area-averaging to the HadAT2 grid (a resolution of 5° latitude by 10° longitude) and masked with the coverage of HadAT2 before zonal averages were taken. The following set of pressure levels common to all data sets was used: 850, 700, 500, 300, 200, 150, 100, 50, and 30 hPa.

2.2. Analysis

[11] Most changes in this study compared to previous studies [*Tett et al.*, 1996; *Allen and Tett*, 1999; *Thorne et al.*, 2002, 2003; *Jones et al.*, 2003] were due to the greater quantity of data now available. For example, the previous exclusion of the regions south of 30°S in *Jones et al.* [2003] was due to a lack of radiosonde data in that region. However, with the advent of the new data sets detailed in section 2.1, not only is there more data, but also with the RICH ensembles the uncertainties should be well-represented. Consequently the three latitude bands analyzed here are an exclusively tropical zone (20°S to 20°N) and north and south extra-tropical zones (60°S to 20°S

Table 1. CMIP5 Models Used for This Study and the Number Runs With Each Forcing[a]

Modeling Center (or Group)	Model(s)	Members Included		
		ALL	NAT	GHG
Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence	CSIRO-Mk3.6.0	10	5	5
Centre National de Recherches Météorologiques/Centre Européen de Recherche et Formation Avancées en Calcul Scientifique	CNRM-CM5	10	5	5
NASA Goddard Institute for Space Studies	GISS-E2-R	5	5	5
	GISS-E2-H	5	5	5
Canadian Centre for Climate Modelling and Analysis	CanESM2	5	5	5
Met Office Hadley Centre	HadGEM2-ES	4	4	4
Norwegian Climate Centre	NorESM1-M	3	1	1
Beijing Climate Center, China Meteorological Administration	BCC-CSM1.1	3	1	1

[a]Further details of these models can be found in the supporting information.

and 20°N to 60°N), which were then used for detection and attribution analysis. For information, later figures will also show the average over the whole studied area (i.e., 60°S to 60°N). It was considered that there were insufficient data beyond 60° for analysis. Note also that the surface is omitted from this investigation, as a separate surface analysis is presented in another companion paper [*Jones et al.*, 2013].

[12] One criticism of previous free atmosphere studies [*Legates and Davis*, 1997] has been that this type of detection and attribution study is critically dependent on the inclusion of the stratosphere and that the detectable signal component is effectively driven by opposite trends in the stratosphere and troposphere and no other aspect of climate change. It was claimed that this can result in erroneous detection. In response, analyses were also carried out in the troposphere separately to investigate whether detection relies upon stratospheric changes, as was performed in *Thorne et al.* [2002]. Note this was using an old observation data set and two versions of the same climate model. It is therefore of substantial value to revisit and reassess this issue with the broader range of observational and model tools now available.

[13] Here the study of *Liu et al.* [2005] was used to establish the range of positions of the tropopause in the zones under study. Zonal levels where there is likely to be a mix of troposphere and stratosphere were omitted. Consequently, the troposphere was considered to include the 150 hPa level and below in the tropics and the 300 hPa level and below in the extratropics. Using the same tropopause data, only the 50 and 30 hPa levels are left for a stratosphere-only study. This proved insufficient for an optimal detection analysis, so the effect of the stratosphere was instead determined by comparing results of the troposphere alone to that of the whole free atmosphere up to 30 hPa height.

3. Results

3.1. Zonal Temperature Trends by Pressure Level

[14] To examine temperature data varying over time and height, the trends at each pressure level were calculated using a median pair-wise algorithm (as this is less affected by outliers and end-point effects than a conventional linear fit)

[*Lanzante*, 1996]. These trends were plotted against pressure level, for all models and forcings within them (Figure 1).

[15] The first point to notice is that the observations (OBS) have their largest spread in the southern extra-tropics, with very little uncertainty in the northern extra-tropics. In all zones, this uncertainty grows larger with height, with HadAT showing the least stratospheric cooling and RAOBCORE the most. In the tropical upper troposphere, the spread of the RICH ensembles shows that this region is also uncertain, but provides scope for there to be rather more warming than is apparent in other data sets.

[16] The ALL model simulations are generally closer to the observations than either GHG or NAT simulations. Both ALL and OBS show a warming in the troposphere and a cooling in the stratosphere, with RICH observations showing closest agreement with the set of ALL model runs, including a greater warming in the tropical upper troposphere (individual model-OBS trend comparisons are shown in the supporting information). Whilst agreeing well with OBS in the troposphere, GHG shows much less cooling in the stratosphere consistent with ozone depletion (which is excluded from the GHG simulations) having made a substantial contribution to trends there. This conclusion is supported by observing that CNRM-CM5 alone of all the analyzed ALL runs has apparently unrealistic stratospheric ozone depletion, with ozone changes generated with its internal chemistry scheme [*Santer et al.*, 2012] instead of specified from *Cionni et al.* [2011]. The apparent consequence is that the CNRM-CM5 simulation has the least negative temperature trend of all the ALL simulations in the stratosphere globally (and indeed less negative than the observations) while agreeing with observations and other models throughout the troposphere. Ozone changes are also included in the GHG run, leaving its trends indistinguishable from those of its ALL run (see the supporting information). Note that the ozone database of *Cionni et al.* [2011] is not itself free of uncertainty [*Solomon et al.*, 2012].

