HAC-ROBUST TREND COMPARISONS AMONG CLIMATE SERIES WITH POSSIBLE INTERCEPT SHIFTS

Ross McKitrick*
Department of Economics
University of Guelph
Guelph ON Canada
N1G 2W1
rmckitri@uoguelph.ca
519-824-4120 x52532

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

This version February 28, 2012
*Corresponding author
HAC-ROBUST TREND COMPARISONS AMONG CLIMATE SERIES WITH POSSIBLE INTERCEPT SHIFTS

Abstract: Comparisons of trends across climatic data sets, as well as univariate trend significance tests, are complicated by the presence of serial correlation and possible step-changes in the mean. We build on heteroskedasticity and autocorrelation (HAC) robust methods, specifically the Vogelsang-Franses (VF) nonparametric variance estimator, to allow for a step-change in the mean at a known or unknown date. Our application evaluates climate models by comparing observed trends in the tropical lower- and mid-troposphere since 1958 to model simulations. Inclusion of a level shift regressor to capture the Pacific Climate Shift of 1977 causes apparently significant observed trends to become statistically insignificant. Model over-estimation of warming is significant whether or not we account for a level shift, although null rejections are much stronger when the level shift is included.

The VF approach is a HAC estimator for the linear trend model in which the bandwidth is set equal to the sample size. It provides a powerful multivariate trend estimator robust to unknown serial correlation up to but not including unit roots, making it ideal for applications on climate data. We show that the critical values of the VF statistic change when the mean shift occurs at a known or unknown date. The latter case is useful in “data mining” situations where the researcher suspects a discontinuity exists but is not able to date it based on exogenous information. If the conclusions are upheld in the data mining framework, as is the case in our application, then the result can be considered both conservative and robust. We derive an asymptotic approximation that can be used to simulate critical values, and we outline a simple bootstrap procedure that generates valid critical values and p-values. Our application is relevant to ongoing debates about the whether climate models over-predict warming in the tropical troposphere, an important region for diagnosing model accuracy.

Keywords: Autocorrelation; trend estimation; HAC variance matrix; global warming; model comparisons
HAC-Robust Trend Comparisons Among Climate Series With Possible Intercept Shifts

1 Introduction

Referring to Figure 1, suppose we want to determine whether two time series have the same trend slopes. In the top panel a comparison of the simple linear trend slope 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 at the break point $T_b$. By contrast, in the bottom panel a failure to account for the shift would overstate the difference between the trend slopes. In each case, the role of the shift term is illustrated by the fact that if the trend slope comparisons were conducted over the pre-shift or post-shift intervals, they might indicate opposite results to those based on the entire sample (with the shift term omitted).

If an $nx1$ vector of trend slope coefficients across $n$ time series is denoted $b$, multivariate hypotheses can be expressed using restriction matrices $R$ and $r$ (respectively $nxq$ and $qx1$) in the usual form $H_0$: $Rb=r$ where $q$ denotes the number of restrictions. This encompasses comparisons among data sets as well as univariate trend function inference (where $n=1$). Vogelsang and Franses (2005, herein VF) derive a class of heteroskedasticity and autocorrelation robust (HAC) tests of $H_0$: $Rb=r$ for time series with a simple linear deterministic trend function. The VF statistic is similar in form to the familiar regression $F$-type statistics but the VF statistic remains valid under serial dependence up to but not including unit roots in the time series. For treatments of the theory behind HAC estimation and inference see Andrews (1991), Kiefer and Vogelsang (2005), Newey and West (1987), Sun, Phillips and Jin (2008) and White and Domowitz (1984) among others. Like many HAC approaches, the VF approach is nonparametric with respect to the serial dependence structure and does not require a specific model of serial correlation to be implemented. Unlike most
nonparametric approaches, the VF approach avoids sensitivity to bandwidth selection by setting the bandwidth equal to the entire sample. McKitrick et al. (2010) used the VF tests to compare climate model projections of temperatures and observed temperatures in the tropical troposphere over the 1979-2009 interval and concluded that climate models significantly over-predict warming trends.

The VF tests as originally proposed are not robust to intercept shifts in the trend functions or the presence of additional trend regressors in general. We show herein that inclusion of an intercept shift dummy in the model changes the critical values of the VF statistics. In particular, the critical values depend on the date of the intercept shift and depend on whether the shift date is treated as known or unknown. More generally we show that the critical values of the VF statistic depend on the specific trend regressors included in the model.

Our empirical application is a comparison of climate model-generated temperature data to observed weather balloon temperature data in the tropical troposphere over the interval 1958-2010. The comparison is akin to that in the top panel of Figure 1 due to an exogenous step-like change in the observed temperature data around 1978 called the Pacific Climate Shift (PCS). The PCS has been documented as 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 and extensive references therein). A possible explanation of the PCS is that it was a bifurcation in the nonlinear chaotic system governing the coupling of the ocean and atmosphere (Tsonis et al. 2007). For our purposes we do not need a specific physical explanation of the PCS; it is sufficient to note that it is an exogenous event at a known date. While we view the exogeneity assumption about the date of the PCS as reasonable, we also present results where we treat the PCS date as unknown. We develop a version of the VF statistic that corresponds to a data-mining approach with respect to choice of shift date and we derive “data-mining robust” critical values for the VF statistic that explicitly control for the data-mining approach. The data-mining robust critical values deliver trend tests that are very conservative with regards to uncertainty regarding the shift date.
Taking the date of the PCS as given and exogenous, we find that the models project significantly more warming in both the lower- and mid-troposphere than are found in weather balloon records over the interval. This finding is robust if we treat the date of the PCS as unknown and apply the conservative data-mining approach. In fact it is robust whether or not we include a level shift in the regression model: we reject equivalence of the trend slopes between the observed and model-generated temperature series either way. The evidence against equivalence is simply stronger when we control for a level shift and this is true whether we treat the date of the shift to be known or unknown.

We also find that if the date of the PCS is assumed to be known then: a) the appearance of positive and significant trend slopes in the individual observed temperature series vanishes once we control for the effect of the level shift, and b) we find statistical evidence for a level shift in the observed temperature series in the mid-troposphere but not in the lower-troposphere series. If the date of the PCS is assumed to be unknown, statistical evidence for a level shift in the observed mid-troposphere series disappears, but this is not surprising given that we use the data-mining robust critical which decreases the power of detecting such a shift.

While the empirical focus of this paper is an application with trends similar to those depicted in the top panel of Figure 1, there are other potential applications to climate data more akin to the trends in the bottom panel of Figure 1. For example, many long observational records are interrupted by equipment and/or sampling changes, changes in monitoring locations and so forth (see Hansen et al. 1999, Brohan et al. 2006 for examples in the land record; Folland and Parker 1995, Thompson et al. 2008 for examples in sea surface data). A typical method for detecting and removing intercept shifts is to construct a reference series which is not expected to exhibit the discontinuity, such as the mean of other weather station records in the vicinity, and then look for a jump in one series relative to the reference series. The question of whether a jump is real or not can strongly affect the resulting trend calculations.

The remainder of the paper is organized as follows. Section 2 describes the model used in the empirical investigation and provides some discussion about the general issue of trend inference in temperature data. Section 3 lays out a more general trend function model and extends the VF testing framework. Asymptotic null distributions are derived. Section 4 examines the role of the
break date for the level shift and describes methods appropriate under the assumption that the break date is unknown. Section 5 gives a brief discussion of testing for a level shift in a univariate series using the VF statistic. Section 6 focuses on the empirical application where trends in observed temperatures are compared to trends in climate model generated temperature series. Section 7 concludes. We include an Appendix that provides some motivation for the VF approach that will be useful for readers not familiar with nonparametric HAC approaches to inference in deterministic trend models.

2 Empirical Specification

We begin with the simplest linear trend model of the form:

$$y_{it} = a_i + b_i t + u_{it},$$

where \(i = 1, \ldots, n\) denotes a particular time series and \(t = 1, \ldots, T\) denotes the time period. \(y_{it}\) is a temperature series that, in our data, is either generated by a climate model or is an observed weather balloon series. The random part of \(y_{it}\) is given by \(u_{it}\) which is assumed to be covariance-stationary (in which case \(y_{it}\) is labeled a trend stationary series, that is, stationary around a linear time trend, if one is present). The primary focus of the empirical analysis is comparison of the trend slopes, \(b_i\), between observed and climate model generated temperature series.

