Modeling, Forecasting, and Nowcasting U.S. CO_2 Emissions Using Many Macroeconomic Predictors*

Mikkel Bennedsen[†], Eric Hillebrand[‡], Siem Jan Koopman[§]

July 9, 2020

Abstract

We propose a structural augmented dynamic factor model for U.S. CO_2 emissions. Variable selection techniques applied to a large set of annual macroeconomic time series indicate that CO_2 emissions are best explained by industrial production indices covering manufacturing and residential utilities. We employ a dynamic factor structure to explain, forecast, and nowcast the industrial production indices and thus, by way of the structural equation, emissions. We show that our model has good in-sample properties and out-of-sample performance in comparison with univariate and multivariate competitor models. Based on data through September 2019, our model nowcasts a reduction of about 2.6% in U.S. per capita CO_2 emissions in 2019 compared to 2018 as the result of a reduction in industrial production in residential utilities.

Keywords: CO_2 emissions; macroeconomic variables; dynamic factor model; variable selection; forecasting; nowcasting.

JEL Codes: C01; C13; C32; C51; C52; C53; C55; C82; Q43; Q47.

*The authors would like to thank participants at the fourth Econometric Models of Climate Change Conference (EMCC-IV, 2019) for useful comments on an earlier version of this paper. MB and EH acknowledge financial support from the Independent Research Fund Denmark for the project "Econometric Modeling of Climate Change."

[†]Department of Economics and Business Economics and CREATES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: mbennedsen@econ.au.dk

[‡]Department of Economics and Business Economics and CREATES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: ehillebrand@econ.au.dk

[§]Department of Econometrics, School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, and CREATES. E-mail: s.j.koopman@vu.nl

1 Introduction

In this paper, we propose a structural augmented dynamic factor model for United States (U.S.) CO_2 emissions, where emissions are explained contemporaneously by industrial production (IP) variables, which, in turn, are modeled by macroeconomic factors obtained from a large data set. The literature on modeling the relation between macroeconomic activity and CO_2 emissions discusses a range of effects: the scale effect, by which increased IP increases emissions, changes in input and output mixes, changes in production efficiency, and changes in energy intensity (Stern, 2017, p.10). A large body of literature discusses a possible tipping point effect under the label of an environmental Kuznets curve, or EKC (e.g. Grossman and Krueger, 1991; Arrow et al., 1995; Schmalensee et al., 1998; Millimet et al., 2003; Brock and Taylor, 2005; Wagner, 2008, 2015).

The pertinent models that are commonly considered in climate science, integrated assessment models (IAMs), focus in their macroeconomic modules on highly aggregated measures of economic activity, such as global gross domestic product (GDP) and global population (Blanco et al., 2014; Bosetti et al., 2006; Calvin et al., 2019; Fujimori et al., 2017; Gambhir et al., 2019; Luderer et al., 2015; Messner and Strubegger, 1995; Nordhaus and Sztorc, 2013; Stehfest et al., 2014). The aggregated economic output then implies, by way of an energy intensity of GDP and a greenhouse gas (GHG) intensity of energy, an evolution of emissions (Blanco et al., 2014; Raupach et al., 2007). The close connection between economic activity, broadly measured by growth in GDP, and emissions has been extensively studied in the energy economics literature (e.g. Stern, 1993, 2000; Oh and Lee, 2004; Lee, 2005; Zhang and Cheng, 2009; Ozturk, 2010). At this point in time, IAMs are predominantly non-statistic, and parameters are set or calibrated to produce plausible output. In contrast to this, econometric theory for EKC estimation is well developed (Wagner, 2015).

Given this state of the literature, manifold causal channels with different magnitudes and signs of effects, and increasing availability of macroeconomic and emissions data, there is a need for a statistical model class that can capture the cumulative effects of economic activity on CO_2 emissions and that can be used to solve very practical problems, such as forecasting and nowcasting emissions. These problems arise, for example, in the annual update of the Global Carbon Budget (Le Quéré et al., 2018b).

To arrive at such a statistical model class, we start out by identifying a set of contemporaneous explanatory variables for U.S. CO_2 emissions from a large data set of 226 annual macroeconomic time series, using a range of variable selection techniques. The emissions data set is constructed from energy statistics which are compiled by the United Nations using the procedure as set out in Marland and Rotty (1984). We expect that emissions data are closely connected to macroeconomic data. Indeed, we demonstrate that two variables in particular, the IP index and the IP: Residential Utilities index, explain CO_2 emissions to an extent that further information contained in the macroeconomic data set has only negligible explanatory power for emissions. However, the macroeconomic data set does have explanatory power for the IP indices. We draw on the extensive literature on macroeconomic forecasting and specify a dynamic factor model (DFM) in order to exploit the forecast power of the macroeconomic data set for the IP indices (Elliott et al., 2006; Elliott and Timmermann, 2013, 2016; Hillebrand and Koopman, 2016).

The resulting model is labeled a structural augmented dynamic factor model (SADFM). The model is structural because it contains a contemporaneous relation between CO_2 emissions and IP index time series, augmented because individual IP indices and factors that summarize large amounts of data are modeled jointly, and dynamic factor model because a dynamic state equation for the factors harnesses the forecasting power of many macroeconomic predictors for the IP indices. We demonstrate that our proposed SADFM describes the data well and has good in-sample properties. We present a pseudo-out-of-sample forecast exercise and show that the model performs better than a set of univariate and multivariate competitors, such as ARMA, vector autoregressions (VAR), structural VAR, principal components regressions, and standard DFMs. Finally, we show how the model can be used in nowcasting problems such as those faced in the Global Carbon Budget.

The remainder of the paper is organized as follows: Section 2 describes the data used in this study. Section 3 presents the SADFM. Section 4 discusses the selection of the contemporaneous explanatory variables for U.S. CO_2 emissions. Section 5 discusses the estimation of the model, and the in-sample fit. Section 6 considers forecasting and nowcasting. Section 7 concludes.