[17] The NAT simulations show small trends in comparison to the ALL and GHG simulations. While aerosols injected into the stratosphere by explosive volcanic eruptions cool the troposphere and warm the stratosphere, they remain in place for only a few years at most and their effect on the climate is difficult to see in a multi-decadal trend such as shown in Figure 1.
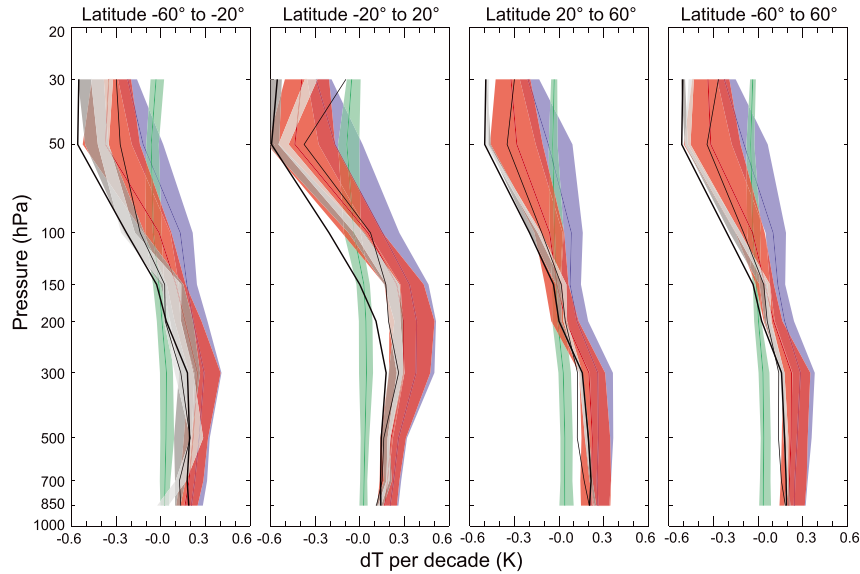
Figure 1. The 1961–2010 multimodel temperature trend by pressure level for the three studied zonal bands, plus the whole studied area for comparison. The shaded region shows the 5–95% range of the multimodel ensemble, with the central line showing the ensemble mean. Red represents all-forcings runs, green shows natural forcings, and blue is greenhouse-gas-forced only. The thick black line is HadAT2, thin black line is RAOBCORE 1.5, while the dark grey band is the RICH-obs 1.5 ensemble range, and light grey is the RICH-τ 1.5 ensemble range. Each band is displayed 25% translucent to better distinguish where forcings and observations overlap. Versions of this plot for each model are shown separately in the supporting information.

3.2. Temperature Time Series at 50 hPa, 500 hPa, and 200 hPa

[18] In this section, time series of observed and modeled temperatures at levels in the lower stratosphere (a sample at 50 hPa, show in Figure 2), the mid-troposphere (500 hPa, Figure 3), and the upper troposphere (200 hPa, Figure 4) are plotted and compared qualitatively. Quantitative comparison of models and observations is carried out in section 4 using optimal detection. Single-model equivalent plots are shown in supporting information. The time series plots confirm the trends seen in Figure 1—a warming at 500 hPa and a cooling at 50 hPa. Of all the data, the southern extra-tropics at 500 hPa is the region with the most ambiguity as to which of the three simulated forcing combinations provide the best match to the observations (Figure 3, bottom panel). Given the large radiosonde uncertainty in this region, the ranges of time series from all three sets of runs overlap with the spread of OBS (as do their trends), although NAT is more of an outlier. For other zones at 500 hPa, only the ranges of time series from ALL and GHG show much overlap with the observed range (Figure 3). At 50 hPa, only the ALL range overlaps with the observational range from the 1990s onward (Figure 2). This excludes CNRM-CM5, which has no simulations of any forcing agreeing with observations. This again indicates the importance of ozone depletion at this height, in line with previous studies [*Santer et al.*, 1995; *Tett et al.*, 1996].

[19] The effects of volcanic activity in Figure 2 (and to a lesser degree Figure 3) are more obvious than in Figure 1 and warming episodes of a time scale of a few years can

be seen at 50 hPa in the ALL and NAT simulations but not GHG (which does not include volcanic aerosols). The warming perturbation associated with the eruption of Pinatubo in the early 1990s acts to offset the overall cooling trend in that decade, giving a step change in the temperatures after the eruption in many models. This feature is also seen in the observations. As has previously been discussed by *Ramaswamy et al.* [2006] and *Thompson and Solomon* [2009], this step can be accentuated by the effect of volcanic emissions on ozone. Note that this effect is not included in the ozone data of *Cionni et al.* [2011], yet the ALL simulations which rely on its forcing data set still appear to reproduce the step observed (see also [*Santer et al.*, 2012]). The step change is seen less reliably in the northern hemisphere, and is not seen at all in the CNRM model with unrealistic ozone depletion. A corresponding cooling at 500 hPa following these volcanic eruptions is also seen in the NAT and ALL simulations.

[20] In previous studies [*Fu et al.*, 2011; *Thorne et al.*, 2011], it has been found that while there is agreement between models and radiosonde observations in much of the atmosphere, discrepancies between the two are found in the tropical upper troposphere. As described in section 3.1 and Figure 1, ALL temperature trends in this layer show greater agreement with the more recently developed RICH ensembles than with the trend calculated by the older approach of HadAT2. This is reinforced by the plot of the 200 hPa tropical temperature time series seen in Figure 4, with the RICH time series being within the ALL model spread for the majority of the years considered. It should also be noted, however, that the
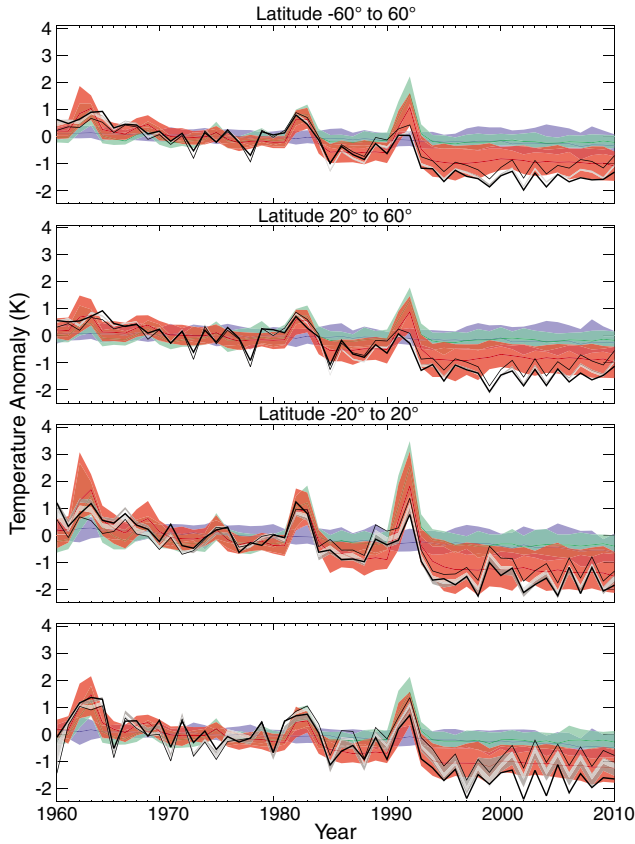
Figure 2. Time series of temperature anomaly in the 50 hPa level, averaged over three zonal bands and the whole region (top). Color codes for the individual model forcing and observations as in Figure 1. Each model is shown separately in the supporting information.

