Multivariate trend comparisons between autocorrelated climate series with general trend regressors

Ross McKitrick
Department of Economics
University of Guelph
rmckitri@uoguelph.ca

Timothy J. Vogelsang
Department of Economics
Michigan State University
tjv@msu.edu

August 16, 2011

Abstract

Inference regarding trends in climatic data series, including comparisons across different data sets as well as univariate trend significance tests, is complicated by the presence of serial correlation and step-changes in the mean. We review recent developments in the estimation of heteroskedasticity and autocorrelation robust (HAC) covariance estimators as they have been applied to linear trend inference, with focus on the Vogelsang-Franses (2005) nonparametric approach, which provides a unified framework for trend covariance estimation robust to unknown forms of autocorrelation up to but not including unit roots, making it especially useful for climatic data applications. We extend the Vogelsang-Franses approach to allow general deterministic regressors including the case where a step-change in the mean occurs at a known date. Additional regressors change the critical values of the Vogelsang-Franses statistic. We derive an asymptotic approximation that can be used to simulate critical values. We also outline a simple bootstrap procedure that generates valid critical values and p-values. The motivation for extending the Vogelsang-Franses approach is an application that compares climate model generated and observational global temperature data in the tropical lower- and mid-troposphere from 1958 to 2010. Inclusion of a mean shift regressor to capture the Pacific Climate Shift of 1977 causes apparently significant observed trends to become statistically insignificant, and rejection of the equivalence between model generated and observed data trends occurs for much smaller significance levels (i.e. is more strongly rejected).

JEL Codes: C14; C32; C52; Q54
Keywords: Autocorrelation; trend estimation; HAC variance matrix; global warming; model comparisons

1 Introduction

Referring to Figure 1, do the two series have the same trend? A comparison of the simple linear trend coefficients might suggest they do, but clearly y_1 differs from y_2 in that the former is steadily trending while the latter is a trendless series with a single discrete step. If the comparisons were conducted over the pre-*s* or post-*s* intervals, they might indicate a significant difference of trends. In cases where a series is known to have undergone a step-change at a specific point in time, failure to account for it in the trend comparison model might lead to biased conclusions. Figure 1 shows a case in which such a test would overstate the agreement between the series, but other conceptual examples could be constructed in which the failure to account for a step-change could overstate the difference.

A key requirement for valid trend comparison methods is that they account for the autocorrelation properties of time series data and correlation between series. McKitrick et al. (2010) critique some standard methods that rely on a first order autoregression (AR1) specification. They recommend the multivariate trend method of Vogelsang and Franses (2005) (VF05) as a robust alternative and apply it to a model-observation comparison in the tropical troposphere. The trend model in that case is the simple form:

$$y_{it} = a_i + b_i t + u_{it}, \qquad (1)$$

where u_{it} , the random errors, is assumed to be covariance-stationary (in which case y_{it} is labeled a trend stationary series, that is, stationary upon removal of a linear trend, if one is present), i=1,…,n denotes the number of time series (different data sources), t = 1,…,T is the time period. Here we are interested in an extension that allows for a shift in the mean:

$$y_{it} = a_i + g_i DU_t + b_i t + u_{it}, \qquad (2)$$

where DU_t is an indicator variable that takes the value 0 prior to some cut-off date, T_b (the break date), and 1 thereafter. Hence, for series *i*, OLS estimation of (2) yields an estimated intercept of $\hat{a}_i$ prior to T_b and $\hat{a}_i + \hat{g}_i$ thereafter. Model (2) may be an appropriate specification for time series subject to changes in measuring equipment at a known point in time, such as when weather stations are moved from one location to another, or mercury thermometers are replaced with electronic ones; or when satellites in remote sensing applications are replaced, etc.

An important feature to note about model (2) is that we are assuming T_b is known ahead of time and is not estimated using the data set itself. When T_b is a parameter to be estimated, a search algorithm may be used in which a criterion such as sum of squared residuals is minimized. However it is not straightforward to determine if g_i is significant, since under the null hypothesis (represented by equation (1)) the parameter T_b is not identified. Andrews (1993) and Hansen (1996) discuss the asymptotic distribution of test statistics when the parameter measuring the change point is only identified under the alternative. Likewise we are assuming that there is known to be only a single value of T_b in the sample. If there can be many change points in the data and the break dates are known, then we simply add additional mean shift dummy variables to the model. If one thinks mean shifts occur frequently and with randomness, then there would be additional difficulties because the range of possible specifications could, in principle, include the case in which the mean changes by a random amount at each time step, which is equivalent to a random walk, or unit root process. If y_{it} has a unit root component, inference in models (1) and (2) becomes more complicated. More importantly, it is difficult to give a physical interpretation to a unit root component of a temperature series. See Mills (2010) for a discussion of temperature trend estimation when a random walk is a possible element of the specification.

The trend estimator used in this application is not applicable to data with unit root components, and we are only considering cases where there is known to be a single step-like change in the data of unknown magnitude. The particular example herein uses the Hadley and RICH radiosonde records (see next section) for the lower- and mid-troposphere levels in the tropics. The exogenous event is the 1977-78 Pacific Climate Shift, an oceanic circulatory system change during which basin-wide wind stress and sea surface temperature anomaly patterns reversed, causing an abrupt step-like change in many weather observations, including in the troposphere, as well as in other indicators such as fisheries catch records (see Seidel and Lanzante 2004, Powell Jr. and Xu 2011).

Estimation of trends using (1) is a routine calculation, and it may hardly seem possible that there is something new to be said on the subject, but in fact the last few years has seen some very useful methodological innovations for the purpose of computing robust confidence intervals, trend significance and trend comparisons in the presence of autocorrelation of unknown form.

Ordinary least squares (OLS) is the appropriate method for estimating the slope coefficient b_i. A common method to obtain valid coefficient variances is to fit an autoregressive-moving average (ARMA) error model with respective lag parameters *p* and *q* (e.g. Hamilton 1994). When applied carefully, the ARMA approach yields uncorrelated residuals and supports valid inferences. However there are some practical and theoretical limitations. First, even for a single series, the search process can be extremely cumbersome. There are, altogether, *p*+*q* possible lag coefficients, requiring the evaluation of 2^{p+q} models. Using daily or monthly data, where significant lags can extend over 12 months or more, this quickly becomes computationally infeasible. Second, a trend comparison may involve dozens of individual series, compounding the dimension problem. Imposing a simplifying assumption (such as AR1) as a practical remedy may lead to misspecification. Third, estimation of a complete ARMA(*p*,*q*) error model uses up *p*+*q* degrees of freedom, and if test statistics are near the significance boundary this may distort the results. Finally, the ARMA(*p*,*q*) model imposes a specific structure on the autocovariances, and in cases where the data exhibit long or complex forms of dependence, this may be restrictive.

Hence there is considerable benefit for researchers to become more familiar with the nonparametric variance estimator approach derived from spectral representations. These methods have found wide application in econometrics and finance, but are less used in applied climatic or geophysical papers although nonparametric approaches were proposed by Bloomfield and Nychka (1992) and further examined by Woodward and Gray (1993) and Fomby and Vogelsang (2002) for the univariate case. As far as we know, McKitrick et al. (2010) is the first empirical climate paper to use nonparametric variance estimation methods in multivariate settings.

The nonparametric approach turns out be relatively simple computationally, despite being based on rather complex underlying theory: for full treatments see Andrews (1991), Kiefer and Vogelsang (2005), Newey and West (1987), Sun, Phillips and Jin (2008) and White and Domowitz (1984). The main advantage is that a single specification is robust to general forms of autocorrelation (and heteroskedasticity) up to, but not including, nonstationarity.

2 Statistical Background and Motivation

To provide some background and intuition for those unfamiliar with robust inference methods in the presence of serial correlation in a multivariate setting, we focus on a model even more simple than model (1):

$$y_{it} = a_i + u_{it}. \tag{3}$$

It is assumed that u_{it} is a mean zero time series in which case $a_i = E(y_{it})$. For the purpose of matrix representations the natural organization is to denote rows by the time index *t* and columns by the data source index *i*. However the matrix representation of the statistical theory becomes easier if we transpose the data so that the columns represent time. We can then refer to time series of column vectors: $y_t = (y_{1t}, y_{2t},..., y_{nt})'$, $a = (a_1, a_2,..., a_n)'$ and $u_t = (u_{1t}, u_{2t},..., u_{nt})'$.

Rewrite the model as

$$y_t = a + u_t,$$

and suppose we are interested in testing linear restrictions about the means, a, of the form:

$$H_0 : Ra = r, \quad H_1 : Ra \neq r,$$

where R and r are, respectively, $q \times n$ and $q \times 1$ matrices of known constants. We require that $q \leq n$ and that R have full rank ($rank(R) = q$). The natural estimator of *a* is the vector of sample averages, i.e. the OLS estimator given by $\hat{a} = \bar{y} = T^{-1}\sum_{t=1}^{T} y_t$.

To understand the statistical properties of $\hat{a}$ and to derive robust tests of H_0, some assumptions about the mean zero time series vector, u_t, are needed. Assume that u_t is covariance stationary with $n \times n$ autocovariance matrices given by

$$\Gamma_j = E(u_t u'_{t-j}).$$

It is well known that $\Gamma_{-j} = \Gamma'_j$. In the case of $j = 0$, Γ_0 is the cross-section variance covariance matrix of the u_{it}.