2 Data

2.1 U.S. CO_2 Emissions Data

Let $CO2_t$ denote annual U.S. CO_2 emissions data at year t, i.e. the sum of annual emissions from fossil fuels and cement production in the U.S., obtained from The Global Carbon Project (Le Quéré et al., 2018b), which in turn is compiled from Boden et al. (2018) and UNFCCC (2018). These data sets are constructed as a result from the replies of the U.S. to the Annual Questionnaire on Energy

Statistics[1] conducted by the U.N. Statistics Division and released in the annual Energy Statistics Yearbook[2]. Hence, the raw data consists of energy statistics, which are converted into estimates of CO_2 emissions using the approach developed in Marland and Rotty (1984). This approach tallies energy use from different sources and multiplies an appropriate physical conversion constant on each source to convert it from energy use to CO_2 emissions (we refer to Marland and Rotty, 1984, for the exact details). The resulting time series represents data on U.S. CO_2 emissions from fossil fuels and from cement production; they do not include CO_2 equivalents from other greenhouse gas emissions. Since the emissions data are constructed from energy statistics they have no direct relation to national accounting, GDP, Industrial Production, and other macroeconomic variables considered in our data set; see Section 2.2.

The variable $E_t = CO2_t/pop_t$, where pop_t denotes total U.S. population in year t, is defined as CO_2 emissions per capita in the U.S., in tons of carbon emitted per capita, and $e_t = \log E_t$ is referred to as log per-capita emissions. Population data, pop_t, are obtained from the World Bank.[3] Finally, Δe_t is taken as a good approximation of the growth rate of CO_2 emissions per capita.

The variables E_t and e_t are observed at an annual frequency, from 1960 to 2017. Figure 1 presents these data series as well as their first differences, ΔE_t and Δe_t, respectively. The results of tests for stationarity and for unit roots in these time series are given in Table 1; we conclude that there is evidence for a unit root in both E_t and e_t and that their differences appear to be stationary. Table 2 contains further descriptive statistics of the time series.

2.2 U.S. Macroeconomic Data

We have collected economic data representative of the U.S. macroeconomy, with a particular emphasis on variables that can plausibly influence the amount of CO_2 emitted to the atmosphere. We consider $N = 226$ variables in our study, of which 126 are macroeconomic variables from the "FRED-MD" database, compiled and maintained by the Federal Reserve Bank of St. Louis.[4] The individual data series are indexed by $i = 1, 2, \ldots, 126$, which can be found in the Online Appendix of McCracken and Ng (2016), see footnote 4. The remaining 100 time series variables are related

[1] See https://unstats.un.org/unsd/energystats/questionnaire/documents/Energy-Questionnaire-Guidelines.pdf for the guidelines to the Annual Questionnaire.

[2] https://unstats.un.org/unsd/energystats/pubs/yearbook/.

[3] https://data.worldbank.org, downloaded on May 7, 2019.

[4] https://research.stlouisfed.org/econ/mccracken/fred-databases/, downloaded on May 7, 2019, see McCracken and Ng (2016).

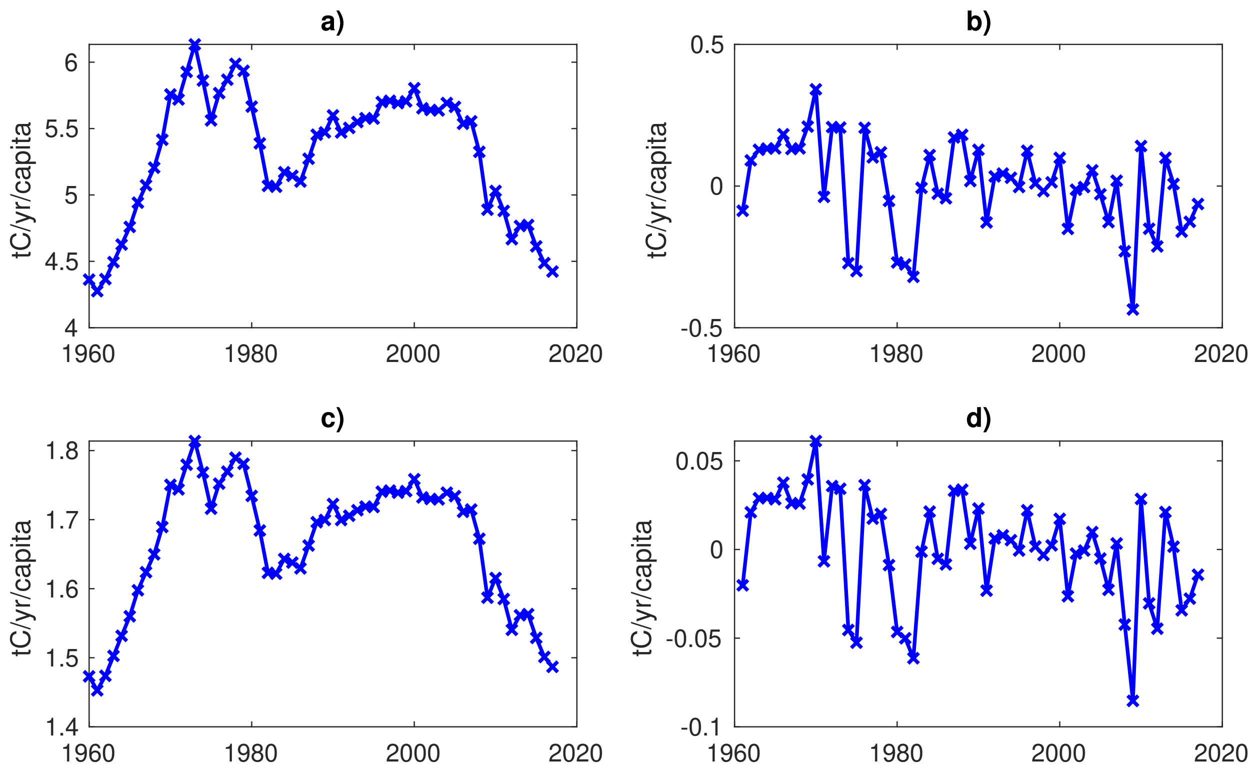

Figure 1: *Per capita U.S. CO_2 emissions from $t = 1960$ to $t = 2017$. a): Levels, E_t. b): Differences, ΔE_t. c): Log-levels, e_t. d): Log-differences (growth rates), Δe_t.*

Table 1: *Tests for stationarity and unit roots: p-values for the KPSS test of the null of stationarity (left) and for the Augmented Dickey-Fuller (ADF) test for the null of a unit root (right). The number of autoregressive lags to include in the linear regressions is a nuisance parameter; we consider 0 to 4 lags as indicated in the top of the table. The alternative of the ADF test is stationarity; results for a trend-stationary alternative are similar and not presented here.*