Figure 3. Temperature time series as in Figure 2, but for the 500 hPa level.

$$\mathbf{T}_{\text{obs}} - \boldsymbol{\varepsilon}_{\text{obs}} = \sum_{i=1}^{n} \beta_i (\mathbf{T}_i - \boldsymbol{\varepsilon}_i) \tag{1}$$

observations fall within the spread of NAT simulations (though the ALL and NAT trends systematically differ), and that once again the model uncertainty is considerably broader than the observational uncertainty. Outside of the tropical upper troposphere, at 200 hPa there are only small differences in trends between the different forcing runs. Thus with the current generation of models and observations, there is little that can be said for the causes of climate change at this level alone. However, by including the full set of levels, the need for all forcings (including those which are anthropogenic) to be included in model simulations for them to be consistent with observations becomes clear.

4. Optimal Detection

4.1. Detection Method

[21] This study applies the Total Least Squares regression approach to detection [*Allen and Stott*, 2003]. This assumes that the observations ($\mathbf{T}_{\text{obs}}$ in equation (1)) can be represented as a linear sum of simulated signals $\mathbf{T}_i$ (which are the estimated responses of the climate system to different forcing factors, each represented by a different subscript i) and internal climate variability $\boldsymbol{\varepsilon}$.
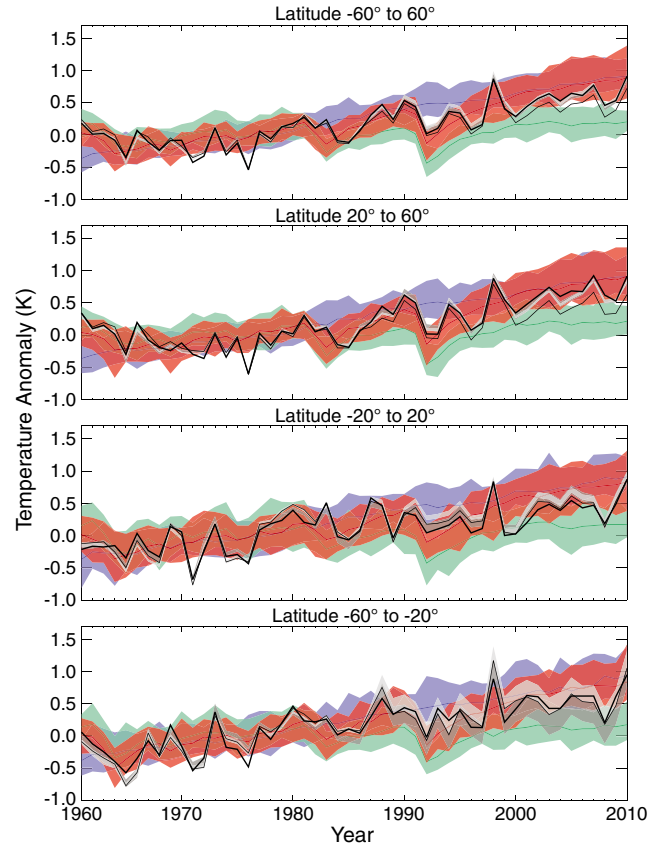
[22] This regression equation is then solved, to estimate the scaling factors β_i and their uncertainties (typically their 5 to 95 percentiles) according to the methodology outlined in *Allen and Stott* [2003]. The regression solution seeks to minimize the distance from the best fit regression line to the points in the regression, taking account both of noise in the observed signal $\boldsymbol{\varepsilon}_{\text{obs}}$ (the observed evolution being made up of a forced component and the effects of internal variability noise) and noise in the model signals $\boldsymbol{\varepsilon}_i$ (which can be reduced by averaging ensembles of simulations with different initial conditions but including identical forcing factors).

[23] This technique was applied by taking the time series for each pressure level and each latitude band, reducing the temporal resolution (with a time average of duration to be determined in section 4.2), then converting it into a one-dimensional column vector. This then represents the spatio-temporal pattern in temperature produced for the forcings present in a given model run. This is known as the fingerprint of a forcing, as it is this pattern which is sought in the observations as evidence of the presence of that forcing in the real world. To remove some of the noise arising from natural internal climate variations (as depicted in the model) associated with any given model run, the fingerprint is created from an ensemble average of simulations with that forcing.

[24] Regression coefficients derived by regressing onto the ALL, GHG, and NAT patterns are transformed into GHG,
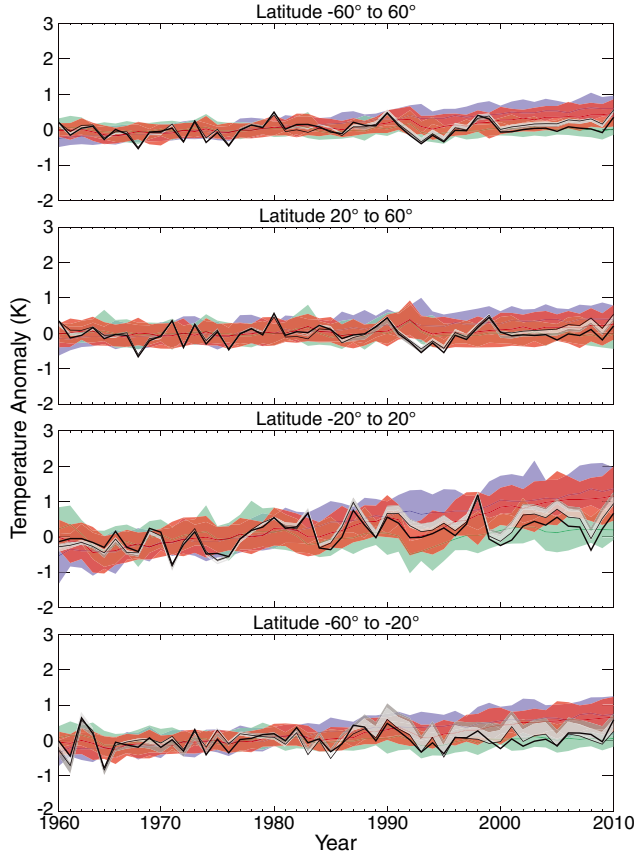
Figure 4. Temperature time series as in Figure 2, but for the 200 hPa level.