Allowing for a shift in the level yields the model

$$y_{it} = a_i + g_i DU_i + b_i t + u_{it},$$

where \(DU_i\) is an indicator variable that takes the value 0 up to the shift date \(T_b\) (the break date), and 1 thereafter. Hence, for series \(i\), OLS estimation of (2) yields an estimated intercept of \(\hat{a}_i\) up to date \(T_b\) and \(\hat{a}_i + \hat{g}_i\) thereafter. Inclusion of \(DU_i\) allows us to control for a one-time jump in the observed temperature series caused by the PCS. When the break date is assumed known, estimation and inference in model (2) is straightforward. However, when the break date is assumed unknown, it is less straightforward to determine whether \(g_i\) is significantly different from zero.
because under the null hypothesis of no level shift (represented by equation (1)), the parameter $T_b$ is not identified whereas $T_b$ is identified under the alternative of a level shift. Davies (1987), Andrews (1993), Andrews and Ploberger (1994), and Hansen (1996) among others discuss inference approaches when a parameter is only identified under the alternative.

In model (2) we are also assuming that there is only a single level shift in the sample, which is a reasonable assumption for our empirical application. However, if there can be many level shifts in the data and the break dates are known, then we can simply add additional level shift dummy variables to the model. Our generalization and extension of the VF approach includes this case. Treating multiple breaks as occurring at unknown dates would greatly complicate the analysis from a computational perspective and is beyond the scope of this paper. In addition, if one thinks level 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 level changes by a random amount at each time period, which is equivalent to having a random walk, or unit root component in $u_it$. 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. In the end, our view is that it is reasonable to assume that our observed temperatures series are well characterized by at most one level shift (with a known shift date) and that the errors are covariance stationary.

The literature on estimation and inference in model (1) is by now well established, and it may hardly seem possible that there is something new to be said on the subject. In fact the last decade or so 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. Many of these robust estimators use the nonparametric HAC approach which is now widely used in econometrics and empirical finance literatures. In constrast the nonparametric HAC approach is used less in applied climatic or geophysical papers although nonparametric approaches have been 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
3 EXTENSION AND GENERALIZATION OF THE VF APPROACH

3.1 STATISTICAL MODEL AND TEST STATISTICS

Vogelsang and Franses (2005) propose and analyze robust tests of hypotheses involving the trend slope parameters, \( b_i \), in model (1). As we shall demonstrate, the null critical values of the the VF test change when additional deterministic trend regressors are included of which model (2) is an example. The methodological contribution of this paper is to extend the VF approach to a general specification that includes models (1) and (2) as special cases.

Consider the more general deterministic trend model

\[
y_{it} = \beta_0 + \delta_1 d_{1t} + u_{it}
\]

where \( d_{0t} \) is a single deterministic regressor and \( d_{1t} \) is a \( k \times 1 \) vector of additional deterministic regressors. Model (1) is obtained for \( d_{0t} = t, \beta_0 = b \), and \( d_{1t} = 1, \delta = a \), whereas model (2) is obtained for \( d_{0t} = t, \beta_0 = b \) and \( d_{1t} = (1, DU)^\prime, \delta = (a, g)^\prime \).

Notice 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 conditional heteroskedasticity and serial correlation. In some applications it might be reasonable to model some of the series as having different trend functions. For example, we know that the climate model series in the application do not have level shifts because level shifts are not part of the climate model structures. When we think series could have different functional forms for the trend, we can simply include in \( d_{1t} \) the union of trend regressors across all the series. While this will result in a loss of degrees of freedom, in many applications the regressors will be similar across series, so the loss in degrees of freedom will often be small. We view this loss of degrees of freedom as a small price to pay for robustness to unknown forms of conditional heteroskedasticity and autocorrelation.
We estimate model (3) using OLS equation by equation. OLS equation by equation has some nice properties in our set up. Because the regressors are the same for each equation, we have the well known exact equivalence between OLS and generalized least squares (GLS) estimators that account for cross series correlation. Because we have covariance stationary errors, the well known Grenander and Rosenblatt (1957) result applies in which case OLS is also asymptotically equivalent to GLS estimators that account for serial dependence in the data.

Defining the $k \times n$ matrix $\delta = (\delta_1, \delta_2, \ldots, \delta_n)$, model (3) can be written in vector notation as

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

(4)

Because the parameters of interest are the vector $\beta$, we express the OLS estimator of $\beta$ 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 $\beta$ can be expressed as

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

(5)

and it directly follows that

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

Using (4) the OLS residuals can be written as

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

(6)

where $\hat{\beta}$ and $\hat{\delta}$ are the OLS estimators of $\beta$ and $\delta$ using OLS equation by equation.

We are interested in testing null hypotheses of the form

$$H_0 : R\beta = r,$$

(7)
against alternatives $H_1 : R \beta \neq r$, where $R$ and $r$ are known restriction matrices of dimension $q \times n$ and $q \times 1$ respectively where $q$ denotes the number of restrictions being tested. The matrix $R$ is assumed to have full row rank. Robust tests of $H_0$ need to account for correlation across time, correlation across series, and conditional heteroskedasticity as summarized by the long run variance of $u_t$ defined as

$$\Omega = \Gamma_0 + \sum_{j=1}^{\infty} (\Gamma_j + \Gamma'_j),$$

where $\Gamma_j = E(u_t u'_{t-j})$ is the matrix autocovariance function of $u_t$. Those familiar with the time series literature will notice that $\Omega$ is proportional to the spectral density matrix of $u_t$ evaluated at frequency zero.

The VF statistic is constructed using the following estimator of $\Omega$:

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

which is the Bartlett kernel nonparametric estimator of $\Omega$ using a bandwidth (truncation lag) equal to the sample size. The VF statistic for testing $H_0 : R \beta = r$ is given by

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

In the Appendix we provide a finite sample motivation for the form of $\hat{\Omega}_T$ and we note that $\hat{\Omega}_T$ was originally proposed by Keifer, Vogelsang and Bunzel (2000, 2001) although in the different but computationally identical form:

$$\hat{\Omega}_T = 2T^{-2} \sum_{t=1}^{T-1} \hat{S}_t \hat{S}_t',$$  

where of $\hat{S}_t = \sum_{j=1}^{t-1} \hat{u}_j$. See Kiefer and Vogelsang (2002) for a formal derivation of the exact equivalence between (8) and (10).
3.2 Asymptotic Approximations

In this section we derive the asymptotic limit of \( VF \) as given by (9) which will provide an approximation that can be used to generate critical values. The key assumption to our result is that the partial sums of \( u_i, S_j = \sum_{j=1}^{i} u_j \), follow a functional central limit theorem (FCLT) of the form

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

where \( c \in [0,1], \lfloor cT \rfloor \) denotes the integer part of \( cT \), the symbol \( \Rightarrow \) denotes weak convergence in distribution, \( \Lambda \) is the matrix square root of \( \Omega \), i.e. \( \Omega = \Lambda \Lambda' \) and \( W_n(c) \) is an \( n \times 1 \) vector of standard Wiener processes that are independent of each other. Because Wiener processes are mean zero and Gaussian, it obviously follows from (11) that

\[
RT^{-1/2} S_{\lfloor cT \rfloor} \Rightarrow R\Lambda W_n(c) = \Lambda' W_q(c) = \Lambda' \int_0^c dW_q(s),
\]

where \( \Lambda' \) is the matrix square root of \( R\Lambda \Lambda' R' \) and \( W_q(s) \) is a \( q \times 1 \) vector of independent standard Wiener processes.

