Model Selection and Shift Detection: General to Specific Modelling in Climatology

—

– Conference Draft –

David F. Hendry and Felix Pretis*

Institute for New Economic Thinking at the Oxford Martin School, University of Oxford

Abstract

Empirical research in climatology regularly deals with complex systems and non-stationary data, making it near-impossible to correctly specify an appropriate model *a-priori*. We introduce a new empirical approach to modelling in climatology using automatic model selection. The methodology is based on an extended general to specific approach which allows for more variables than observations. This enables non-stationarity to be tackled both via any unit roots and through the simultaneous detection of outliers and structural breaks in the form of impulses and step-shifts without forcing any to be significant or to be excluded. The methodology is applied to empirically estimate anthropogenic contributions to atmospheric CO_2 and to determine long run relations based on ice core data while controlling for un-modelled structural breaks.

1 Introduction

Empirical research in climatology regularly deals with complex systems and non-stationary data, making it near-impossible to correctly specify an appropriate model *a-priori*. As a result models often use statistical methods that are ill–fitted and inappropriate given the time-series characteristics of the data. Further, the impact of un-modelled structural

*This research was supported in part by grants from the Open Society Foundations and the the Oxford Martin School. Contact details: david.hendry@nuffield.ox.ac.uk and felix.pretis@nuffield.ox.ac.uk.

changes in time series on model estimates is rarely considered. Modern econometric methods based on model selection can provide a comprehensive solution to these challenges, and provide an agnostic data-driven methodology.

These methods can be seen as a complement to the large scale coupled models and can contribute by providing methods to enable learning from data and to test whether hypothesised relationships can be identified in the observational record. In terms of attribution of variation or forecasting, often simple empirical models outperform those that are simulation based, large scale climate models regularly struggle to match current observations. Naturally there are many inherent simplifications within these empirical models that can lead to pitfalls (e.g. discussed in Hendry & Pretis 2013b). Models thus need to be carefully specified taking into account the complex dynamics and interrelations.

We present a new empirical approach to modelling in climatology using automatic model selection. The methodology is based on an extended general to specific approach which allows for more variables than observations. This enables non-stationarity to be tackled both via any unit roots and through the simultaneous detection of outliers and structural breaks in the form of impulses and step-shifts without forcing any to be significant or to be excluded. As a result, the main relevant explanatory variables are determined and their magnitudes estimated, while irrelevant factors are dropped from the model.

Section 2 introduces model selection with more variables than observations. These methods are then applied in section 3 to estimate the human contribution to atmospheric carbon dioxide measured at Mauna Loa (section 3.1), and to model long term interactions between temperature and other factors using ice core records over approximately the past four hundred thousand years (section 3.2). Section 4 concludes.

2 Methodology

We introduce a modelling methodology that can handle more variables (N) than observations (T). General to specific modelling relies on the theory of reduction in which the basic principle is to reduce a very general model to a specific one (see Hendry, 1995). First, we define a set of N variables that ideally nest the underlying (local) data generating process. Second, starting with that general model as a good approximation of the overall properties of the data, reduce its complexity by removing insignificant variables through an automatic tree search, while checking that at each reduction the validity of the model is preserved.

Empirical models often face a large number of potential unknown unknowns. For a given series of (possibly) non-stationary data, there might be an unknown number of location shifts for unknown durations. Therefore, we employ two new methods of

detecting and modelling previously unknown shifts as a direct result of being able to handle more variables than observations: Impulse Indicator Saturation (IIS) and Step Indicator Saturation (SIS).

The literature for selecting models using more variables than observations is relatively new and can be split into two categories:[1] general to specific (GETS) modelling, and models based on penalised shrinkage estimators. GETS (which is outlined in more detail below) relies on the theory of reduction in which the basic principle is to reduce a very general model to a specific one (see Campos et al. 2005).

The alternative are models based on penalised shrinkage estimators in the sparse modelling literature (see Tibshirani 1996, and Stodden 2006). A model is said to be sparse, if for a large number of variables, the associated coefficient vector only has few non-zero rows. Penalised shrinkage estimators such as the LASSO (Least absolute shrinkage and selection operator, Tibshirani 1996), least-angle-regression (LARS, Stodden 2006) or Stagewise Orthogonal Matching Pursuit (StOMP, Donoho et al. 2006) therefore rely on a penalisation parameter on coefficients. The models are estimated subject to a constraint that some of the entries in the coefficient vector are smaller than a penalisation parameter. While many theoretical results exist for this literature, in practice LASSO and LARS are closely related to forward stepwise selection (Stodden 2006) in which variables are selected after each other, starting with the highest correlated ones. This approach suffers from well known problems associated with forward-step-wise regression, in particular these methods often fail in selection when negative correlation between regressors is present (see e.g. Castle et al. 2011a).

2.1 Model Selection in General to Specific Modelling

GETS modelling is based on the theory of reduction which attempts to explain the discovery of econometric models based on the unobserved underlying data generating process (DGP) (see Hendry 1995, Ch. 9). The theory describes the process of moving from the unobserved DGP to the local DGP (LDGP), which is a simplified admissible version of the underlying DGP that can be modelled using the observed variables $\mathbf{y}_t, \mathbf{x}_t$. The choice of variables $\mathbf{y}_t, \mathbf{x}_t$ will define the properties of the LDGP to be modelled. The primary aim is to achieve a final model that is congruent – defined as matching the empirical properties of the LDGP. From the theory of reduction (Hendry 1995, p.365), there are five criteria that a final model should fulfil to be congruent: 1) homoskedastic innovation errors, 2) weakly exogenous conditioning variables for the estimated parameters of interest, 3) constant invariant parameters (e.g. tested using the Chow (1960) test. 4) theory consistent estimates 5) data admissible formulations on accurate observations – impulse

[1] In the field of machine learning there also exist other approaches dealing with $N > T$, such as Random Forests (Siroky 2009), though these are not directly applicable to regression analysis which is the focus here.

and step indicator saturation (see section 2.3.1) are used to check for outliers. Additional to the steps mentioned here, congruence is evaluated by a wide range of further diagnostic and misspecification tests and through the interpretation of selected indicators. Further, to be un-dominated, a model should encompass all other valid competing sub-models (see e.g., Mizon & Richard 1986, Hendry & Richard 1989, and Bontemps & Mizon 2008). The final aim for a congruent model is to be encompassing of other models, that is to say able to explain outcomes of other valid models within its own framework.

2.2 Mis-specification Testing

Using more variables than observations, in particular IIS and SIS (further outlined in the next section) allows for a new approach to control for model misspecification. By not being restricted in the number of variables to include initially, the idea is to be agnostic about the first formulation and use model reduction steps to achieve a well specified model. Impulse and step indicators can be an especially useful source of information on misspecification – a large number of outliers (selected through indicators) provide evidence for potential functional form misspecification.

2.3 More Variables than Observations: $N > T$

Within the general to specific framework model selection with more variables than observations was first introduced through impulse indicator saturation, and has recently been extended to step indicators and the general case of any number of explanatory variables.

2.3.1 Impulse indicator saturation

For a given series of data, there might be an unknown number of location shifts for unknown durations for unknowingly omitted variables. Therefore, an agnostic approach to detecting these breaks and outliers should be applied. Impulse indicator saturation (IIS) was developed to solve this problem. IIS adds to the set of candidate variables a zero/one indicator variable for every observation in the sample, such that for T observations there are T variables added that correspond to $1_{\{j=t\}}$ indicators (Hendry et al., 2008). Using model selection, only indicators that deviate significantly from the estimated model will be retained. The variables are equal to one when $j = t$ for each $j = 1...T$ and zero otherwise. For example, this could capture the un-modelled effects of sudden release of stratospheric aerosols from volcanic eruptions. By definition (and since an intercept is always included), this leads to the situation of facing more variables than observations, something that could not be handled in the past. The postulated solution to this problem is to split the sample of indicators into blocks and estimate the model on these partitions of variables. Consider the basic case of estimating the mean of a sample using a split-half

approach proposed in Hendry et al. (2008). In step one, only the first $T/2$ indicators are included, this is equivalent to dummying out the first part of the sample and estimating the mean on half of the sample. Indicators that deviate significantly from this estimated mean will be retained. This step is repeated for the other half of indicators, similarly significant indicators are retained. In the final step all previously retained indicators are included (which will generally lead to $N < T$), and the mean is estimated on the full sample while simultaneously selecting significant indicators. Under the null hypothesis, for a significance level of selection α, the expected rate of retention for irrelevant indicators is αT (Hendry & Mizon 2011).

The resulting estimate of the mean is then based on joint information from two subsamples. Intuitively, this approach is valid as the indicator variables are mutually orthogonal. Distributional properties of IIS are analysed by Hendry et al. (2008) for estimating a sample mean under the null-hypothesis of no breaks or outliers and generalise to more than two blocks. Johansen & Nielsen (2009) extend this analysis for the inclusion of regressors for both stationary and unit-root autoregressions, similarly under the null hypothesis of no breaks in the DGP.

IIS may appear surprising, though some commonly used econometric techniques are variants of the general IIS approach. Many techniques that "dummy out" observations are related. Recursive estimation of regression coefficients is identical to IIS over future samples with the number of indicators reduced in each recursion. Similarly, the Chow (1960) test for parameter constancy can be seen as IIS over sub-samples of the data for $T - k + 1$ to T (without selection) as Salkever (1976) showed.

