Earth Syst. Dynam., 4, 375–384, 2013
www.earth-syst-dynam.net/4/375/2013/
doi:10.5194/esd-4-375-2013

Earth System Dynamics
Open Access

Comment on "Polynomial cointegration tests of anthropogenic impact on global warming" by Beenstock et al. (2012) – some hazards in econometric modelling of climate change

F. Pretis and D. F. Hendry

Department of Economics, and Institute for New Economic Thinking at the Oxford Martin School, University of Oxford, Oxford, UK

Correspondence to: F. Pretis (felix.pretis@nuffield.ox.ac.uk) and D. F. Hendry (david.hendry@nuffield.ox.ac.uk)

Received: 16 January 2013 – Published in Earth Syst. Dynam. Discuss.: 6 February 2013
Revised: 28 August 2013 – Accepted: 9 September 2013 – Published: 22 October 2013

Abstract. We outline six important hazards that can be encountered in econometric modelling of time-series data, and apply that analysis to demonstrate errors in the empirical modelling of climate data in Beenstock et al. (2012). We show that the claim made in Beenstock et al. (2012) as to the different degrees of integrability of CO_2 and temperature is incorrect. In particular, the level of integration is not constant and not intrinsic to the process. Further, we illustrate that the measure of anthropogenic forcing in Beenstock et al. (2012), a constructed "anthropogenic anomaly", is not appropriate regardless of the time-series properties of the data.

1 Introduction

Global temperature records and radiative forcing of greenhouse gases (GHGs) are non-stationary time series, the statistical properties of which invalidate standard inference procedures that seek to detect relationships between them. Cointegration analysis can be used to overcome the inferential difficulties resulting from stochastic trends when that is the only source of non-stationarity, and is applied to test whether there exist combinations of non-stationary variables that are themselves stationary (see Hendry and Juselius, 2001). Cointegration analysis crucially relies on the time-series properties of the available data, and while tests can be performed on an estimated equation's residual in bi-variate models, cases with more than 2 variables require testing in a system setting (see Engle and Granger, 1987).

In their empirical statistical study of temperature and radiative forcing of greenhouse gases, Beenstock et al. (2012) present statistical tests that purport to show that these variables have different integrability properties, and hence cannot be related unless they polynomially cointegrate. Beenstock et al. (2012) then show that their constructed measure of anthropogenic forcing, an "anthropogenic anomaly", does not cointegrate with observed temperature, presenting this as evidence against anthropogenic global warming.

Beenstock et al. (2012) address an interesting question, to do so they rely heavily on the time-series properties of the data to reach their conclusions. We show that, consistent with the existing literature, the claim that all anthropogenic forcing variables are only stationary in second differences is erroneous. In particular, this level of integration is not constant and not intrinsic to the process. Further, we show that the measure of anthropogenic forcing in Beenstock et al. (2012), a constructed "anthropogenic anomaly", is inappropriate regardless of the time-series properties of the data.

The literature on stochastic trends and cointegration differentiates between series being stationary, trend stationary, first difference stationary, denoted $I(1)$, and stationary in second differences, $I(2)$. Stationary variables are drawn from distributions that are invariant over time. Trend stationarity implies that a series is stationary once a linear trend component is removed. Integrability refers to one aspect of the stationarity properties of a time series. Series that are integrated of

order one, $I(1)$, and two, $I(2)$, are stationary if differenced once or twice respectively.[1] An $I(1)$ process contains a unit-root, that is, the characteristic polynomial of the process has a root at 1. The presence of more than one unit root is indicative of higher order integration (e.g. $I(2)$). The level of integration is traditionally determined using unit-root tests.[2] Cointegration methods as applied in Beenstock et al. (2012) require that series first exhibit the same degree of integration before a meaningful relationship between them can be established. Therefore, unit-root tests to determine the level of integration often play a major part in single-equation cointegration analyses.

In simple zero-dimensional energy balance models, the time-series properties of radiative forcing should be transferred onto temperature. This may not hold for a limited number of observations, sufficiently noisy time series, or if the model is inappropriate. However, Kaufmann et al. (2013) provide evidence that this result generalizes to more complex climate models: the stochastic trend in model temperature is driven by the stochastic trends in the anthropogenic forcing series. Thus, in a zero-dimension energy balance model, if radiative forcing of greenhouse gases were $I(2)$, and there is a long-run relationship (cointegration) between such forcing and temperature, then temperature should be $I(2)$ as well. There could be a long-run relation (cointegration) between greenhouse gases that are hypothesized to be $I(2)$ and temperature (hypothesized to be $I(1)$) if the anthropogenic forcing series themselves cointegrate to an $I(1)$ process and this process then cointegrates with temperature. Beenstock et al. (2012) show in their work that anthropogenic greenhouse gas forcings are all $I(2)$, which cointegrate to an $I(1)$ variable, an "anthropogenic anomaly". This anthropogenic variable, according to Beenstock et al. (2012), does not cointegrate with the global temperature anomaly.

The physics of greenhouse gases are reasonably well understood, and date from insights in the late 19th century by Arrhenius (1869), who showed that atmospheric temperature change was proportional to the logarithmic change in CO_2. Myhre et al. (2001) provide an extensive overview of the different radiative forcing of various greenhouse gases. In highly simplified terms (i.e. using a model with an atmosphere made up of a single isothermal slab), heat enters the Earth's atmosphere as short-wave radiation from the Sun, and is radiated in the form of long-wave radiation from the warmed surface to the atmosphere, where greenhouse gases absorb some of that heat. This heat is re-radiated, so some radiation is directed back towards the Earth's surface. Thus, greater concentrations of greenhouse gases increase the amount of absorption and hence re-radiation.