We also need to make some assumptions about the deterministic trend regressors. To that end, assume that there is a scalar, \( \tau_{0T} \), and a \( k \times k \) matrix, \( \tau_{1T} \), such that

\[
T^{-1} \tau_{0T} \sum_{i=1}^{\lfloor cT \rfloor} d_{0i} \to \int_0^c f_0(s) ds, \quad T^{-1} \tau_{1T} \sum_{i=1}^{\lfloor cT \rfloor} d_{1i} \to \int_0^c f_1(s) ds.
\]

For example, in model (2) \( d_{0t} = t \), \( \tau_{0T} = T \), \( f_0(s) = s \) and \( d_{1t} = (1, DU, \lambda)' \), \( \tau_{1T} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \), \( f_1(s) = (1, 1(s > \lambda))' \) where \( \lambda = T_b/T \), \( 1(s > \lambda) \) equals 1 for \( s > \lambda \) and 0 otherwise. We follow the standard approach in the change point literature and assume that \( \lambda \) remains fixed as \( T \) increases. Define the function

\[
\tilde{f}_0(c) = f_0(c) - \left( \int_0^c f_0(s) f_1(s) ds \right) \left( \int_0^c f_1(s) f_1(s) ds \right)^{-1} f_1(c).
\]
Using (13) it is easy to show that
\[ T^{-1} \tau_{oT} \sum_{r=1}^{[cT]} \tilde{d}_{0r} \to \int_0^T \tilde{f}_0(s) ds, \quad T^{-1} \tau_{oT}^2 \sum_{r=1}^{[cT]} \tilde{d}_{0r}^2 \to \int_0^T \tilde{f}_0^2(s) ds. \]  

Using (12) and (14) it immediately follows that
\[ RT^{-1/2} \sum_{r=1}^{T} \tau_{oT} \tilde{d}_{0r} u_r \Rightarrow \int_0^T \tilde{f}_0(s) \Lambda^* dW_q(s). \]  

We are now in position to derive the asymptotic behavior of \( R \hat{\beta} - r \) under \( H_0 \):
\[ \sqrt{T} \tau_{oT}^{-1} (R \hat{\beta} - r) = \sqrt{T} \tau_{oT}^{-1} R(\hat{\beta} - \beta) = \sqrt{T} \tau_{oT}^{-1} R \left( \sum_{r=1}^{T} \tilde{d}_{0r}^2 \right)^{-1} \sum_{r=1}^{T} \tilde{d}_{0r} u_r, \]
\[ = \left( T^{-1} \sum_{r=1}^{T} \tau_{oT}^2 \tilde{d}_{0r}^2 \right)^{-1} RT^{-1/2} \sum_{r=1}^{T} \tau_{oT} \tilde{d}_{0r} u_r \to \left( \int_0^T \tilde{f}_0^2(s) ds \right)^{-1} \left( \int_0^T \tilde{f}_0(s) \Lambda^* dW_q(s) \right), \]  

where (16) follows from (14) and (15).

We next derive the limit of \( \hat{\Omega}_r \). In deriving this limit it is convenient to stack the deterministic regressors into a single column vector \( d \) where \( d' = (d_{0r}, d_{1r}) \). Define the combined scaling matrix
\[ \tau_r = \begin{bmatrix} \tau_{oT} & 0_k \ 0_{k \times 1} & \tau_{1T} \end{bmatrix}, \]
\[ \text{It immediately follows that} \]
\[ T^{-1} \sum_{r=1}^{[cT]} \tau_r d_r \to \int_0^T f(s) ds, \quad T^{-1} \sum_{r=1}^{[cT]} \tau_r d'_r \tau_r \to \int_0^T f(s) f(s') ds, \]  

\[ RT^{-1/2} \sum_{r=1}^{T} u_r d'_r \tau_r \Rightarrow \int_0^T \Lambda^* dW_q(s) f(s'), \]  

where \( f(s') = (f_0(s), f_1(s')). \) The next step is to derive the limit of \( RT^{-1/2} \tilde{\hat{S}}_{[cT]} \).
\[ RT^{-1/2} \hat{S}_{[\tau]} = RT^{-1/2} \sum_{r=1}^{[\tau]} \hat{u}_r = RT^{-1/2} \sum_{r=1}^{[\tau]} \left( u_r - \sum_{j=1}^{r} u_j d' \left( \sum_{j=1}^{r} d_j d'_j \right)^{-1} \right) \]

\[ = RT^{-1/2} S_{[\tau]} - RT^{-1/2} \sum_{j=1}^{T} u_j d'_j \tau_T \left( T^{-1} \sum_{j=1}^{T} \tau_j d_j d'_j \right)^{-1} T^{-1} \sum_{r=1}^{[\tau]} \tau_r d_r. \]

\[ \Rightarrow \int_0^c \Lambda^* dW_q(s) - \left( \int_0^c \Lambda^* dW_q(s) f(s) \right) \left( \int_0^c f(s) f(s)' ds \right)^{-1} \int_0^c f(s) ds = \Lambda^* B^f_q(c), \] (19)

where

\[ B^f_q(c) = \int_0^c dW_q(s) - \left( \int_0^c dW_q(s) f(s) \right) \left( \int_0^c f(s) f(s)' ds \right)^{-1} \int_0^c f(s) ds, \]

and the limit in (19) follows from (12), (17) and (18). Using (19) it immediately follows that

\[ R \hat{\Omega}_T R' = R 2T^{-2} \sum_{r=1}^{T} \hat{S}_r \hat{S}'_r R' = 2T^{-1} \sum_{r=1}^{T} RT^{-1/2} \hat{S}_r \hat{S}'_r T^{-1/2} R' \xrightarrow{d} 2 \Lambda^* \int_0^c B^f_q(s) B^f_q(s)' ds \Lambda^*. \] (20)

Combining (14), (16) and (20) gives the limit of \( VF : \)

\[ VF = \sqrt{T} \tau_{0 \tau}^{-1} (R \hat{\beta} - r) \left( T^{-1} \tau_{0 \tau}^2 \sum_{r=1}^T \tilde{d}_r^2 \right)^{-1} \sqrt{T} \tau_{0 \tau}^{-1} (R \hat{\beta} - r) / q \]

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

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

Using well known properties of Wiener processes, it follows that

\[ \]
\[
\left( \int_0^1 \tilde{f}_0(s)^2 \, ds \right)^{-1} \left( \int_0^1 \tilde{f}_0(s) \, dW_q(s) \right) = Z_q \sim N(0, I_q),
\]
which allows us to write
\[
VF \rightarrow Z_q' \left[ 2 \int_0^1 B_q^l(s) B_q^l(s)' \, ds \right]^{-1} Z_q/q = VF_q^{-\infty}.
\]

It can be shown that the normal vector, \( Z_q \), is independent of the random matrix,
\[
2 \int_0^1 B_q^l(s) B_q^l(s)' \, ds.
\]
Therefore, \( VF_q^{-\infty} \) is similar to an \( F \) random variable but is nonstandard and
depends on the deterministic regressors in the model via the stochastic process \( B_q^l(s) \). The critical
values of \( VF_q^{-\infty} \) depend on the regressors included in \( d_i \) and depend on the break date through \( \lambda \)
but the critical values do not depend on which regressor is placed in \( d_{0t} \) (the regressor of interest
for hypothesis testing). In other words, 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
\[
VF_t = \frac{R\hat{\beta} - r}{\sqrt{\sum_{r=1}^T \tilde{d}_{0r}^2} \, R\hat{\Omega}_T R'}
\]
and its limit is given by
\[
VF_t \rightarrow Z_1 \frac{Z_1}{\sqrt{2 \int_0^1 B_t^l(s) B_t^l(s)' \, ds}} = VF_t^{-\infty}.
\]
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 (21) and (22) 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 level shift, \( \lambda \), affects the form of \( d_t \) and hence the form of \( f(s) \). Therefore, the location of the level shift affects the asymptotic critical values of \( VF \) and \( VF_t \).

In the application, when we take the date of the Pacific Climate Shift to be exogenously given at 1977:12 and this yields a value of \( \lambda = 0.3774 \). For model (2) with \( \lambda = 0.3774 \) we simulated the asymptotic critical values of \( VF \) and \( VF_t \) for testing one restriction \( (q=1) \) which we tabulate in Table 1a. 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 \( (1, 1(t > 0.3774T), t/T)^T \) for \( t=1,2,...,T \). The integrals are approximated by simple averages. 50,000 replications were used. We see from Table 1a 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 a fatter tail than a \( \chi^2_1 \) random variable.

### 3.3 Bootstrap Critical Values and P-values

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 (6) and sample with replacement from \( \hat{u}_{i1}, \hat{u}_{i2}, ..., \hat{u}_{iT} \) 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 (3) by OLS using \( y_{it}^* \) in place of \( y_{it} \). Let \( \hat{\beta}_i^* \) and \( \hat{\delta}_i^* \) denote the OLS estimators and let \( \hat{e}_{it}^* = y_{it}^* - \hat{\beta}_i^* d_{it} + \hat{\delta}_i^* d_{it} \) denote the OLS residuals. Let \( \hat{\epsilon}_i^* \) denote the \( n \times 1 \) vector \( \hat{\epsilon}_i^* = (\hat{\epsilon}_{i1}^*, \hat{\epsilon}_{i2}^*, ..., \hat{\epsilon}_{in}^*)' \) 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 (8) with $\hat{\epsilon}_t^*$ in place of $\hat{u}_t$. The equivalent form given by (10) can also be used and is faster to compute.