	KPSS					ADF				
Number of lags:	0	1	2	3	4	0	1	2	3	4
E_t = per capita emissions										
E_t :	0.01	0.01	0.01	0.01	0.02	0.62	0.62	0.60	0.53	0.51
ΔE_t :	0.10	0.10	0.10	0.10	0.10	0.00	0.00	0.01	0.00	0.01
e_t = log-per capita emissions										
e_t :	0.01	0.01	0.01	0.01	0.02	0.62	0.60	0.65	0.72	0.76
Δe_t :	0.10	0.10	0.10	0.10	0.10	0.00	0.00	0.01	0.01	0.01

Table 2: *Descriptive statistics and diagnostics of per capita emissions data. N is the Jarque-Bera test statistic (Jarque and Bera, 1987): the null hypothesis of Gaussianity is rejected if N is larger than the* 95% *critical value of* 5.99*. DW is the Durbin-Watson test statistic (Durbin and Watson, 1971): $DW < 2$ indicates positive serial correlation, $DW > 2$ negative serial correlation; $DW = 2$ indicates no serial correlation. LB is the Ljung-Box Q test statistic (Ljung and Box, 1978) for autocorrelation with a* 95% *critical value of* 31.41*.*

	Descriptive statistics					Diagnostics		
	NumObs	Mean	Std.	Skew	Kurt	N	DW	LB
E_t = per capita emissions								
E_t :	58	5.30	0.48	−0.56	2.21	4.52	0.11	188
ΔE_t :	57	0.00	0.16	−0.57	2.93	3.10	1.38	25
e_t = log-per capita emissions								
e_t :	58	1.67	0.09	−0.68	2.36	5.49	0.10	181
Δe_t :	57	0.00	0.03	−0.60	2.94	3.43	1.36	27

to different aspects of the U.S. macroeconomy: 18 variables represent U.S. agricultural production; 76 variables represent U.S. foreign trade[5]; 2 time series represent U.S. cement production; and 4 represent U.S. transport[6]. These data series are indexed by $i = 127, 128, \ldots, 226$ as shown in the tables given in Appendix A. We have followed the procedure described in McCracken and Ng (2016) to transform the non-stationary time series to a set of stationary time series and to studentize all resulting time series to mean zero and unit variance. Further details of the data set, including how we collected and transformed each individual data series, are given in Appendix A.

The emissions time series $CO2_t$ is constructed from energy statistics which are not part of our macroeconomic data set. However, it is expected that many macroeconomic variables are closely linked to energy use and therefore to the energy statistics. It is the key aim of this paper to study these links and to utilize econometric model-based analyses, which have previously been applied successfully to macroeconomic time series, for the nowcasting and forecasting of CO_2 emissions data.

The sample period of the annual economic data set is from 1960 until 2018. Given that, at the

[5]Collected from the World Bank website, `https://data.worldbank.org`, downloaded on September 19, 2019.

[6]Collected from the website of the U.S. Federal Reserve Bank, `https://fred.stlouisfed.org`, downloaded on October 10, 2019.

time of writing, the official emissions variable $CO2_t$ is only available until 2017[7], we only consider economic data up to 2017. Many of the time series start in 1961 after first differencing, while some variables start even later. We are therefore faced with an *unbalanced data set* of economic variables. Hence the economic data matrix X, with $T = 57$ rows and $N = 226$ columns, contains a number of missing entries. We impute these missing values using a factor model approach, following Stock and Watson (2002) and McCracken and Ng (2016). As a result, in our empirical study we work with a $T \times N$ balanced data set of economic data X. The details of the imputation method are given in Appendix B.

3 Statistical Model for Growth in CO_2 Emissions

The variable of interest $y_t = e_t$ in our study is the log-difference (i.e., growth rate) of U.S. CO_2 per-capita emissions. Our analysis is based on the statistical dynamic model

$$y_t = \alpha + \sum_{j=1}^{k} \beta_j x_t^{(i_j)} + \beta_f' f_t + \gamma' z_t + \epsilon_t^y, \tag{3.1}$$

where $x_t^{(i_j)}$, for $j = 1, \ldots, k$, are pre-selected individual economic variables, $i_j \in \{1, 2, \ldots, N\}$ indicates the column number in the data matrix X, β_j is the regression coefficient for the i_jth variable, f_t is an $r \times 1$ vector of economic factors or principal components, β_f is the $r \times 1$ corresponding loading coefficient vector, z_t is an $l \times 1$ vector of exogenous variables not contained in X, such as dummy variables, γ is the corresponding $l \times 1$ vector of regression coefficients, and ϵ_t^y is the disturbance term. For any disturbance term in our modelling framework, we assume it is an independent and identically distributed (IID) random sequence with mean zero and a positive variance. All $2 + k + r + l$ unknown coefficients are collected in the parameter vector $\theta_y = (\alpha, \beta_1, \ldots, \beta_k, \beta_f', \gamma', \sigma_y)'$ with $\sigma_y^2 = \mathrm{Var}(\epsilon_t^y)$ for all t. The statistical model (3.1) can be regarded as a standard regression model with the addition of latent dynamic factors in f_t. A similar modeling framework is considered in many macroeconomic studies, for example, in Stock and Watson (2002), Giannone et al. (2008), and Stock and Watson (2010).

As outlined in Section 2, the CO_2 emissions variable e_t is compiled via energy statistics. To avoid having energy variables on both sides of our model equation (3.1), we do not include energy variables in the economic data vector x_t. Hence our model directly links CO_2 emissions to the macroeconomy, which is a perspective that is taken earlier in the energy economics literature; see,

[7] UNFCCC inventories are available at https://di.unfccc.int/time_series, accessed on November 18, 2019.

for example, Stern (1993). The Global Carbon Project[8], in preparing the annual Global Carbon Budget, has projected world-wide CO_2 emissions, until the 2017 edition, using estimates of World GDP (Le Quéré et al., 2018a, pp. 405–448, in particular p. 430). Our model-based approach is similar in its purpose but it offers a higher resolution of the macroeconomic sphere. Instead of focusing on GDP only, we consider a wide range of potentially relevant macroeconomic variables.