There are various reasons why the observed record of temperature and other climate variables may not exhibit the warming patterns suggested by theory or large-scale coupled models. Empirical modelling can play a role in investigating these underlying issues and test whether observations exhibit the relationships implied by theory. However, it is dangerous to draw hasty conclusions, especially given the large number of problems that can distort conclusions from all forms of empirical statistical analyses. We illustrate these with an uncontroversial example, then show errors in Beenstock et al.'s (2012) approach, such that the paper's starting point is incorrect, and the analysis provides little evidential basis for the strong conclusions. Our paper is not an attempt to provide a complete climate model, but merely show that the statistical modelling approach of Beenstock et al. (2012) does not stand up to scrutiny.

Section 2 uses an uncontroversial example to highlight the dangers of approaches that fail to address all the complications inherent in statistical analyses of observational data, listing six important difficulties facing empirical analyses that lead to fallacious inferences if not handled correctly. Section 3 applies that reasoning to the apparently more controversial case of the relationship between greenhouse gases and temperature, and highlights where the six problems can be encountered in Beenstock et al. (2012).

2 A case study

To highlight hazards that can be encountered in statistical analyses, and to illustrate the points we make in Sect. 3 with reference to Beenstock et al. (2012), we first use an example where the analysis is completely uncontroversial: road fatalities are due to people killed by or in moving vehicles.[3]

Consider Fig. 1 that records total vehicle distances driven in billions of km p.a. (denoted X_t) and road fatalities (Y_t), both for the UK.[4]

The four panels labeled a, b, c, d respectively show X_t, Y_t, $X_t - X_{t-1} = \Delta X_t$ and ΔY_t. It is manifest from the graphs that X_t and Y_t are highly non-stationary (do not have constant means and variances), and have strong opposite trends. Thus, it might seem that the further vehicles drive, the fewer the number of deaths. We can establish that finding "rigorously" by a statistical analysis as follows, but however sophisticated such an analysis may appear to be, the implications that road

[1] In general a series that is $I(k)$ needs to be differenced k times to be stationary.

[2] The term unit root stems from the case when unity is a solution to the characteristic polynomial of a process. As an example, let y_t be the first-order autoregressive process: $y_t = \alpha\, y_{t-1} + \epsilon_t$ where ϵ_t is a white-noise process. Note that y_t is a random walk, which is an $I(1)$ process, and has a unit root if $\alpha = 1$. To see this, using the lag-operator L (see Hendry 1995, Ch. 4), the process can be expressed as $y_t = \alpha\, L\, y_t + \epsilon_t$, so re-arranged to $(1 - \alpha\, L)\, y_t = \epsilon_t$ where $(1 - \alpha\, L)$ is a first order polynomial in L. This polynomial has a root of $1/\alpha$, and thus a unit-root if $\alpha = 1$.

[3] All results are obtained using *Autometrics* (see Doornik, 2009).

[4] Fatalities are only available continuously from 1979 onwards and interpolated from intermittent data for 1930–1979, an issue of some importance below.

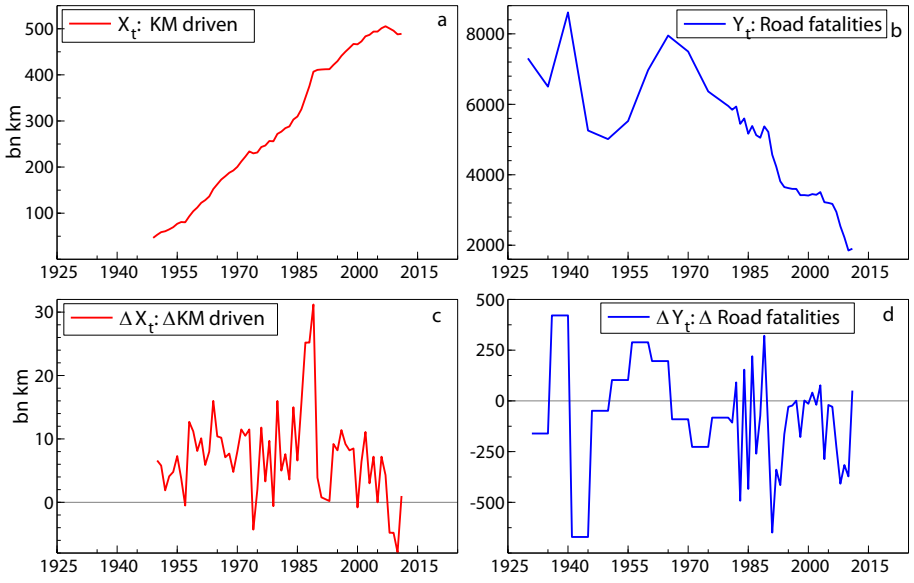

Fig. 1. Vehicle kilometers driven and road fatalities p.a. in the UK: in levels **(a, b)** and first differences **(c, d)**

fatalities are not due to moving vehicles, or can even be reduced by more driving, both remain absurd.

A simple first-order autoregressive-distributed lag model, commonly used to capture autoregressive dynamics while allowing for contemporaneous and lagged covariates, (ADL: see Hendry, 1995, Ch. 7) estimated by regressing Y_t, on a constant, X_t, X_{t-1} and Y_{t-1}, delivers the long-run solution of predicted road fatalities (standard errors in parentheses):

$$\widehat{Y} = \underset{(636)}{8257} - \underset{(2.4)}{15.7}\, X, \tag{1}$$

where the test for a unit root in the model rejects at 1 %, $t_{\rm ur} = -4.08^{**}$ (see Ericsson and MacKinnon, 2002), apparently confirming cointegration – with a negative sign. This suggests that there is long-run stationary relation between the non-stationary series of road fatalities and vehicle kilometers, such that road fatalities decrease with vehicle kilometers driven. To assess short-run effects, we then estimate an equilibrium-correction model using the derived long-run (cointegrating) solution in Eq. (1), modelling the changes of road fatalities in terms of changes of vehicle kilometers driven and deviations from the long-run equilibrium:

$$\widehat{\Delta Y_t} = \underset{(2.48)}{12.3}\, \Delta X_t - \underset{(0.012)}{0.088}\, (Y_{t-1} - 8257 + 15.7\, X_{t-1}), \tag{2}$$

where the residual standard deviation is $\widehat{\sigma} = 160$. Equation (2) shows a short-run increase in deaths as vehicle kilometers driven increases, but a long-run decrease.