2.3.2 Step Indicator Saturation

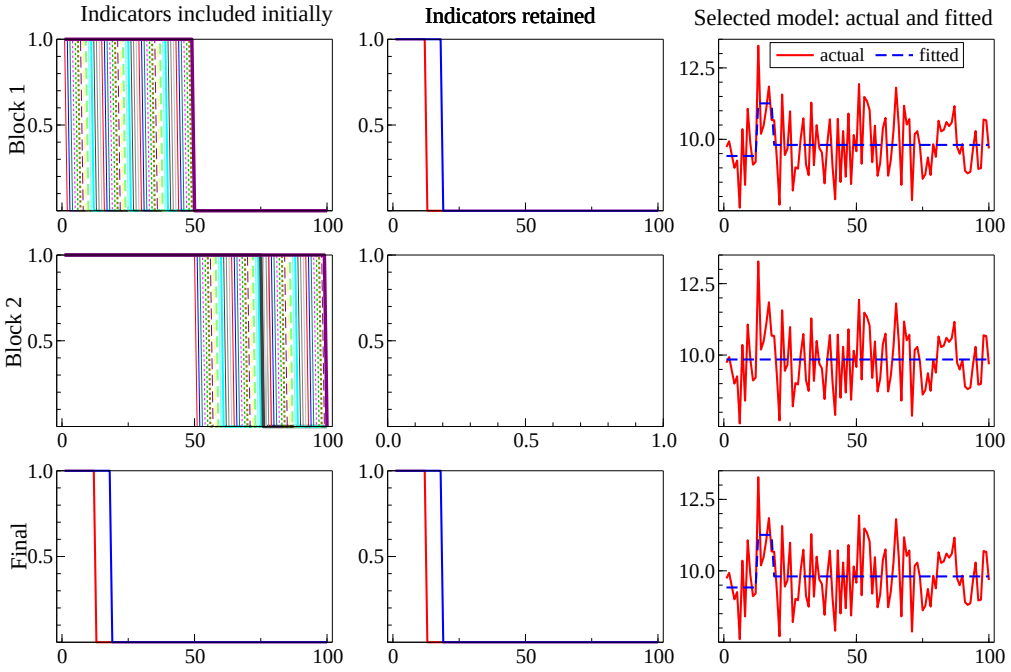
SIS extends the methodology introduced by IIS to cover step-shifts – we consider selecting significant step indicators to capture longer location shifts. Doornik et al. (2013) derive the theory properties of SIS and investigate the performance using Monte Carlo simulations. A short overview is provided here. By including a complete set of step indicators $S = \left\{1_{\{t \leq j\}}, j = 1, \ldots, T\right\}$, a step shift of any magnitude at any point in time can be detected without prior specification. The sudden onset of anthropogenic greenhouse gas emissions could be considered a step-shift increase.

The initial specification of step indicators $\left\{1_{\{t \leq j\}}, j = 1, \ldots, T\right\}$ entering a basic linear model is given by equation (1):

$$y_t = \sum_{j=1}^{T} \delta_j 1_{\{t \leq j\}} + u_t \tag{1}$$

where $u_t \sim \mathsf{IN}\left[0, \sigma_u^2\right]$. It is infeasible to estimate (1), but the split-half approach to understanding IIS applies to SIS. For T indicators, add the first $T/2$ and select at significance

Figure 1: SIS under the null of no break

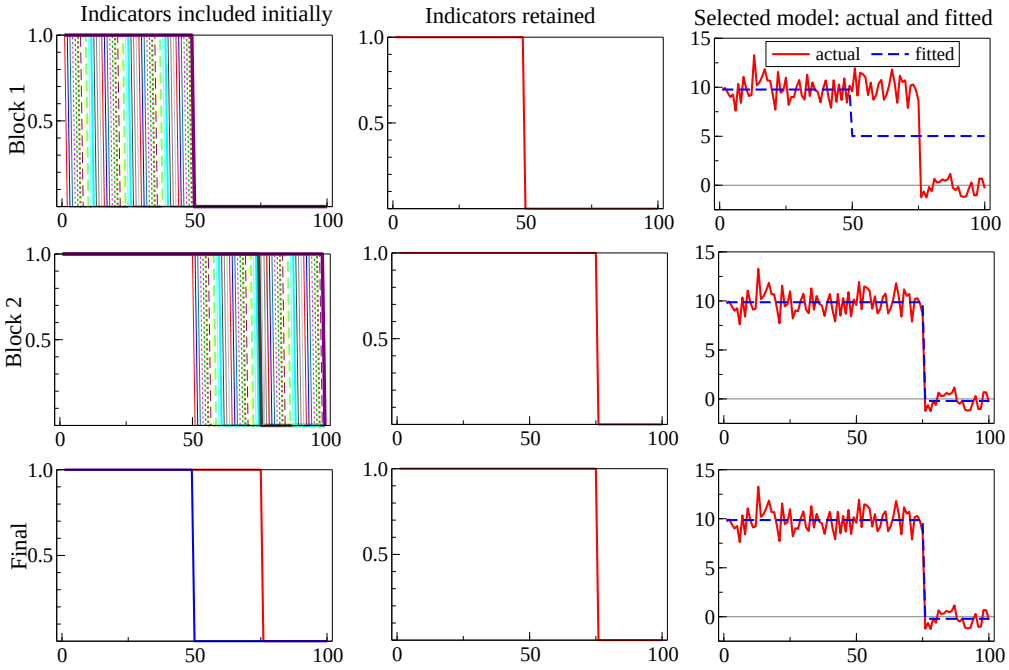


level α, recording which indicators have significant coefficients. Eliminate those, and instead add the second block of $T/2$ to the original model, again selecting at significance level α, and recording which are significant in that subset. Finally, combine the recorded variables (if any) from the two stages, and select again at significance level α. Under the null, setting $\alpha = 1/T$, on average at both sub-steps, $\alpha T/2$ (namely 1/2 an indicator) will be retained by chance, so on average αT indicators will be retained from the combined stage, so one degree of freedom is lost on average. When there are more regressors plus indicators than T, the procedure can be extended by dividing the total set of N candidate variables into smaller sub-blocks setting $\alpha = 1/N$ overall. In the presence of one (or multiple breaks), Doornik et al. (2013) show that SIS provides unbiased estimators of an unknown break point at any time within the sample.

Figures 1 and 2 outline the two step procedure of using SIS under the null of no break and under a specific alternative of a single step break. In the example, under the null two step indicators out of 100 are spuriously retained at $\alpha = 0.01$. The power of this split half approach is obvious in Figure 2 where the break is identified without prior knowledge of timing, magnitude or length.

Despite some similarities between the procedures in IIS and SIS, there are important differences necessitating additional analysis. First, while impulse indicators are mutually-orthogonal candidate regressors, step indicators are far from orthogonal, overlapping increasingly as the second index increases. Second, for a location shift that is not at either end, say from T_1 to T_2, two indicators are required to characterize it: $1_{\{t \leq T_2\}} - 1_{\{t < T_1\}}$. Third, the ease of detection must be affected by whether or not T_1 and T_2 lie in the

Figure 2: SIS under the alternative of a single break, $\lambda_1 = 10$SD



same split. Next, location shifts may occur in both halves. Consequently, instead of just the split-half approach, the combination of expanding and contracting multiple block searches will need to be used in practice.

To see IIS and SIS operating in practice when faced with locations shifts, consider a trending DGP (2) with a seasonal cycle not unlike many climate data series:

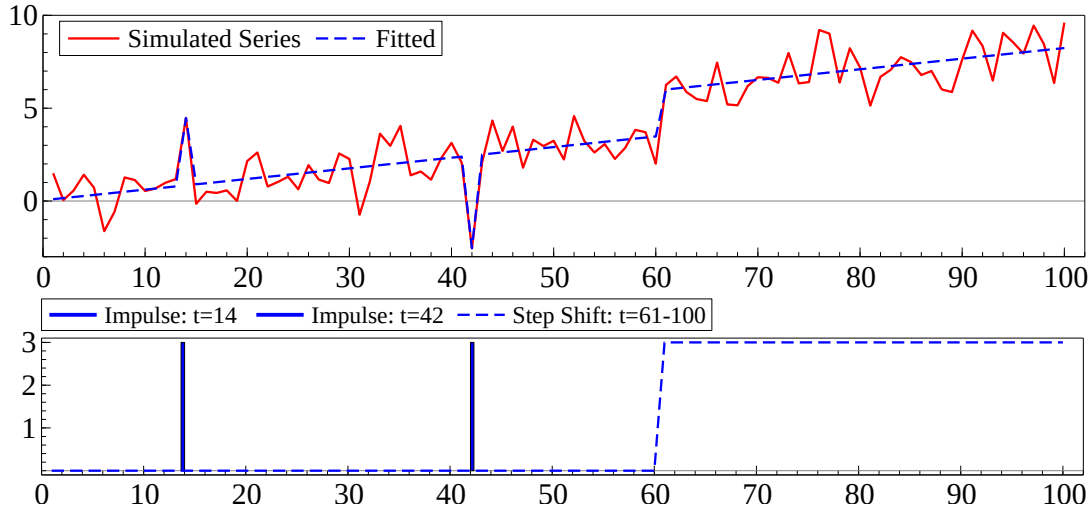
$$y_t = \mu + \beta_1 t + \beta_2 sin(t) + \lambda_1 1_{\{t=T_1\}} + \lambda_2 1_{\{t=T_2\}} + \lambda_3 1_{\{t>T3\}} + \epsilon_t \tag{2}$$

where $sin(t)$ simulates a simple seasonal cycle. This series is shown in Figure 3 for $\mu = 1$, $\beta_1 = 0.05$, $\beta_2 = 0.6$ and three breaks: two impulses ($T_1 = 14, T_2 = 42$) and one step shift ($T_3 = 60$) of 3 standard deviations of the error term ($\lambda_{1,3} = 3, \lambda_2 = -3$). As Figure 3 shows, IIS and SIS successfully identifies all three breaks even in the presence of a trend and seasonal component. The breaks are identified without prior knowledge of the magnitude, timing or duration of it taking place. This is a significant advantage over many other break or outlier detection techniques. These large breaks here are shown for illustration purposes, IIS and SIS has been shown to detect breaks of much smaller magnitude with high power (Castle et al. 2011b, Doornik et al. 2013).

Over-fitting is no concern: when no breaks occur, the expected rate of incorrectly retained break indicators in IIS and SIS can easily be controlled: it is equal to the significance level of selection. For example, for a time series of 100 observations and significance of $\mathsf{p} = 0.01$, on average only a single indicator will be spuriously retained.

There are two clear purposes of IIS and SIS. First, as the examples above demon-

Figure 3: Automatic Detection of Impulses and Step Shifts



strated, IIS and SIS are successful in identifying location shifts, breaks, and outliers. Second, as mentioned in section 2.2, IIS and SIS can be used as an indicator of model misspecification. A large number of retained dummies is indicative of model misspecification in the sense that a large fraction of the data cannot be explained by the estimated model, this could apply to non-linearities, omitted variables or other forms of misspecification.