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

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

5. Repeat Steps 1 through 4 $N_B$ times where $N_B$ is a relatively large integer. This generates $N_B$ random draws from $VF^*$ or $VF_i^*$.

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 $VF_i$. For the $VF$ statistic the right tail critical value for a test with significance level $\alpha$ 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 $VF_i$, the critical value is given by $VF_i^*[\alpha N_B]$ and for a right tail test the critical value is given by $VF_i^*[(1-\alpha)N_B]$.

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 $\beta_i^*$ is zero. Therefore, $R\beta^* = 0$, i.e. $r = 0$ in the bootstrap samples and $VF^*$ and $VF_i^*$ are computed using $r = 0$ to ensure that the null holds for $VF^*$ and $VF_i^*$. 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 $VF_i^*$ are based on HAC estimators and their asymptotic null distributions do not depend on unknown correlation parameters, $VF^*$ and $VF_i^*$ 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

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

In other words, the bootstrap statistics have the same limits as the \(VF\) and \(VF_i\) statistics under the null hypothesis. Therefore, the bootstrap critical values are equivalent to the approximations given by (21) and (22).

## 4 Treating the Break Date as Unknown

In the application we think it is reasonable to treat the break date of the level shift as known because the PCS was an exogenous event with respect to comparisons of climate model generated temperature series and observed temperature series. As a robustness check to the exogeneity assumption about the date of the PCS, we also report results where we treat the date of the level shift as unknown. We take a “data-mining” approach which has a long history in the change point literature. For a given hypothesis, we compute the \(VF\) statistic for a grid of possible break dates and determine the break date that gives the largest \(VF\) statistic. In other words, we search for the break date that gives the strongest evidence against the null hypothesis. If the effect of searching over break dates is unaccounted for, this approach is a "data-mining" exercise that could give potentially misleading inference: the level of the test will be inflated above the nominal level compared to the case where the break date is assumed to be known. Fortunately, it is easy to obtain critical values that take into account the search over break dates.

For a given potential break date, \(T_b\), let \(VF(T_b)\) denote the \(VF\) statistic for testing a given null hypothesis. The limiting random variable give by (21) depends on \(\lambda\) through the level shift regressor and we now label the limit by \(VF^*_q(\lambda)\) to explicitly acknowledge the dependence on the break date used to estimate the model. Consider a grid of potential break dates given by

17
Define the “data-mined” VF statistic as
\[ \sup VF = \sup_{T_b} VF(T_b) \text{ for } T_b \in [T^*_b + 1, T^*_b + 2, \ldots, T - T^*_b] \].

Under the null hypothesis (7) and under the assumption there is no level shift in the data, we have
\[ \sup VF \rightarrow^d \sup_{\lambda \in (\lambda^*, \lambda^*)} VF^\infty_q (\lambda) \] (23)
where the limit follows from (21) and application of the continuous mapping theorem. Using simulation methods identical to those used for the known break date case, we computed asymptotic critical values for \( \sup VF \) for \( \lambda^* = 0.1 \) and \( q=1 \) for testing hypotheses about the trend slope parameters in model (2). These critical values are given in Table 1b. Using the \( \sup VF \) statistic along with the critical values given by (23) provides a very conservative test with regard to the break date.

5 Testing for a Shift in Level of a Univariate Time Series

As part of the empirical application we provide visual evidence that the observed temperature series exhibit level shifts around the time of the PCS. Some formal statistical evidence regarding these level shifts can be provided by application of the VF statistic to an individual time series. Consider model (2) for the case of \( n=1 \) and place the model in the general framework (3) with \( d_{0t} = DU_t, \beta_1 = g_1 \) and \( d_{1t} = (1,t)' \), \( \delta_1 = (a_1, b_1)' \). If we take the break date as known, then the VF statistic for testing for no level shift, \( H_0 : g_1 = 0 \) can be computed as before using (9) with \( R=1 \) and \( r=0 \). The asymptotic null critical values are still given by Table 1a.

If we treat the break date as unknown, we can apply the \( \sup VF \) statistic although the asymptotic critical values depend on which regressor is placed in \( d_{0t} \). While it is true that for a given value of
\( \lambda \), the distribution of \( VF^\infty_q (\lambda) \) is the same regardless of the regressor placed in \( d_{0i} \), the covariance structure of \( VF^\infty_q (\lambda) \) across \( \lambda \) depends on which regressor is placed in \( d_{0i} \). Therefore, the sup\( VF \) statistic for testing zero trend slope has different asymptotic critical values than the sup\( VF \) statistic for testing a zero level shift. We simulated the asymptotic critical values of sup\( VF \) for testing for a zero level shift for the case of \( \lambda^* = 0.1 \) and \( q = 1 \) and provide those critical values in Table 1b.

6 Data and Methods

6.1 Observations and Model Data

Our application mainly focuses on trend slopes and comparisons of trend slopes across series in which case we set \( d_{0i} = t \) and therefore \( \beta_i = b_i \). For model (1) \( d_{ii} = 1 \), and for model (2) \( d_{ii} = (1, DU_i)' \) with the level shift set at 1977:12, implying \( \lambda = 0.3774 \). We also provide some results on the level shift parameters themselves in which case \( d_{0i} = DU_i \), \( \beta_i = g_i \) and \( d_{ii} = (1, t)' \).

Let \( \hat{\beta}_i \) denote the OLS estimator of \( \beta_i \) for a given parameter of interest for a given time series using either model (1) or model (2). The VF standard error of \( \hat{\beta}_i \) is given by

\[
se(\hat{\beta}_i) = \left( \sum_{i=1}^{T} \hat{d}_{0i}^2 \right)^{-1/2} \hat{\Omega}_i^t, 
\]

where \( \hat{\Omega}_i^t \) is computed with (8) (or equivalently (10)) using \( \hat{a}_{it} \) from the respective models. Let \( cv_{0.025} \) denote the 2.5% right tail critical value of the asymptotic distribution of \( VF_0 \). For model (1) \( cv_{0.025} = 6.482 \) (see Table 1 of Vogelsang and Franses, 2005; their \( t_2^* \) statistic) and for model (2) \( cv_{0.025} = 6.965 \) (see Table 1a). A 95% confidence interval (CI) is computed as \( \hat{\beta}_i \pm se(\hat{\beta}_i) \cdot cv_{0.025} \).

Our empirical application uses data from the tropical 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 level breaks were
not warranted. Soden and Held (2005), Karl et al. (2006), Douglass et al. (2007) and Santer et al. (2008) discuss the particular importance of examining the tropical troposphere for assessing GCM performance.

The tropics are defined as 20°N to 20°S. The GCM runs were compiled for McKitrick et al. (2010). There were 57 runs from 23 models for each of LT and MT layers. 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 level 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 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.

Figure 2 displays the observed LT and MT trends with the least squares trend lines shown. The estimated trends are 0.14 and 0.16 °C/decade in the LT and 0.09 and 0.11 °C/decade in the MT. The effect of allowing for a level shift (step-change) in 1977:12 is shown in Figure 3. The two radiosonde series are averaged and the broken trend is plotted along with the broken trend (with break located at the same place) of the GCM average. Individual monthly observations from three GCMs are also shown for illustrative purposes. Using a break date of 1977:12 the observed LT trends fall to 0.06 and 0.09 °C/decade and the MT trends fall to -0.001 and 0.04 °C/decade. Thus about half of the positive LT trend in Figure 2 can be attributed to the one-time change at 1977:12 and essentially all the MT change is accounted for by the step-change.

\(^1\)http://www.metoffice.gov.uk/hadobs/hadat/msu/anomalies/hadat_msu_tropical.txt.
Figure 4 plots all the estimated trend slopes along with their CIs. The top row leaves out the level shift and the bottom row includes it. The model-generated trends are grouped on the left with the 95% 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. With or without the break term the range of model runs and their associated CI’s overlaps 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 4 (spanning roughly 0.1 to 0.4 C/decade in each layer) 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 physical theory. Using a range comparison, the fact that, in Figure 4, 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.

6.2 Multivariate Trend Comparisons in the No-break and Known Break Cases

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 \( d_{t0} = t \). We are interested in testing the null hypothesis that the weighted average of the trend slopes in the 23 climate model generated series is the same as the average trend slope of the observed series. The weight coefficient \( w_i \) equals the number of runs in model \( i \)'s ensemble mean, to adjust for the reduction in variance in multi-run ensemble means. Placing the observed series in positions \( i=24,25 \), the restriction matrices for this null hypothesis are
\[
R = \left[ \frac{w_1}{57}, \frac{w_2}{57}, \ldots, \frac{w_{23}}{57}, -\frac{1}{2}, -\frac{1}{2} \right], \quad r = 0
\]

where the \(w_i\) terms sum to 57.