OAnt, and NAT regression coefficients following the approach of *Allen and Tett* [1999]. Regression coefficients derived by regressing onto ALL and NAT are similarly transformed into ANT and NAT regression coefficients, where ANT refers to all anthropogenic forcings, while OAnt is "other anthropogenic," i.e., anthropogenic with greenhouse gas forcings excluded. A version of part of the regression, simplified for clarity, is shown in equations (2) and (3), where **T** is a column-array of temperatures and the subscript obs denotes observational temperatures, while other subscripts denote the model forcings. β values are the scaling factors required for these forcings, while ε_{obs} represents the internal variability in the temperature observation. The other ε terms are the noise in the model runs, which should tend to zero as the number of ensemble members tends to infinity.

$$\mathbf{T}_{\text{obs}} = \beta_1(\mathbf{T}_{\text{ALL}} - \boldsymbol{\varepsilon}_{\text{ALL}}) + \beta_2(\mathbf{T}_{\text{NAT}} - \boldsymbol{\varepsilon}_{\text{NAT}})$$
$$+ \beta_3(\mathbf{T}_{\text{GHG}} - \boldsymbol{\varepsilon}_{\text{GHG}}) + \boldsymbol{\varepsilon}_{\text{obs}} \qquad (2)$$

$$\text{but} \quad \mathbf{T}_{\text{ALL}} = \mathbf{T}_{\text{GHG}} + \mathbf{T}_{\text{NAT}} + \mathbf{T}_{\text{OAnt}}$$

simplify, rearrange

$$\mathbf{T}_{\text{obs}} \approx (\beta_1 + \beta_3)\mathbf{T}_{\text{GHG}} + (\beta_1 + \beta_2)\mathbf{T}_{\text{NAT}} + \beta_1\mathbf{T}_{\text{OAnt}} + \boldsymbol{\varepsilon}_{\text{obs}}$$

$$i.e., \quad \beta_{\text{GHG}} = \beta_1 + \beta_3, \ \ \beta_{\text{NAT}} = \beta_1 + \beta_2, \ \ \beta_{\text{OAnt}} = \beta_1 \qquad (3)$$

$$\text{similarly if} \quad \mathbf{T}_{\text{ALL}} = \mathbf{T}_{\text{ANT}} + \mathbf{T}_{\text{NAT}}$$

$$\text{then} \quad \beta_{\text{ANT}} = \beta_1 \text{ and } \beta_{\text{NAT}} = \beta_1 + \beta_2$$

[25] The noise terms are usually estimated from control simulations without external influences on the climate. However, the quantity of control data available within the CMIP5 archive was lacking, so the following alternative technique was employed, with the control reserved for significance testing.

[26] The CMIP5 model runs with external forcings cover the period 1850–2012, but this study concentrates on changes during 1961–2010, being five complete decades available within the radiosonde period. This enabled a much more extensive estimate of internal variability using the technique known as intra-ensemble variability (IEV) [*Tett et al.*, 2002]. For each model ensemble with a given forcing, the data from each ensemble member can be subtracted from that of the ensemble average. Whilst taking account of the effect on the number of degrees of freedom and adjusting the variance accordingly [*Tett et al.*, 2002], many overlapping segments (of the same length as the period of interest, in this case 50 years) can be extracted to represent an ensemble of possible modes of variability. This can then be used to calculate the noise terms, with the model used to derive all ε terms, with each ε_i from equation (1) scaled appropriately, so that the noise contamination of an ensemble mean is reduced (the variance of the noise reducing by a factor m for the mean of an m-member ensemble).

[27] All data are projected onto empirical orthogonal functions (EOFs) of the data assembled from the IEV segments, and each component is weighted by the inverse standard deviation of each EOF. This weights the signal toward directions in this phase space in which the variance is low, following a methodology originally proposed by *Hasselmann* [1979], later shown by *Allen and Tett* [1999] to be a form of multivariate regression (as shown in equation (1)). This is equivalent to a filter which maximizes signal to noise. Because, as is usual, there are not enough data to accurately estimate the inverse covariance matrix needed to solve equations (1) and (2), its inverse is estimated from a truncated form based on the projection onto the leading EOFs [see, e.g., *Tett et al.*, 2002 for further discussion].

[28] Care needs to be taken in choosing the truncation. In particular, a truncation should not be chosen which is higher than the number of degrees of freedom in the data set, but the truncation should be high enough that a sufficiently high fraction of variance of the original data is explained by the truncated version. In addition, care must be taken not to include EOFs that would be given unrealistically large weight through the optimization, as these EOFs sample unrealistic aspects of model variability. This issue can be addressed through consideration of the residuals of regression as outlined by *Allen and Tett* [1999].

[29] The degrees of freedom offered by the amount of noise data, the fraction of observed variance explained, and the consistency between observation and control run variability as a function of truncation are all considered, to gain a greater understanding of the data and to determine a suitable truncation level. Equation (1) can be rearranged in terms of ε_{obs}, the residual variability in the observed temperatures. In this way, it can be the compared to the model variability exhibited in control runs using an F test, to determine if the two variances are significantly different and therefore if the solution of the regression equation is not a satisfactory explanation of the data [*Allen and Tett*, 1999]. For this test, the control runs

are processed into slices by the same method described above for IEV. Note that control run variability is used to maintain its independence from the previous calculation of observed variability, which uses IEV data. It thereby avoids artificial skill [*Hegerl et al.*, 1996]. Note also that although a significant quantity of data is required for the *F* test, it was found not to be as critical as the variability calculation. Consequently IEV data are used for the variability "noise," while control data are used for the *F* test, except where there is only one member of certain forcings (NorESM1-M, BCC-CSM1.1). In that case, insufficient IEV slices can be generated, so the roles of IEV and control are reversed.