Trivial algebra gives the relationship $\hat{a} = a + \bar{u}$ where $\bar{u} = T^{-1}\sum_{t=1}^{T} u_t$. Because u_t is mean zero, it obviously follows that $\hat{a}$ is an unbiased estimator: $E(\hat{a}) = a$. What is the variance of $\hat{a}$? Computing the variance-covariance matrix of $\hat{a}$ is straightforward given the covariance stationarity assumption:

$$Var(\hat{a}) = E[(\hat{a} - a)(\hat{a} - a)'] = E[\bar{u}\bar{u}'] = T^{-2}E[(\sum_{t=1}^{T} u_t)(\sum_{t=1}^{T} u_t)'].$$

The product of the sums inside the expectation can be organized by how far apart in time elements of the first sum are from elements in the second sum. There are T terms of the form $u_t u_t'$, which have expectation $E(u_t u_t') = \Gamma_0$, $T-1$ terms of the form $u_t u_{t-1}'$ (for $t = 2,3,...,T$) with expectation $E(u_t u_{t-1}') = \Gamma_1$, $T-1$ terms of the form $u_t u_{t+1}'$ with expectation $E(u_t u_{t+1}') = \Gamma_{-1} = \Gamma_1'$. In general for j $j = 1,2,...,T-1$, Γ_j and Γ_j' appear $T-j$ times in the variance formula leading to

$$\begin{aligned}
Var(\hat{a}) &= T^{-2}E[(\sum_{t=1}^{T} u_t)(\sum_{t=1}^{T} u_t)'] \\
&= T^{-2}\left[T\Gamma_0 + (T-1)(\Gamma_1 + \Gamma_1') + (T-2)(\Gamma_2 + \Gamma_2') + ... + (T-(T-1))(\Gamma_{T-1} + \Gamma_{T-1}')\right] \\
&= T^{-1}\left[\Gamma_0 + (1-\frac{1}{T})(\Gamma_1 + \Gamma_1') + (1-\frac{2}{T})(\Gamma_2 + \Gamma_2') + ... + (1-\frac{T-1}{T})(\Gamma_{T-1} + \Gamma_{T-1}')\right] \\
&= T^{-1}\left[\Gamma_0 + \sum_{j=1}^{T-1}(1-\frac{j}{T})(\Gamma_j + \Gamma_j')\right].
\end{aligned}$$

Letting $\Omega_T = \Gamma_0 + \sum_{j=1}^{T-1}(1-\frac{j}{T})(\Gamma_j + \Gamma_j')$ we have the more compact expression

$$Var(\hat{a}) = T^{-1}\Omega_T. \tag{4}$$

If one were willing to make the strong assumption that u_t is a vector of normally distributed random variables, then it directly follows that $\hat{a}$ is normally distributed:

$$\hat{a} \sim N(a, T^{-1}\Omega_T),$$

and under H_0 it follows that

$$R\hat{a} - r = R\hat{a} - Ra = R(\hat{a} - a) \sim N(0, T^{-1}R\Omega_T R').$$

One could test H_0 using the *infeasible* F statistic

$$F_{\inf} = (R\hat{a} - r)'[T^{-1}R\Omega_T R']^{-1}(R\hat{a} - r)/q.$$

This F-statistic is infeasible because Ω_T is unknown. Because the numerator of $F_{\inf}$ is a quadratic form involving a $q \times 1$ vector of mean zero normal random variables and the inverse of the vector's variance-covariance matrix, we obtain the result that under H_0, $F_{\inf} \sim \chi_q^2/q$ where

χ_q^2 is a chi-square random variable with q degrees of freedom. The null hypothesis would be rejected at the α significance level if $F_{\inf} > cv_\alpha$ where cv_α is the right tail critical value from a χ_q^2/q random variable.

To make this F-statistic feasible, a proxy (or estimator) is needed for Ω_T. A natural estimator of Ω_T is given by

$$\hat{\Omega}_T = \hat{\Gamma}_0 + \sum_{j=1}^{T-1}(1-\frac{j}{T})(\hat{\Gamma}_j + \hat{\Gamma}_j'), \quad \hat{\Gamma}_j' = T^{-1}\sum_{t=j+1}^{T}\hat{u}_t\hat{u}_{t-j}', \tag{5}$$

where $\hat{u}_t = y_t - \hat{a}.$ Using $\hat{\Omega}_T$ in place of Ω_T leads to the F-statistic proposed by VF05:

$$VF = (R\hat{a}-r)'[T^{-1}R\hat{\Omega}_T R']^{-1}(R\hat{a}-r)/q. \tag{6}$$

Because $\hat{\Omega}_T$ is constructed without assuming a specific model of serial correlation, $\hat{\Omega}_T$ is in the class of nonparametric spectral estimators of Ω.

Obviously, $\hat{\Omega}_T$ is a relatively complicated function of the data, and it is very difficult to characterize the exact distribution of $\hat{\Omega}_T$ or VF even if one is willing to make the strong assumption that u_t is normally distributed. Instead, asymptotic theory is used to generate an approximation for $\hat{\Omega}_T$ and the null distribution of VF. The key tool in obtaining an asymptotic approximation for VF is a functional central limit theorem (FCLT) for the scaled partial sums of u_t. A FCLT is an extension of the ideas behind the more familiar central limit theorem (CLT). Recall that $\hat{a} = a + \bar{u}$ in which case we have $\hat{a} - a = \bar{u}$. Scaling by $\sqrt{T}$ gives

$$\sqrt{T}(\hat{a}-a) = \sqrt{T}\bar{u}.$$

Under some regularity conditions, if $\sum_{j=0}^{\infty}\left|\Gamma_j^{(lm)}\right| < \infty$ where $\Gamma_j^{(lm)}$ is the l,m element of the matrix Γ_j, a CLT holds for $\bar{u}$:

$$\sqrt{T}\bar{u} = T^{-1/2}\sum_{t=1}^{T}u_t \xrightarrow{d} N(0,\Omega), \tag{7}$$

where $\xrightarrow{d}$ denote convergence in distribution and $\Omega = \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j')$. The matrix Ω is the asymptotic variance of $\sqrt{T}\bar{u}$ and is often called the long run variance of u_t. Ω is also directly related to the zero frequency spectral density matrix of u_t. Using the CLT delivers a useful result for $(R\hat{a} - r)$ when H_0 is true:

$$\sqrt{T}(R\hat{a} - r) = \sqrt{T}(R\hat{a} - Ra) = R\sqrt{T}(\hat{a} - a) = R\sqrt{T}\bar{u} \xrightarrow{d} RN(0,\Omega) \sim N(0, R\Omega R')$$

This result in turn leads to the approximation

$$(R\hat{a} - r) \approx N(0, T^{-1}R\Omega R').$$

If it were the case that $\hat{\Omega}_T$ were a consistent estimator of Ω, then VF would converge in distribution to a χ_q^2/q random variable and the same critical value would be used for VF as for $F_{\inf}$. It turns out that $\hat{\Omega}_T$ is not a consistent estimator of Ω and at first glance this would seem make the VF statistic useless in practice. However, it is relatively easy to show that while $\hat{\Omega}_T$ is not a consistent estimator of Ω, it does converge in distribution to a random matrix that is proportional to Ω but otherwise does not depend on unknown quantities. This property of $\hat{\Omega}_T$ means that the VF statistic can be used to test H_0 because VF can be approximated by a random variable that does not depend on unknown parameters.

It is in establishing the limit of $\hat{\Omega}_T$ that the FCLT plays a key role. A FCLT is, intuitively, a collection of CLTs for sums of u_t indexed by the proportion of data used to construct the sums. Define the partial sum time series as the summation of u_t up to time *t*:

$$S_t = \sum_{j=1}^{t} u_j.$$

Take a real number c from the interval $[0,1]$ and let $[cT]$ denote the integer part of cT. The observations $t = 1,2,...,[cT]$ comprise the first c^{th} proportion of the data set. If we evaluate S_t at $t = [cT]$, we have

$$S_{[cT]} = \sum_{t=1}^{[cT]} u_t,$$

which is the sum of the first c^{th} proportion of the data. For a given value of *c*, the quantity $[cT] \to \infty$ as $T \to \infty$; therefore if we scale $S_{[cT]}$ by $[cT]^{-1/2}$ we can apply the CLT to obtain

$$[cT]^{-1/2} S_{[cT]} \xrightarrow{d} N(0,\Omega).$$

Alternatively, we if scale by $T^{-1/2}$ we obtain the result

$$T^{-1/2} S_{[cT]} = \left(\frac{[cT]}{T} \right)^{1/2} [cT]^{-1/2} S_{[cT]} \xrightarrow{d} c^{1/2} N(0,\Omega) = N(0, c\Omega).$$

For a given value of *c*, the scaled partial sums of u_t satisfy a CLT. These limits hold pointwise in *c*. The FCLT is a stronger statement that says this collection of CLTs, as indexed by *c,* hold jointly and uniformly in *c* and that the family of limiting normal random variables are in fact a well known stochastic process called a Wiener process (or standard Brownian motion). Not surprisingly, the FCLT requires slightly stronger assumptions for u_t than a CLT. For example, the condition $\sum_{j=0}^{\infty} \left| \Gamma_j^{(lm)} \right| < \infty$ is strengthened to $\sum_{j=1}^{\infty} j \left| \Gamma_j^{(lm)} \right| < \infty$ which requires the autocovariances to shrink faster to zero as j increases.