Table 4 presents the \(VF\) statistics for the test of trend equivalence between 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, the magnitudes of which (°C/decade) are indicated in parentheses beside the series name. Asymptotic critical values are provided in the table captions 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 10,000 bootstrap replications.

In the trend model without level shifts (top panel of Table 4), the zero trend-hypothesis is rejected at the 1% significance level for all 4 observed series, indicating strong evidence of a significant warming trend in the tropical troposphere 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. Table 5 repeats the model-observation trend equivalence test for each of the 23 models individually. In the LT layer the differences are significant at 5% or lower in 5 cases and 14 cases in the MT layer. (Not reported are the single-model tests of trend significance, which are significant at <1% in both layers for 22 models, at <10% for one model in the LT layer and insignificant for one model in the MT layer.) So while, on average, the model trends are significantly different from observations, in the LT layer it can at least be said that if we ignore the step-change at 1977:12, almost four-fifths of the models have trends that individually do not significantly differ from the observations.

When we add the level shift dummy at 1977:12 (middle block of Table 4), the trend magnitudes and values of the \(VF\) statistics for testing the zero trend-hypothesis drop considerably. 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 intercept shift is left out, the increase in the series is spuriously associated with a trend slope. But the entire trend is explained by a jump in the data around 1977.
The VF test of equivalence of trends between the climate models and observed data is more strongly rejected when the level 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 level shift dummy is included, the estimated trend slopes of the climate model series are not systematically affected by the level shift dummy. Therefore, there is a greater discrepancy between the climate model trends and the observed trends. Table 5 confirms this in the model-specific tests (see “Known Break Date” columns). The number of rejections in the LT layer jumps from 5 to 14 out of 23, and in the MT layer from 14 to 17 out of 23.

The VF scores on the level shift magnitudes for the individual time series are shown in Tables 2 and 3 for the case where the break date is 1977:12 and is treated as known. Not surprisingly, the break terms are not significant for the model generated series. Things are different for the observed series. The Hadley and Rich series yield break terms that are significant at 10% in the LT and 5% in the MT. It might seem surprising that the effect of the break dummy is so dramatic on the estimated trend slope parameters, yet the break coefficients themselves are not strongly significant. This is not surprising because, in general, trend slopes are estimated more efficiently than level shifts (or intercepts) and therefore it is more difficult to conduct inference about level shifts than about trend slopes. While there is sufficient noise in the data to make inference about level shifts difficult, the noise is not so large to mask information about the trend slopes. In addition, unmodeled level shifts also induce spurious noise into the model when conducting inference about trend slopes. Therefore, controlling for a level shift makes inferences about trend slopes more informative.

6.3 Multivariate Trend Comparisons in the Unknown-Break Case

The climate-models do not explicitly model the Pacific Climate Shift and so the level shift coefficient has no special meaning for the climate model data. Not surprisingly, the estimated level shift coefficients were positive in 11 cases and negative in 12 of the climate model series.
As a robustness check regarding the exogeneity assumption about the break date of the PCS, we report results where we treat the break date as unknown and use the supVF statistic. The supVF statistic is calculated by computing the VF score with the break term $T_b$ sequentially set across the middle 80% of the data set then selecting the maximum value. We ignore the 10% at each end of the sample since the break test is not reliable near the sample boundaries.

For the tests of model-observational equivalence allowing for a level shift, Figure 5 shows the sequence of VF scores, with 10%, 5% and 1% critical values shown. The supVF occurs at 1979:5 in the LT layer and at 1978:10 in the MT layer. These break dates are close to, but not the same as, 1977:12 and all of the VF scores for dates near and around 1977:12 time interval far exceed the 1% critical values for the SupVF statistic. Since the supVF test is conservative regarding the choice of break date, this provides strong additional support for the results in Section 6.2.

The SupVF scores of tests of model-observational equivalence are reported in Table 4 for model averages and in Table 5 for individual models. A pattern emerges particularly clearly in Table 5 that when we apply the data mining approach, the test scores get larger, as expected, but so do the critical values. The net effect is to reduce the significance of the test scores when the break date is treated as unknown. Had we not used the critical values as given by (23), we would have spuriously inflated the significance of our findings. This is a useful lesson in the perils of naïve data-mining, in which a specification is selected that maximizes the chance of rejecting some null hypothesis, without taking into account the effect of the data mining process on the null distribution of the test. The price one pays for being honest and using the conservative critical values implied by (23) is lower power in detecting a deviation from the null hypothesis. Even with lower power, we see in the third panel of Table 4 that the average model is still clearly rejected against the data in both the LT and MT layers. Since this result is not dependent on choosing a particular break date it provides very strong confirmation of the known break date empirical finding.

Regarding the trend magnitudes, we do not report the supVF score for the test of a zero trend in the 3rd block of Table 4. Recall that the search process looks for the location that maximizes the chance of rejecting a null hypothesis. In this case the significance of the trend would be maximized
simply by leaving the break term out altogether, which corresponds to the test scores in the first block of Table 4.

We not report the SupVF scores for the level shift parameters. While the SupVF scores are greater than the VF scores with break date 1977:12, the critical values, as given by Table 1b, increase substantially, and we do not reject the null of no level shift for the observed series. Again the problem is the low power of detecting a level shift in data with substantial noise. Treating the break date as unknown further reduces this already low power. Even if were were to choose a model with no level shifts in the observed temperature series, we would still reject model-observational equivalence as shown in the first block of Table 4.

7 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 around 1978. With respect to the Vogelsang and Franses (2005) approach, this amounts to adding a level 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 level shift term at December 1977. 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 different from zero in either the LT or MT layers. Most climate models predict a significantly larger trend over this interval than is observed in either lower- or mid-tropospheres. We find a statistically significant mismatch between the average climate model trend 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).

The testing method used herein is both powerful and conservative. The power of the test is indicated by the span of test scores in Table 5 in which relatively small changes in modeled trends translate into much higher rejection probabilities. Using the data-mining method provides a check on the extent to which the results depend on the assumption of a known break date.

As such our empirical approach has many other potential applications on climatic and other data sets in which level shifts are believed to have occurred. Examples could include stratospheric temperature trends which are subject to level shifts coinciding with major volcanic eruptions, and land surface trends where it is believed that the measuring equipment changed or was moved.
APPENDIX: MOTIVATION AND BACKGROUND OF VF APPROACH

Motivation for the form of the VF statistic can be developed by considering the very simple model

\[ y_{it} = a_i + u_{it}. \] \hspace{1cm} (A1)

Model (A1) can written in terms of the general model (3) setting \( d_{it} = 1 \), \( \beta_i = a_i \) and \( d_{iu} = 0 \). Because \( u_t \) is a mean zero time series, it follows that \( a_i = E(y_{it}). \) 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: \( y_t = (y_{i1}, y_{i2}, \ldots, y_{im})' \), \( a = (a_1, a_2, \ldots, a_n)' \) and \( u_t = (u_{1t}, u_{2t}, \ldots, u_{mt})' \).

Rewrite the model as

\[ y_t = a + u_t, \]

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

\[ H_0 : Ra = r, \hspace{1cm} H_1 : Ra \neq r, \]

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 \( \hat{a} = \bar{y} = T^{-1} \sum_{t=1}^{T} y_t. \)

The variance of \( \hat{a} \), given the covariance stationarity assumption for \( u_t \), is given by

\[ Var(\hat{a}) = T^{-2} E[(\sum_{i=1}^{T} u_i)(\sum_{i=1}^{T} u_i)'] = T^{-1} \left[ \Gamma_0 + \sum_{j=1}^{T-1} (1 - \frac{j}{T}) (\Gamma_j + \Gamma_j') \right]. \]

Letting \( \Omega_T = \Gamma_0 + \sum_{j=1}^{T-1} (1 - \frac{j}{T}) (\Gamma_j + \Gamma_j') \) we have the more compact expression
\[ \text{Var}(\hat{\alpha}) = T^{-1} \Omega_T. \quad (A2) \]

An \( F \)-statistic constructed using (A2) is infeasible because \( \Omega_T \) is unknown. A natural estimator of \( \Omega_T \) is given by

\[
\hat{\Omega}_T = \Gamma_0 + \sum_{j=1}^{T-1} (1 - \frac{j}{T})(\hat{\Gamma}_j + \hat{\Gamma}'_j), \quad \hat{\Gamma}_j = T^{-1} \sum_{t=j+1}^{T} \hat{u}_t \hat{u}'_{t-j},
\]

where \( \hat{u}_t = y_t - \hat{\alpha} \). Using \( \hat{\Omega}_T \) in place of \( \Omega_T \) leads to the VF statistic:

\[
\text{VF} = (R\hat{\alpha} - r)'[T^{-1}R\hat{\Omega}_T R']^{-1}(R\hat{\alpha} - r)/q.
\]

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 \( \Omega \).