How can "statistical evidence" fly in the face of the obvious? There are at least six key reasons why such a result occurs: data measurement errors (here inaccurate interpolation); unmodelled shifts (when a change in legislation or technology shifts a relationship); mistaken inference; incorrectly modelled relations (when the residuals from the estimated relationship do not satisfy the statistical properties of the assumed error processes, so claimed inferences are invalid); omitted variables' bias (omitting relevant explanatory variables); and aggregation bias (mixing data from very different populations). All six can powerfully distort any empirical statistical study, leading to fallacious conclusions as we now discuss.

2.1 Data measurement errors

Data measurement errors can mislead any form of inference. Empirical relationships then represent correlations between what was incorrectly measured, not what actually happened. Figure 1d, showing the annual changes in road fatalities, illustrates the interpolation over the early sample, with constant periods followed by large jumps, quite unlike any real data. Interpolation, unless perfect, creates measurement errors and measurement errors in explanatory variables induce downwards biases in parameter estimation. Further, interpolation leads to negative error autocorrelation in dynamic relations which invalidates standard statistical inference unless this auto-correlation is handled appropriately. To see this, let $\{Y_t\}$ be a stationary autoregressive process:

$$Y_t = \gamma Y_{t-1} + e_t,$$

where $e_t \sim \mathrm{ID}\,[0, \sigma_e^2]$ (denoting independent sampling from a constant distribution with mean 0 and variance σ_e^2). Suppose $\{Y_t\}$ is only observed with error as $\widetilde{Y}_t = Y_t + v_t$ where $v_t \sim \mathrm{ID}\,[0, \sigma_v^2]$ is a random error of measurement or interpolation, then the observed variable is

$$\widetilde{Y}_t = \gamma \widetilde{Y}_{t-1} + e_t + v_t - \gamma v_{t-1} = \gamma \widetilde{Y}_{t-1} + u_t - \rho u_{t-1},$$

where ρ depends on σ_e^2 and σ_v^2. The presence of u_t and u_{t-1} in the above equation implies that the model will automatically exhibit negative error autocorrelation. In turn, this negative error autocorrelation strongly affects the outcomes of integration and cointegration tests, usually suggesting a lower order of integrability than actually applies (see Schwert, 1987 and Hendry, 1995, Ch. 12).

An additional major concern in the application of modelling road deaths above is that periods of interpolation and regular measurement are combined and treated as if they stemmed from the same measurement process despite changes in error autocorrelation and variance.

2.2 Unmodelled shifts

Unmodelled shifts are unaccounted changes in the distributions of the variables in the model over time. These unmodelled shifts (due to many potential causes, including technological innovations, changes in legislation, wars, major geophysical disturbances, etc.) can play havoc with statistical inference: they add an additional non-stationarity to that induced by integrating forces (such as unit roots). For example, standard inference made under the assumption of stationarity will be invalid if there is a shift in the mean – the underlying distribution of that variable is then not invariant over time. Further, unmodelled shifts distort the relationships between the variables that have been included; lead to residuals with properties that differ from the assumed error processes and thus invalidate inference; and can induce forecast failure out of sample. The impact from the greatly reduced private motoring from petrol rationing during the Second World War is likely the cause of the visible shift in the fatalities graph during the early 1940s (Fig. 1b).

2.3 Mistaken inferences

Even when estimated standard errors used in statistical hypothesis tests correctly reflect the actual sampling standard deviations, there are two well-known mistaken inferences arising from: (a) failing to reject a false null hypothesis; and (b) rejecting a correct null hypothesis using a test with power against more than one alternative hypothesis. We take these in turn.

a. Consider a sample of 100 observations on an accurately measured variable Z_t. The sample mean, $\widehat{\mu}$ is 0.005 and the estimated standard error $\widehat{\sigma}$ is 0.05. Then, under the hypothesis that $Z_t \sim \mathrm{ID}\,[\mu,\, \sigma^2]$, a Student's t test of the null hypothesis that $\mu = 0$ has the value of approximately 1.0. The null is not rejected at any reasonable significance level. But neither is the null that $\mu = 0.0025$ or even $\mu = -0.0025$. When these are quarterly growth rates of real income per capita, there is a dramatic difference between the substantive outcomes not reflected in the statistics, namely no growth ($\mu = 0$) growth of approximately 1 % p.a. ($\mu = 0.0025$); and real incomes falling at 1 % p.a. ($\mu = -0.0025$). Not rejecting the null does not entail it is true, merely that evidence is inconclusive.

b. In the example just discussed, an investigator decides to test the assumption that the $\{Z\}$ are independent draws against the alternative that the series is a first-order autoregression, and strongly rejects the null hypothesis of independence. While that discovery vitiates the analysis used in (a), it does not imply that $\{Z_t\}$ is a first-order autoregression. Indeed, rejection does not even imply that the elements of $\{Z_t\}$ are not drawn independently: the cause of residual autocorrelation could be due to inappropriately using a linear approximation to a non-linear relation (see Hendry, 1995, Ch. 6); or be induced by an unmodelled location shift (see Castle and Hendry, 2013a), defined by μ taking different values at different times (as has happened historically for real income growth: see Castle and Hendry, 2013b).

2.4 Incorrectly modelled relations

Incorrectly modelled relations arise in addition to all the problems just noted, when the wrong functional form is imposed, say linear rather than non-linear; inadequate dynamics are allowed for in the included variables, inducing residual autocorrelation; or heteroskedastic errors are not handled, all of which entail that estimated standard errors (on which tests are based) can be far from the correct sampling uncertainty standard deviations.