4.2. Detection at Different Average Time Periods

[30] Retaining the zonal bands and levels used thus far, and considering the original time series rather than the trend analysis in section 3, the effectiveness of different averaging periods was investigated. Control period overlaps were kept at 10 years throughout and averaging periods were taken to have common factors with this. This led to 1, 2, 5, and 10 year averages being assessed.

[31] To aid in this assessment, the variance in the observations was calculated after it was masked and time averaged. This was then repeated once the observations were projected onto EOFs, for the range of truncations (essentially the cumulative sum of squares over the EOFs included). The resulting plots of the ratio of these variances can be seen in Figure 5. It is found that even with very long truncations, 1 and 2 year averages cannot reproduce much more than half of the observed variance. Thus, only 5 and 10 year averages should be considered.

[32] The next factor to be considered is the consistency of observed internal variability (as estimated from the residuals of the regression in equation (3)) with the control. This also

varies with truncation and can be similarly plotted, such as the example in Figure 6. As can be seen, a shorter averaging period increases the number of EOFs that can be included while still ensuring the residuals of regression are consistent with the model's control. The shorter averaging period also ensures that transient phenomena such as the stratospheric warming coincident with volcanic eruptions are captured. Therefore, 5 year means are used from this point onward.

[33] With an averaging period chosen, the optimal detection analysis can be carried out. Figure 7 shows the example of CSIRO Mk3.6.0 model data used against HadAT2 observations and the effect of using increasing EOF truncation. Examples for each model can be found in the supporting information along with an analysis of the ensemble members of all models together. For this multi-model analysis, there was a
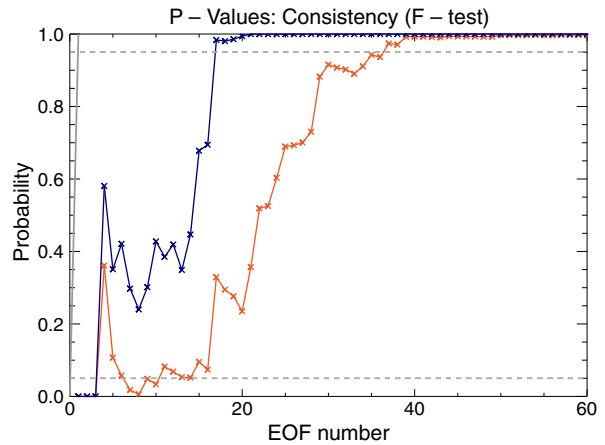

Figure 6. An example consistency test for CSIRO Mk3.6.0 versus HadAT2, performed for 5 year (orange) and 10 year (blue) averaging.
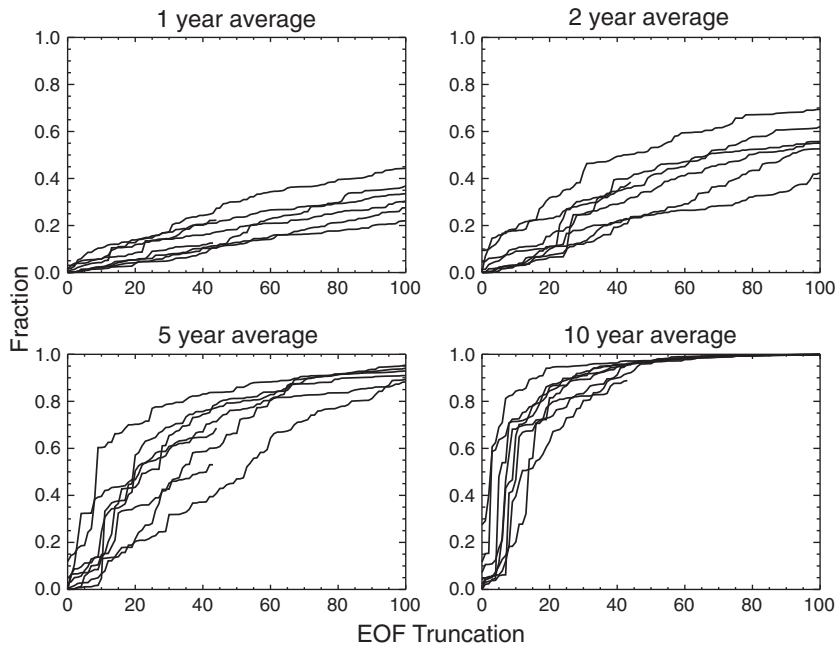

Figure 5. The fraction of observed variance (here from HadAT2) reproduced following projection onto EOFs based on model variability, for a selection of temporal averaging periods. Each trace is for a different model.
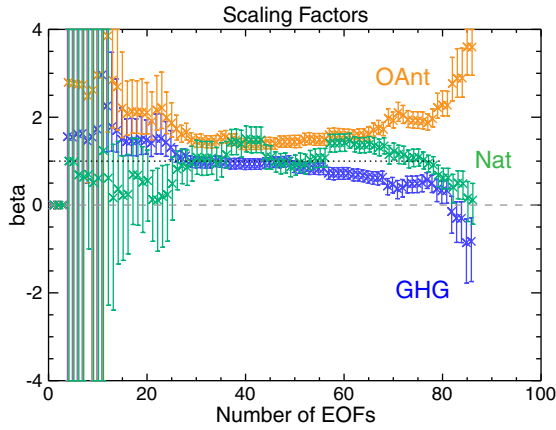
Figure 7. Scaling factors, with 5–95% confidence interval, as a function of EOF truncation for HadAT2 versus CSIRO Mk3.6.0 GHG (blue), NAT (green), and Other Anthropogenic (OAnt, yellow) forcings, all taken with 5 year averages.

choice between two principal methods: to consider all models equally by taking the same number of ensemble members from each or to consider all members equally irrespective of their model of origin [e.g., *Gillett et al.*, 2002]. Given that some of the CMIP5 models have a single ensemble member for certain forcings, it was considered that equally weighting the models would result in insufficient data, such that the exercise would not be worthwhile. Consequently the method of equal ensemble member likelihood was applied for this study, with no weighting given. Considering each member equally also removes the risk of selection bias, although clearly it adds prominence to any behavior seen in models with larger ensembles. In contrast, there were sufficient data to contribute an equal number of slices of control and IEV from each model. In the case where a model had more data than its required contribution, the higher-numbered slices were not contributed.