Asymptotic theory is used to generate an approximation for \( \hat{\Omega}_T \) and the null distribution of \( \text{VF} \) using a functional central limit theorem (FCLT) for the scaled partial sums of \( u_t \). If it were the case that \( \hat{\Omega}_T \) were a consistent estimator of \( \Omega \), then \( \text{VF} \) would converge in distribution to a \( \chi^2_q/q \) random variable. It turns out that \( \hat{\Omega}_T \) is not a consistent estimator of \( \Omega \), however, it is relatively easy to show that \( \hat{\Omega}_T \) does converge in distribution to a random matrix that is proportional to \( \Omega \) but otherwise does not depend on unknown quantities. This property of \( \hat{\Omega}_T \) means that the \( \text{VF} \) statistic can be used to test \( H_0 \) because \( \text{VF} \) can be approximated by a random variable that does not depend on unknown parameters.

Recall the partial sum of \( u_t \) given by \( S_t = \sum_{j=1}^{t} u_j \). Evaluating \( S_t \) at \( t = [cT] \) gives \( S_{[cT]} = \sum_{r=1}^{[cT]} u_t \), which is the sum of the first \( c^T \) proportion of the data. For a given value of \( c \), the quantity \( [cT] \to \infty \) as \( T \to \infty \); therefore if we scale by \( T^{-1/2} \) we obtain the result
For a given value of \( c \), the scaled partial sums of \( u_t \) satisfy a Central Limit Theorem (CLT). These limits hold pointwise in \( c \). The FCLT given by (11) 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 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 \( \sum_{j=0}^{\infty} |\Gamma^{(lm)}_j| < \infty \) is strengthened to \( \sum_{j=1}^{\infty} |\Gamma^{(lm)}_j| < \infty \) which requires the autocovariances to shrink faster to zero as \( j \) increases.

Using the FCLT given by (11) it immediately follows that

\[
\sqrt{T}(\hat{\alpha} - \alpha) = \sqrt{T} u = T^{-\frac{1}{2}} \sum_{t=1}^{T} u_t = T^{-\frac{1}{2}} S_T \Rightarrow \Lambda W_n(1) \sim N(0, \Lambda \Sigma_n \Lambda') = N(0, \Omega).
\]

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 \( \hat{S}_j = \sum_{i=1}^{T} \hat{u}_j \) using (10):

\[
\hat{\Omega}_T = \hat{\Sigma}_0 + \sum_{j=1}^{T-1} (1 - \frac{j}{T})(\hat{\Sigma}_j + \hat{\Sigma}'_j) = 2T^{-2} \sum_{j=1}^{T-1} \hat{S}_j \hat{S}'_j.
\]

Using the FCLT, the limit of \( T^{-\frac{1}{2}} \hat{S}_{[cT]} \) is easy to derive:

\[
T^{-\frac{1}{2}} \hat{S}_{[cT]} = T^{-\frac{1}{2}} \sum_{i=1}^{[cT]} \hat{u}_i = T^{-\frac{1}{2}} \sum_{i=1}^{[cT]} (y_i - \hat{u}) = T^{-\frac{1}{2}} \sum_{i=1}^{[cT]} (a + u_i - \hat{u})
\]

\[
= T^{-\frac{1}{2}} \sum_{i=1}^{[cT]} u_i - T^{-\frac{1}{2}} [cT]|(\hat{u} - \hat{u}) = T^{-\frac{1}{2}} S_{[cT]} - \left( \frac{cT}{T} \right) \sqrt{T}(\hat{\alpha} - \alpha)
\]

\[
\Rightarrow \Lambda W_n(c) - r\Lambda W_n(1) = \Lambda(W_n(c) - c W_n(1)) \equiv \Lambda B_n(c).
\]
The stochastic process, $B_n(c)$, is the well known Brownian bridge. Using this result for $T^{-\frac{1}{2}} \tilde{S}_{|c[T]}$ and the continuous mapping theorem, it follows that

$$
\hat{\Omega}_T = 2T^{-1} \sum_{i=1}^{T-1} (T^{-\frac{1}{2}} \tilde{S}_i)(T^{-\frac{1}{2}} \tilde{S}_i') \Rightarrow 2\Lambda \int_0^1 B_n(c)B_n(c)' dc \Lambda'.
$$

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:

$$
VF = (R\hat{\alpha} - r)'[T^{-1} \hat{\Omega}_T R']^{-1} (R\hat{\alpha} - r)/q = \sqrt{T} (R\hat{\alpha} - r)'[R\hat{\Omega}_T R']^{-1} \sqrt{T} (R\hat{\alpha} - r)/q
$$

$$
= (R\sqrt{T} \hat{\omega})'[T^{-1} \hat{\Omega}_T R']^{-1} R\sqrt{T} \hat{\omega}/q \Rightarrow (R\Lambda W_q(1))'[R2\Lambda \int_0^1 B_n(c)B_n(c)' dc \Lambda']^{-1} R\Lambda W_q(1)/q.
$$

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, $R\Lambda W_q(c)$ is a $q \times 1$ vector of Wiener processes and we can rewrite $R\Lambda W_q(c)$ as $\Lambda^* W_q(c)$ where $\Lambda^*$ 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 $B_q(c) = W_q(c) - c W_q(1)$. Because $R$ is assumed to be full rank, it follows that $\Lambda^*$ is full rank and is therefore invertible. We have

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

$$
= (\Lambda^* W_q(1))'[2\Lambda^* \int_0^1 B_q(c)B_q(c)' dc \Lambda^* ]^{-1} \Lambda^* W_q(1)/q
$$

$$
= W_q(1)[2\int_0^1 B_q(c)B_q(c)' dc ]^{-1} W_q(1)/q,
$$

and the $\Lambda^*$ matrices drop out because $\Lambda^*$ 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 $2\int_{0}^{1} B_q(c)B_q(c)\,dc \cdot$. 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 $R\hat{\Omega}R'$ whereas $W_q(1)$ approximates the randomness of $\sqrt{T}(R\hat{a} - r)$.

REFERENCES


Davies, R.B. (1987) “Hypothesis testing when a nuisance parameter is present only under the alternative,” *Biometrika*, 74, 33-43.


### Tables

#### Table 1a: Asymptotic Critical Values. Model (2), known break date with $\lambda = 0.3774$, $q=1$. Left tail critical values of $VF_i$ follow by symmetry around zero.

<table>
<thead>
<tr>
<th>%</th>
<th>$VF_i$</th>
<th>$VF$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.700</td>
<td>1.668</td>
<td>11.479</td>
</tr>
<tr>
<td>.750</td>
<td>2.166</td>
<td>14.429</td>
</tr>
<tr>
<td>.800</td>
<td>2.728</td>
<td>18.312</td>
</tr>
<tr>
<td>.850</td>
<td>3.388</td>
<td>23.676</td>
</tr>
<tr>
<td>.900</td>
<td>4.279</td>
<td>32.178</td>
</tr>
<tr>
<td>.950</td>
<td>5.673</td>
<td>48.514</td>
</tr>
<tr>
<td>.975</td>
<td>6.965</td>
<td>67.987</td>
</tr>
<tr>
<td>.990</td>
<td>8.639</td>
<td>97.928</td>
</tr>
<tr>
<td>.995</td>
<td>9.896</td>
<td>123.92</td>
</tr>
</tbody>
</table>