3.1 Structural augmented dynamic factor model

The main motivation for model (3.1) is that we have a large number of macroeconomic time series available for our analysis, in our case $N = 226$. The formulation of a statistical model with all individual time series present in X is not feasible, given the relatively small number of observations, in our case $T = 57$. Our model faces this challenge in two ways. First, we make a pre-selection of relevant variables for y_t and only include the $k << N$ most important variables $x_t^* := (x_t^{(i_1)}, \ldots, x_t^{(i_k)})'$ as regressors in (3.1). The selection of these k "most important" variables can be done in different ways, depending on selection strategies and criteria; see the discussion in Section 4. Second, we include r latent dynamic *common factors* f_t, with $r << N$, that are extracted from all economic variables in X. For this purpose, we consider the $N \times 1$ vector of economic time series x_t, being the t-th row of X, and specify the dynamic factor model as

$$x_t = \Lambda f_t + \epsilon_t^x, \tag{3.2}$$

where Λ is an $N \times r$ matrix of *factor loadings*, and ϵ_t^x is an idiosyncratic disturbance term with $\mathrm{Var}(\epsilon_t^x) = \Sigma_x$, an $N \times N$ matrix. We assume that the mild conditions of Stock and Watson (2002) apply to our setting. The dynamic properties of the common factors f_t are given by a vector autoregressive process of order one, or VAR(1), as

$$f_t = \Phi f_{t-1} + \eta_t, \tag{3.3}$$

where Φ is an $r \times r$ autoregressive coefficient matrix, and η_t is the disturbance vector with $r \times r$ variance matrix $\mathrm{Var}(\eta_t) = \Sigma_\eta$. We assume that the roots of the characteristic polynomial $|I_r - \Phi z|$ lie outside the unit circle, where I_r denotes the $r \times r$ identity matrix. This condition ensures that the VAR(1) model for f_t is stable.[9] We further restrict the variance matrix Σ_η such that the factors

[8] https://www.globalcarbonproject.org, see also Section 6.2 for a discussion of emissions projections made by the Global Carbon Project.

[9] If a stable VAR process is initialized according to its stationary distribution, then the resulting process is stationary and ergodic.

are orthogonal and normalized to have unit variance: $\text{Var}(f_t) = I_r$.

In case $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ in equation (3.1), the model for y_t reduces to the "standard" dynamic factor model (DFM) as studied in Doz et al. (2012), Jungbacker and Koopman (2015), Stock et al. (2016), among others. Stock and Watson (2002) have suggested to improve forecast performance of the DFM by including lags of selected variables, $x_{t-1}^{(i_j)}$, in the forecast equation for y_t. In this spirit, we include *contemporaneous* individual variables, the selection $x_t^{(i_j)}$, for $j = 1, \ldots, k$, in the equation for y_t, and we consider $\beta_j \neq 0$, for $j = 1, \ldots, k$, in (3.1). This alternative specification provides a more structural interpretation of the model and hence we refer to the resulting model as a *structural augmented dynamic factor model* (SADFM).[10]

3.2 State space model representation

The SADFM model (3.1)–(3.3) can be written jointly in matrix form by

$$B \begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} \beta_f' \\ \Lambda \end{pmatrix} f_t + \begin{pmatrix} \gamma' \\ 0 \end{pmatrix} z_t + \begin{pmatrix} \epsilon_t^y \\ \epsilon_t^x \end{pmatrix}, \qquad f_{t+1} = \Phi f_t + \eta_{t+1},$$

where B is the $(N+1) \times (N+1)$ selection matrix

$$B = \begin{pmatrix} 1 & -\beta^{s\prime} \\ 0 & I_N \end{pmatrix},$$

where β^s is the $N \times 1$ vector of zeros except for the entries $i_1, \ldots, i_k$ which are equal to β_j, respectively, for $j = 1, \ldots, k$. Given that the inverse of B is

$$B^{-1} = \begin{pmatrix} 1 & \beta^{s\prime} \\ 0 & I_N \end{pmatrix},$$

it follows that the model (3.1)–(3.3) is equivalent to

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} \beta_f' + \beta^{s\prime}\Lambda \\ \Lambda \end{pmatrix} f_t + \begin{pmatrix} \gamma' \\ 0 \end{pmatrix} z_t + u_t, \qquad f_{t+1} = \Phi f_t + \eta_{t+1}, \tag{3.4}$$

with

$$u_t = \begin{pmatrix} \epsilon_t^y + \beta^{s\prime}\epsilon_t^x \\ \epsilon_t^x \end{pmatrix}, \qquad \text{Var}(u_t) = \begin{pmatrix} \sigma_y^2 + \beta^{s\prime}\Sigma_x\beta^s & \beta^{s\prime}\Sigma_x \\ \Sigma_x\beta^s & \Sigma_x \end{pmatrix}.$$

[10] This model is related to the structural dynamic factor model (SDFM) as studied by Stock et al. (2016), among others. The lagged dependence of $x_t^{(i_j)}$, for $j = 1, \ldots, k$, in the SADFM is implicitly specified through the dynamics of the factors in f_t.

The $(N+1) \times 1$ vector $(y_t, x_t')'$ is the observed data vector for the complete model, the $r \times 1$ vector f_t is a latent dynamic factor, the $l \times 1$ vector z_t is exogenous (fixed covariates), and the $(N+1) \times 1$ vector u_t is the disturbance. The system of equations (3.4) represents a linear state space model, with the initial moment conditions $\mathrm{E}(f_1) = 0$ and $\mathrm{Var}(f_1) = I_r$. See Durbin and Koopman (2012) for a complete treatment of this class of models. The model is subject to a number of choices, including the composition of the economic data matrix X, the selection of indices $i_1, \ldots, i_k$, and the number of factors r. In the empirical part of our study we discuss these choices further and show how they have been made for our data set.

3.3 Parameter estimation

The unknown parameters in (3.4) are the collection of θ_y, Λ, Σ_x, and Φ.[11] All parameters can be estimated consistently (as $T, N \to \infty$) in a two-step iterative procedure as proposed in Doz et al. (2011, 2012). The procedure requires computation of principal components from the data matrix X as in Stock and Watson (2002), where the principal components (or diffusion indices) are used as regressors for a variable of interest, for the purpose of forecasting. In our analysis, we consider the r principal components from X that are associated with the r largest eigenvalues of the sample variance matrix of X. The principal components are collected in the $r \times 1$ vector F_t and are regarded as proxies of f_t, for $t = 1, \ldots, T$.

First, we carry out two regressions: one to estimate matrix Φ by regressing the principal components F_{t+1} on F_t (equation by equation), the other to estimate matrix Λ and diagonal matrix Σ_x by regressing x_t on F_t. We obtain the estimate Σ_η using its definition $\Sigma_\eta = I_r - \Phi\Phi'$ and replacing Φ by its estimate. Next, we take these estimates to replace their unknown counterparts in (3.4) while the remaining parameters in vector θ_y are estimated by the method of maximum likelihood, for which the Kalman filter is used to evaluate the loglikelihood function based on the prediction error decomposition; see Jungbacker and Koopman (2015) for further details in the context of the DFM. After this first stage of our estimation process, we consider (3.4), with the parameter estimates, to signal extract the dynamic factors f_t, for $t = 1, \ldots, T$, using the Kalman filter and the associated smoother (KFS) method; see Durbin and Koopman (2012, Chapter 4). We denote these smoothed estimates by $\widetilde{f}_t$, for $t = 1, \ldots, T$.