2.3.3 Performance of Indicator Saturation

The main theory results of IIS are covered in Johansen & Nielsen (2009) and Hendry et al. (2008), SIS theory is analysed in Doornik et al. (2013). Here we provide a brief Monte Carlo simulation to demonstrate the performance. We first compare the success of identifying step shifts using IIS relative to IIS. Second, we assess the situation when IIS and SIS are combined and selection takes place over both impulses and step indicators. The following simulations are based on the stationary DGP given in equation (3) with the a single step shift set from $T_1 = 25$ to $T_2 = 35$.

$$y_t = \lambda \left(1_{\{t \leq T_2\}} - 1_{\{t \leq T_1\}}\right) + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right] \tag{3}$$

The simulated model for step indicators is given by equation (1). Selection of indicators is based on multiple block splits and conducted at $\alpha = 0.01$ (1% significance).

To evaluate the success of identifying step shifts three measures are used. First, the retention frequency of the step indicator and impulse indicator at the specified break timing is measured. High retention frequencies (close to 1) are preferred. Second, potency, the model selection equivalent of power, is defined as the proportion of relevant variables (featured in the DGP) selected in the final model. For a given (linear) DGP, let N be the total number of potential explanatory variables, of which the first $q + 1$ are relevant

$(\delta_0, \delta_1, ...\delta_q \neq 0)$, and $N-(q+1)$ are irrelevant $(\delta_{q+1}, \delta_{q+2}, ...\delta_N = 0)$. Let $\hat{\delta}_{k,i}$ denote the coefficient of the selected variable k in replication i. For M replications, $i = 1, 2, ..., M$, potency is thus defined as:

$$\text{Potency} = \frac{1}{q+1}\frac{1}{M}\sum_{k=1}^{q+1}\sum_{i=1}^{M} 1_{\{\hat{\delta}_{k,i}\neq 0\}} \tag{4}$$

Gauge, the model selection equivalent of size, measures the fraction of spuriously retained irrelevant variables:

$$\text{Gauge} = \frac{1}{N-(q+1)}\frac{1}{M}\sum_{k=q+2}^{N}\sum_{i=1}^{M} 1_{\{\hat{\delta}_{k,i}\neq 0\}} \tag{5}$$

Low gauge (close to zero) and high potency (close to 1) are preferred.

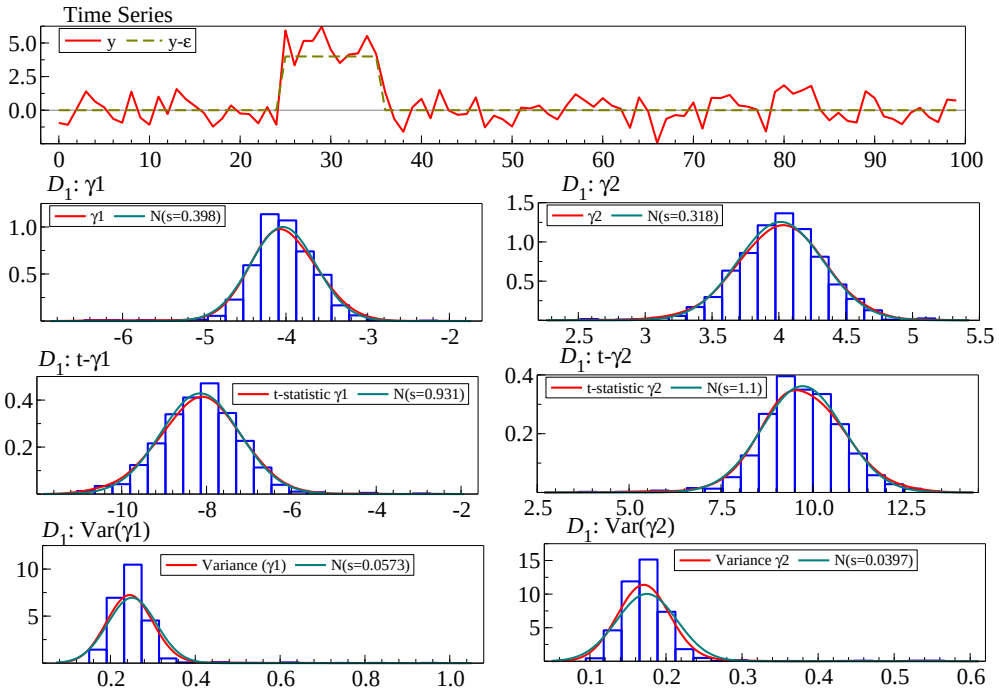
IIS compared to SIS in presence of a step shift: Table 1 summarize the simulation outcomes showing the high retention frequencies and potencies of using step indicators and impulse indicators in the presence of a small step shift. Gauge is easily controlled and close to the significance level of selection ($\alpha = 0.01$). Step shifts of two standard deviations are retained approximately 50% of the time, this increases to around 90% for a shift of 4 standard deviations. Figure 4 shows the simulated DGP together with the simulation densities for the step indicators at the break point. The step shift estimates γ_1, γ_2 are unbiased with low variance without prior knowledge of the break point or magnitude.

Table 1: IIS compared to SIS in presence of a step shift ($\alpha = 0.01$)

Retention freq.				
λ_1	IIS T_1 Step	SIS T_1 Step	IIS T_2 Step	SIS T_2 Step
2 SD	0.25	0.51	0.22	0.56
4 SD	0.90	0.93	0.90	0.93
	IIS Gauge	**SIS Gauge**	**IIS Potency**	**SIS Potency**
λ_1				
2 SD	0.007	0.032	0.25	0.98
4 SD	0.01	0.022	0.91	0.99

IIS combined with SIS in presence of a step shift: Combining IIS with SIS (what Ericsson & Reisman (2012) refer to as super saturation) allows for simultaneous selection over both. IIS in theory can recover step shifts by selecting a group of impulses and vice versa a linear combination of step indicators in SIS can identify a single impulse. Here

Figure 4: Step Indicator Saturation in the presence of a step shift, $\lambda_1 = 4$SD



we consider the case of a single step shift as before ($T_1 = 25$ and $T_2 = 35$). Table 2 shows the simulation outcomes.

Comparing the results to using SIS only as in the previous section, it appears that the presence of impulse indicators does not have a detrimental effect on the high power of the step indicators. For a step shift of 4 standard deviations using SIS alone (see table 1) yields average retention frequency of 0.93 (for T_1 and T_2), while combined with IIS (see table 2) the average retention frequency is 0.91 (for T_1 and T_2).

Table 2: IIS combined with SIS in presence of a step shift ($\alpha = 0.01$)

	Retention Freq.			
	IIS T_1 Step	SIS T_1 Step	IIS T_2 Step	SIS T_2 Step
λ_1				
2 SD	0.014	0.51	0.014	0.53
4 SD	0.021	0.91	0.018	0.91
	IIS Gauge	**SIS Gauge**	**IIS Potency**	**SIS Potency**
λ_1				
2 SD	0.008	0.025	0.014	0.97
4 SD	0.007	0.013	0.012	0.99

Using IIS and SIS together with retained regressors: So far the simulations have considered only indicators on their own without general explanatory variables. In line

with the theoretical work of Johansen & Nielsen (2009) on IIS, we assess by means of simulations the use of SIS with $n < T$ general regressors by including the $T \times n$ matrix $\mathbf{Z}$ as independent variables. For a single step shift with unknown timing requiring two indicators the DGP is then given by:

$$y_t = \beta_1' \mathbf{z}_t + \lambda \left(1_{\{t \leq T_2\}} - 1_{\{t \leq T_1\}}\right) + \epsilon_t \text{ where } \epsilon_t \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right] \tag{6}$$

For the present simulation we set $\sigma_\epsilon^2 = 1$ and $n = 10$. For each of the $i = 1, ..., n$ regressors the associated non-centralities are set to $\mathsf{E}[t_i] = \psi_i = 4$. The individual z_i are orthogonal in expectation and not selected over, thus present in every selection iteration of the step indicators. The break timing is set as before to $T_1 = 25$ and $T_2 = 35$. Table 3 and Figure 5 display the simulation outcomes and properties of the step indicators. With the inclusion of 10 relevant independent variables the densities of the two break estimators are centered around the true value of $\lambda_1 = 4SD$. The high power of SIS seems unaffected by the presence of additional regressors – the retention frequencies are close to values in experiments without regressors.

Table 3: SIS with General Regressors ($\alpha = 0.01$)

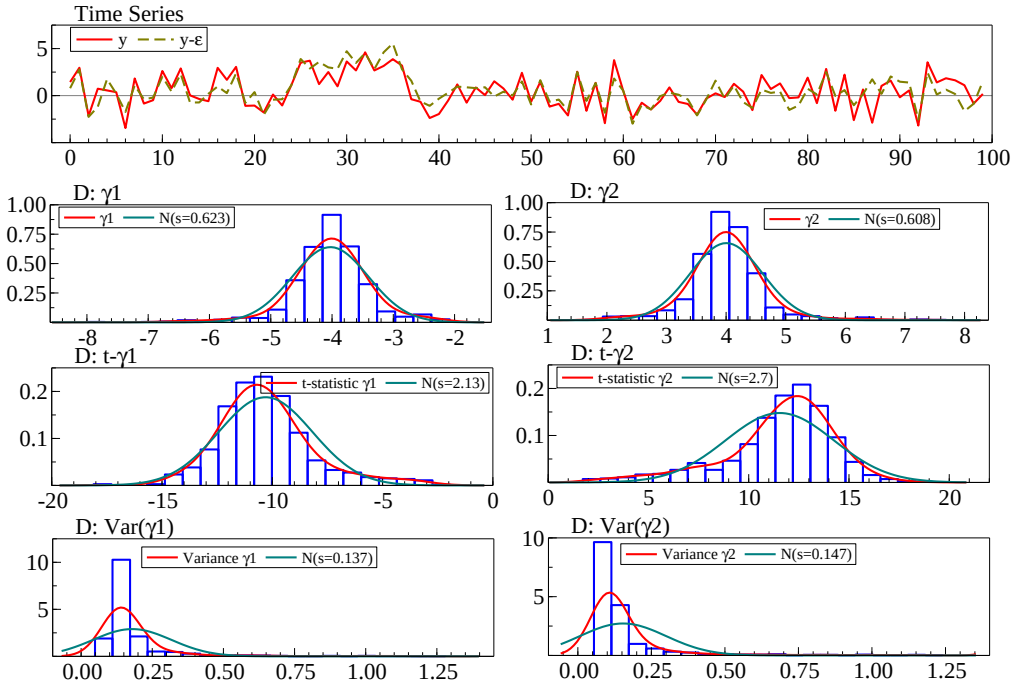
	Gauge	**Retention Frequency**	
λ_1		T_1 Step	T_2 Step
2 SD	0.035	0.5	0.62
4 SD	0.024	0.91	0.94

The Monte Carlo experiments support the theory for the feasibility of detecting location shifts using IIS and SIS. In the case of static DGPs with specific location shifts, the step indicators exhibit high retention frequencies – around 50% in the case of a shift equal to 2 standard deviations and around 90% for shifts of 4 standard deviations. These results appear constant across single shifts, multiple shifts, spanning both halves and even including additional regressors and impulses.