For the remainder of the paper, we assume that a FCLT holds for $T^{-1/2} S_{[cT]}$ which we write as

$$T^{-1/2} S_{[cT]} \Rightarrow \Lambda W_n(c), \tag{8}$$

where $\Rightarrow$ denotes weak convergence in distribution, Λ is matrix square root of Ω, i.e. $\Omega = \Lambda\Lambda'$ and $W_n(c)$ is an $n \times 1$ vector of Wiener processes that are independent of each other. For a given value of *c*, $W(c) \sim N(0, cI_n)$ where I_n is an $n \times n$ identity matrix. Wiener processes are correlated across *c* but have independent increments (non-overlapping differences in $W(c)$ are independent). Essentially $W(c)$ is a vector of continuous time random walks. Because the FCLT is a stronger result than the CLT, the result in (7) directly follows from the FCLT:

$$\sqrt{T}(\hat{a} - a) = \sqrt{T}\,\overline{u} = T^{-1/2} \sum_{t=1}^{T} u_t = T^{-1/2} S_T \Rightarrow \Lambda W_n(1) \sim N(0, \Lambda I_n \Lambda') = N(0,\Omega). \tag{9}$$

Using the FCLT, it is straightforward to determine the asymptotic behavior of $\hat{\Omega}_T$. The first step is to write $\hat{\Omega}_T$ as a function of $\hat{S}_t = \sum_{j=1}^{t} \hat{u}_j$. It has been shown by Kiefer and Vogelsang (2002) that equation (5) can be simplified as

$$\hat{\Omega}_T = \hat{\Gamma}_0 + \sum_{j=1}^{T-1}(1 - \frac{j}{T})(\hat{\Gamma}_j + \hat{\Gamma}_j') = 2T^{-2}\sum_{t=1}^{T-1}\hat{S}_t\hat{S}_t' \ . \qquad (10)$$

Note that formula (10) requires that $\hat{S}_T = 0$ which holds as long intercepts are included in the model. Using the FCLT, the limit of $T^{-1/2}\hat{S}_{[cT]}$ is easy to derive:

$$\begin{aligned}
&T^{-1/2}\hat{S}_{[cT]} = T^{-1/2}\sum_{t=1}^{[cT]}\hat{u}_t = T^{-1/2}\sum_{t=1}^{[cT]}(y_t - \hat{a}) = T^{-1/2}\sum_{t=1}^{[cT]}(a + u_t - \hat{a}) \\
&= T^{-1/2}\sum_{t=1}^{[cT]}u_t - T^{-1/2}[cT](\hat{a} - a) = T^{-1/2}S_{[cT]} - \left(\frac{[cT]}{T}\right)\sqrt{T}(\hat{a} - a) \\
&\Rightarrow \Lambda W_n(c) - r\Lambda W_n(1) = \Lambda(W_n(c) - cW_n(1)) \equiv \Lambda B_n(c).
\end{aligned}$$

The stochastic process, $B_n(c)$, is the well known Brownian bridge. Using this result for $T^{-1/2}\hat{S}_{[cT]}$ and the continuous mapping theorem, it follows that

$$\hat{\Omega}_T = 2T^{-1}\sum_{t=1}^{T-1}(T^{-1/2}\hat{S}_t)(T^{-1/2}\hat{S}_t') \Rightarrow 2\Lambda\int_0^1 B_n(c)B_n(c)'dc\Lambda'.$$

We see that while $\hat{\Omega}_T$ does not converge to $\Omega = \Lambda\Lambda'$, it does converge to a random matrix that is proportional to $\Lambda\Lambda'$.

Establishing the limit of VF is now simple:

$$\begin{aligned}
VF &= (R\hat{a} - r)'[T^{-1}R\hat{\Omega}_T R']^{-1}(R\hat{a} - r)/q \\
&= \sqrt{T}(R\hat{a} - r)'[R\hat{\Omega}_T R']^{-1}\sqrt{T}(R\hat{a} - r)/q \\
&= (R\sqrt{T}\bar{u})'[T^{-1}R\hat{\Omega}_T R']^{-1}R\sqrt{T}\bar{u}/q \\
&\Rightarrow (R\Lambda W_n(1))'[R2\Lambda\int_0^1 B_n(c)B_n(c)'dc\Lambda'R']^{-1}R\Lambda W_n(1)/q.
\end{aligned}$$

While not obvious at first glance, the restriction matrix, R, drops from the limit. Because Wiener processes are Gaussian (normally distributed), linear combinations of Wiener processes are also Wiener processes. Therefore, $R\Lambda W_n(c)$ is a $q\times 1$ vector of Wiener processes and we can rewrite $R\Lambda W_n(c)$ as $\Lambda^* W_q(c)$ where Λ^* is the $q\times q$ matrix square root of $R\Lambda\Lambda'R'$, i.e. $\Lambda^*\Lambda^{*'} = R\Lambda\Lambda R'$. Similarly, we can rewrite $R\Lambda B_n(c)$ as $\Lambda^* B_q(c)$ where $B_q(c) = W_q(c) - cW_q(1)$. Because R is assumed to be full rank, it follows that Λ^* is full rank and is therefore invertible. We have

$$VF \Rightarrow (R\Lambda W_n(1))'[R2\Lambda\int_0^1 B_n(c)B_n(c)'dc\Lambda'R']^{-1}R\Lambda W_n(1)/q$$
$$= (\Lambda^* W_q(1))'[2\Lambda^*\int_0^1 B_q(c)B_q(c)'dc\Lambda^{*'}]^{-1}\Lambda^* W_q(1)/q$$
$$= W_q(1)'[2\int_0^1 B_q(c)B_q(c)'dc]^{-1}W_q(1)/q,$$

and the Λ^* matrices drop out because Λ^* is invertible.

The limit of VF does not depend on unknown parameters. The limit is a quadratic form involving a vector of independent standard normal random variables, $W_q(1)$, and the inverse of the random matrix $2\int_0^1 B_q(c)B_q(c)'dc$. Because $W_q(1)$ is independent of $B_q(c)$ for all *c*, $W_q(1)$ is independent of $2\int_0^1 B_q(c)B_q(c)'dc$ and the limit of VF is similar in spirit to an F random variable but its distribution is nonstandard. The random matrix $2\int_0^1 B_q(c)B_q(c)'dc$ can be viewed as an approximation to the randomness of $R\hat{\Omega}_T R'$ whereas $W_q(1)$ approximates the randomness of $\sqrt{T}(R\hat{a}-r)$. Because the asymptotic distribution of VF is nonstandard, asymptotic critical values need to be computed using numerical methods. We discuss two methods in the next section.

3 Extension of the VF Approach

3.1 Statistical Model and Test Statistics

As will become clear in the subsequent discussion, the limiting behavior of $\hat{\Omega}_T$, and hence the VF statistic, depends on the deterministic trend regressors included in the model. VF05 analyzed model (1) but those results do not directly apply to models (2) or (3). In this section we extend the VF05 approach to a more general setting that include models (1), (2) and (3) as special cases.

We consider the model

$$y_{it} = \beta_i d_{0t} + \delta_i' d_{1t} + u_{it}, \qquad (11)$$

where d_{0t} is a single deterministic regressor and d_{1t} is a $k \times 1$ vector of additional deterministic regressors. Defining the $k \times n$ matrix $\delta = (\delta_1, \delta_2, ..., \delta_n)$, model (7) can be written in vector notation as

$$y_t = \beta d_{0t} + \delta' d_{1t} + u_t. \tag{12}$$

Notice that model (1) is obtained for $d_{0t} = t$, $\beta_i = b_i$ and $d_{1t} = 1$, $\delta_i = a_i$, model (2) is obtained for $d_{0t} = t$, $\beta_i = b_i$ and $d_{1t} = (1, DU_t)'$, $\delta_i = (a_i, g_i)'$ and model (3) is obtained for $d_{0t} = 1$, $\beta_i = a_i$ and $d_{1t} = 0$.

Note that we are assuming that each time series has the same deterministic regressors. This is needed for the VF statistic to be robust to unknown heteroskedasticity and serial correlation. In some applications it might be reasonable to model some of the series as having different trend functions. In that case, we can simply include in d_{1t} the union of trend regressors across all the series. This will result in a loss of degrees of freedom but in many applications the regressors will be similar across series. So, the loss in degrees of freedom will often be small and this is a small price to pay for robustness to heteroskedasticity and autocorrelation.

As before, the model is estimated by OLS equation by equation. Because the parameters of interest are the vector β, we express the OLS estimator of β using the "partialling out" result, aka the Frisch-Waugh result (see Davidson and MacKinnon, 1993 and Wooldridge, 2005) as follows. Let $\tilde{d}_{0t}$ denote the OLS residuals from the regression of d_{0t} on d_{1t}. The OLS estimator of β can be expressed as

$$\hat{\beta} = \left(\sum_{t=1}^{T} \tilde{d}_{0t}^{2} \right)^{-1} \sum_{t=1}^{T} \tilde{d}_{0t} y_t, \tag{13}$$

and it follows directly that

$$\hat{\beta} - \beta = \left(\sum_{t=1}^{T} \tilde{d}_{0t}^{2} \right)^{-1} \sum_{t=1}^{T} \tilde{d}_{0t} u_t.$$

The OLS residuals for model (12) can be written as

$$\hat{u}_t = y_t - \hat{\beta} d_{0t} - \hat{\delta}' d_{1t}, \tag{14}$$

where $\hat{\beta}$ and $\hat{\delta}$ are the OLS estimators of β and δ using OLS equation by equation. Let $\hat{\Omega}_T$ be defined as before using (10) but with $\hat{u}_t$ given by (14). The VF statistic for testing $H_0 : R\beta = r$ is given by

$$VF = (R\hat{\beta}-r)'\left[\left(\sum_{t=1}^{T}\tilde{d}_{0t}^{2}\right)^{-1} R\hat{\Omega}_T R'\right]^{-1}(R\hat{\beta}-r)/q. \tag{15}$$

3.2 Asymptotic Approximations

In this section we derive the asymptotic limit of *VF* which will provide an approximation that can be used to generate critical values. We continue to assume that the scaled partial sums of u_t follow the FCLT given by (8) and we need to make some assumptions about the deterministic trend regressors. To that end, assume that there is a scalar, τ_{0T}, and a $k\times k$ matrix, τ_{1T}, such that

$$T^{-1}\tau_{0T}\sum_{t=1}^{[cT]} d_{0t} \to \int_0^c f_0(s)ds, \qquad T^{-1}\tau_{1T}\sum_{t=1}^{[cT]} d_{1t} \to \int_0^c f_1(s)ds.$$