2.5 Omitting relevant explanatory variables

There are many potentially relevant explanatory variables omitted: a partial list would include improved driving standards from more stringent driving tests; better road safety training; safer cars with improved impact designs and better brakes (abs); seat belts and air bags (see the analysis of the impact of the former in Harvey and Durbin, 1986); separation of opposite direction traffic flows on motorways; reductions in drunk driving; and so on. Converse effects come from faster driving, driver overconfidence, driving after taking drugs, or while using mobile phones, etc.

2.6 Aggregation bias

Aggregation bias is due to the total data comprising distinct sub-populations with different characteristics, here, across both age and sex; geographical location; urban and rural; vehicle types (motorcycles, cars, trucks etc.) and road structures. For example, replacing a "one lane each way" road with a motorway (still desperately needed in the northeast of both England and Scotland) would increase kilometers driven yet probably reduce deaths.

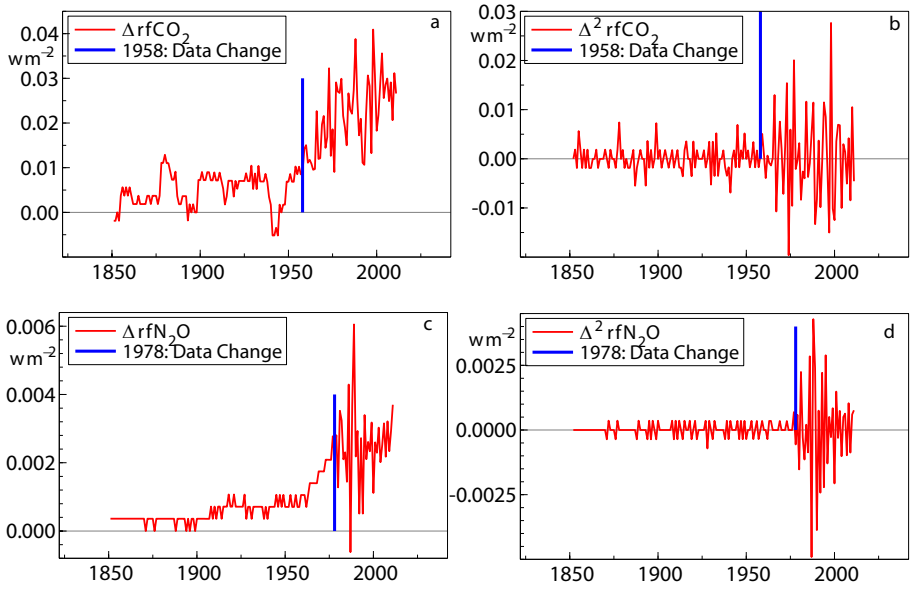

Fig. 2. Time series of the first and second differences of $rfCO_2$ and rfN_2O.

2.7 Implications

The substantive occurrence of any of these problems precludes establishing the genuine presence or absence of any meaningful causal relationship. Despite fatalities and distance driven having very different statistical properties, and decidedly opposite trends summarized in the "cointegrating relation" (Eq. 2), it is fallacious to conclude that moving vehicles do not cause road deaths.[5]

3 Statistical hazards in "polynomial cointegration tests"

All six problems just discussed are relevant in the analysis in Beenstock et al. (2012), although five of them stand out when just viewing the data series, namely data measurement and unmodelled shifts, mistaken inference, incorrectly modelled relations, and omitted variables. This section follows the structure of hazards listed above in Sect. 2 and roughly the analysis of Sects. 3.1 to 3.3 in Beenstock et al. (2012). The paper by Beenstock et al. (2012) first shows uni-variate unit root tests to determine the time-series properties of the data (see their Sect. 3.1), then a measure of anthropogenic forcing is constructed (see their Sect. 3.2). This is followed by the test for cointegration between the measure of anthropogenic forcing and the observed temperature anomaly (see their Sect. 3.3). Subsequently Beenstock et al. (2012) conduct robustness checks on their results, provide model extensions, and estimate a short-run model (Sects. 3.4–3.9).

We first investigate the time-series properties of the data (Sect. 3.1 in Beenstock et al., 2012) in the following sections on data measurement (Sect. 3.1), shifts (Sect. 3.2), and mistaken inferences (Sect. 3.3). These are relevant to all sections of Beenstock et al. (2012) which rely on the use of anthropogenic forcing time series (these are Sects. 3.1–3.9). We address the construction of the measure of anthropogenic forcing, "the anthropogenic anomaly", (Sect. 3.2 in Beenstock et al., 2012) in our Sect. 3.4 on incorrectly modelled relations. We then investigate the cointegration test (Sect. 3.3 in Beenstock et al., 2012) in our Sect. 3.5 on omitted variables. These two sections refer to the main argument made in Beenstock et al. (2012) (in Sects. 3.2 and 3.3) which presents evidence against anthropogenic global warming. Aggregation bias is raised as a general point in our Sect. 3.6.

3.1 Data measurement errors

We obtained the data on greenhouse gases used by Beenstock et al. (2012) (see their Sect. 3.1) using the values provided by Myhre et al. (1998) to convert the series into their radiative forcing equivalents. The fact that the measured series of GHGs come from a variety of different sources is omitted from Beenstock et al. (2012). If the measurements were identical in all sources, this would not be an issue: however, our graphs reveal sharp differences in the data properties. Consider the CO_2 and N_2O series. Both are initially based on ice core data (up to the dates indicated by the vertical lines in Fig. 2: 1850 until 1958 for CO_2 and 1850 until 1978 for N_2O) followed by flask and other measurements thereafter. Figure 2b shows that up until approximately the point when the

[5]Indeed, allowing for just one of these six general problems, namely unmodelled shifts using step-indicator saturation (see Doornik et al., 2013, available in *Autometrics*), reveals many location shifts in the ADL relationship leading to Eq. (1) – after which X ceases to be significant.