2.3.4 Model Selection generalization to more variables than observations

A natural extension of IIS and SIS is to move from allowing more variables than observations with indicators to the general case of all forms of independent variables (see Hendry & Krolzig 2005 and Hendry & Johansen 2013). Suppose there are N total regressors partitioned into J blocks of n_j, where $N = \sum_{j=1}^{J} n_j$ such that $N > T$ and $n_j < T$ for all j. Consequently the total number of variables N exceeds the number of observations T but total variables can be partitioned into J blocks n_j each smaller than T. One approach is then to randomly partition the set of variables into blocks of n_j, applying a selection

Figure 5: SIS with General Regressors, $\lambda_1 = 4SD$



algorithm to each block retaining the selected variables and crossing the groups to mix variables. The next step is to use the union of selected variables from each block to form a new initial model and repeat the process until the final union of selected variables is sufficiently small.

2.4 Implementation of the Selection Algorithm

Automated general to specific modelling using IIS, SIS, and allowing for more variables than observations via block estimation is currently implemented under the name of *Autometrics* (see Doornik 2009) in the *OxMetrics* (Doornik 2010b) software package.

We briefly outline the main components, the algorithm is described in detail in Doornik (2009). The algorithm is based on the following main components: The general unrestricted model (GUM) is the starting point of the search, it should be specified broadly based on theoretical considerations to nest the local data generating process. This includes a full set of impulse and step indicators to be searched over. Starting from the GUM using a tree-search, the algorithm removes the least significant variable as determined by the lowest absolute t-ratio. Each removal constitutes one branch of the tree. For every reduction, there is a unique sub-tree which is then followed; each removal is back-tested against the initial GUM using likelihood ratio tests (equivalent to an F-test). If back-testing fails, no sub-nodes of this branch are considered (though different variants of this removal exist). Branches are followed until no further variable can be removed at the pre-specified level of significance α. If no further variable can

be removed, the model is considered to be terminal. Each terminal model is subjected to a range of diagnostic tests based on a separately chosen level of significance. These tests include tests for normality (based on skewness and kurtosis), heteroskedasticity (for constant variance using squares), the Chow test (for parameter constancy in different samples), and residual autocorrelation and autoregressive conditional heteroskedasticity. Parsimonious encompassing of the feasible general model by sub-models both ensures no significant loss of information during reductions, and maintains the null retention frequency of the algorithm close to α: see Doornik (2008). Both congruence and encompassing are checked by the algorithm when each terminal model is reached after path searches, and it backtracks to find a valid less reduced earlier model on that path if any test fails. This repeated re-use of the original mis-specification tests as diagnostic checks on the validity of reductions does not affect their distributions (see Hendry & Krolzig 2003). As a result of the tree search, multiple valid terminal models can be found. As a tiebreaker to select a unique model, the likelihood-based Schwarz (1978) information criterion (SIC) is used, though other methods are also applicable, and terminal models should be considered individually.

The main calibration decision in the search algorithm is the choice of significance level α at which selection occurs. Selection continues until retained variables are significant at α, though it can be the case that variables in the final model are also retained at a level above α if removal leads to diagnostic tests failing. As described for IIS and SIS, α is approximately equal to the gauge of selection. Further, the choice of diagnostic tests and lag length selection for residual autocorrelation and autoregressive conditional heteroskedasticity need to be set.

In the broad case of more variables than observations (IIS, SIS and general $N > T$), the algorithm groups variables into two categories: selected and not selected (Doornik 2010a). Not currently selected variables are split into sub-blocks and the algorithm proceeds by alternating between two steps: first, the expansion step, selection is run over not-selected sub-blocks to detect omitted variables. Second, the reduction step, a new selected set is found by running selection on the system augmented with the omitted variables found in step one. This is repeated until the dimensions of the terminal model are small enough and the algorithm converges, so the final model is unchanged by further searches for omitted variables. For a comparison to other automatic model selection methods see Castle et al. (2011a).

Given the complexities in climate time series, large scale extended general to specific model selection together with impulse and step indicator saturation thus leads to an agnostic data-driven modelling methodology.

3 Applications

We consider two applications to demonstrate the model selection methodology. First, using a large set of potential variables, based on Hendry & Pretis (2013a), we empirically estimate contributions to atmospheric CO_2. Second, we estimate a long-run climate system based on ice core data while accounting for structural breaks.

3.1 Mauna Loa: Atmospheric Carbon Dioxide

Empirically estimating the determinants of atmospheric CO_2 is traditionally a challenge due to the complex systems of data involved. Atmospheric levels of carbon dioxide are a highly autocorrelated, non-stationary time series, and globally there exist a large number of potential carbon sources and sinks. There is mixed evidence in the literature on human contributions to atmospheric CO_2: the long-term trend is widely attributed to human factors, while the main seasonal fluctuations are thought to be driven by the biosphere. However, the statistical measures applied are often somewhat unsatisfactory due to the complexities of dealing with large numbers of variables. Without being restricted by *a priori* selection of explanatory variables, our approach uses high frequency (monthly) data and selects over a number of natural carbon sources and sinks: vegetation, temperature, weather phenomena, as well as accounting for dynamic transport. This allows for an estimate of the human contribution to CO_2 as measured by industrial output indices and fossil fuel use for different geographical areas. The resulting estimates describe the direct effects on CO_2 growth within the estimated model and the proportional contribution of each factor. We find that natural factors alone cannot explain either the trend or all the variation in CO_2 growth – industrial production components driven by business cycles and shocks are highly significant contributors. The present application is based on the work of Hendry and Pretis (2013) which provides a detailed analysis.

Data

The aim of this analysis is to empirically estimate the determinants of atmospheric background CO_2 using high frequency (monthly) observations. The CO_2 series is taken from Keeling's measurements at Mauna Loa, available from Tans & Keeling (2013). As a measure for vegetation the normalized difference vegetation index (NDVI) (Tucker et al., 2010) in the form of principal components (as well as Winter/Summer weighted components) covering the northern hemisphere are used. NDVI is available from 1981:7 until 2002:12, this defines the model estimation period. To capture oceanic effects, the southern oscillation index (SOI) from the Australian Bureau of Meteorology (2011) is included. There is evidence for an approximate linear feedback of temperature on CO_2 Scheffer et al. (2006). Using the Northern hemisphere land and sea surface temperature

anomaly (NASA Goddard Institute for Space Studies (GISS) 2011) these temperature measurements are included for the potential effect of decreased oceanic uptake with higher sea-surface temperatures and higher release from soil and the biosphere with increased temperature. Naturally the feedback of CO_2 onto temperature is a concern, in terms of estimation potentially inducing a problem of endogeneity. Since no variable is a-priori enforced to enter the model, if contemporaneous temperature were selected, we would proceed by instrumenting it using stratospheric aerosols and lagged SOI. However, any lagged temperature can for this short time period of analysis be seen as pre-determined and thus treated as given.

The main high frequency indicators for anthropogenic contributions to CO_2 used here are monthly industrial production indices (IP) due to a lack of high frequency emissions data. Industrial production is taken from multiple regions and re-parametrised using principal components due to high collinearity.[2] The series include US Industrial production (Federal Reserve 2011), UK production (Office of National Statistics 2011), German and Japanese production (OECD 2011) as well as Indian production (Government of India, Ministry of Statistics, 2011). While industrial production reflects the intensity of economic activity associated with higher emissions, it does not account for changing emission intensity. More efficient processes could lead to an increase in industrial production without an equivalent increase in emissions. This is a crucial missing measure and difficult to control for. By including overall long-run emissions estimates in addition to production we make an attempt to control for this shortcoming, together with the lack of high frequency production data for some regions (e.g. China). We thus include interpolated annual low frequency variables of total CO_2 emissions in thousand metric tons of carbon for North America, Western and Eastern Europe, Central Asia and the Far East (Marland et al., 2011). Due to the high collinearity these again are combined through the use of principal components.

Methodology

Atmospheric background CO_2 here is approximated as a stock variable, we model the change in CO_2 as a function of sources and sinks. This basic model does not intrinsically ensure that the property of CO_2 being a long-lived gas is preserved.

In terms of model specification, there are two main groups of variables: the main variable of interest $\Delta \mathrm{CO}_{2,t}$, and the economic and natural factors which are grouped in the (20×1) vector $\mathbf{X}_t = (\mathrm{IP}_{1-3,t}, \mathrm{Emission}_{1-3,t}, \mathrm{NDVI}_{1-3\ \mathrm{Eur,\ NA},t}, \mathrm{Seasonal\ NDVI}_{1-3\ \mathrm{Eur,\ NA},t}, \mathrm{Temp}_t, \mathrm{SOI}_t)'$. The general unrestricted model is given in equation (7):

$$\Delta CO_{2,t} = a_j \sum_{j=1}^{s} \Delta CO_{2,t-j} + \sum_{j=0}^{s} \mathbf{X}'_{t-j}\mathbf{B}_j + \mathbf{Z}_t{}'\phi + \epsilon_{\Delta CO_2 \cdot x,t} \tag{7}$$

[2]Due to a lack of high frequency data crucially Chinese production is omitted.