For example, for model (2) $d_{0t}=t$, $\tau_{0T}=T^{-1}$, $f_0(s)=s$ and $d_{1t}=(1,DU_t)'$, $\tau_{1T}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$, $f_1(s)=(1,\mathbf{1}(s>\lambda))'$ where $\lambda = T_b/T$ and $\mathbf{1}(s>\lambda)$ equals 1 for $s>\lambda$ and 0 otherwise. Define the function

$$\tilde{f}_0(c) = f_0(c) - \left(\int_0^1 f_0(s)f_1(s)'ds\right)\left(\int_0^1 f_1(s)f_1(s)'ds\right)^{-1} f_1(c).$$

It it easy to show that

$$T^{-1}\tau_{0T}\sum_{t=1}^{[cT]}\tilde{d}_{0t} \to \int_0^c \tilde{f}_0(s)ds, \qquad T^{-1}\tau_{0T}^{2}\sum_{t=1}^{T}\tilde{d}_{0t}^{2} \to \int_0^1 \tilde{f}_0^{2}(s)ds.$$

In writing down the limit of $\hat{\Omega}_T$ it is covenient to stack the deterministic regressors into a single column vector d_t where $d_t' = (d_{0t}, d_{1t}')$. Define the combined scaling matrix

$$\underset{(k+1)\times(k+1)}{\tau_T} = \begin{bmatrix}\tau_{0T} & \underset{1\times k}{\mathbf{0}}\\ \underset{k\times 1}{\mathbf{0}} & \tau_{1T}\end{bmatrix}.$$

It immediately follows that

$$T^{-1}\tau_T\sum_{t=1}^{[cT]} d_t \to \int_0^c f(s)ds,$$

where $f(s)' = (f_0(s), f_1(s)')$.

Using similar but slightly more complicated algebra as in the previous section, we obtain the following results:

$$\sqrt{T}\tau_{0T}^{-1}(R\hat{\beta}-r)\xrightarrow{d}\Lambda^*\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1}\left(\int_0^1\tilde{f}_0(s)dW_q(s)\right)$$

and

$$RT^{-1/2}\hat{S}_{[cT]}\Rightarrow\int_0^c\Lambda^* dW_q(s)-\left(\int_0^1\Lambda^* dW_q(s)f(s)'\right)\left(\int_0^1 f(s)f(s)'ds\right)^{-1}\int_0^c f(s)ds\equiv\Lambda^* B_q^f(c),$$

where

$$B_q^f(c)=\int_0^c dW_q(s)-\left(\int_0^1 dW_q(s)f(s)'\right)\left(\int_0^1 f(s)f(s)'ds\right)^{-1}\int_0^c f(s)ds,$$

and

$$R\hat{\Omega}_T R'\xrightarrow{d}2\Lambda^*\int_0^1 B_q^f(s)B_q^f(s)'ds\Lambda^{*'}.$$

Combining these results gives the limit of VF :

$$VF=\sqrt{T}\tau_{0T}^{-1}(R\hat{\beta}-r)'\left[\left(T^{-1}\tau_{0T}^2\sum_{t=1}^{T}\tilde{d}_{0t}^2\right)^{-1}R\hat{\Omega}_T R'\right]^{-1}\sqrt{T}\tau_{0T}^{-1}(R\hat{\beta}-r)/q$$

$$\xrightarrow{d}\left(\Lambda^*\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1}\left(\int_0^1\tilde{f}_0(s)dW_q(s)\right)\right)'\left[\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1}2\Lambda^*\int_0^1 B_q^f(s)B_q^f(s)'ds\Lambda^{*'}\right]^{-1}$$

$$\times\left(\Lambda^*\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1}\left(\int_0^1\tilde{f}_0(s)dW_q(s)\right)\right)/q$$

$$=\left(\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1/2}\left(\int_0^1\tilde{f}_0(s)dW_q(s)\right)\right)'\left[2\int_0^1 B_q^f(s)B_q^f(s)'ds\right]^{-1}\left(\int_0^1\tilde{f}_0(s)^2ds\right)^{-1/2}\left(\int_0^1\tilde{f}_0(s)dW_q(s)\right)/q.$$

Using well known properties of Wiener processes, it follows that

$$\left(\int_0^1 \tilde{f}_0(s)^2 ds\right)^{-1}\left(\int_0^1 \tilde{f}_0(s) dW_q(s)\right) = Z_q \sim N(\mathbf{0}, I_q),$$

which allows us to write

$$VF \xrightarrow{d} Z_q' \left[2\int_0^1 B_q^f(s) B_q^f(s)' ds\right]^{-1} Z_q/q. \tag{16}$$

It can be shown that the normal vector, Z_q, is independent of the random matrix, $2\int_0^1 B_q^f(s) B_q^f(s)' ds$. Therefore, the limit of VF is similar to an F random variable but is nonstandard and depends on the deterministic regressors in the model via the $B_q^f(s)$ stochastic process. The critical values of VF depend on what regressors are included in d_t but do not depend on which regressor is placed in d_{0t} (the regressor of interest for hypothesis testing). For example, one uses the same critical values for testing the equality of trend slopes or testing the equality of intercepts or testing the equality of intercept shifts in model (2).

In the case where one restriction is being tested, $q = 1$, we can define a t-statistic as

$$VF_t = \frac{R\hat{\beta} - r}{\sqrt{\left(\sum_{t=1}^{T} \tilde{d}_{0t}^2\right)^{-1} R\hat{\Omega}_T R'}} \tag{17}$$

and its limit is given by

$$VF_t \xrightarrow{d} \frac{Z_1}{\sqrt{2\int_0^1 B_1^f(s) B_1^f(s)' ds}}. \tag{18}$$

The VF_t statistic can be used to test one-sided hypotheses. Using VF_t to test two-sided hypotheses is exactly equivalent to using VF .

Obtaining the critical values of the nonstandard asymptotic random variables defined by (16) and (18) is straightforward using Monte Carlo simulation methods that are widely used in the econometrics and statistics literatures. In the case of model (2), the location of the mean shift, λ, affects the form of d_t and hence the form of $f(s)$. Therefore, the location of the mean shift affects the asymptotic critical values of VF and VF_t. In the application $\lambda = 0.358$. For this case we simulated the asymptotic critical values of VF and VF_t for testing one restriction ($q = 1$)

which we tabulate in Table 1. The Wiener process that appears in the limiting distribution is approximated by the scaled partial sums of 1,000 i.i.d. N(0,1) random deviates. The vector $f(s)$ is approximated using $(1,\ \mathbf{1}(t > 0.358T),\ t/T)'$ for t=1,2,…,T. The integrals are approximated by simple averages. 50,000 replications were used. We see from Table 1 that the tails of the VF_t statistic are fatter than the tails of a standard normal random variable and the right tail of the VF statistic has fatter tails than a χ_1^2 random variable.

3.3 Bootstrap Critical Values and p-values

What should an empirical practitioner do when critical values are needed for other specifications of the trend function? If carrying out simulations of the asymptotic distributions is not easily accomplished using standard statistical packages, an alternative is to use a simple bootstrap approach as follows:

1. For each i take the OLS residuals, $\hat{u}_{it}$, from (11) (see (14)) and sample with replacement from $\hat{u}_{i1}, \hat{u}_{i2}, \ldots, \hat{u}_{iT}$ to generate a bootstrap series $\hat{u}_{i1}^*, \hat{u}_{i2}^*, ..., \hat{u}_{iT}^*$. Let $y_{it}^* = \hat{u}_{it}^*$ denote a bootstrap resampled series for y_{it}.

2. For each i, estimate model (11) by OLS using y_{it}^* in place of y_{it}. Let $\hat{\beta}_i^*$ and $\hat{\delta}_i^*$ denote the OLS estimators and let $\hat{\varepsilon}_{it}^* = y_{it}^* - \hat{\beta}_i^* d_{0t} + \hat{\delta}_i^{*}{}' d_{1t}$ denote the OLS residuals. Let $\hat{\varepsilon}_t^*$ denote the $n \times 1$ vector $\hat{\varepsilon}_t^* = (\hat{\varepsilon}_{1t}^*, \hat{\varepsilon}_{2t}^*, ..., \hat{\varepsilon}_{nt}^*)'$ and let $\hat{\beta}^*$ denote the $n \times 1$ vector $\hat{\beta}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, ..., \hat{\beta}_n^*)'$.

3. Compute $\hat{\Omega}_T^*$ using (5) with $\hat{\varepsilon}_t^*$ in place of $\hat{u}_t$. The equivalent form given by (10) can also be used and may be faster to compute.

4. Compute the bootstrap versions of VF or VF_t as follows:

$$VF^* = (R\hat{\beta}^*)' \left[\left(\sum_{t=1}^{T} \tilde{d}_{0t}^2 \right)^{-1} R\hat{\Omega}_T^* R' \right]^{-1} (R\hat{\beta}^*)/q, \quad VF_t^* = \frac{R\hat{\beta}^*}{\sqrt{\left(\sum_{t=1}^{T} \tilde{d}_{0t}^2 \right)^{-1} R\hat{\Omega}_T^* R'}}$$

5. Repeat Steps 1 through 4 N_B times where N_B is a relatively large (and usually odd) integer. This generates N_B random draws from VF^* or VF_t^*.

6. Sort the N_B values of VF^* from smallest to largest and let $VF^*[1], VF^*[2], ..., VF^*[N_B]$ indicate the sorted values. Do likewise for VF_t. For the VF statistic the right tail critical value for a test with significance level α is given by $VF^*[(1-\alpha)N_B]$ where the integer part of $(1-\alpha)N_B$ is used if $(1-\alpha)N_B$ in not an integer. For a left tail test using VF_t, the critical value is given by $VF_t^*[\alpha N_B]$ and for a right tail test the critical value is given by $VF_t^*[(1-\alpha)N_B]$.

7. Bootstrap p-values can be computed by computing the frequency of VF^* values that exceed the value of VF from the actual data.