Table 1. ADF unit-root tests on $\Delta rfCO_2$.

1850–1957 constant			1958–2011 constant and trend		
D-lag	t-adf	Reject H_0	D-lag	t-adf	Reject H_0
5**	−3.737	**	5	−4.089	*
4	−2.910	*	4	−3.807	*
3	−2.948	*	3	−3.383	
2	−3.146	*	2	−4.197	**
1	−2.706		1	−5.365	**
0	−3.544	**	0	−6.563	**

ADF unit-root tests: the null hypothesis H_0 is that the series has a unit root so is non-stationary. Rejecting the null hypothesis suggests no unit-root non-stationarity. D-lag specifies the number of lags included in the ADF unit root test, where * indicates that longest lag is significant at 5 % and ** at 1 %. If no lags are significant, the model with zero lags is appropriate. Unit root test outcome: ** indicates rejection of the null hypothesis at 1 % and * at 5 %.

Table 2. ADF Unit-root Tests on ΔrfN_2O.

1850–1978 const. and trend			1978–2011 constant		
D-lag	t-adf	Reject H_0	D-lag	t-adf	Reject H_0
5	2.098		5	−3.832	**
4	1.864		4	−3.347	*
3*	1.427		3	−3.636	**
2**	0.801		2	−4.048	**
1**	−0.619		1	−4.793	**
0	−3.87		0	−7.845	**

ADF unit-root tests: the null hypothesis H_0 is that the series has a unit root and is non-stationary. Rejecting the null hypothesis suggests no unit-root non-stationarity. D-lag specifies the number of lags included in the ADF unit-root test, where * indicates that longest lag is significant at 5 % and ** at 1 %. If no lags are significant, the model with zero lags is appropriate. Unit root test outcome: ** indicates rejection of the null hypothesis at 1 % and * at 5 %.

switch from ice-core to non-ice core data was made, many of the changes have precisely the same magnitude, revealing an artificial pattern different to the latter half of the sample. Moreover, there are large changes in the variances of the second differences of both series at the measurement system switch. Despite the well-established problems for unit-root tests described in Sect. 2.1, the data are analyzed in Beenstock et al. (2012) as if they come from the same populations.

We worry that Beenstock et al.'s (2012) Fig. 2 – reproduced here as Fig. 3 – camouflages the serious problem of measurement-regime shifts.

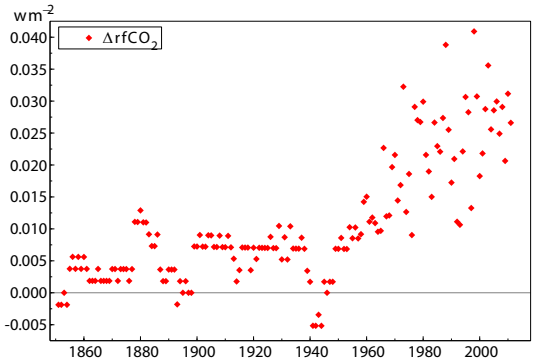

Fig. 3. Time series of the first differences of $rfCO_2$ from Beenstock et al. (2012).

3.2 Unmodelled shifts

Interacting with unmodelled shifts, measurement errors can can lead to false interpretations of the stationarity properties of data. In the presence of these different measurements exhibiting structural changes, a unit-root test on the entire sample could easily not reject the null hypothesis of $I(2)$ even when the data are in fact $I(1)$. Indeed, once we control for these changes, our results contradict the findings in Beenstock et al. (2012) (see their Sect. 3.1 and Table 1). Our results are presented here in Tables 1 and Table 2 below.

Sub-sample unit-root tests

Unit-root tests are used to determine the level of integration of time series. Rejection of the null hypothesis provides evidence against the presence of a unit-root and suggests that the series is $I(0)$ (stationary) rather than $I(1)$ (integrated). As is to be expected from the data in Fig. 2, but based on augmented Dickey–Fuller (ADF) tests (see Dickey and Fuller, 1981), the first difference of annual radiative forcing of CO_2 is stationary initially around a constant (over 1850–1957), then around a linear trend (over 1958–2011). Although these tests are based on sub-samples corresponding to the shift in the measurement system, there is sufficient power to reject the null hypothesis of a unit root. In a similar manner, unit-root tests reject non-stationarity of the first difference of N_2O for the second set of observations (1978–2011). Unit-root non-stationarity cannot be rejected for 1850–1978 for N_2O: however, given the manifestly artificial appearance of the data such a result should be interpreted with extreme caution. Particularly, the ΔN_2O series appears to exhibit a step shift in the early 1900s, which also leads to spurious results in unit-root tests. The split points for our sub-sample unit-root tests are given by the extraneous dates of change in the measurement system, so the tests do not need to allow for that choice.

Given these time-series properties including apparent shifts and changes in variance, assuming that all annual anthropogenic GHGs are $I(2)$ is an incorrect starting point, and the findings in Beenstock et al. (2012) appear to be an artefact of pooling data with very different measurement systems and behaviour in the two sub-samples.

3.3 Mistaken inferences

In line with our analysis suggesting that the presence of measurement changes affects the unit root test outcomes, Stern and Kaufmann (2000) show that when using univariate unit

root tests, the results can vary considerably, depending on which type of test is used and also vary across different anthropogenic gases. For example, CO_2 appears to be $I(1)$ in two out of their four tests, and $I(2)$ in the others (Table 1 in their paper). Similarly, N_2O appears $I(1)$ in three out of the four test types. Due to these conflicting results they therefore then employ a structural time-series approach. Second, Kaufmann and Stern (2002) test the time-series properties of the aggregate of the radiative forcing of all the major greenhouse gases (CO_2, CH_4, N_2O, CFC-11, CFC-12) and find them to be $I(1)$.