#### Table 1b: Asymptotic Critical Values. Model (2), unknown break date, $q=1$. 10% Trimming ($\lambda^* = 0.1$).

<table>
<thead>
<tr>
<th>%</th>
<th>supVF trend slope</th>
<th>supVF intercept shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>.700</td>
<td>79.765</td>
<td>95.455</td>
</tr>
<tr>
<td>.750</td>
<td>88.184</td>
<td>109.94</td>
</tr>
<tr>
<td>.800</td>
<td>98.532</td>
<td>116.20</td>
</tr>
<tr>
<td>.850</td>
<td>111.78</td>
<td>130.76</td>
</tr>
<tr>
<td>.900</td>
<td>131.92</td>
<td>150.99</td>
</tr>
<tr>
<td>.950</td>
<td>166.41</td>
<td>188.68</td>
</tr>
<tr>
<td>.975</td>
<td>205.15</td>
<td>225.78</td>
</tr>
<tr>
<td>.990</td>
<td>261.39</td>
<td>279.85</td>
</tr>
<tr>
<td>.995</td>
<td>301.94</td>
<td>322.48</td>
</tr>
<tr>
<td>Data Series</td>
<td>Model/ Obs Name Extra Forcings; No. runs</td>
<td>Simple Trend</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>1</td>
<td>BCCR BCM2.0</td>
<td>0.173</td>
</tr>
<tr>
<td>2</td>
<td>CC;CMA3.1-T47 NA; 5</td>
<td>0.347</td>
</tr>
<tr>
<td>3</td>
<td>CCCMA3.1-T63 NA; 1</td>
<td>0.373</td>
</tr>
<tr>
<td>4</td>
<td>CNRM3.0 O; 1</td>
<td>0.249</td>
</tr>
<tr>
<td>5</td>
<td>CSIRO3.0 1</td>
<td>0.139</td>
</tr>
<tr>
<td>6</td>
<td>CSIRO3.5 1</td>
<td>0.242</td>
</tr>
<tr>
<td>7</td>
<td>GFDL2.0 O, LU, SO, V; 1</td>
<td>0.186</td>
</tr>
<tr>
<td>8</td>
<td>GFDL2.1 O, LU, SO, V; 1</td>
<td>0.109</td>
</tr>
<tr>
<td>9</td>
<td>GISS_AOM 2</td>
<td>0.171</td>
</tr>
<tr>
<td>10</td>
<td>GISS_EH 0, LU, SO, V; 6</td>
<td>0.193</td>
</tr>
<tr>
<td>11</td>
<td>GISS_ER 0, LU, SO, V; 5</td>
<td>0.178</td>
</tr>
<tr>
<td>12</td>
<td>IAP_FGOALS1.0 3 O, LU, SO, V; 5</td>
<td>0.198</td>
</tr>
<tr>
<td>13</td>
<td>ECHAM4 1</td>
<td>0.210</td>
</tr>
<tr>
<td>14</td>
<td>INMCM3.0 SO, V; 1</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>Model</td>
<td>Ozone Depletion</td>
</tr>
<tr>
<td>---</td>
<td>---------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>15</td>
<td>IPSL_CM4</td>
<td>0.189</td>
</tr>
<tr>
<td>16</td>
<td>MIROC3.2_T106</td>
<td>0.141</td>
</tr>
<tr>
<td>17</td>
<td>MIROC3.2_T42</td>
<td>0.210</td>
</tr>
<tr>
<td>18</td>
<td>MPI2.3.2a</td>
<td>0.205</td>
</tr>
<tr>
<td>19</td>
<td>ECHAM5</td>
<td>0.204</td>
</tr>
<tr>
<td>20</td>
<td>CCSM3.0</td>
<td>0.217</td>
</tr>
<tr>
<td>21</td>
<td>PCM_B06.57</td>
<td>0.176</td>
</tr>
<tr>
<td>22</td>
<td>HADCM3</td>
<td>0.190</td>
</tr>
<tr>
<td>23</td>
<td>HADGEM1</td>
<td>0.226</td>
</tr>
<tr>
<td>24</td>
<td>HadAT</td>
<td>0.135</td>
</tr>
<tr>
<td>25</td>
<td>RICH</td>
<td>0.157</td>
</tr>
</tbody>
</table>

**Table 2: Summary of Lower Troposphere data series.** 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>
<thead>
<tr>
<th>Data Series</th>
<th>Model/ Obs Name Extra Forcings; No. runs</th>
<th>Simple Trend</th>
<th>Trend + Level shift</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Trend (C/decade)</td>
<td>Std Error</td>
</tr>
<tr>
<td>1</td>
<td>BCCR BCM2.0 O; 2</td>
<td>0.176</td>
<td>0.0060</td>
</tr>
<tr>
<td>2</td>
<td>CC;CMA3.1-T47 NA; 5</td>
<td>0.372</td>
<td>0.0046</td>
</tr>
<tr>
<td>3</td>
<td>CCCMA3.1-T63 NA; 1</td>
<td>0.399</td>
<td>0.0093</td>
</tr>
<tr>
<td>4</td>
<td>CNRM3.0 O; 1</td>
<td>0.311</td>
<td>0.0072</td>
</tr>
<tr>
<td>5</td>
<td>CSIRO3.0 1</td>
<td>0.108</td>
<td>0.0086</td>
</tr>
<tr>
<td>6</td>
<td>CSIRO3.5 1</td>
<td>0.229</td>
<td>0.0097</td>
</tr>
<tr>
<td>7</td>
<td>GFDL2.0 O, LU, SO, V; 1</td>
<td>0.174</td>
<td>0.0117</td>
</tr>
<tr>
<td>8</td>
<td>GFDL2.1 O, LU, SO, V; 1</td>
<td>0.103</td>
<td>0.0198</td>
</tr>
<tr>
<td>9</td>
<td>GISS_AOM 2</td>
<td>0.163</td>
<td>0.0081</td>
</tr>
<tr>
<td>10</td>
<td>GISS_EH O, LU, SO, V; 6</td>
<td>0.180</td>
<td>0.0114</td>
</tr>
<tr>
<td>11</td>
<td>GISS_ER O, LU, SO, V; 5</td>
<td>0.162</td>
<td>0.0127</td>
</tr>
<tr>
<td>12</td>
<td>IAP_FGOALS1.0 3</td>
<td>0.185</td>
<td>0.0125</td>
</tr>
<tr>
<td>13</td>
<td>ECHAM4 1</td>
<td>0.200</td>
<td>0.0131</td>
</tr>
<tr>
<td>14</td>
<td>INMCM3.0 SO, V; 1</td>
<td>0.183</td>
<td>0.0100</td>
</tr>
<tr>
<td>Model</td>
<td>Variables</td>
<td>Temp</td>
<td>Precip</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>IPSL_CM4</td>
<td>O, LU, SO, V; 1</td>
<td>0.195</td>
<td>0.0081</td>
</tr>
<tr>
<td>MIROC3.2_T106</td>
<td>O, LU, SO, V; 1</td>
<td>0.147</td>
<td>0.0113</td>
</tr>
<tr>
<td>MIROC3.2_T42</td>
<td>O, LU, SO, V; 3</td>
<td>0.211</td>
<td>0.0143</td>
</tr>
<tr>
<td>MPI2.3.2a</td>
<td>SO, V; 5</td>
<td>0.182</td>
<td>0.0124</td>
</tr>
<tr>
<td>ECHAM5</td>
<td>O; 4</td>
<td>0.202</td>
<td>0.0059</td>
</tr>
<tr>
<td>CCSM3.0</td>
<td>O, SO, V; 7</td>
<td>0.201</td>
<td>0.0132</td>
</tr>
<tr>
<td>PCM_B06.57</td>
<td>O, SO, V; 4</td>
<td>0.161</td>
<td>0.0048</td>
</tr>
<tr>
<td>HADCM3</td>
<td>O; 1</td>
<td>0.170</td>
<td>0.0059</td>
</tr>
<tr>
<td>HADGEM1</td>
<td>O, LU, SO, V; 1</td>
<td>0.221</td>
<td>0.0108</td>
</tr>
<tr>
<td>HadAT</td>
<td></td>
<td>0.089</td>
<td>0.0088</td>
</tr>
<tr>
<td>RICH</td>
<td></td>
<td>0.109</td>
<td>0.0075</td>
</tr>
</tbody>
</table>

Table 3: Summary of Mid-Troposphere data series. Notes same as for Table 2.
<table>
<thead>
<tr>
<th>Trend Coef</th>
<th>Null Hypothesis</th>
<th>Test Score</th>
<th>Bootstrap p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Level shift</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hadley LT (0.135)</td>
<td>trend = 0</td>
<td>213.2***</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>RICH LT (0.157)</td>
<td>trend = 0</td>
<td>356.1***</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Hadley MT (0.089)</td>
<td>trend = 0</td>
<td>102.5***</td>
<td>0.007</td>
</tr>
<tr>
<td>RICH MT (0.109)</td>
<td>trend = 0</td>
<td>212.4***</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

| LT Models = Observed | 58.8** | 0.033 |
| MT Models = Observed | 123.9*** | 0.004 |