[34] Typically in optimal detection, the applicable scaling factors are chosen once they are found to vary little with truncation, while still retaining consistent *p*-values. Thus, in Figure 7, a truncation of around 35 would be selected, giving β values of around 1 for all three forcings. In the case of this paper however, the combination of eight models with four observational data sets (two of which are ensembles) means the technique to choose truncation should be automated to make it more objective. This was achieved by to the following criteria.

[35] 1. The default is the longest truncation where the *F* test probability (as in the example in Figure 6) is within the 5–95% range while exceeding a third of the number of degrees of freedom.

[36] 2. If no truncation meets these criteria, the truncation is taken as where observed variance explained by the EOFs (seen in Figure 5) exceeds 75%.

[37] 3. If these criteria are still not met, the truncation is selected as the number of degrees of freedom in the IEV.

[38] In this way, the very short and long truncations which are typically inconsistent with stable truncations (as seen in Figure 7) are automatically avoided. Any *F* test failures which are carried though in this algorithm are marked as

such in subsequent figures. Results in such cases should be treated with caution.

[39] For ensembles of observations, a detection analysis was conducted for each ensemble member in turn against the model in question, and its distribution of scaling factors output for the selected truncation. A composite distribution of all the members together was then created, giving the spread of scaling factors across all realizations of RICH. Each member is considered equally likely. The median and 5–95 percentiles of the composite distribution of all members together were computed as would normally be done for a single observation set. The above technique was applied to every combination of model and observation (be they a single set or an ensemble), then plot in Figure 8, annotated to show consistency and explained variance.

[40] With the forcings now separated into GHG, NAT, and OAnt (other anthropogenic and ozone), a number of points become clear. GHG forcings are almost always detected once sufficient ensemble members are included. Only NorESM1-M and BCC-CSM1.1 have any failure to detect, which is likely caused by poorly defined forcing signals due to only having one member for each GHG and NAT. Four models consistently attribute a change due to greenhouse gases (i.e., they have GHG scaling factors consistent with 1). Seven out of eight models also detect OAnt, although only four have combinations with observations with scaling factors consistent with 1, and always the minority of observation sets. CSIRO mk3.6.0 and HadGEM2-ES attribute OAnt with the RICH ensembles and HadAT2, while all the other model-OBS combinations
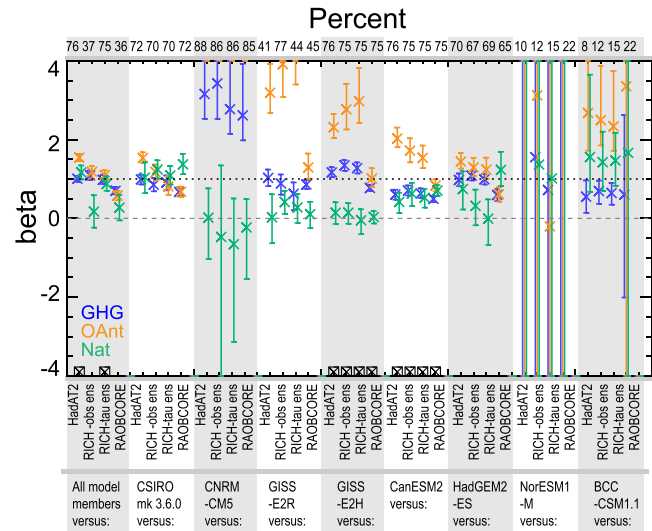

Figure 8. Scaling factors as in Figure 7 for each model (in order of decreasing number of ensemble members, as in Table 1), and for a multi-model analysis, against each observation data set. EOF truncation was automatically selected to be in a stable region with observed variability well represented. Model versus observation combinations marked with a boxed cross fail the consistency *F* test at that truncation. Explained variance is marked as a percentage above the graph.

(where OAnt fingerprints are detected but not attributed) need scaling up. This suggests the effect of other anthropogenic forcings is not strong enough in most models. Combined with the results of previous figures, realistic ozone forcing appears to be required. For example, when ozone depletion is included within the GHG runs in CNRM-CM5, a large scaling of both GHG and OAnt is required. In any case, the larger error bars on OAnt than on GHG are perhaps to be expected given that aerosol forcings have greater uncertainty than greenhouse gas forcing [Forster et al., 2007].

[41] The effect of natural forcings is more difficult to detect, although when a signal is detected, it is also consistent with 1 (with the exception of CanESM2). This difficulty is likely due to the variability being insufficiently represented, such that spikes due to volcanism are often obscured, either by temporal smoothing or by truncation of EOFs. Thus where, in half the models, natural effects can be detected, the detection is due to volcanic aerosols, not solar forcings. That said, solar forcing cannot be ruled out as a contributing factor. Just as an averaging period other than 5 years renders volcanic forcing invisible, it is possible that the 5 year period obscures solar forcing through aliasing of the solar cycle. However, given the length of the solar cycle is variable, it is difficult to choose an averaging period that does not have this problem.

[42] As might be expected, the analysis of the combined multi-model ensemble yields scalings with much smaller error bars than any of the individual simulations. It shows GHG scalings consistent with 1 against all observations excepting RAOBCORE, with two of the four analyses each showing attribution of NAT and/or OAnt. Conversely it is interesting to note that two have NAT consistent with zero at the chosen truncation. It is possible that the mixture of control runs from multiple simulations makes passing the consistency F test more difficult due to the variety of internal variability.

4.3. Isolated Troposphere

[43] One question that remains from the above analysis is whether the detection of an anthropogenic influence requires the pattern of contrast between a cooling stratosphere and a warming troposphere. To address this, the above analysis was repeated whilst masking the stratosphere as described in section 2.2, so that Figure 9 is based on tropospheric data alone.