Crucially, the level of integration, and thus stationarity, of data is not intrinsic to its process and can change over time. There is nothing inherent in the physical data generating process that makes anthropogenic forcings, or other variables, $I(1)$ or $I(2)$. The observational data may, over some period, be consistent with a process that is $I(1)$ or $I(2)$, but this is not an intrinsic property that cannot change. There are many examples of changes, two being that the level of CO_2 emissions is related to economic activity, which has varied over time, and emissions of CFCs which only arose in the latter part of the 20th century. Both of these may be stationary in second differences from the 1950s onwards, but because of the underlying processes, they may well have been stationary in first differences or levels before then, or in the case of CFCs non-existent before their discovery and declining after the Montreal Protocol of 1989 (Myhre et al., 2001, provide some examples concerning CFCs). The claim that all greenhouse gases are always $I(2)$ is incorrect. Such a result is also inconsistent with the tests conducted in the previous literature and with our analysis.

3.4 Incorrectly modelled relations

The main message of Hendry (1995) is that before any statistical inferences can be conducted, a model must be congruent, or well-specified in that it satisfies the assumptions on which the statistical analysis relies. The unit root tests in Table 1 in Beenstock et al. (2012) make many untested assumptions, including accurate data, that there is a single measurement regime, and that no location shifts occurred.

Beenstock et al.'s (2012) conclusion that anthropogenic forcings do not cointegrate with the observed temperature anomaly relies on their constructed measures of anthropogenic forcing, the "anthropogenic anomaly" (see their Sect. 3.2). In light of model specification, we assess this method of constructing the measure of anthropogenic forcing. The measures of anthropogenic forcing, the "anthropogenic anomalies" in in Beenstock et al. (2012) (given by Eqs. 9 and 10 in Beenstock et al., 2012, reproduced here as Eqs. 3 and 4) are the residuals g_1 and g_2 of a single regression of radiative forcing of CO_2 on the forcing of other greenhouse gases:

$$\mathrm{rfCO_2} = 10.972 + 0.046\,\mathrm{rfCH_4} + 10.134\,\mathrm{rfN_2O} + g_1 \quad (3)$$

$$\mathrm{rfCO_2} = 12.554 + 0.345\,\mathrm{rfCH_4} + 9.137\,\mathrm{rfN_2O} + 1.029\,\mathrm{BC} + 0.441\,\mathrm{ReflAer} + g_2, \quad (4)$$

where BC is their radiative forcing of black-carbon concentration, and ReflAer is their radiative forcing of all reflective aerosols. Such regressions are a variant for possibly $I(2)$ variables of the approach in Engle and Granger (1987). Banerjee et al. (1986) demonstrated that this type of test imposed "common-factor" restrictions of the form criticized by Hendry and Mizon (1978) and Mizon (1995), as a consequence of which the test often lacks power and is substantively inferior to the systems approach in Johansen (1988). We now consider their "anthropogenic anomaly" in two cases.

First, despite the above ADF unit root test outcomes, suppose one accepted the starting point of Beenstock et al. (2012) that all anthropogenic variables are $I(2)$. They state that Eqs. (3) and (4) are to test for cointegration between the anthropogenic series. However, cointegration is a system property, so the variables need to be treated as such. To establish cointegration between the variables in Eq. (3) ($\mathrm{rfCO_2}$ regressed on $\mathrm{rfCH_4}$ and $\mathrm{rfN_2O}$), the full system of three variables needs to be considered (see Hendry and Juselius, 2001). This is further complicated here as the variables are assumed to be $I(2)$, so an $I(2)$ cointegration analysis is required (see Juselius, 2006): the system has at most full rank (= 3), or if there is (polynomial) cointegration the system may exhibit reduced rank of one or two, and if no (polynomial) cointegration, rank zero. If there is reduced rank (which has to be tested in an $I(2)$ procedure), the system needs to be decomposed into the cointegrating relations (which are $I(0)$ and therefore stationary) and the common underlying stochastic trends (of which some may be $I(1)$ and some $I(2)$ trends). Given the assumed $I(2)$ property, if the three variables cointegrate, there may be up to two $I(1)$ cointegrating relations between the three series, and thus two potential anthropogenic anomaly measures. The single anthropogenic anomaly given in their Eq. (3) is then a linear combination of these measures of anthropogenic forcing. The same problem generalizes to their Eq. (4), with there being five variables in the system and a much larger set of potential cointegrating relations. The system of five variables may have full rank (= 5), rank zero, or reduced rank between one and four if there is cointegration, implying up to four $I(1)$ cointegrating relations and up to four measures of the anthropogenic anomaly. Thus, even if their starting point that all anthropogenic are $I(2)$ is accepted, then their measure of the anthropogenic anomaly is likely only one of many, given the large number of potential cointegrating relations. There could well be a residual (anthropogenic anomaly) that does cointegrate with temperature and solar irradiance.

Second, given that the starting point of assuming all anthropogenic variables are $I(2)$ is incorrect, the measures of the anthropogenic anomaly (g_1 and g_2) are inappropriate. The measures of anthropogenic forcing are the residuals of

regressions of $rfCO_2$ on radiative forcing of the other greenhouse gases. This means that the measure of anthropogenic forcing used in Beenstock et al. (2012) is really the variation in radiative forcing of CO_2 that is unexplained by the variation in other greenhouse gases. In a basic energy balance model, radiative forcings are mostly considered additively. The total effect of all forcings together (while taking feedbacks into account) is what is important. Taking the unexplained variation in radiative forcing of CO_2 as a measure of anthropogenic forcing (i.e. the "anomaly") is then incorrect and does not measure what Beenstock et al. (2012) state it does. Their main test of anthropogenic global warming (the regression of temperature on solar irradiance and the anthropogenic anomaly in their Table 3) is then a regression of temperature on solar irradiance and a residual, which need not capture any anthropogenic component at all, and does not capture the main anthropogenic forcing component.