The smoothed factor estimates $\widetilde{f}_t$ will be different from the principal components F_t, for $t =$

[11] For computational convenience, we restrict Σ_x to a diagonal matrix. The variance matrix $\Sigma_\eta = I_r - \Phi\Phi'$ is restricted and not explicitly estimated.

$1, \ldots, T$. Based on these new factor estimates, the procedure described above can be repeated by replacing F_t by $\widetilde{f}_t$. The two regressions can be repeated and, based on the resulting new estimates for Φ, Λ and Σ_x, the maximum likelihood estimation for θ_y can also be repeated. In turn, based on these new parameter estimates, new smoothed estimates can be obtained for f_t, for $t = 1, \ldots, T$, using KFS. This iterative procedure is close to what is described in Doz et al. (2012, p. 1018). Let $\mathcal{L}_m$ be the likelihood function value for the parameter estimates in iteratiom m. We define the statistic

$$c_m = \frac{\mathcal{L}_m - \mathcal{L}_{m-1}}{(\mathcal{L}_m + \mathcal{L}_{m-1})/2},$$

and compute it at the end of each iteration. This iterative procedure can stop at iteration M, where M is the first integer such that $c_M < 10^{-4}$. In our empirical study it turns out that the number of iterations M is as low as 4 or 5.

3.4 Collapsed structural augmented dynamic factor model

From an econometric viewpoint, we can base our empirical analysis on the model formulation (3.4). However, in case the model needs to cope with a high-dimensional $N \times 1$ data vector x_t, or with many consecutive estimations in a forecast exercise, various computational steps in the estimation procedure become somewhat cumbersome. In our empirical study below, we consider the analysis on $N = 226$ economic variables: it requires many regression computations in each iteration. To alleviate the computational burden somewhat, we consider a dimension reduction by employing the *collapsed dynamic factor model*, as proposed by Bräuning and Koopman (2014).

In this collapsed approach, we consider the principal components in F_t as data proxies of the corresponding latent dynamic factors in f_t, for $t = 1, \ldots, T$. In particular, we assume that

$$F_t = f_t + \epsilon_t^f,$$

where ϵ_t^f is a disturbance vector with $\mathrm{Var}(\epsilon_t^f) = \Sigma_f$. We further treat the principal component vector F_t as a data vector that can replace the original large data vector x_t, since the principal components provide a sufficiently accurate description of the data matrix X. This specification leads to a much lower-dimensional model with a much smaller number of parameters. This eases computation and may potentially provide more accurate forecasts, given that the model is more parsimonious; see Bräuning and Koopman (2014) and Hindrayanto et al. (2016).

The collapsed SADFM is given by

$$\begin{pmatrix} y_t \\ x_t^* \\ F_t \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \beta_f' + \beta^{*\prime}\Lambda^* \\ \Lambda^* \\ I_r \end{pmatrix} f_t + \begin{pmatrix} \gamma' \\ 0 \\ 0 \end{pmatrix} z_t + u_t^*, \qquad f_{t+1} = \Phi f_t + \eta_{t+1}, \tag{3.5}$$

where $x_t^* = (x_t^{(i_1)}, \ldots, x_t^{(i_k)})'$ is a $k \times 1$ sub-vector of x_t, β^* is the corresponding regression coefficient vector for x_t^*, Λ^* is the loading matrix only with the rows of Λ corresponding to $i_1, \ldots, i_k$, and u_t^* is the same as the disturbance vector u_t with ϵ_t^x replaced by $(\epsilon_t^{x*\prime}, \epsilon_t^{f'})'$, for $t = 1, \ldots, T$, where ϵ_t^{x*} is the $k \times 1$ sub-vector of ϵ_t^x corresponding to $i_1, \ldots, i_k$. While the SADFM has an observation vector of dimension $N + 1$, the collapsed SADFM has a reduced observation vector of dimension $k + r + 1$ where both $k << N$ and $r << N$.

We employ the same iterative procedure for the estimation of the parameters in the collapsed SADFM as described above for the non-collapsed model. First, we carry out two regressions: one to estimate matrix Φ by regressing the principal components F_{t+1} on F_t (equation by equation); the other to estimate matrix Λ^* by regressing x_t^* on F_t. The remaining parameters in vector θ_y are estimated by the method of maximum likelihood. After this first stage, we signal extract the dynamic factors f_t, for $t = 1, \ldots, T$, using KFS, and we denote these by $\widetilde{f}_t$, for $t = 1, \ldots, T$. The iterative process can start by replacing F_t with $\widetilde{f}_t$, for $t = 1, \ldots, T$. The iterations can be stopped on the basis of the same convergence criterion c_m as suggested above.

4 Selection of macroeconomic variables for CO_2 emissions

To select economic variables to be included in the sub-vector x_t^*, we carry out a preliminary statistical analysis of per-capita CO_2 emission growth y_t and the economic data matrix consisting of the $N \times 1$ data vectors x_t, with $N = 226$. The description of the data matrix is provided in Section 2.

First, we run N static separate regressions for y_t on each variable in the data vector x_t in simple univariate regressions with intercept. The resulting N coefficients of determination (R^2) are displayed in Figure 2. It shows that many individual economic variables in x_t can provide a considerable goodness-of-fit statistic for the growth in CO_2 emissions. The variables producing a particularly high R^2 are macroeconomic variables related to the real economy, such as production indices ($i = 1, 2, \ldots, 20$) and employment ($i = 21, \ldots, 49$). Some of the trade-related variables ($i = 145, \ldots, 220$) also display a high R^2, as do the cement-production-related variables ($i = 221, 222$).