The vector $\mathbf{Z}_t$ consists of non-lagged deterministic terms such as a linear time trend, centred seasonal variables and impulse indicators. Let q denote the total number of explanatory variables that appear on the right-hand side of equation (7). Then the change in atmospheric CO_2 in the conditional model is modelled as a function of past values of the change in CO_2, current and past values of selected independent variables $\mathbf{X}_{t-j}$, and deterministic components. Table 4 lists the variables included in the general unrestricted model, lag lengths are chosen based on earlier estimates in Hendry & Pretis (2013a). Selection occurs at the conservative $\alpha = 0.001$ (0.1%) significance level.

Table 4: General Unrestricted Model

Variables included	lag length
Temperature	6
NDVI PC1 Eurasia (Eur) + Winter Interaction	12
NDVI PC2 Eurasia (Eur) + Winter Interaction	12
NDVI PC3 Eurasia (Eur) + Winter Interaction	12
NDVI PC1 North America (NA) + Summer Interaction	12
NDVI PC2 North America (NA) + Summer Interaction	12
NDVI PC3 North America (NA) + Summer Interaction	12
SOI	6
Industrial Production Comp. 1	6
Industrial Production Comp. 2	6
Industrial Production Comp. 3	6
CO_2 Emissions Comp. 1	6
CO_2 Emissions Comp. 2	6
CO_2 Emissions Comp. 3	6
Constant	yes
Trend	yes
Centred Seasonal Variables	yes
Impulse Indicators	yes
Total variables	483
Number of Observations	246

Results

The model search algorithm estimated 571 models reducing the initial potential 483 variables down to a model of 19 variables given in equation (8). Unit roots in the residual are rejected for the model residuals at the 1% level using ADF tests covering up to 12 lags. There is only a single impulse indicator selected (1990:7), suggesting no major breaks or mis-specification.

First: as is to be expected from the theory, controls for natural factors are selected in the final model. Temperature anomalies enter the model with a positive coefficient, likely capturing the decreased effect of oceanic uptake and increased soil activity. Vegetation

controls through the principal components of NDVI are selected, as is the control for Southern Oscillation. However, a key finding is that in both terminal models, natural controls are insufficient to account for the variation in the change of atmospheric CO_2 in this conditional model. Anthropogenic factors captured through components of industrial output indices are consistently selected. Selection of these is robust to the addition of emissions components which are not selected, suggesting that the high frequency measures provide a better approximation for monthly anthropogenic emissions measured at Mauna Loa.

Second: most selected variables enter the model in lagged form. Only the third principal component of production and the first component of Eurasian NDVI have an estimated immediate effect on the growth of CO_2. Most anthropogenic emissions and vegetation growth "lead" measured atmospheric CO_2 by suggested time periods of 1 to 12 months. This suggests that no instrumental variable approach is required since lagged temperature can be seen as pre-determined for this very short time period. Relative to the initial sizes of the GUM, few variables are retained, yet relative to the tight significance levels, many more are retained than could be attributed to chance (less than 1 on average).

Third: the final model appears to be well specified. As a result of the algorithm, all models pass tests for normality, heteroskedasticity, residual autocorrelation and autoregressive conditional heteroskedasticity. The number of selected indicators from IIS is low. There is only one indicator selected for 1990:7. This suggests that the model is correctly specified and there appear to be no real structural breaks or shifts in the change of atmospheric CO_2. No deterministic variables are selected: no constant, time trend or centred seasonals appear in the final models. This suggests that changes in CO_2 are well approximated by the selected final variables covering anthropogenic and natural factors. Linearity is not rejected (p=0.619) using a non-linear factor based test (Castle & Hendry 2010). If constants are included post-selection (though not statistically different from zero), R^2 can be used as a rough measure of goodness of fit. The final model exhibits a high goodness of fit (R^2= 0.978), in comparison a random walk model with deterministic seasonal indicators results in a much lower goodness of fit, $\mathsf{R}^2 = 0.753$. High goodness of fit of the models is not a straight result of selection as the selection algorithm does not directly maximize the goodness of fit. Moreover, R^2 measures should not be attributed much weight when assessing models, as there are preferred likelihood-based measures that also account for the number of parameters included.

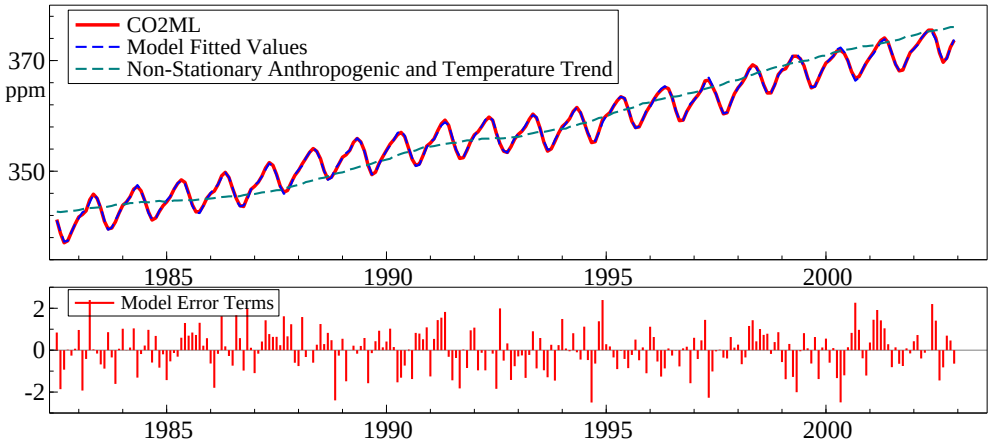
$$
\begin{aligned}
\widehat{\Delta \mathrm{CO}}_{2,t} = & \underset{(0.051)}{-0.57}\ \Delta \mathrm{CO}_{2,t-2} - \underset{(0.050)}{0.22}\ \Delta \mathrm{CO}_{2,t-4} + \underset{(0.062)}{0.27}\ \mathrm{IP}_{1_{t-1}} \\
& - \underset{(0.063)}{0.34}\ \mathrm{IP}_{1_{t-4}} - \underset{(0.039)}{0.28}\ \mathrm{IP}_{2_{t-4}} + \underset{(0.035)}{0.16}\ \mathrm{IP}_{3_t} - \underset{(0.21)}{0.74}\ \mathrm{I}_{1990(7)} \\
& + \underset{(0.0006)}{0.004}\ \mathrm{Temp}_{t-4} - \underset{(0.001)}{0.007}\ \mathrm{SOI}_{t-5} - \underset{(0.008)}{0.047}\ \mathrm{NDVI}_{1,\,\mathrm{Eur}_t} \\
& - \underset{(0.006)}{0.044}\ \mathrm{NDVI}_{1,\,\mathrm{Eur}_{t-3}} + \underset{(0.006)}{0.033}\ \mathrm{NDVI}_{1,\,\mathrm{Eur}_{t-11}} - \underset{(0.009)}{0.042}\ \mathrm{NDVI}_{1,\,\mathrm{Eur}_{t-12}} \\
& + \underset{(0.008)}{0.030}\ \mathrm{NDVI}_{2,\,\mathrm{Eur}_{t-11}} - \underset{(0.007)}{0.029}\ \mathrm{w\_NDVI}_{2,\,\mathrm{Eur}_{t-7}} - \underset{(0.006)}{0.022}\ \mathrm{w\_NDVI}_{2,\,\mathrm{Eur}_{t-8}} \\
& + \underset{(0.007)}{0.035}\ \mathrm{w\_NDVI}_{3,\,\mathrm{Eur}_{t-8}} - \underset{(0.008)}{0.039}\ \mathrm{s\_NDVI}_{1,\,\mathrm{NA}_{t-8}} + \underset{(0.007)}{0.034}\ \mathrm{s\_NDVI}_{1,\,\mathrm{NA}_{t-9}} \qquad (8) \\
& \widehat{\sigma} = 0.199 \ \ T = 246 \ \ n = 19 \ \ \mathsf{SIC} = -0.048 \ \ \mathsf{F}_{\mathsf{ar}}(8, 219) = 1.62 \\
\chi^2_{\mathsf{nd}}(2) = 0.044 \ \ & \mathsf{F}_{\mathsf{reset}}(2, 225) = 0.042 \ \ \mathsf{F}_{\mathsf{arch}}(8, 230) = 0.57 \ \ \mathsf{F}_{\mathsf{het}}(36, 208) = 1.03
\end{aligned}
$$

The model in its present specification describes the change in background atmospheric carbon dioxide, $\Delta \mathrm{CO}_2$. It is straight forward to recover the the estimated level. Level estimates for the model are based on $\widehat{\mathrm{CO}}_{2,t} = \widehat{\Delta \mathrm{CO}}_{2,t} + \mathrm{CO}_{2,t-1}$ are given in Figure 6 and show the close fit. The next step is to attribute the components explaing the long-run trend. To do so, we derive the relation after all dynamics from lagged variables have been solved out (the 'long-run solution': see Hendry 1995, p.212).

We divide the variables into two different groups, $\mathbf{x}_{q,t}$ with stationary cumulative sums (non-trending) and $\mathbf{x}_{q-s,t}$ with non-stationary cumulative sums (trending). Stationary variables by nature cannot drive the trend. Only explanatory variables with non-stationary cumulative sums determine the underlying trend.

Out of the selected variables in model (8), only a sub-set exhibit trending cumulative sums, which are all the anthropogenic factors and the temperature anomaly. The natural controls of NDVI remain approximately stationary around zero over time, SOI appears to be slightly increasing over the time frame considered, but its effect is small in magnitude when weighted by its long run coefficient. Importantly, neither final model includes a deterministic intercept or trend, which on summation would become a linear or a quadratic time trend. However, summed variables do not have a straight-forward interpretation in the case of PCs of industrial output and temperature. The trending temperature anomaly is likely a mutually supporting feedback effect. Overall, the trend in the levels of CO_2 in these conditional models is derived from the trends in the independent variables in both estimated models, so is attributed primarily to the PCs of production and partly by temperature. Specifically, the empirical specification of the non-stationary component in the level for the model is given in equation (9).