3.5 Omitted variables

We briefly list a few of these variables that have been omitted and may play an important role, though they should be considered with caution as some of these gases are not as well mixed (but spatially varied) and thus may not be appropriate in a zero-dimensional model. Myhre et al. (2001) provide a good overview of available time series for the historical period used; these include CFCs (Chlorofluorocarbons), as used by Stern and Kaufmann (2000), which together with tropospheric ozone likely exhibit a positive forcing, as well as stratospheric ozone (see Myhre et al., 2001) which likely acts as negative forcing.

Omitted variables induce biases in general (unless orthogonal to all included variables), and in Beenstock et al.'s (2012) analysis based on cointegration, the negative effects of having omitted important factors may be even more pronounced. Suppose one accepts their analysis of integration properties (see above in Sect. 3.1), and the anthropogenic anomaly, (see above in Sect. 3.4), the main equation to test for cointegration between the anthropogenic component and temperature in Beenstock et al. (2012) is a regression of temperature on the anthropogenic anomaly and solar irradiance (reproduced here as Eq. 5, see Sect. 3.3 and Table 3 of cointegration tests in Beenstock et al., 2012):

$$\begin{aligned}\text{Temperature}_t &= \beta_0 + \beta_1 \,\text{Solar Irradiance}_t \\ &\quad + \beta_2 \,\text{Anthropogenic Anomaly}_t + \epsilon_t. \end{aligned} \tag{5}$$

Their main conclusion stems from the fact that, using unit-root tests, the error term ϵ_t of this regression is non-stationary, $I(1)$, suggesting that there is no cointegration between the three series. Any omitted $I(1)$ variable in this equation will induce the error term to appear $I(1)$, and lead to spurious rejection of cointegration. Ocean heat uptake is one of many factors missing in this equation, and, according to Beenstock et al.'s (2012) analysis is $I(1)$. Beenstock et al. (2012) state that the omission of ocean heat is not a concern as ocean heat content and temperature do not cointegrate (Table 4 in Beenstock et al., 2012), however, this cointegration test is undertaken using temperature, ocean heat and water vapour alone, rather than considering the full system where the anthropogenic measure and solar radiation are also included. These individual regressions do not capture the important system property of cointegration (emphasized in Sect. 3.4). Further, any of the above mentioned forcing series (CFCs, ozone) may also be $I(1)$ and were omitted. Any omitted $I(1)$ variable in their main test of cointegration between temperature, the anthropogenic anomaly and solar radiation will induce an $I(1)$ stochastic trend in the residual. Given that the regression only consists of a constructed measure of anthropogenic forcing and solar irradiance, there are many factors that will lead to the error term appearing $I(1)$ and thus, to spurious rejection of cointegration.

3.6 Aggregation bias

The time-series literature studying radiative forcing and its effect on temperature primarily relies on the global temperature anomaly as a single temperature series based on zero-dimensional energy balance models. In practice this is a common and often useful simplification, but temperature trends vary spatially, suggesting there may be unmodelled heterogeneity. Finding no cointegration between a global aggregate and global anthropogenic forcing then does not imply there does not exist a relationship overall. To illustrate some of this spatial variation, Fig. 4 shows the global anomaly together with approximate Arctic (averaged over 64–90° N latitude) and close to Antarctic (averaged over 90–64° S latitude) anomalies (data from NASA Goddard Institute for Space Studies , 2011). As can be seen, temperature has risen far faster in the Arctic region than globally.

4 Conclusions

A complete analysis of this data would require separate models of, or controlling for, the pre and post ice-core measurements, taking account of the myriad influences impinging on the climate, temperature, and different greenhouse gases. The system nature of cointegration needs to be taken into account when the analysis (such as Beenstock et al.'s, 2012) relies solely on the time-series properties of different series.

The aim of this paper is merely to demonstrate that the conclusions claimed by Beenstock et al. (2012) about the different degrees of integrability of temperature and CO_2 are rejected once the regime-shift nature of the measurement system is taken into account. We emphasize that the time-series properties and degrees of integrability of data can change over time. Further, we note that the measure of anthropogenic forcing in Beenstock et al. (2012), a constructed "anthropogenic anomaly", is inappropriate regardless of the time-series properties of the data. Any one of those missteps by

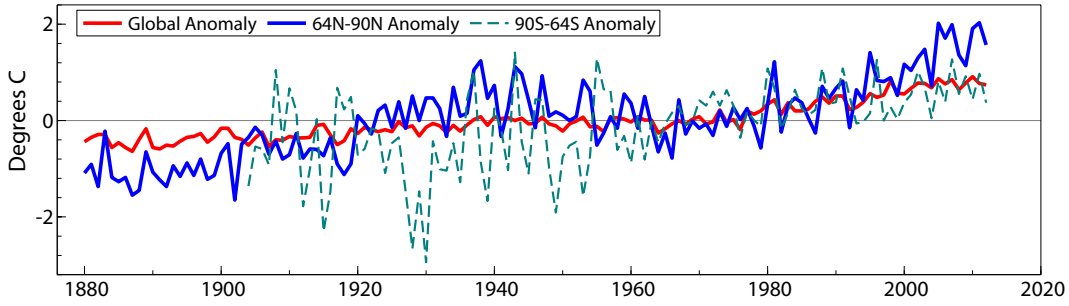

Fig. 4. Temperature anomalies relative to 1951–1980 average.

itself is sufficient to cast serious doubt on the conclusions of Beenstock et al. (2012).