[44] In general, Figure 9 shows similar results to Figure 8 although uncertainties increase with the loss of stratospheric information. GHG forcing is still detected for three models, NAT in four models, while OAnt remains almost always detected. The fact that only some models retain GHG detection without the stratosphere might suggest that those other models do indeed owe their detection to differences between stratosphere and troposphere. However, it is not true in all cases as might be implied by Legates and Davis [1997]. It is nonetheless evident that the addition of the stratosphere to the overall analysis gives a clearer picture (and thus, detection). Without the stratosphere, most models are more poorly constrained. This is consistent with results found over the satellite era by Santer et al. [2012].
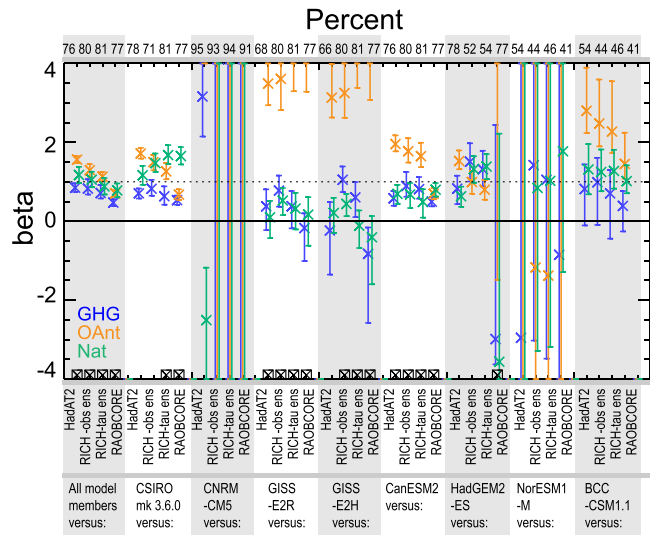

Figure 9. Scaling factors with 5 year averaging as in Figure 8 but for troposphere only.

5. Discussion and Conclusions

[45] Using the latest generation of climate models and observational data sets, this study has comprehensively explored the effects of modeling and observational uncertainty and detected the effects of anthropogenic greenhouse gases on free atmosphere temperature trends since 1961. The effects of other anthropogenic forcings are also detected, dominated by stratospheric ozone depletion, which can explain much of the lower stratospheric cooling observed over the last 50 years. While changes due to greenhouse gases and other anthropogenic forcings are detected in purely tropospheric data, detection of human influence is considerably strengthened on the additional inclusion of the stratosphere. The spread in possible temperature profiles across the latest observation data sets intersects at all levels of the atmosphere with that of the model runs with all forcings, with a much greater spread amongst the models than the observations.

[46] By eye it is unclear if there is any influence from solar forcing in the trends and time series, while the temperature responses to volcanic eruptions are far more clearly visible and are well reproduced by models. Using formal attribution techniques however, the combined effect of natural forcings can be detected in relatively few model-observation combinations, possibly due to the short duration of volcanic forcings and the difficulty in representing them with the combination of temporal averaging and EOF decomposition. Distinguishing between the effects of volcanic eruptions and changes in solar irradiance (as analyzed by Jones et al. [2003] with earlier simulations) requires the use of separate volcanic and solar forcing simulations in any models intended for detection and attribution in successor experiments to CMIP5. Expansion of the model archive could also help investigate (in addition to anthropogenic ozone changes) the different effects of natural ozone variability in the stratosphere, including response to solar irradiance variability [Ineson et al., 2011] and volcanic aerosol [Thompson and Solomon, 2009] which may not be included in the current generation of simulations.

[47] The CMIP5 experiments offer an opportunity to compare climate models and different approaches to their setup by their corresponding research groups. The absence of realistic ozone changes in CNRM-CM5, and the consequent mismatch between the trends in that model and the observations, is useful for emphasizing the importance of including realistic stratospheric ozone forcing in climate simulations. Elsewhere there are indications in the residuals of regression from optimal detection analyses that some models represent variability inadequately. This implies that detection of external forcings in these models should be treated with caution and that further attention should be given to diagnosing unrealistic atmospheric dynamics or chemistry. To investigate the importance of stratospheric processes to temperature trends lower in the atmosphere, *Mitchell et al.* [2013] compare between models such as those in this paper and those that extend higher into the stratosphere. It is found that significant differences in temperature trends manifest in the tropical lower stratosphere, and that, in the case of Met Office Hadley Centre models at least, the detection of external forcings is more robust in a model that extends into the stratosphere.

[48] For the study of phenomena outside the original remit of this paper, it would be useful to include the surface in these studies as in *Jones et al.* [2003]. Although this would cause some complications with interfacing free atmosphere and surface observations (especially for ensemble data sets), it would make it possible to compare not only temperature trends but also lapse rates (i.e., the ratio between free atmosphere and surface temperature changes). This analysis would be aided greatly by a substantial increase in the length of available control data, as would all further detection analyses of free atmosphere temperature trends. More simulations for each forcing also could substantially reduce uncertainties from internal variability. Analyses should also improve with time as the observational record extends. However, the uncertainty remains greater in the models than in observations, and it seems this must be the focus of development.

[49] In summary, this study has revisited the detection and attribution of radiosonde data, and its conclusions are in line of those of previous studies, with improvements in both models and observations having led to improved agreement at all levels of the free atmosphere, with uncertainties increasingly well quantified. This shows a clearly identifiable human influence in the temperature series as recorded by radiosondes since the mid-20th century.

References

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

Allen, M. R., and P. A. Stott (2003), Estimating signal amplitudes in optimal fingerprinting, part I: Theory, *Clim. Dyn.*, 21, 477–491, doi:10.1007/s00382-003-0313-9.

Cionni, I., *et al.* (2011), Ozone database in support of CMIP5 simulations: Results and corresponding radiative forcing, *Atmos. Chem. Phys.*, 11, 11267–11292. doi:10.5194/acp-11-11267-2011.

Cubasch, U., *et al.* (1997), Simulation of the influence of solar radiation variations on the global climate with an ocean-atmosphere general circulation model, *Clim. Dyn.*, 13, 757–767.

Dee, D. P., *et al.* (2011), The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Quart. J. Roy Meteor. Soc.*, 137, 553–597. doi:10.1002/qj.828.

Eyring, V., *et al.* (2013), Long-term ozone changes and associated climate impacts in CMIP5 simulations, *J. Geophys. Res.*, doi:10.1002/jgrd.50316.