Figure 6: Modelled CO_2: Fitted values, non-stationary anthrop. trend and error terms



$$\mathbf{x}'_{q-s,t}\beta_{q-s} = +0.0345 \sum_{j=1}^{t} \mathrm{IP}1_j - 0.154 \sum_{j=1}^{t} \mathrm{IP}2_j + 0.088 \sum_{j=1}^{t} \mathrm{IP}3_j + 0.0021 \sum_{j=1}^{t} \mathrm{Temp}_j \quad (9)$$

Figure 6 shows the resulting coefficient-weighted cumulative sums of the combined anthropogenic components (IP_1 to IP_3) and temperature trend for the estimated model together with the recorded level of CO_2. The industrial production components and temperature approximate the level of CO_2 well, marking a slight slow down in the trend around 1991–1993. Both cumulative stationary components (NDVI and SOI) vary over a small range so contribute little to the long-run changes. Even though the model is estimated in net inflows to atmospheric carbon dioxide, in re-parametrized form it can explain the long-run trend—and, within the conditional model, attributes it primarily to anthropogenic emissions. This is an outcome of the data analysis and is not enforced.

The model treats CO_2 as a stock variable and while nothing intrinsically guarantees that modelled CO_2 is long-lived, the model does exhibit this property – the level of CO_2 does not drop to zero when emissions are set to zero, consistent with the properties of CO_2 being a long-lived gas (Archer et al., 2009).

3.2 Paleoclimate: Structural Breaks and Cointegration in the Ice Core Record

The aim of this second application is to investigate statistical long run relationships between multiple series of the Antarctic ice core record while accounting for unknown structural changes induced by breaks and measurement error. The ice core time series faces high measurement uncertainty (Jansen et al. 2007) and likely multiple structural breaks, which if un-modelled lead to inconsistent parameter estimates. Kaufmann & Juselius (2013) estimate a cointegrated vector auto-regression model to untangle the long

run relations in a system based on Antarctic measurements. We closely follow their approach (with a smaller set of endogenous variables), while introducing IIS and SIS to control for un-modelled shifts.

Thus, we employ IIS and SIS to estimate a co-integrated system of temperature, carbon dioxide, sea surface temperature, ice volume, sea-level and variation in Earth's orbit over approximately the past four hundred thousand years. A large number of previously un-modelled location shifts are detected and controlled for through the use of impulse and step indicators.

Data

Antarctic land surface temperature proxies are taken from Jouzel et al. (2007), currently the only variable measuring greenhouse gases included in the model is amtospheric CO_2 obtained from Luethi et al. (2008). Ice volume estimates ($\delta_{18}O$) obtained from Lisiecki & Raymo (2005) are included to capture effects on the sea level and potential surface albedo effects (Lea 2004). Sea surface temperature (SST) from Martinez-Garcia et al. (2009) is added to the model to capture oceanic CO_2 uptake and interactions with land surface temperature, sea level data based on sediments is obtained from Siddall et al. (2003). All observations are adjusted to the common EDC3 time scale and linearly interpolated for missing observations to bring all observations on a 1000 year time interval. Solar variables are given by a specific cumulative insolation at 65 degrees South (June 21). To capture orbital variations, eccentricity (deviations from perfectly circular orbit to elliptical orbit), obliquity (angle of axial tilt) and precession (orientation change of the rotational axis) are included (Paillard et al., 1996). Each data series $x_{i,t}$ is standardized prior to estimation such that each variable has a zero mean and standard deviation of one: $y_{i,t} = (x_{i,t} - \bar{x}_i)/\mathrm{SD}(x_i)$. The total sample size is T=376 from -376k years to -1k years before present. This initial analysis omits any effect that dust concentrations, biological activity, other greenhouse gases and aerosols might have. The interaction of these together with detection of structural breaks will be investigated in further research.

Methodology

In line with the analysis of Kaufmann and Juselius (2013) we model the system as a cointegrated vector auto-regression (CVAR) (Johansen 1995, Juselius 2006, Hendry & Juselius 2001). This allows us to identify stationary long-run equilibrium relations and the relative adjustment out of dis-equilibrium. While the methodology of VARs suggests a highly simplified approximation of the climate system, it appears consistent with a simple zero-dimensional energy balance model (Kaufmann et al. 2013, Kaufmann & Juselius 2013). However, there are a few concerns with this methodology. First, while radiative forcing of CO_2 on temperature is logarithmic, Scheffer et al. (2006) and our analysis in section 3.1 suggest that temperature has a linear feedback onto CO_2. Nevertheless the

CVAR is limited to either a log or linear specification. We follow the general literature and use a linear approximation, acknowledging that the functional form of CO_2 is thus slightly mis-specified from the outset. Second, CO_2 in a VAR setting is generally modelled in levels which suggests that the lifetime of the gas is under-estimated and leads to rapid adjustment in the model. This is a concern, however, given that the time-interval of estimation here is 1000 years, even a rapid-adjustment within a few periods actually corresponds to a slow process. The model predicted response (in particular reduction) is of the order of magnitude of thousands of years, approximately consistent with the lifetime of atmospheric CO_2 described by Archer et al. (2009).

Specifying our CVAR, we assume that all solar variables are weakly exogenous, there are no feedbacks of climate variables onto solar variables. This allows us to model the relations as a partial system, taking the solar variables as exogenous. Treating the 4 solar and orbital variables (insolation, eccentrictiy, obliquity and precession) as non-stationary in this partial system induces common trends into all modelled endogenous relations. When weak exogeneity is imposed by assumption rather than tested, each weakly exogenous variable is by assumption equivalent to a common trend and corresponds to a unit root in the full system (see Juselius 2006).

The partial system is set up such that the $(p \times 1)$ vector $\mathbf{Y}_t$ (surface temperature, atmospheric CO_2, ice volume, sea surface temperature and sea level) is modelled endogenously, while the $(z \times 1)$ vector $\mathbf{Z}_t$ (insolation, eccentricity, obliquity and precession) is taken as exogenous. The associated cointegrated VAR model is given by equation (10).

$$\Delta \mathbf{Y}_t = \mathbf{A}_0 \mathbf{Z}_t + \mathbf{A}_1 \Delta \mathbf{Z}_{t-1} + \mathbf{\Gamma}_1 \Delta \mathbf{Y}_{t-1} + \Pi(\mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}) + \mu + \mathbf{\Phi} \mathbf{D}_t + \epsilon_t \qquad (10)$$

where $\mathbf{A}_0$, $\mathbf{A}_1$ are $(p \times z)$, $\mathbf{\Gamma}_1$ is $(p \times 1)$, Π is $(p \times (z+p))$, μ is $(p \times 1)$ and $\mathbf{\Phi}$ is $(p \times l)$ where l denotes the total number of indicators selected.

Under cointegration, the coefficient matrix Π has reduced rank and can be expressed as the product $\Pi = \phi\beta'$, where β is the matrix determining the long-run stationary cointegrating relationships and ϕ are the adjustment coefficients to this long run equilibrium.

A major concern when working with paleoclimatic ice core observations is the high uncertainty associated with the data. We therefore employ impulse and step indicator saturation to detect otherwise un-modelled shifts and outliers, as well as to provide evidence for model mis-specification. These indicators enter our main model (10) through the $(p \times l)$ matrix $\mathbf{D}_t$. Indicators are selected by estimating each equation of the VAR individually and retaining significant impulses using the multiple split approach outlined in section 2. For the present estimation, a lag length of 4 is chosen based on the Akaike information criterion (AIC). To control for the risk of spuriously retained indicators the level of significance of selection of indicators is set extremely conservatively to $\alpha = 0.0001$ (equivalent to significance at 0.01%). Given the sample size of 376, the total number of

potential indicators[3] is $2T - 1 = 751$. At the chosen level of significance $\alpha = 0.0001$, we expect on average to retain $\alpha(2T - 1) = 0.075$ impulses spuriously. The joint set of indicators is then entered in the CVAR formulation (for an example of dummies in VAR and cointegration, see Juselius 2006). The next section discusses the large number of retained indicators. Following the inclusion of indicators into the system, we test the cointegrating rank, noting that the weakly exogenous variables induce 4 underlying common trends. Thus, under cointegration the full system of 9 variables will exhibit at most 5 stationary long-run (cointegrating) relations.

Results

Indicators: Using impulse and step indicator saturation at $\alpha = 0.0001$ a large number of outliers and structural breaks are identified in the model. Table 5 and Figure 7 provide details on the selected indicators. The surface temperature and ice volume equations appear well approximated by the estimated linear model and there is little evidence of structural change with only a single indicator selected. Sea surface temperature, CO_2 and sea level exhibit a larger number of structural breaks. It is difficult to differentiate retained indicators between measurement uncertainty and actual structural breaks but the method identifies time periods that warrant further investigation and crucially accounts for them when the cointegrated model is estimated. Further research will investigate attribution of retained indicators to paleo-climatic events. An additional concern are omitted variables, so a further stage of analysis is to broaden the set of variables to include dust concentration, measures for biological activity and additional greenhouse gases.

Cointegrating Relations: Following the identification of structural breaks within the model, the partial system is estimated to determine the cointegration rank. In the partial system model of five endogenous variables reduced rank of 4 of the Π matrix is rejected ($p \simeq 0.000$), suggesting five cointegrating relationships among the complete system of 9 variables (where the exogenous are treated as non-stationary and due to weak exogeneity impose common trends on the other variables). Thus the system is estimated under reduced rank of 5 being imposed to estimate the cointegrating relations and adjustment coefficients. We impose identifying restrictions to determine the cointegrating relations based on: physical theory, tests for the validity of restrictions, and individual significance (see Table 6). The imposed restrictions are not rejected at the 1% level (p=0.0221).

Table 6 provides the estimated long run stationary relations (as given by the beta coefficients) together with the imposed restrictions and the adjustment towards this equilibrium (as given by the ϕ coefficients)[4]. Negative values of ϕ suggest that the system returns to a state of equilibrium when perturbed. Each cointegrating relation is normal-

[3]Total potential indicators: 376 impulses + 376 step indicators - 1 to control for the intercept.

[4]Full set of estimated ϕ coefficients and standard errors available from the authors upon request.