To handle the hazards associated with statistical models of temperature and greenhouse gases one could consider the following. First, one could model the pre-break and post-break periods separately. Second, given the uncertainties about univariate tests of time-series properties and degrees of integration, one could follow Stern and Kaufmann (2000) who focus on a structural time-series approach, or Kaufmann and Stern (2002) who work with aggregates of radiative forcing. Third, if one accepts that GHG forcings are $I(2)$, a full $I(2)$ system approach is required as outlined in Juselius (2006) testing that the degrees of integration (and relationships between series) have not changed over time. Fourth, one can account for the changes in measurement (the most basic approach being indicator variables for the time periods before or after the break) while potentially also accounting for other unknown unmodelled breaks (see e.g. Hendry and Pretis, 2013). As emphasized in Hendry (2009), to draw substantive conclusions from a statistical or econometric analysis requires a complete, comprehensive and constant model; and to draw causal conclusions further requires that such a model is invariant to changes in all other variables.

Acknowledgements. This research was supported in part by grants from the Open Society Foundations and the Oxford Martin School. We are grateful to Myles Allen, Vanessa Berenguer-Rico, Margaret Ziriax, an editor and four anonymous referees for comments on a previous version.

Edited by: S. Smith

References

Arrhenius, S. A.: On the influence of carbonic acid in the air upon the temperature of the ground, London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science (fifth series), 41, http://www.globalwarmingart.com/images/1/18/Arrhenius.pdf, last access: 29 May 2013, 237–275, 1896.

Banerjee, A., Dolado, J., Hendry, D. F. and Smith, G.: Exploring equilibrium relationships in econometrics through statistical models: some monte carlo evidence, Oxford Bull. Econom. Stat., 48, 253–277, 1986.

Beenstock, M., Reingewertz, Y., and Paldor, N.: Polynomial cointegration tests of anthropogenic impact on global warming, Earth Syst. Dynam., 3, 173–188, doi:10.5194/esd-3-173-2012, 2012.

Castle, J. L. and Hendry, D. F.: Model selection in under-specified equations with breaks, J. Econometr., doi:10.1016/j.jeconom.2013.08.028, in press, 2013a.

Castle, J. L. and Hendry, D. F.: Semi-automatic non-linear model selection, in: Essays in Nonlinear Time Series Econometrics, edited by: Haldrup, N., Meitz, M., and Saikkonen, P., Oxford University Press, Oxford, 2013b.

Dickey, D. A. and Fuller, W. A.: Likelihood ratio statistics for autoregressive time series with a unit root, Econometrica, 49, 1057–1072, 1981.

Doornik, J. A.: Autometrics, in: The Methodology and Practice of Econometrics, edited by: Castle, J. L. and Shephard, N., Oxford University Press, Oxford, 2009.

Doornik, J. A., Hendry, D. F., and Pretis, F.: Step-indicator saturation, Discussion paper 658, Economics Department, Oxford University, Oxford, 2013.

Engle, R. F. and Granger, C. W. J.: Co-integration and error correction: Representation, estimation, and testing, Econometrica, 55, 251–276, 1987.

Ericsson, N. R. and MacKinnon, J. G.: Distributions of error correction tests for cointegration, Econometr. J., 5, 285–318, 2002.

Harvey, A. C. and Durbin, J.: The effects of seat belt legislation on British road casualties: A case study in structural time series modelling, J. Roy. Stat. Soc. B, 149, 187–227, 1986.

Hendry, D. F.: Dynamic Econometrics, Oxford University Press, Oxford, 1995.

Hendry, D. F.: The methodology of empirical econometric modeling: Applied econometrics through the looking-glass, in: Palgrave Handbook of Econometrics, edited by: Mills, T. C. and Patterson, K. D., Palgrave MacMillan, Basingstoke, 3–67, 2009.

Hendry, D. F. and Juselius, K.: Explaining cointegration analysis: Part II, Energy J., 22, 75–120, 2001.

Hendry, D. F. and Mizon, G. E.: Serial correlation as a convenient simplification, not a nuisance: A comment on a study of the demand for money by the bank of England, Economic J., 88, 549–563, 1978.

Hendry, D. F. and Pretis, F.: Anthropogenic Influences on Atmospheric CO_2, in: Handbook on Energy and Climate Change, edited by: Fouquet, R., Edward Elgar, Cheltenham, 287–323, 2013.

Johansen, S.: Statistical analysis of cointegration vectors, J. Econom. Dynam. Contr., 12, 231–254, 1988.

Juselius, K.: The Cointegrated VAR Model: Methodology and Applications, Oxford University Press, Oxford, 2006.

Kaufmann, R. K. and Stern, D. I.: Cointegration analysis of hemispheric temperature relations, J. Geophys. Res., 107, ACL 8-1–ACL 8-10, doi:10.1029/2000JD000174, 2002.

Kaufmann, R. K., Kauppi, H., Mann, M. L., and Stock, J. H.: Does temperature contain a stochastic trend: linking statistical results to physical mechanisms, Climatic Change, 118, 729–743, 2013.

Mizon, G. E.: A simple message for autocorrelation correctors: Don't, J. Econometr., 69, 267–288, 1995.

Myhre, G., Highwood, E. J., Shine, K., and Stordal, F.: New estimates of radiative forcing due to well mixed greenhouse gases, Geophys. Res. Lett., 25, 2715–2718, 1998.

Myhre, G., Myhre, A., and Stordal, F.: Historical evolution of radiative forcing of climate, Atmos. Environ., 35, 2361–2373, 2001.

NASA Goddard Institute for Space Studies – GISS: GISS – Surface Temperature Analysis, available on-line: http://data.giss.nasa.gov/gistemp/ (last access: 29 May 2013), 2011.

Schwert, G. W.: Effects of model specification on tests for unit roots in macroeconomic data, J. Monet. Econom., 20, 73–103, 1987.

Stern, D. I. and Kaufmann, R. K.: Detecting a global warming signal in hemispheric temperature series: A structural time series analysis, Climatic Change, 47, 411–438, 2000.