Note that by construction, the true value of β_i^* is zero. Therefore, $R\beta^* = \mathbf{0}$, i.e. $r = \mathbf{0}$ in the bootstrap samples and VF^* and VF_t^* are computed using $r = \mathbf{0}$ to ensure that the null holds for VF^* and VF_t^*. Those familiar with bootstrap methods will notice that the resampling scheme used in Step 1 does not reflect the serial correlation with a series or the correlation across series because an i.i.d. resampling method is being used. Because VF^* and VF_t^* are based on HAC estimators and their asymptotic null distributions do not depend on unknown correlation parameters, VF^* and VF_t^* fall within the general framework considered by Gonçalves and Vogelsang (2011) where it was shown that the simple, or naive, i.i.d. bootstrap will generate valid critical values. No special methods, such as blocking, are required here. The formal results implied by the theory of Gonçalves and Vogelsang (2011) are that

$$VF^* \xrightarrow{d} Z_q' \left[2\int_0^1 B_q^f(s) B_q^f(s)' ds \right]^{-1} Z_q / q, \quad VF_t^* \xrightarrow{d} \frac{Z_1}{\sqrt{2\int_0^1 B_1^f(s) B_1^f(s)' ds}}.$$

In other words, the bootstrap statistics have the same limits as the VF and VF_t statistics under the null hypothesis. Therefore, the bootstrap critical values are equivalent to the approximations given by (16) and (18).

4 Data and Methods

The application here is to data from the lower- and mid-troposphere (LT, MT respectively), where we will compare trends from a large suite of general circulation models (GCMs) to those observed in two radiosonde records over the 1958-2010 interval using monthly data. McKitrick et al. (2010) present results from the post-1979 interval where mean breaks were not warranted. Karl et al. (2006) and Soden and Held (2005) discuss the particular importance of examining the tropical troposphere for assessing GCM performance. We will show that the inclusion of the mean shift dummy variable causes the trend slopes to become insignificant at the 5% level. In comparison of trends between GCM generated data (climate model data) and observed data (LT, MT) we find significantly different trends with or without the inclusion of the dummy variable but the significance levels where we can reject the null of equal trends is much smaller (more significant) with the mean shift dummy included.

Throughout this section the trend slopes are the parameters of interest. Therefore, for both models (1) and (2) we always set $d_{0t} = t$, in which case $\beta_i = b_i$. For model (1) $d_{1t} = 1$, and for model (2) $d_{1t} = (1, DU_t)'$ with the mean shift set at January 1978. Let $\hat{\beta}_i$ denote the OLS estimator of β_i for a given time series using either model (1) or model (2) and let $\hat{\delta}_i$ denote the OLS estimator of δ_i. The *VF* standard error is given by

$$se(\hat{\beta}_i) = \sqrt{\left(\sum_{t=1}^{T} \tilde{d}_{0t}^{2}\right)^{-1} \hat{\Omega}_T^i}\,,$$

where $\hat{\Omega}_T^i$ is computed with (5) (or equivalently (10)) using $\hat{u}_t$ from (14). Let $cv_{0.025}$ denote the 2.5% right tail critical value of the asymptotic distribution of VF_t. For model (1) $cv_{0.025} = 6.482$ (see Table 1 of VF05; their t_2^* statistic) and for model (2) $cv_{0.025} = 7.032$ (see Table 1). A 95% confidence interval (CI) is computed as $\hat{\beta}_i \pm se(\hat{\beta}_i) \cdot cv_{0.025}$.

4.1 Climate Model Series and Observation Data Series: Trends

All data are averages over the tropics (20N to 20S). The GCM runs were compiled for McKitrick et al. (2010). There were 57 runs from 23 models for each of the lower troposphere (LT) and mid-troposphere (MT). Each model uses prescribed forcing inputs up to the end of the 20th century climate experiment (20C3M, see Santer et al. 2005), and most models include at least one extra forcing such as volcanoes or land use. Projections forward after 2000 use the A1B emission scenario. Tables 2 and 3 report, for the LT and MT layers respectively, the climate models, the extra forcings, the number of runs in each ensemble mean, estimated trend slopes in the cases with and without mean shifts, and *VF* standard errors.

We used two observational temperature series. The HadAT radiosonde series is a set of MSU-equivalent layer averages published on the Hadley Centre web site[1] (Thorne et al. 2005). We use

[1] http://www.metoffice.gov.uk/hadobs/hadat/msu/anomalies/hadat_msu_tropical.txt.

the 2LT layer to represent the GCM LT-equivalent and the T2 layer to represent the GCM MT-equivalent. The Radiosonde Innovation Composite Homogenization (RICH) series is due to Haimberger et al. (2008) and was supplied by John Christy (pers. comm.) in LT- and MT-equivalent forms. The last two lines of Tables 2 and 3 report the estimated trend slopes and VF05 standard errors for the two observed temperature series.

Figures 2 and 3 display the observed LT and MT trends, respectively, with the least squares trend lines shown. The estimated trends are 0.13 and 0.16 C/decade in the LT and 0.09 and 0.11 C/decade in the MT. Allowing for a mean shift (step-change) at 1977 yields Figures 4 and 5. The LT trends fall to 0.06 and 0.09 C/decade and the MT trends fall to -0.01 and 0.04 C/decade. Thus about half of the positive LT trend in Figure 3 can be attributed to the one-time change at 1977-78 and nearly all the MT change is likewise accounted for by the step-change. Consequently, the trend comparison between models and observations needs to take into account the discontinuity.

Figure 6 plots all the estimated trend slopes along with their CIs. The top panel (a) leaves out the mean shift and the bottom panel (b) includes it. The model-generated trends are grouped on the left with the CI's shown as the shaded region. The trends are ranked from smallest to largest and the numbers beside each marker refer to the GCM number (see Table 2 for names). The two trends on the right edge are, respectively, the Hadley and RICH series. The range of model runs and their associated CI's clearly overlap with those of the observations. In that sense we could say there is a visual consistency between the models and observations. However, that is too weak a test for the present purpose, since the range of model runs can be made arbitrarily wide through choice of parameters and internal dynamical schemes, and even if the reasonable range of parameters or schemes is taken to be constrained on empirical or physical grounds, the spread of trends in Figure 6 (spanning roughly 0.1 to 0.4 C/decade) indicates that it is still sufficiently wide as to be effectively unfalsifiable. Also, if we base the comparison on the range of model runs rather than some measure of central tendency it is impossible to draw any conclusions about the models as a group, or as an implied methodology. Using a range comparison, the fact that in Figure 6a models 8, 5 and 16 are reasonably close to the observational series does not provide any support for models 2, 3 and 4, which are far away. We want to pose the trend comparison in a form that tells us something about the behaviour of the models as a group, or as a methodological genre, and this requires a multivariate testing framework.

4.2 Multivariate Trend Comparisons

For each layer we now treat the 23 climate model generated series and the 2 observational series as an n=25 panel of temperature series. We estimate models (1) and (2) using the methods described in Section 3. The parameters of interest are the trend slopes ($d_{0t} = t$). We are interested in testing the null hypothesis that the *average* of the trend slopes in the 23 climate model generated series is the same as the *average* trend slope of the observed series. Placing the observed series in positions *i=24,25*, the restriction matrices for this null hypothesis are[2]

$$R = \left[\frac{1}{23}, \frac{1}{23}, ..., \frac{1}{23}, -\frac{1}{2}, -\frac{1}{2} \right], \qquad r = 0\,.$$

Table 4 presents the *VF* statistics for the equivalence of trends in the climate models and observed data. Also reported are the *VF* statistics for testing the significance of the individual trends of the observed temperature series. Asymptotic critical values are provided in the table and significance is indicated as described in the table. We also compute bootstrap p-values for the tests using the method outlined in Section 3.3. We used 1499 bootstrap replications.

In the trend model without mean shifts, the zero trend-hypothesis is rejected at the 1% significance level for all 4 observed series, indicating strong evidence of a significant warming trend over the 1958-2010 interval. A test that the climate models, on average, predict the same trend as the observational data sets is rejected in the LT layer at 5% and in the MT layer at 1% significance.

But when we add the mean-shift term at 1978, the values of the *VF* statistics for testing the zero trend-hypothesis drop substantially. The critical values for *VF* are slightly larger than in the case without the mean-shift dummy. We see that only one of the observed series has a significant trend, and only at the 10% level. When the one-time jump is not modeled, the increase in the series from the jump is spuriously associated with the trend slope. This spurious association is no longer present when the mean shift dummy is included.

The test of equivalence of trends betweens the climate models and observed data is more strongly rejected when the mean shift dummy is included. Notice that bootstrap p-values drop to essentially zero in this case. This finding is not surprising because, as is clear in Tables 2 and 3, while the estimated trend slopes decrease for the observed series when the mean shift dummy is included, the estimated trend slopes of the climate model series are not systematically affected by the mean shift dummy.[3] Therefore, there is a greater discrepancy between the climate model trends and the observed trends.

[2] This form weights each model equally, even though some models supplied more than one run. Adjusting the *R* matrix so that models are weighted according to the number of runs does not change our conclusions, and in fact makes the model-observation equivalence test reject more strongly.

[3] The climate-models do not explicitly model the Pacific Climate Shift and so the mean shift coefficient has no special meaning for the climate model data. Not surprisingly, the estimated mean shift coefficients were positive in 11 cases and negative in 12 for the climate model series.

5 Conclusions

Heteroskedasticity and autocorrelation robust (HAC) covariance matrix estimators have been adapted to the linear trend model, permitting robust inferences about trend significance and trend comparisons in data sets with complex and unknown autocorrelation characteristics. Here we extend the multivariate HAC approach of Vogelsang and Franses (2005) to allow more general deterministic regressors in the model. We show that the asymptotic (approximating) critical values of the test statistics of Vogelsang and Franses (2005) are nonstandard and depend on the specific deterministic regressors included in the model. These critical values can be simulated directly. Alternatively, we outline a simple bootstrap method for obtaining valid critical values and p-values.