Table 5: Retained indicators from estimated vector auto-regression

	Impulse Indicators	Step Indicators
Temperature	none	none
CO_2	I_{-264}	S_{-341}, S_{-339}, S_{-318}, S_{-316}, S_{-308}, S_{-305}, S_{-252}, S_{-251}, S_{-242}, S_{-241}, S_{-128}, S_{-127}, S_{-45}
Ice Volume	I_{-129}	none
SST	I_{-374}, I_{-347}, I_{-342}, I_{-333}, I_{-282}, I_{-280}, I_{-147}, I_{-134}, I_{-126}, I_{-29}, I_{-15}, I_0	S_{-349}, S_{-347}, S_{-343}, S_{-341}, S_{-276}, S_{-269}, S_{-267}, S_{-227}, S_{-222}, S_{-151}, S_{-150}, S_{-147}, S_{-135}, S_{-125}, S_{-45}, S_{-31}, S_{-29}
Sea Level	I_{-336}	S_{-239}, S_{-194}, S_{-193}, S_{-132}, S_{-125}, S_{-12}, S_{-8}, S_{-7}, S_{-6}, S_{-5}, S_{-3}

Indicators selected at $\alpha = 0.0001$. I indicates single impulse, S indicates step indicator, subscripts denote k-years before present.

Figure 7: Standardized Observations and Retained Indicators (matched by means and ranges)

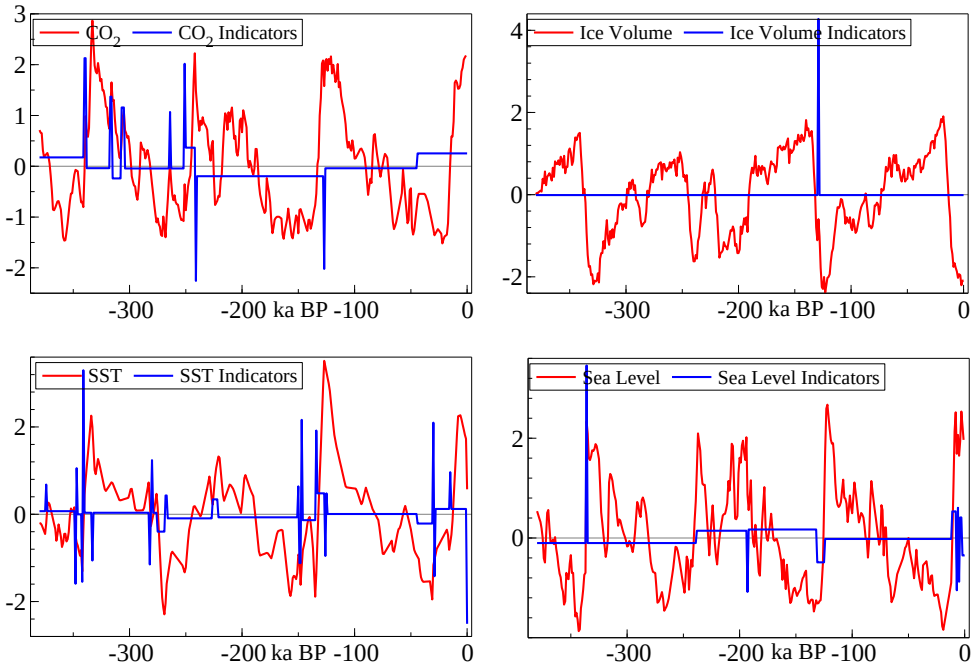


Table 6: Long run relations given by β and equilibrium adjustment given by ϕ

	LR1	LR2	LR3	LR4	LR5
ϕ	-0.40**	-0.29**	-0.07	-0.07	-0.24**
	(0.11)	(0.04)	(0.08)	(0.04)	(0.07)
Temperature	1	-	1.76**	-1.96**	-
			(0.086)	(0.084)	
CO_2	-0.39**	1	-	-	-
	(0.047)				
Ice Volume	-	-	1	-	1.67**
					(0.083)
SST	-	-1.4**	-	1	0.45**
		(0.086)			(0.061)
Sea level	-	-	-	0.55**	1
				(0.045)	
Eccentricity	-	-	-	-0.21**	-
				(0.044)	
Obliquity	-	-	0.07**	-	-
			(0.023)		
Precession	-	-	-	-	-
Solar	-0.13**	-	-	-	-
	(0.023)				

Parameter estimates of the cointegrating vectors β and equilibrium adjustment coefficients ϕ. Each long run relation (LR) is normalized on an endogenous variable. The coefficients β and ϕ are asymptotically normal. Standard errors given in parentheses. ** indicates significant at 1%. - indicates coefficient is restricted.

ized on a single endogenous variable, this yields five long run or cointegrating relations that are stationary over time and can be interpreted relative to the normalized variable. While stationary, they are not causal by default.

Equation LR1, normalized on temperature, provides evidence for a significant stationary relationship of Antarctic temperature increasing with measured atmospheric CO_2 and incoming solar radiation. Re-arranging results from LR1 in Table 6, this long run relation is given by: $\text{Temp}_t = 0.39\text{CO}_{2,t} + 0.13\text{Solar} + u_t$, where u_t is stationary. Temperature adjusts to this equilibrium relation with the estimated coefficient ϕ of -0.40. The second cointegrating vector suggests a long run relationship between atmospheric CO_2 and sea surface temperature, consistent with results based on our CO_2 model in section 3.1 and the evidence that oceanic CO_2 uptake decreases with sea surface temperature (Watson et al. 1995). The third cointegrating vector suggests a negative relationship between ice volume and surface temperature. Contrary to Kaufmann & Juselius (2013) we find a negative relationship between ice volume and obliquity. LR4 describes a positive relationship between sea surface temperature and land surface temperature. The fifth estimated long run relation shows the statistical relation of sea level being positively associated with a decrease in ice volume. All equilibrium adjustment coefficients are negative and in magnitude less than 1, however, for LR3 and LR4 appear not to be significantly different from zero.

System Simulation: The estimated system model based on a CVAR and indicator saturation allows us to conduct a simple simulation of the five endogenous variables over glacial and inter-glacial cycles. This enables a comparison of the model with indicators to a model without the use of indicators. The simulations are initialized with the given starting values in the dataset and then computed based on the model estimates alone together with the exogenous solar drivers. Unlike the model fitted values this only takes the initial observations and aside from exogenous solar variables determines climate dynamics endogenously other than the ex-post determined values of indicators.

Figure 8 graphs the simulated (standardized) glacial and inter-glacial record for both the model with indicators and without indicators against the observed values. Based on visual inspection it can be seen that this basic statistical model is able to reproduce, to an extent, the observed glacial and inter-glacial cycles, and that the inclusion of indicators provides a higher fit. This is supported by the respective R^2 values (Table 7) of a simple regression of the observations against the simulated values.[5] Note though that this is not a result of over-fitting, step indicators individually cannot remove single observations and only few impulses are included. Without the inclusion of indicators it is likely that the long run relations are not estimated correctly, uncontrolled structural breaks and measurement noise bias the estimates.

Kaufmann & Juselius (2013) provide simulation evidence of the glacial cycles without

[5]This is only a basic measure and only one of many ways to establish goodness of fit.

Figure 8: Observations and Simulations with and without Indicator Saturation (IS)

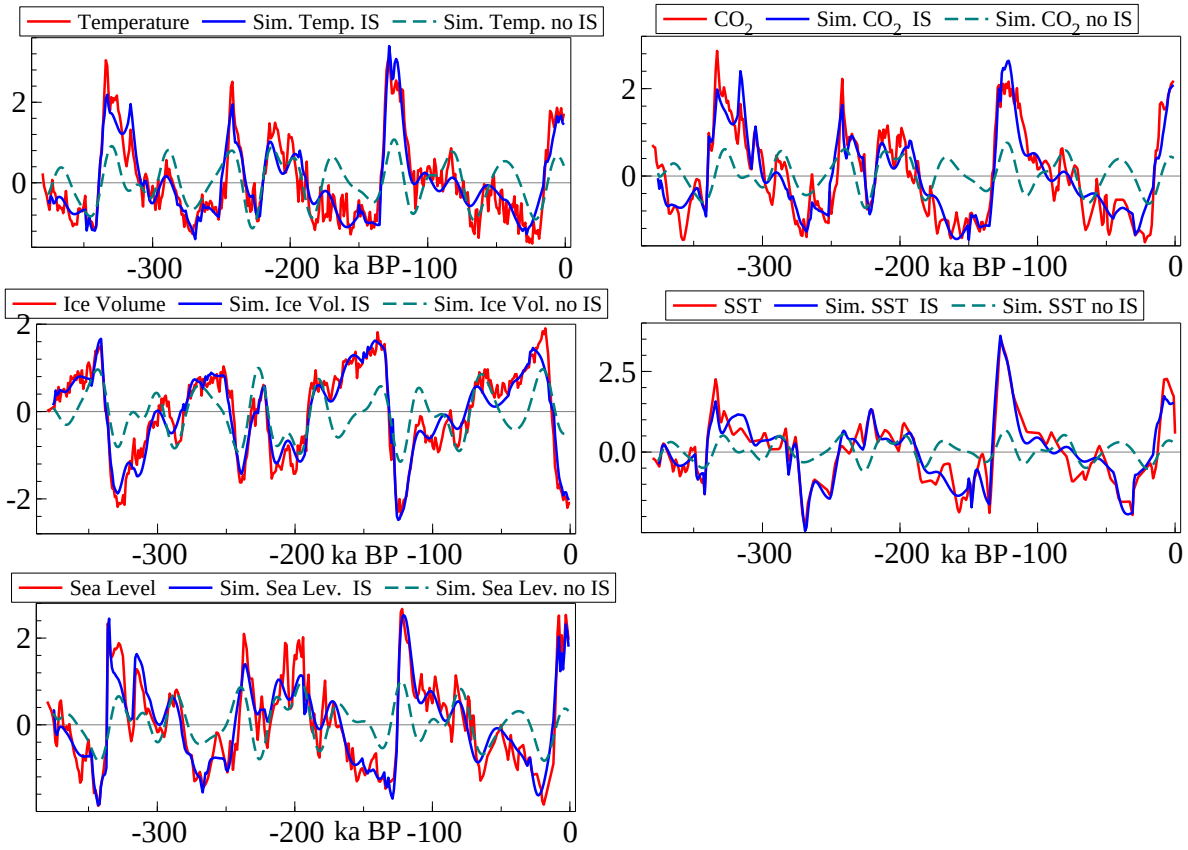


Table 7: R^2 of Observed Record against Simulations with and without Indicators

R^2	With Indicators	Without Indicators
Temperature	0.86	0.38
CO_2	0.85	0.19
Ice Volume	0.92	0.41
SST	0.88	0.12
Sea Level	0.86	0.36

the use of indicators. In relation to their work, two points are worth noting. First, by inspection their simulated cycles without impulses provide a closer fit than the model without indicators presented here. This suggests that some of the poor fit without indicators is driven by the omission of relevant variables. Second, however, in certain time periods the use of indicators leads to better performance than with relevant variables included. In particular, the time period spanning from -180ka until -140ka exhibits a low goodness of fit without the use of indicators. The improved fit that indicators provide during this time period is not directly driven by the indicators "dummying out" the observations. As Table 5 shows, there are few impulses that would allow for a perfect fit during this time period. Instead, the indicators throughout the rest of the model ensure unbiased estimates and correctly estimated cointegrating relations by removing previously un-modelled shocks that induce biases (see Castle & Hendry 2013). This then leads to better simulation accuracy during the remaining time period. This suggests that, without over-fitting, using automatic detection of structural changes through indicator saturation can lead to higher accuracy in empirically simulating glacial & inter-glacial dynamics, and is likely to reduce bias in estimated long run relations.