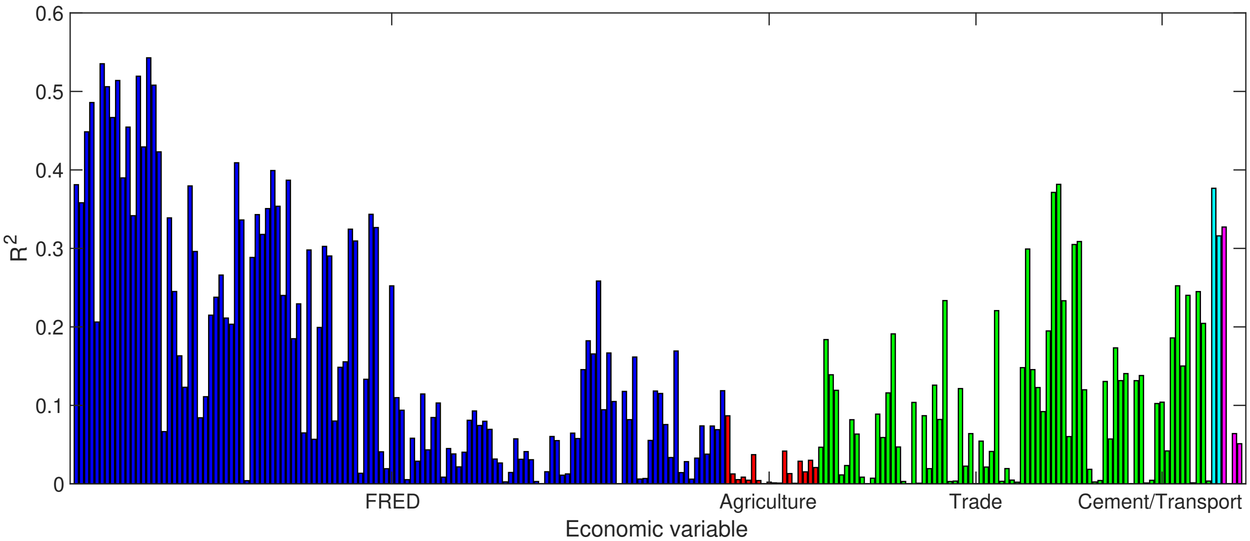

Figure 2: *Coefficient of determination (R^2) obtained from simple regressions of CO_2 growth on an intercept and a single explanatory economic variable (one at a time). The economics variables are grouped (and coloured) into those from the FRED-MD database (blue), agricultural data series (red), trade-related data series (green), cement-related data series (cyan), and transport-related data series (magenta).*

The results so far only show how a *single* economic variable explains growth in CO_2 emissions. We are interested in selecting *multiple* economic variables that provide the best possible fit for CO_2 emissions growth. For this purpose, we run multiple regressions (with intercept) based on a subset of economic variables indexed by a subset of indices $I \subseteq \{1, 2, \ldots, N\}$, with $|I|$ being the cardinality of the set I. For example, in case of $I = \{1, 5, 10\}$, the multiple regression contains y_t (CO_2 growth), the intercept, and the vector of regressors $(x_t^{(1)}, x_t^{(5)}, x_t^{(10)})'$. Depending on the size of the permissible set, there is a vast range of possible sets I with all possible combinations. There are various strategies for selection. Well-known methods are the LASSO of Tibshirani (1996) and the Elastic Net (EN) of Zou and Hastie (2005). For our data set with many series that are highly correlated, we find that both LASSO and EN include implausibly many variables in their preferred sets I.[12] Another regression selection method is advocated by Doornik and Hendry (2015) and is referred to as "AutoMetrics"; it is an automated "general-to-specific" (GETS) method that chooses the best set I on the basis of multiple criteria including goodness-of-fit, model diagnostics, and exogeneity tests. We have adopted the AutoMetrics implementation in the PcGive module of the

[12] Using a 25-fold cross validation to minimize the mean squared error of the LASSO/Elastic Net regression, they both retain the variables $I = \{3, 6, 11, 13, 15, 17, 23, 38, 46, 55, 102, 159, 185, 213, 222\}$.

OxMetrics software; see Doornik (2009); Doornik and Hendry (2018). The documentation provides the details of the automated selection method. AutoMetrics has the option to specify a prior belief on the size of I. We report results based on sizes that are indicated as "Tiny" and "Small". The AutoMetrics results are presented in the top panel of Table 3. The "Tiny" setting provides $I = \{6, 17, 48, 51, 56\}$, while the "Small" setting yields $I = \{6, 17, 48, 51, 55, 89\}$. Furthermore, the procedure detects two outliers in $t = 1970$ and $t = 1990$.[13]

Although the selected set I obtained from the AutoMetrics procedure is smaller than those from the LASSO and EN procedures, it is still large for our purposes. Therefore, we consider an alternative route. Given a pre-set maximum size $s \geq 1$ for I, and subject to $|I| \leq s$, we determine which variables need to be selected to provide the best goodness-of-fit. Table 3 presents these results for a range of s values and for a variety of selection procedures. At this stage, all regressions include the intercept and two outlier dummy variables, for the time indices $t = 1970, 1990$. When s is small, it is computationally feasible to search over all possible combinations of variables in x_t to be included in I, subject to $|I| \leq s$. For each combination, we record the realized values for the R^2, the maximized loglikelihood value (LogL), the Akaike Information Criterion (AIC), its corrected version (AICc), and the Bayes Information Criterion (BIC). We rank the sets according to each criterion separately. Furthermore, we restrict the LASSO and EN procedures by choosing the sets with the smallest penalizing parameter such that the number of variables retained in I are $|I| \leq s$. The results from these different selection procedures (and for different s) are presented in the bottom panel of Table 3.[14]

In case of the complete search, for a given s, and with the different rankings based on R^2, LogL, AIC, AICc, and BIC, the selections all agree on the set of variables to include in I. The LASSO and EN methods choose similar variable sets, although their selections differ from those of the complete search strategy. All methods agree that the variable corresponding to $i = 6$ is the most important variable to include in the model. This variable represents the general level of industrial production (IP) in the U.S. economy. A regression with this single variable and an intercept,

[13] An outlier around $t = 1970$ has been found in similar data; see, for example, Auffhammer and Steinhauser (2012) and Schmalensee et al. (1998). The interpretation of the outlier in $t = 1990$ is not clear. The estimate of its corresponding coefficient δ_{1990} in Table 4 has a positive sign; it implies that the emissions in 1990 are higher than implied by Industrial Production through the econometric model.

[14] We have also carried out the selection procedures with the inclusion of ten principal components from X that correspond to the ten largest eigenvalues of the sample variance matrix of X. In this case, the principal components have never been selected as part of the best model.