The empirical focus of the paper is a comparison of trends in climate model-generated temperature data and corresponding observed temperature data in the tropical troposphere. Our empirical innovation is to model a level shift in the observed data corresponding to the Pacific Climate Shift that occurred in 1977-78. With respect to the Vogelsang Franses (2005) approach, this amounts to adding a mean shift dummy to the model which requires a new set of critical values which we provide.

As our empirical findings show, the detection of a trend in the tropical lower- and mid-troposphere data over the 1958-2010 interval is contingent on the decision of whether or not to include a mean-shift term at January 1978. If the term is included, a time trend regression with error terms robust to autocorrelation of unknown form indicates that the trend observed over the 1958-2010 interval is not statistically significant in either the LT or MT layers. Most climate models predict a larger trend over this interval than is observed in the data. We find a statistically significant mismatch between climate model trends and observational trends whether the mean-shift term is included or not. However, with the shift term included the null hypothesis of equal trend is rejected at much smaller significance levels (much more significant).

References

Andrews, D.W.K. (1991) “Heteroskedasticity and autocorrelation consistent covariance matrix estimation,” *Econometrica* 59, 817–854.

Andrews, D. W. K. (1993): “Tests for Parameter Instability and Structural Change with Unknown Change Point,” *Econometrica*, 61, 821-856.

Bloomfield, P. and D. Nychka (1992) “Climate spectra and detecting climate change,” *Climate Change*, 21, 1-16.

Davidson, R. and J.G. MacKinnon (2004) *Econometric Theory and Methods*. Toronto: Oxford.

Fomby, T. and T.J. Vogelsang (2002) “The application of size-robust trend statistics to global warming temperature series,” *J. Climate*, 15, 117-123.

Gonçalves, S. and T.J. Vogelsang (2011) “Block bootstrap HAC robust tests: The sophistication of the naïve bootstrap,” *Econometric Theory*, 27, 745-791.

Haimberger, L., C. Tavolato, S. Sperka, (2008) “Towards the elimination of the warm bias in historic radiosonde temperature records - some new results from a comprehensive intercomparison of upper air data.” *J. Climate*. 21, 4586—4606 DOI 10.1175/2008JCLI1929.1.

Hamilton, James D. (1994) *Time Series Analysis* Princeton: Princeton University Press.

Hansen, Bruce (1996) “Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis” Econometrica, Vol. 64, No. 2 (Mar., 1996), pp. 413-430.

Held, I. and B. J. Soden (2000) “Water Vapor Feedback and Global Warming” *Annual Review of Energy and Environment* 25:441—75.

Karl, T. R., Susan J. Hassol, Christopher D. Miller, and William L. Murray (2006). *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences.* Synthesis and Assessment Product. Climate Change Science Program and the Subcommittee on Global Change Research. http://www.climatescience.gov/Library/sap/sap1-1/finalreport/sap1-1-final-all.pdf. Accessed August 3 2010.

Kiefer, N.M and T.J. Vogelsang (2002) “Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation,” *Econometrica,* 70, 2093-2095.

Kiefer, N.M and T.J. Vogelsang (2005) “A new asymptotic theory for heteroskedasticity-autocorrelation robust tests,” *Econometric Theory*, 21, 1130-1164.

Kiefer, N.M., T.J. Vogelsang & H. Bunzel (2000) “Simple robust testing of regression hypotheses” *Econometrica* 68, 695–714.

McKitrick, Ross R., Stephen McIntyre and Chad Herman (2010) "Panel andMultivariate Methods for Tests of Trend Equivalence in Climate Data Sets." *Atmospheric Science Letters*, 11(4) pp. 270-277, October/December 2010 DOI: 10.1002/asl.290.

Mills T. C. (2010) Skinning a cat: alternative models of representing temperature trends. *Climatic Change* 101: 415-426, DOI 10.1007/s10584-010-9801-1.

Newey, W.K. & K.D. West (1987) "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." *Econometrica* 55, 703–708.

Powell Jr., Alfred M. and Jianjun Xu (2011) "Abrupt Climate Regime Shifts, Their Potential Forcing and Fisheries Impacts" *Atmospheric and Climate Sciences* 1, 33-47. doi:10.4236/acs.2011.12004

Santer, B.D., T. M. L. Wigley, C. Mears, F. J. Wentz, S. A. Klein, D. J. Seidel, K. E. Taylor, P. W. Thorne, M. F. Wehner, P. J. Gleckler, J. S. Boyle, W. D. Collins, K. W. Dixon, C. Doutriaux, M. Free, Q. Fu, J. E. Hansen, G. S. Jones, R. Ruedy, T. R. Karl, J. R. Lanzante, G. A. Meehl, V. Ramaswamy, G. Russell, G. A. Schmidt (2005) Amplification of Surface Temperature Trends and Variability in the Tropical Atmosphere. *Science* Vol. 309. no. 5740, pp. 1551 – 1556 DOI: 10.1126/science.1114867

Seidel, D. J. and J. R. Lanzante (2004) "An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes." *Journal of Geophysical Research* VOL. 109, D14108, doi:10.1029/2003JD004414, 2004.

Sun, Y., P.C.B. Phillips, and S. Jin (2008) "Optimal bandwidth selection in heteroskedasticity--autocorrelation robust testing," *Econometrica*, 76, 175-194.

Thorne, P. W., D. E. Parker, S. F. B. Tett, P. D. Jones, M. McCarthy, H. Coleman, P. Brohan, and J. R. Knight, (2005) "Revisiting radiosonde upper-air temperatures from 1958 to 2002." *J. Geophys. Res.*, 110, D18105, doi:10.1029/2004JD005753.

Vogelsang, T.J. and P. H. Franses (2005) "Testing for Common Deterministic Trend Slopes," *Journal of Econometrics* 126, 1-24.

White, H., and I. Domowitz (1984) "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143-161.

Woodward, W. A. and H.L. Gray (1993) "Global warming and the problem of testing for trend in time series data," *J. Climate*, 6, 953-962.

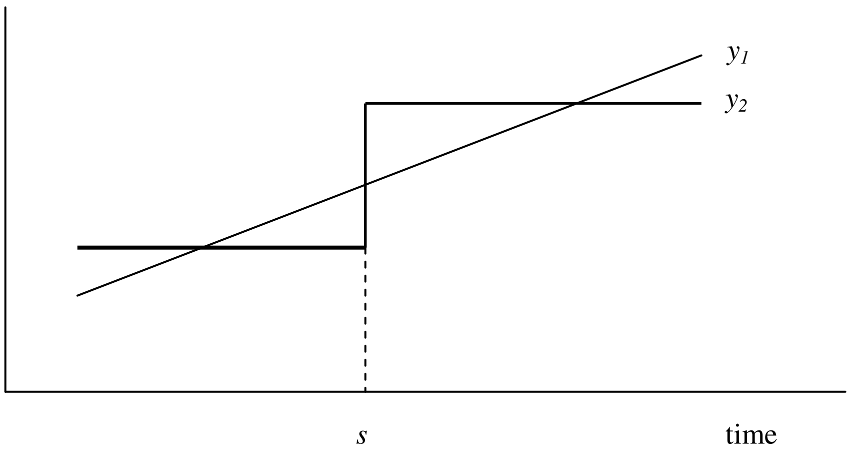

Figure 1. Schematic of two series to be compared.

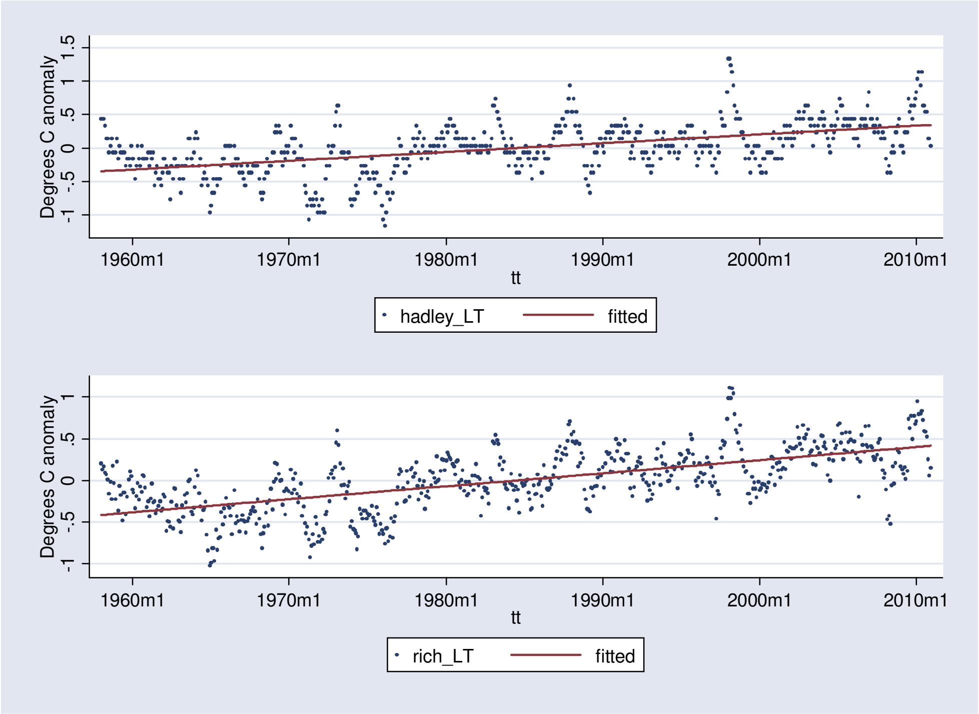

Figure 2 Monthly Hadley (top) and RICH (bottom) observations in the LT layer, 1958 to 2010.