Forster, P., *et al.* (2007), 2: Changes in atmospheric constituents and in radiative forcing, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to AR4 of the IPCC edited by S. Solomon, *et al.* pp. 663–745, CUP, Cambridge UK.

Fu, Q., *et al.* (2011), On the warming in the tropical upper troposphere: Models versus observations, *Geophys. Res. Lett.*, 38, L15704, doi:10.1029/2011GL048101.

Gillett, N. P., *et al.* (2002), Detecting anthropogenic influence with a multi-model ensemble, *Geophys. Res. Lett.*, 29, 1970, doi:10.1029/2002GL015836.

Gray, L. J., *et al.* (2010), Solar influences on climate, *Rev. Geophys.*, 48, RG4001, doi:10.1029/2009RG000282.

Hasselmann, K. (1979), On the signal-to-noise problem in atmospheric response studies, in Meteorology of Tropical Oceans, edited by D. B. Shaw, pp. 251–259, RMetS, London UK.

Haimberger, L., *et al.* (2012), Homogenization of the Global Radiosonde Temperature Dataset through Combined Comparison with Reanalysis Background Series and Neighboring Stations, *J. Clim.*, 25, 8108–8131, doi:10.1175/JCLI-D-11-00668.1.

Hegerl, G. C., *et al.* (1996), Detecting Greenhouse-Gas-Induced Climate Change with an Optimal Fingerprint Method, *J. Climate*, 9, 2281–2306.

Hegerl, G. C., *et al.* (2007), 9: Understanding and attributing climate change, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to AR4 of the IPCC, edited by S. Solomon, *et al.*, pp. 663–745, CUP, Cambridge UK.

Ineson, S., *et al.* (2011), Solar forcing of winter climate variability in the Northern Hemisphere, *Nat. Geosci.*, 4, 753–757.

Jones, G. S., *et al.* (2003), Causes of atmospheric temperature change 1960–2000: A combined attribution analysis, *Geophys. Res. Lett.*, 30, 1228.

Jones, G. S., *et al.* (2013), Attribution of observed historical near surface temperature variations to anthropogenic and natural causes using CMIP5 simulations, *J. Geophys. Res.*, (in press), doi:10.1002/jgrd.50239.

Karoly, D. J., *et al.* (1994), An example of fingerprint detection of greenhouse climate change, *Clim. Dyn.*, 10, 97–105.

Lanzante, J. R. (1996), Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data, *Int. J. Climatol.*, 16, 1197–1226.

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

Liu, X., *et al.* (2005), Ozone profile and tropospheric ozone retrievals from the Global Ozone Monitoring Experiment: Algorithm description and validation, *J. Geophys. Res.*, 110, D20307.

Mears, C. A., *et al.* (2011), Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique, *J. Geophys. Res.*, 116, D08112 doi:10.1029/2010JD014954.

Meehl, G. A., *et al.* (2007), The WCRP CMIP3 Multimodel dataset, a new era in climate change research, *Bull. Amer. Meteor. Soc.*, doi:10.1175/BAMS-88-9-1383.

Mitchell, D., *et al.* (2013), The Impact of Stratospheric Resolution on the Detectability of Climate Change Signals in the Free Atmosphere, *Geophys. Res. Lett.*, (in press), doi:10.1002/grl.50177.

Ramaswamy, V., *et al.* (2006), Anthropogenic and natural influences in the evolution of lower stratospheric cooling, *Science*, 311, 1138–1141, doi:10.1126/science.1122587.

Santer, B. D., *et al.* (1995), Towards the detection and attribution of an anthropogenic effect on climate, *Clim. Dyn.*, 12, 77–100.

Santer, B. D., *et al.* (2003), Contributions of anthropogenic and natural forcing to recent tropopause height changes, *Science*, 301, 479–483, doi:10.1126/science.1084123.

Santer, B. D., *et al.* (2011), Separating signal and noise in atmospheric temperature changes: The importance of timescale, *J. Geophys. Res.-Atmos.*, 116, D22105, doi:10.1029/2011JD016263.

Santer, B. D., *et al*. (2012), Identifying human influences on atmospheric temperature, *P. N. A. S*, (in press), doi:10.1073/pnas.1210514109.

Solomon, S., *et al*. (2012), Uncertainties in the evolution of stratospheric ozone and implications for recent temperature changes in the tropical lower stratosphere, *Geophys. Res. Lett.*, *39*, L17706, doi:10.1029/2012GL052723.

Taylor, K. E., *et al*. (2011), An overview of CMIP5 and the experiment design, *Bull. Amer. Meteor. Soc.*, doi:10.1175/BAMS-D-11-00094.1.

Tett, S. F. B., *et al*. (1996), Human Influence on the atmospheric vertical temperature structure: detection and observations, *Science*, *247*, 1170–1173, doi:10.1126/science.274.5290.1170.

Tett, S. F. B., *et al*. (2002), Estimation of natural and anthropogenic contributions to twentieth century temperature change, *J. Geophys. Res.*, *107*, 4306, doi:10.1029/2000JD000028.

Thompson, D. W. J., and S. Solomon (2009), Understanding recent stratospheric climate change, *J. Climate*, *22*, 1934–1943, doi:10.1175/2008JCLI2482.1.

Thorne, P. W., *et al*. (2002), Assessing the robustness of zonal mean climate change detection, *Geophys. Res. Lett.*, *29*, 1920, doi:10.1029/2002GL015717.

Thorne, P. W., *et al*. (2003), Probable causes of late twentieth century tropospheric temperature trends, *Clim. Dyn.*, *21*, 573–591, doi:10.1007/s00382-003-0353-1.

Thorne, P. W., *et al*. (2005), Revisiting radiosonde upper air temperatures from 1958 to 2002, *J. Geophys. Res.*, *110*, D18105, doi:10.1029/2004JD005753.

Thorne, P. W., *et al*. (2011), Tropospheric temperature trends: history of an ongoing controversy, *WIREs Climate Change*, *2*, 66–88, doi:10.1002/wcc.80.

Titchner, H. A., *et al*. (2009), Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments, *J. Clim.*, *22*, 465–485, doi:10.1175/2008JCLI2419.1.

Uppala, S. M., *et al*. (2005), The ERA-40 Reanalysis, *Quart. J. Roy Meteor. Soc.*, *131*, 2961–3012.