4 Conclusion

Automatic model selection with extended general to specific modelling, as well as impulse and step-indicator saturation, can provide tools to successfully model complex non-stationary relationships in empirical climate research. Modelling CO_2, we find that, without prior restrictions, natural factors are necessary but not sufficient in explaining CO_2 growth – industrial production components are highly significant and consistently selected in estimated models. In turn, while accounting for un-modelled location shifts, in a long-run system of approximately the past four hundred thousand years, we find stationary relations between atmospheric CO_2 concentrations and the temperature record. Applying automatic detection of structural breaks through the use of indicators leads to improved accuracy in empirically simulating glacial and inter-glacial cycles and estimating cointegrating relations.

References

Archer, D., Eby, M., Brovkin, V., Ridgwell, A., Cao, L., Mikolajewicz, U., et al. (2009). Atmospheric lifetime of fossil fuel carbon dioxide. *Annual Review of Earth and Planetary Sciences*, *37*(1), 117-134.

Australian Bureau of Meteorology. (2011). *Southern Oscillation Index Archive.* Available on-line [http://www.bom.gov.au/climate/current/soihtm1.shtml].

Bontemps, C., & Mizon, G. E. (2008). Encompassing: Concepts and implementation. *Oxford Bulletin of Economics and Statistics*, *70*, 721–750.

Campos, J., Ericsson, N. R., & Hendry, D. F. (2005). Editors' introduction. In J. Campos, N. R. Ericsson, & D. F. Hendry (Eds.), *Readings on general-to-specific modeling* (pp. 1–81). Cheltenham: Edward Elgar.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011a). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3 (1)*, DOI: 10.2202/1941-1928.1097.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011b). Model selection when there are multiple breaks. *Journal of Econometrics*, forthcoming.

Castle, J. L., & Hendry, D. F. (2010). A low-dimensional portmanteau test for non-linearity. *Journal of Econometrics*, *158*, 231-245.

Castle, J. L., & Hendry, D. F. (2013). Model selection in under-specified equations with breaks. *Journal of Econometrics*, –, forthcoming.

Castle, J. L., & Shephard, N. (Eds.). (2009). *The methodology and practice of econometrics.* Oxford: Oxford University Press.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, *28*, 591–605.

Donoho, D. L., Tsaig, Y., Drori, I., & Starck, J. (2006). *Sparse solution of undetermined linear equations by stagewise orthogonal matching pursiut.* Available on-line [http://www.cs.tau.ac.il/ idrori/StOMP.pdf].

Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics*, *70*, 915–925.

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), (pp. 88–121). Oxford: Oxford University Press.

Doornik, J. A. (2010a). *Econometric model selection with more variables than observations* (Working paper). University of Oxford: Economics Department.

Doornik, J. A. (2010b). *Oxmetrics software package.* Timberlake Consultants [http://www.timberlake.co.uk/software/?id=64].

Doornik, J. A., Hendry, D. F., & Pretis, F. (2013). *Step indicator saturation.* (Working Paper)

Ericsson, N. R., & Reisman, E. L. (2012). Evaluating a global vector autoregression for forecasting. *International Advances in Economic Research*, *18*, 247–258.

Federal Reserve. (2011). *Industrial Production and Capacity Utilization.* Available on-line [http://www.federalreserve.gov/releases/g17].

Government of India, Ministry of Statistics. (2011). *Index of Industrial Production (IIP).* Available on-line [http://mospi.nic.in/].

Hendry, D. F. (1995). *Dynamic econometrics.* Oxford: Oxford University Press.

Hendry, D. F., & Johansen, S. (2013). Model discovery and trygve haavelmo's legacy. *Econometric Theory*, *forthcoming*.

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. , *33*, 317-335.

Hendry, D. F., & Juselius, K. (2001). Explaining cointegration analysis: Part II. *Energy Journal*, *22*, 75–120.

Hendry, D. F., & Krolzig, H.-M. (2003). New developments in automatic general-to-specific modelling. In B. P. Stigum (Ed.), *Econometrics and the philosophy of economics* (pp. 379–419). Princeton: Princeton University Press.

Hendry, D. F., & Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, *115*, C32–C61.

Hendry, D. F., & Mizon, G. E. (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, *3 (1)*, DOI: 10.2202/1941-1928.1100.

Hendry, D. F., & Pretis, F. (2013a). Anthropogenic influences on atmospheric co_2. In R. Fouquet (Ed.), (chap. 12). Cheltenham: Edward Elgar Ltd.

Hendry, D. F., & Pretis, F. (2013b). Some fallacies in econometric modelling of climate change. *Earth System Dynamics Disc.*, *4*, C112–C117.

Hendry, D. F., & Richard, J.-F. (1989). Recent developments in the theory of encompassing. In B. Cornet & H. Tulkens (Eds.), *Contributions to operations research and economics. the xxth anniversary of core* (pp. 393–440). Cambridge, MA: MIT Press.

Jansen, J., E., Overpeck, K., Briffa, J.-C., Duplessy, F., Joos, V., Masson-Delmotte, D., et al. (2007). *Palaeoclimate. climate change 2007: The physical science basis* (Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change). Cambridge University Press, Cambridge.

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. Oxford: Oxford University Press.

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In J. L. Castle & N. Shephard (Eds.), (pp. 1–36). Oxford: Oxford University Press.

Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., et al. (2007). Orbital and millennial antarctic climate variability over the past 800,000 years. *Science*, *317*, 793–797.

Juselius, K. (2006). *The cointegrated var model*. Oxford: Oxford University Press.

Kaufmann, R. K., & Juselius, K. (2013). Testing hypotheses about glacial cycles against the observational record. *Paleoceanography*, *28*, 175–184.

Kaufmann, R. K., Kauppi, H., Mann, M. L., & Stock, J. H. (2013). Does temperature contain a stochastic trend: linking statistical results to physical mechanisms. *Climatic Change*, 1-15.

Lea, D. W. (2004). The 100,000-yr cycle in tropical sst, greenhouse forcing and climate sensitivity. *Journal of Climate*, *17*, 2170–2179.

Lisiecki, L. E., & Raymo, M. E. (2005). A pliocine-pleistocene stack of 57 globally distributed benthic $\delta^1 8$ records. *Paleoceanography*, *20*, doi:10.1029/2004PA001071.

Luethi, D., Le Floch, M., Bereiter, B., Blunier, T., Barnola, J. M., Siegenthaler, U., et

al. (2008). High-resolution carbon dioxide concentration record 650,00-800,000 years before present. *Nature*, *453*, 379–382.

Marland, G., Boden, T. A., & Andres, R. (2011). *Global, regional, and national fossil fuel CO2 emissions.* In Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., U.S.A. (Available on-line [http://cdiac.ornl.gov/trends/emis/overview.html])

Martinez-Garcia, A., Rosell-Mele, A., Geibert, W., Gersonde, R., Masque, P., Gaspari, V., et al. (2009). Links between iron supply, marine productivity, sea surface temperature, and co2 over the last 1.1 ma. *Paleoceanography*, *24*, doi:10.1029/2008PA001657.

Mizon, G. E., & Richard, J. F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, *54*, 657–678.

NASA Goddard Institute for Space Studies (GISS). (2011). *GISS - Surface Temperature Analysis.* Available on-line [http://data.giss.nasa.gov/gistemp/].

OECD. (2011). *Composite Leading Indicators: MEI.* Available on-line [http://stats.oecd.org/Index.aspx].

Office of National Statistics. (2011). *Primary Production.* Available on-line [http://www.statistics.gov.uk/hub/business-energy/].

Paillard, D., L., L., & Yiou, P. (1996). Macintosh program performs time-series analysis. *EOS*, *77:39*, 379.

Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics*, *4*, 393–397.

Scheffer, M., Brovkin, V., & Cox, P. M. (2006). Positive feedback between global warming and atmospheric co_2 concentrations inferred from past climate change. *Geophysical Research Letters*, *33*, doi:10.1029/2005GL025044.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Siddall, M., Rohling, E. J., Almogi-Labin, A., Hemleben, C., Meischner, D., Schmelzer, I., et al. (2003). Sea-level fluctuations during the last glacial cycle. *Nature*, *423*, 853–858.

Siroky, S. D. (2009). Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys*, *3*, 147-163.

Stodden, V. (2006). *Model selection when the number of variables exceeds the number of observations.* Thesis. Available on-line [http://www.stanford.edu/ vcs/Papers.html].

Tans, P., & Keeling, R. (2013). *Mauna Loa, monthly mean carbon dioxide.* Scripps Institution of Oceanography. (scrippsco2.ucsd.edu/) and NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) Available on-line [http://www.esrl.noaa.gov/gmd/ccgg/trends/].

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58:1*, 267-288.

Tucker, C. J., Pinzon, J., Brown, M., & GIMMS/GSFC/NASA. (2010). *ISLSCP II GIMMS Monthly NDVI, 1981-2002.* In Hall, Forrest G., G. Collatz, B. Meeson, S. Los, E. Brown de Colstoun, and D. Landis (eds.). ISLSCP Initiative II Collection. Data set. Available on-line [http://daac.ornl.gov/] from Oak Ridge National Labora-

tory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.

Watson, A. J., Nightingale, P. D., & Cooper, D. J. (1995). Modeling atmosphere ocean co_2 transfer. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, *353*, 41–51.