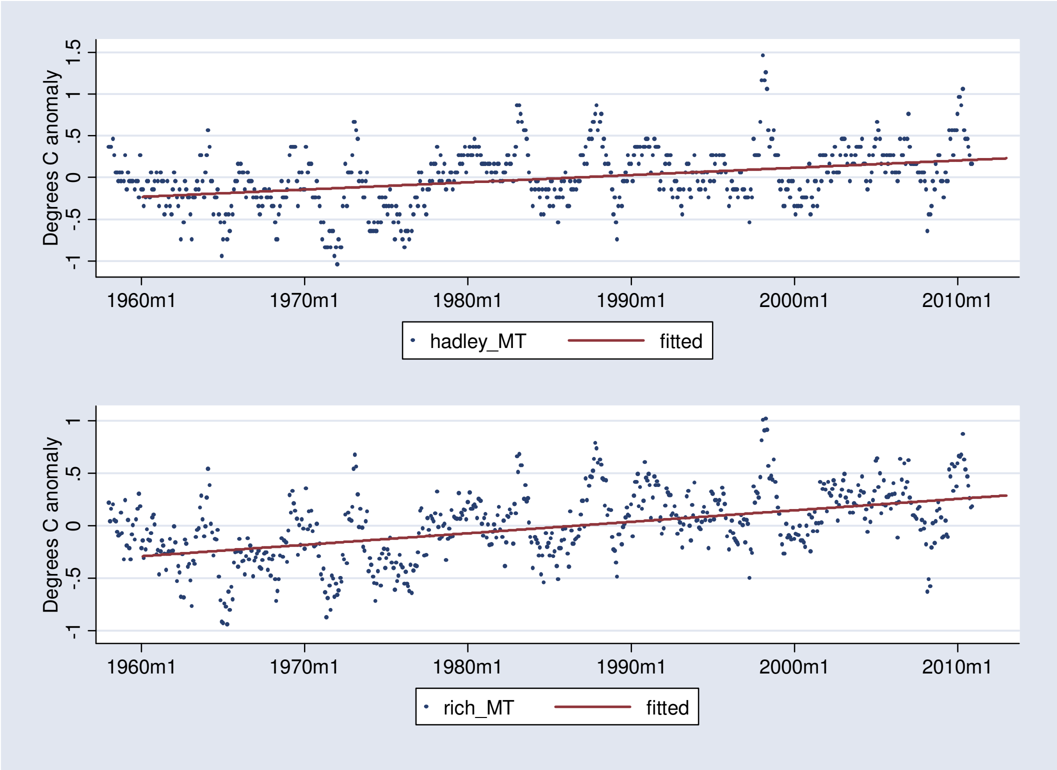

Figure 3 Monthly Hadley (top) and RICH (bottom) observations in the MT layer, 1958 to 2010.