Table 3: *Regression output from the models chosen by the various selection criteria. The AutoMetrics procedure detects two outliers in $t = 1970$ and in $t = 1990$. These have been included in the other models as well.*

	Tiny		*Small*	
AutoMetrics	$\{6, 17, 48, 51, 56\}$		$\{6, 17, 48, 51, 55, 89\}$	
	$s = 1$	$s \leq 2$	$s \leq 3$	$s \leq 4$
R^2	$\{6\}$	$\{4, 17\}$	$\{4, 17, 185\}$	$\{6, 17, 51, 53\}$
logL	$\{6\}$	$\{4, 17\}$	$\{4, 17, 185\}$	$\{6, 17, 51, 53\}$
AIC	$\{6\}$	$\{4, 17\}$	$\{4, 17, 185\}$	$\{6, 17, 51, 53\}$
AICc	$\{6\}$	$\{4, 17\}$	$\{4, 17, 185\}$	$\{6, 17, 51, 53\}$
BIC	$\{6\}$	$\{4, 17\}$	$\{4, 17, 185\}$	$\{6, 17, 51, 53\}$
LASSO	$\{6\}$	$\{6, 9\}$	$\{6, 9\}$	$\{6, 9, 15, 17\}$
Elastic Net	$\{6\}$	$\{6, 9\}$	$\{6, 9\}$	$\{6, 9, 15, 17\}$

produces an R^2 of 0.54. Many other IP related indices produce the same degree of goodness-of-fit. The variable corresponding to $i = 17$ is present in most models. This variable is the industrial production index of residential utilities (IP: Residential Utilities). The variable set $I = \{6, 17\}$ produces an R^2 of 0.71. It is plausible that these two variables explain much of the variation in the growth of CO_2 emissions: variable $i = 6$ represents the production part of the economy while variable $i = 17$ represents the residential part of the economy. When more explanatory power is desired, a trade-related variable ($i = 185$) or a housing-related variable ($i = 51$ or $i = 53$) can be included in the set I.

The results in Table 3 show that for $s \leq 2$ and $s \leq 3$, the variable $i = 4$ (real manufacturing and trade industries sales) is preferred over $i = 6$ (industrial production). However, the correlation between the two variables $x_t^{(4)}$ and $x_t^{(6)}$ is 95%. There is little gain from including further variables. Figure 3 plots the R^2 from the regression from the sets I selected by the various methods. The IP variable $x_t^{(6)}$ clearly contributes most of the explanatory power with $R^2 = 0.54$. When adding an extra variable, we obtain either $R^2 = 0.71$ or $R^2 = 0.73$ when the extra variable is IP: Residential Utilities ($x_t^{(17)}$) and Real Manufacturing and Trade Industries Sales ($x_t^{(4)}$), respectively. The increase in R^2 levels off when further variables are added: we obtain $R^2 = 0.75$, for $s \leq 3$, and $R^2 = 0.77$, for $s \leq 4$. We thus prefer a small set I, set $k = 2$, and select the variables $x_t^{(i_1)}$ and

$x_t^{(i_2)}$ with $i_1 = 6$ and $i_2 = 17$ in the following.

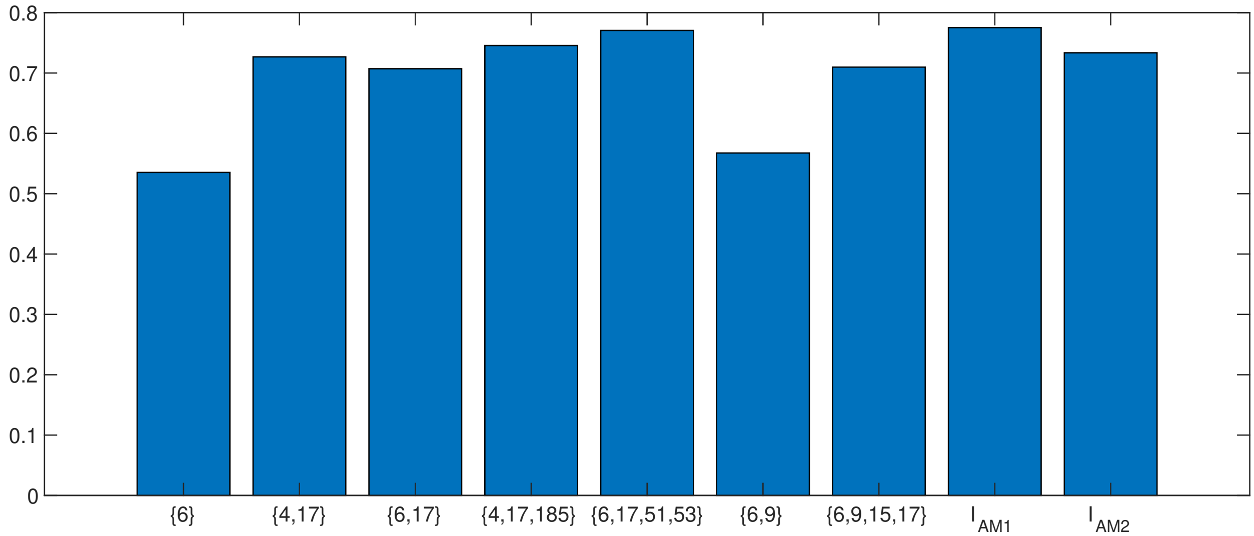

Figure 3: *Coefficient of determination (R^2) for the regression for different sets I. The AutoMetrics sets $I_{AM1} = \{6, 17, 48, 51, 56\}$ and $I_{AM2} = \{6, 17, 48, 51, 55, 58\}$ are the sets obtained from the options "Tiny" and "Small", respectively.*

5 Estimation Results and In-Sample Analysis

In our empirical analysis of U.S. per-capita CO_2 emissions growth, we consider the SADFM where we need to specify k, l and r, and select variables for x_t^* and z_t. As motivated in Section 4, we take $k = 2$, with $i_1 = 6$ and $i_2 = 17$, that is $x_t^* = (x_t^{(6)}, x_t^{(17)})'$, and $l = 2$, with z_t representing two dummy variables to extract the outliers in per-capita CO_2 emissions growth for 1970 and 1990. The dimension of f_t is r, and it is determined by the p_2 criterion of Bai and Ng (2002). In almost all our model settings, we find strong empirical support for $r = 4$. We therefore present the results for the SADFM with these settings.

Furthermore, we consider three variants of the SADFM specification:

(i) DFM: with restriction $\beta_1 = \beta_2 = 0$, leading to the standard DFM model;

(ii) SADFM-U: without any restriction;

(iii) SADFM-R: with restriction $\beta_f = 0$, leading to a model with factors only for x_t.