<table>
<thead>
<tr>
<th>With Level shift at Date (Assumed Known): December 1977</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadley LT (0.063)</td>
<td>trend = 0</td>
<td>9.1</td>
</tr>
<tr>
<td>RICH LT (0.092)</td>
<td>trend = 0</td>
<td>34.8*</td>
</tr>
<tr>
<td>Hadley MT (-0.001)</td>
<td>trend = 0</td>
<td>0.003</td>
</tr>
<tr>
<td>RICH MT (0.043)</td>
<td>trend = 0</td>
<td>13.1</td>
</tr>
</tbody>
</table>

| LT Models = Observed | 361.5*** | < 0.0001 |
| MT Models = Observed | 951.9*** | < 0.0001 |

<table>
<thead>
<tr>
<th>With Level shift at Date Assumed Unknown</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>LT Models = Observed</td>
<td>532.64***</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>MT Models = Observed</td>
<td>1069.96***</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

**Table 4: Results of hypothesis tests using VF statistic with and without level shift term at January, 1977.** Notes: sample period (monthly): January 1958 to December 2010. The bootstrap p-value* is computed using the method described in Section 3.3 using 10000 bootstrap replications. **VF Critical Values:** Without level shift, 20.14 (10%, denoted *) 41.53 (5%, denoted **), 83.96 (1%, denoted ***). With level shift at known date, 31.34 (10%, denoted *), 48.81 (5%, denoted **), 93.56 (1%, denoted ***). With level shift at unknown date, 131.92 (10%, denoted *), 166.41 (5%, denoted **), 261.39 (1%, denoted ***).
<table>
<thead>
<tr>
<th>Model</th>
<th>No Break</th>
<th>Known Break Date</th>
<th>Unknown Break Date</th>
<th>No Break</th>
<th>Known Break Date</th>
<th>Unknown Break Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.99</td>
<td>64.43**</td>
<td>65.24</td>
<td>80.32**</td>
<td>206.48***</td>
<td>206.67**</td>
</tr>
<tr>
<td>2</td>
<td>741.05***</td>
<td>289.41***</td>
<td>598.05***</td>
<td>1497.65***</td>
<td>724.02***</td>
<td>1100.48***</td>
</tr>
<tr>
<td>3</td>
<td>479.95***</td>
<td>1050.09***</td>
<td>1051.79***</td>
<td>636.39***</td>
<td>1898.60***</td>
<td>2132.88***</td>
</tr>
<tr>
<td>4</td>
<td>97.33***</td>
<td>51.67**</td>
<td>100.26</td>
<td>469.47***</td>
<td>192.48***</td>
<td>408.74***</td>
</tr>
<tr>
<td>5</td>
<td>0.22</td>
<td>8.05</td>
<td>18.73</td>
<td>0.43</td>
<td>16.17</td>
<td>18.09</td>
</tr>
<tr>
<td>6</td>
<td>59.65**</td>
<td>115.99***</td>
<td>117.85</td>
<td>93.27***</td>
<td>222.04***</td>
<td>225.39**</td>
</tr>
<tr>
<td>7</td>
<td>23.71</td>
<td>18.47</td>
<td>36.13</td>
<td>58.72**</td>
<td>47.08*</td>
<td>81.52</td>
</tr>
<tr>
<td>8</td>
<td>5.62</td>
<td>7.53</td>
<td>69.53</td>
<td>0.04</td>
<td>24.93</td>
<td>55.14</td>
</tr>
<tr>
<td>9</td>
<td>3.55</td>
<td>7.57</td>
<td>26.33</td>
<td>31.29*</td>
<td>29.52</td>
<td>68.17</td>
</tr>
<tr>
<td>10</td>
<td>16.97</td>
<td>347.40***</td>
<td>454.63***</td>
<td>36.83*</td>
<td>495.44***</td>
<td>512.93***</td>
</tr>
<tr>
<td>11</td>
<td>6.271</td>
<td>145.08***</td>
<td>165.53*</td>
<td>20.40*</td>
<td>192.87***</td>
<td>197.70**</td>
</tr>
<tr>
<td>12</td>
<td>16.00</td>
<td>339.36***</td>
<td>427.13***</td>
<td>35.91*</td>
<td>466.40***</td>
<td>466.40***</td>
</tr>
<tr>
<td>13</td>
<td>28.04</td>
<td>393.54***</td>
<td>604.61***</td>
<td>53.93*</td>
<td>339.05***</td>
<td>665.39***</td>
</tr>
<tr>
<td>14</td>
<td>4.40</td>
<td>8.84</td>
<td>27.97</td>
<td>37.72*</td>
<td>27.26</td>
<td>59.86</td>
</tr>
<tr>
<td>15</td>
<td>21.92</td>
<td>18.19</td>
<td>47.43</td>
<td>116.07***</td>
<td>57.92***</td>
<td>105.87</td>
</tr>
<tr>
<td>16</td>
<td>0.18</td>
<td>2.62</td>
<td>45.24</td>
<td>19.89</td>
<td>15.83</td>
<td>47.27</td>
</tr>
<tr>
<td>17</td>
<td>23.44</td>
<td>231.46***</td>
<td>262.73***</td>
<td>46.86**</td>
<td>231.36***</td>
<td>243.55**</td>
</tr>
<tr>
<td>18</td>
<td>19.14</td>
<td>169.75***</td>
<td>229.78***</td>
<td>36.14*</td>
<td>186.32***</td>
<td>216.73**</td>
</tr>
<tr>
<td>19</td>
<td>47.94*</td>
<td>45.78*</td>
<td>64.54</td>
<td>159.76***</td>
<td>156.99***</td>
<td>162.39*</td>
</tr>
<tr>
<td>20</td>
<td>20.29</td>
<td>232.09***</td>
<td>296.54***</td>
<td>45.43**</td>
<td>372.73***</td>
<td>381.00***</td>
</tr>
<tr>
<td>21</td>
<td>19.68</td>
<td>58.03**</td>
<td>80.73</td>
<td>142.56***</td>
<td>130.34***</td>
<td>154.22**</td>
</tr>
<tr>
<td>22</td>
<td>25.26</td>
<td>39.48*</td>
<td>66.88</td>
<td>64.75**</td>
<td>80.93**</td>
<td>117.04</td>
</tr>
<tr>
<td>23</td>
<td>102.70***</td>
<td>97.86***</td>
<td>125.84*</td>
<td>143.26**</td>
<td>235.49***</td>
<td>315.84***</td>
</tr>
</tbody>
</table>

<=5%  5/23  14/23  8/23  14/23  17/23  14/23

Table 5. VF tests of equivalent trends between individual model (ensemble mean in cases of multiple runs) and average of balloon series. Column 1: model. Column 2: LT layer, No Break case. Columns 3 and 4 (Break1 and Break2, respectively): with break assumed known at 1977:12 and with break at date assumed unknown. Columns 5-7: same for MT layer. VF Critical Values: Without level shift, 20.14 (10%, denoted *) 41.53 (5%, denoted **), 83.96 (1%, denoted ***). With level shift known to be at 1977:12, 32.45 (10%, denoted *), 48.10 (5%, denoted **), 93.15 (1%, denoted ***). With level shift at unknown date, 131.92 (10%, denoted *), 166.41 (5%, denoted **), 261.39 (1%, denoted ***). Last row: fraction of models exhibiting difference from observations significant at <=5%.
Figure 1. Schematics of two series to be compared.
Figure 2. Radiosonde series. Top left: Hadley LT. Top right: RICH LT. Bottom left, Hadley MT. Bottom right: RICH MT. Least squares trends shown.
Figure 3. Model-Observation comparisons allowing for intercept shift after 1977:12. Left: LT layer, dots indicate monthly data from 3 GCMs, black line, trend through all GCMs, gray line, trend through average of two radiosonde series. Right: same for MT layer.
Figure 4. 1958-2010 Trends and 95% CIs for 23 models (shaded region) and two radiosonde series Hadley and RICH (respectively, individual markers at right edge). Top row: Trends computed without allowing for level shift. Bottom row: Level shift term included in model. Left column: LT. Right column: MT.
Figure 5. Grid search of VF scores of model-observational equivalence varying the break point across the sample. Top: LT. Bottom: MT. Horizontal lines show 10%, 5%, and 1% critical values (resp. 131.92, 166.41 and 261.39). Maximum values are at obs 257 532.64 (LT) and obs 250 1069.96 (MT).