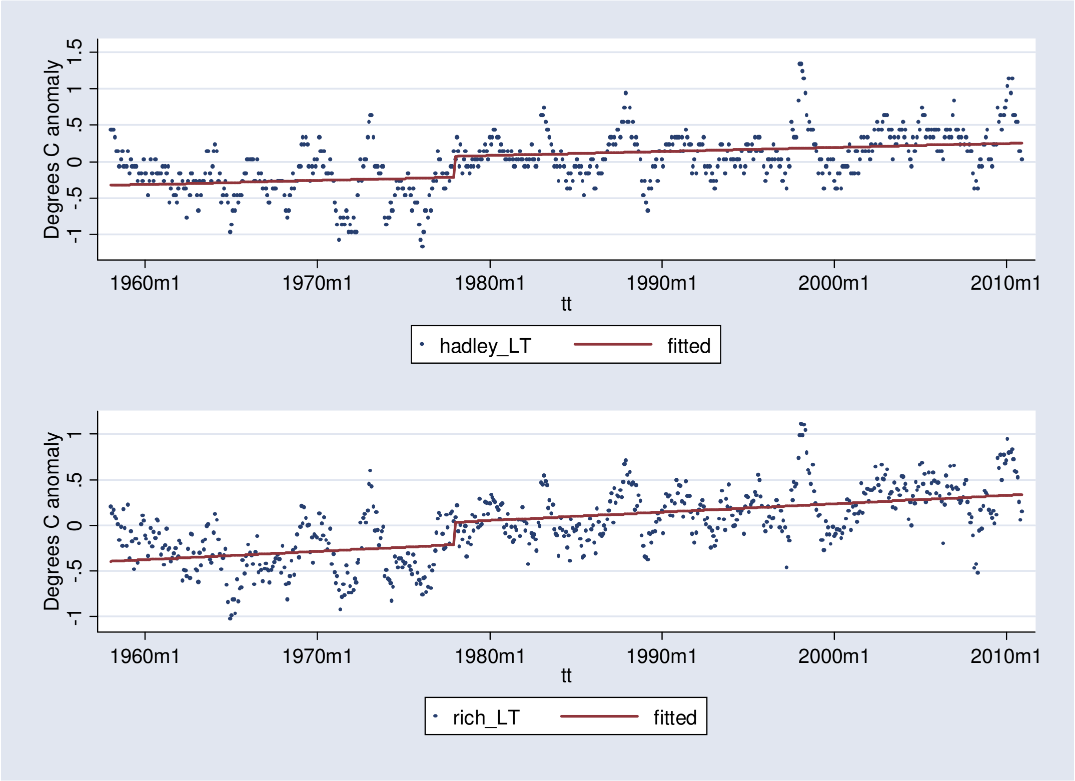

Figure 4: LT Obs series with step changes

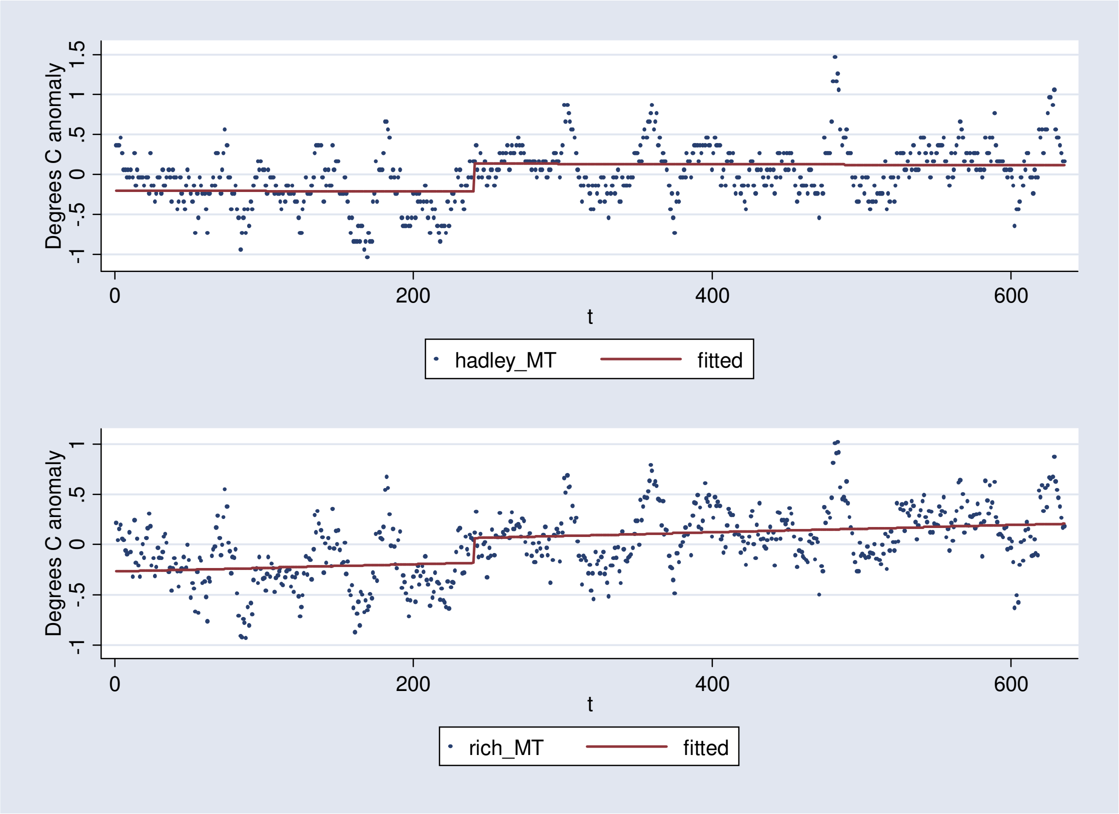

Figure 5: MT Obs series with step changes

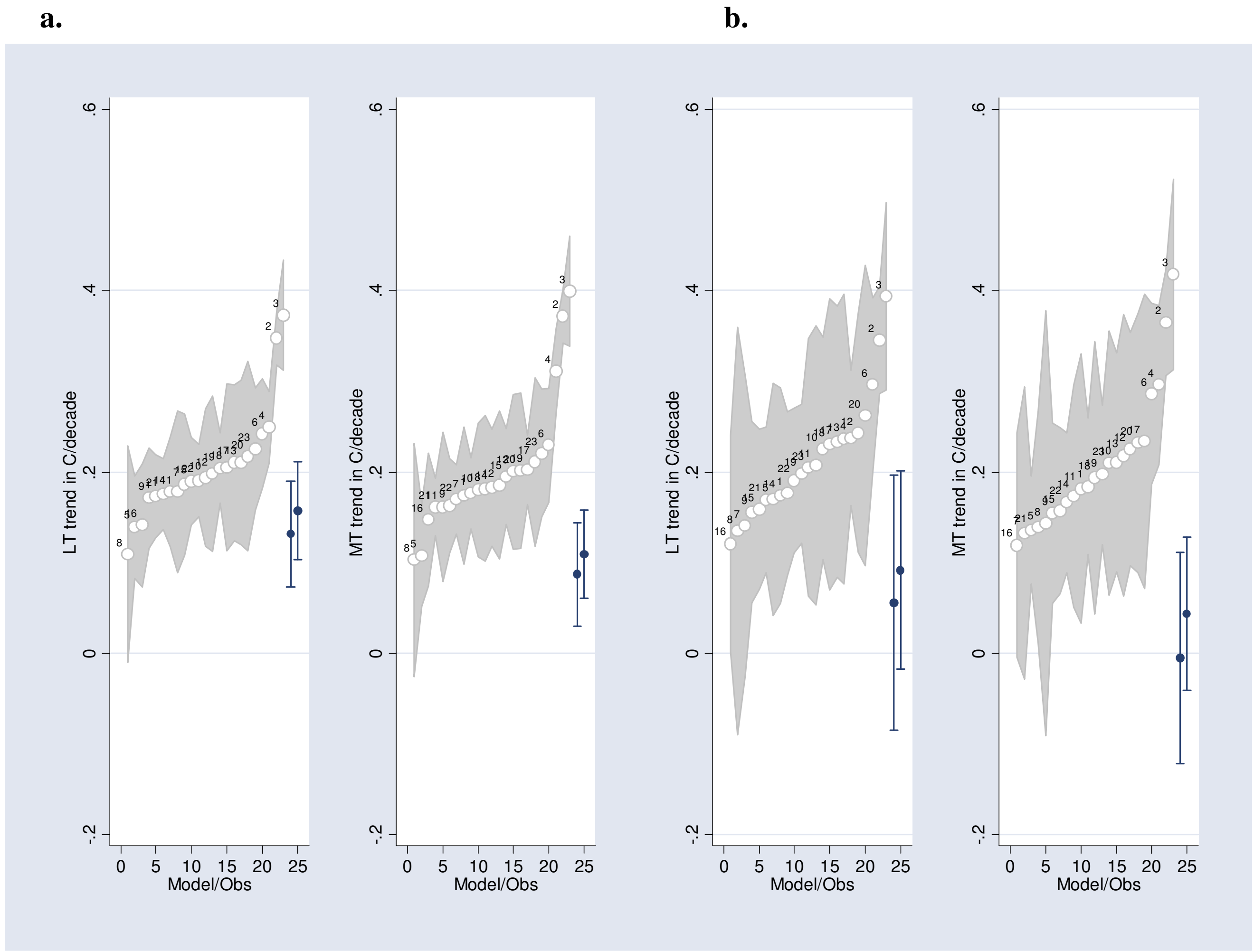

Figure 6. 1958-2010 Trends and 95% CIs for 23 models (shaded region) and two balloon series Hadley and RICH (respectively, individual markers at right edge). a: (Left two panels) Trends computed without allowing for mean shift. b: (Right two panels) Mean shift term included in model.

Table 1: Asymptotic Critical Values.
Model (2), $\lambda = 0.358$, $q = 1$.

%	VF_t	VF
.700	1.678	11.612
.750	2.175	14.534
.800	2.743	18.388
.850	3.408	23.922
.900	4.288	32.385
.950	5.691	49.445
.975	7.032	68.065
.990	8.642	97.901
.995	9.894	123.724

Note: Left tail critical values of VF_t follow by symmetry around zero.

Table 2: Summary of Lower Troposphere data series.

		Simple Trend		Trend + Mean Shift	
	Model/ Obs Name				
Data	Extra Forcings;	Trend		Trend	
Series	No. runs	(C/decade)	Std Error	(C/decade)	Std Error
1	BCCR BCM2.0				
	O; 2	0.173	0.0071	0.176	0.013
2	CC;CMA3.1-T47				
	NA; 5	0.347	0.0046	0.345	0.008
3	CCCMA3.1-T63				
	NA; 1	0.373	0.0094	0.393	0.015
4	CNRM3.0				
	O; 1	0.249	0.0061	0.237	0.011
5	CSIRO3.0				
	1	0.139	0.0087	0.170	0.018
6	CSIRO3.5				
	1	0.242	0.0093	0.296	0.014
7	GFDL2.0				
	O, LU, SO, V; 1	0.186	0.0120	0.141	0.024
8	GFDL2.1				
	O, LU, SO, V; 1	0.109	0.0184	0.135	0.032
9	GISS_AOM				
	2	0.171	0.0085	0.156	0.014
10	GISS_EH				
	O, LU, SO, V; 6	0.193	0.0117	0.226	0.017
11	GISS_ER				
	O, LU, SO, V; 5	0.178	0.0137	0.207	0.022
12	IAP_FGOALS1.0				
	3	0.198	0.0132	0.243	0.019
13	ECHAM4				
	1	0.210	0.0140	0.236	0.023

14	INMCM3.0				
	SO, V; 1	0.178	0.0094	0.174	0.017
15	IPSL_CM4				
	1	0.189	0.0074	0.159	0.013
16	MIROC3.2_T106				
	O, LU, SO, V; 1	0.141	0.0104	0.121	0.017
17	MIROC3.2_T42				
	O, LU, SO, V; 3	0.210	0.0133	0.233	0.021
18	MPI2.3.2a				
	SO, V; 5	0.205	0.0141	0.231	0.023
19	ECHAM5				
	O; 4	0.204	0.0059	0.198	0.011
20	CCSM3.0				
	O, SO, V; 7	0.217	0.0161	0.262	0.024
21	PCM_B06.57				
	O, SO, V; 4	0.176	0.0060	0.169	0.011
22	HADCM3				
	O; 1	0.190	0.0062	0.190	0.011
23	HADGEM1				
	O, LU, SO, V; 1	0.226	0.0104	0.205	0.020
24	HadAT	0.131	0.0090	0.055	0.020
25	RICH	0.157	0.0083	0.092	0.016

Notes: Each row refers to model ensemble mean (rows 1—23) or observational series (rows 24, 25). All models forced with 20th century greenhouse gases and direct sulfate effects. Rows 10, 11, 19, 22 and 23 also include indirect sulfate effects. 'Extra forcing' indicates which models included other forcings: ozone depletion (O), solar changes (SO), land use (LU), volcanic eruptions (V). NA: information not supplied to PCMDI. No. runs: indicates number of individual realizations in the ensemble mean. Trend slopes estimated using OLS, Std Errors computed using VF method (see Section 4).

Table 3: Summary of Mid-Troposphere data series.

		Simple Trend		Trend + Mean Shift	
Data Series	Model/ Obs Name Extra Forcings; No. runs	Trend (C/decade)	Std Error	Trend (C/decade)	Std Error
1	BCCR BCM2.0 O; 2	0.176	0.0060	0.184	0.011
2	CC;CMA3.1-T47 NA; 5	0.372	0.0046	0.365	0.008
3	CCCMA3.1-T63 NA; 1	0.399	0.0093	0.417	0.015
4	CNRM3.0 O; 1	0.311	0.0072	0.296	0.013
5	CSIRO3.0 1	0.108	0.0086	0.139	0.018
6	CSIRO3.5 1	0.229	0.0097	0.286	0.014
7	GFDL2.0 O, LU, SO, V; 1	0.174	0.0117	0.133	0.023
8	GFDL2.1 O, LU, SO, V; 1	0.103	0.0198	0.143	0.033
9	GISS_AOM 2	0.163	0.0081	0.154	0.014
10	GISS_EH O, LU, SO, V; 6	0.180	0.0114	0.210	0.017
11	GISS_ER O, LU, SO, V; 5	0.162	0.0127	0.182	0.021
12	IAP_FGOALS1.0 3	0.185	0.0125	0.225	0.018
13	ECHAM4 1	0.200	0.0131	0.218	0.022

14	INMCM3.0				
	SO, V; 1	0.183	0.0100	0.173	0.017
15	IPSL_CM4				
	1	0.195	0.0081	0.157	0.013
16	MIROC3.2_T106				
	O, LU, SO, V; 1	0.147	0.0113	0.119	0.018
17	MIROC3.2_T42				
	O, LU, SO, V; 3	0.211	0.0143	0.234	0.023
18	MPI2.3.2a				
	SO, V; 5	0.182	0.0124	0.193	0.021
19	ECHAM5				
	O; 4	0.202	0.0059	0.197	0.011
20	CCSM3.0				
	O, SO, V; 7	0.201	0.0132	0.232	0.020
21	PCM_B06.57				
	O, SO, V; 4	0.161	0.0048	0.136	0.008
22	HADCM3				
	O; 1	0.170	0.0059	0.166	0.011
23	HADGEM1				
	O, LU, SO, V; 1	0.221	0.0108	0.210	0.021
24	HadAT	0.087	0.0088	-0.005	0.017
25	RICH	0.109	0.0075	0.043	0.012

Notes same as for Table 2.

Table 4: Results of hypothesis tests using VF statistic with and without mean shift term at January, 1977.

Trend Coef	Null Hypothesis	Test Score	Bootstrap p value
		No Mean Shift	
Hadley LT (0.131)	trend = 0	213.6***	< 0.001
RICH LT (0.157)	trend = 0	356.1***	< 0.001
Hadley MT (0.087)	trend = 0	98.0***	0.006
RICH MT (0.109)	trend = 0	212.4***	< 0.001
LT	Models = Observed	58.5**	0.029
MT	Models = Observed	121.1***	0.005
		With Mean Shift	
Hadley LT (0.055)	trend = 0	7.7	0.379
RICH LT (0.092)	trend = 0	34.8*	0.075
Hadley MT (-0.005)	trend = 0	0.10	0.980
RICH MT (0.043)	trend = 0	13.1	0.266
LT	Models = Observed	402.1***	< 0.001
MT	Models = Observed	999.9***	< 0.001

Notes: sample period (monthly): January 1958 to December 2010. The bootstrap p-value* is computed using the method described in Section 3.3 using 1499 bootstrap replications (N_B = 1499). *VF* Critical Values: Without Mean Shift, 20.14 (10%, denoted *) 41.53 (5%, denoted **), 83.96 (1%, denoted ***). With Mean Shift, 32.39 (10%, denoted *), 49.45 (5%, denoted **), 97.90 (1%, denoted ***).