These specifications have different features. The DFM specification leads to standard dynamic factor analysis with well-known forecasting and nowcasting procedures. SADFM-U combines important individual economic variables with the dynamic features of a large data set for the prediction of CO_2 emissions growth. SADFM-R reduces the equation for y_t to a basic regression model and only the x_t are modeled by factors. The simplicity of a regression model for y_t is an attractive feature, while the forecast power of the dynamic factor model for x_t is retained. These different specifications illustrate the flexibility of our SADFM framework.

5.1 Parameter estimation results

The parameter estimation results are presented in Panel A of Table 4. For all three model specifications, the estimate of the intercept α is estimated to be statistically indistinguishable from zero, while the two regression coefficients in γ, associated with the two dummy variables for 1970 and 1990 are statistically significant. In the case of the DFM specification, y_t loads significantly on the first two economic factors, while the other two factors appear to have less significant support for y_t. In case of the unrestricted model SADFM-U, the regression coefficient for $x_t^{(17)}$ is highly significant, while the loadings on the first three factors are also statistically significant. The regression coefficient for $x_t^{(6)}$ is not significant, but we note that the first principal component in F_t closely resembles the variable $x_t^{(6)}$; the sample correlation between the two variables is as high as -87%. In this light, and given the findings in Section 4, it is not surprising that in the case of the restricted model SADFM-R, we obtain highly significant regression parameter estimates for both $x_t^{(6)}$ and $x_t^{(17)}$.

5.2 In-sample analysis

In Panel B of Table 4 we report the various goodness-of-fit measures for the three SADFM specifications; these measures are introduced in Section 4. In the context of SADFM, the R^2 measures relate only to the fit of y_t, calculated from Equation (3.1) using the estimated parameters and smoothed factors as output from the Kalman filter. This fit is shown in Figure 4 for the SADFMs. The likelihood-based measures are all calculated using the full observation vector with y_t and x_t in equation (3.4). Overall, the goodness-of-fit for the SADFM is high. A basic regression analysis based on $x_t^{(6)}$ and $x_t^{(17)}$ produces an R^2 statistic of 0.71, and with some other regression specifications we may raise it to 0.77; see the discussion in Section 4. The SADFM produces R^2 values of 0.88 and 0.85; these are clearly better fits. The fit measures of LogL and AIC point towards SADFM-U as providing the best fit while AICc and BIC point towards SADFM-R for the best fit. However, the differences are small. We favor the SADFM-R specification, as it is more parsimo-

Table 4: *Estimation results from SADFM model* (3.1). *t-stats in parentheses. N is the Jarque-Bera test statistic for normality. DW is the Durbin-Watson test statistic for serial correlation. LB(1) is the Ljung-Box test statistic for serial correlation with one lag with a* 5% *critical value of* 3.8415. *LB(20) is the Ljung-Box test statistic with* 20 *lags with a* 5% *critical value of* 31.4104.

Panel A: Estimates

	α	β_1	β_2	β_{f1}	β_{f2}	β_{f3}	β_{f4}	δ_{1970}	δ_{1990}
DFM	−0.0020			0.0211	0.0114	0.0026	−0.0030	0.0910	0.0391
	(−0.79)			(7.27)	(4.89)	(1.39)	(−1.39)	(5.58)	(2.40)
SADFM-U	−0.0020	0.0030	0.0136	0.0145	0.0086	0.0038	0.0011	0.0775	0.0505
	(−1.33)	(0.38)	(7.03)	(2.07)	(2.29)	(2.05)	(0.60)	(7.19)	(4.69)
SADFM-R	−0.0019	0.0201	0.0115					0.0770	0.0448
	(−1.14)	(9.11)	(6.65)					(6.52)	(3.79)

Panel B: Diagnostics (fit)

	R^2	R^2_a	logL	AIC	AICc	BIC
DFM	0.72	0.68	−12124.57	24261.15	24262.83	24273.41
SADFM-U	0.88	0.86	−12102.06	24220.12	24223.12	24236.47
SADFM-R	0.85	0.84	−12106.58	24221.16	24221.93	24229.34

Panel C: Diagnostics (standardised residuals)

	std	skew	kurt	N	DW	LB(1)	LB(20)
DFM	0.94	0.01	2.84	0.06	1.56	2.58	16.11
SADFM-U	0.94	0.01	2.85	0.06	1.57	2.46	16.21
SADFM-R	0.95	0.07	2.97	0.05	1.69	1.28	15.12

nious and more convenient for out-of-sample analysis, including forecasting and nowcasting; see the discussions in Section 6.

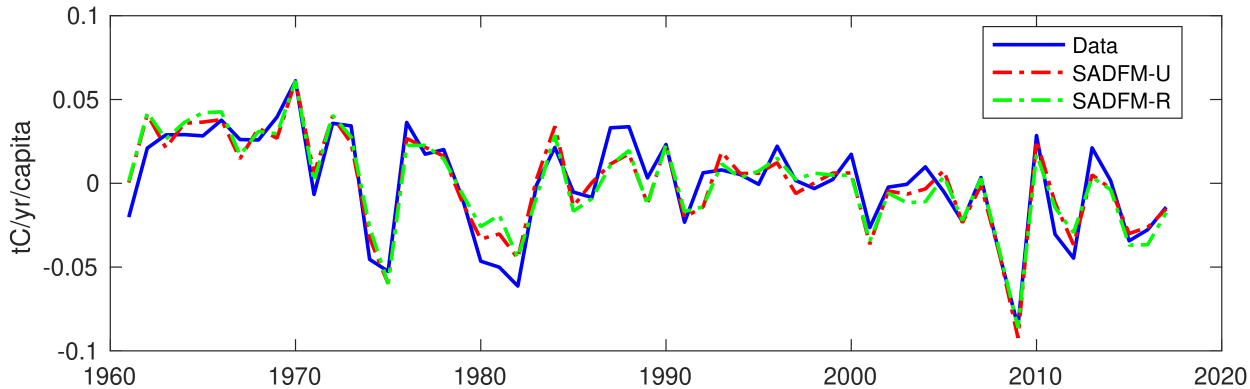

Figure 4: *CO_2 growth data $y_t = e_t$ and in-sample fit of SADFMs from* (3.1) *with the smoothed factors in place of f_t as output from the Kalman smoother.*

Panel C of Table 4 reports diagnostic statistics for the standardised one-step ahead prediction residuals v_t, which are defined as $v_t = (y_t - \hat{y}_{t|t-1})/\sqrt{\hat{F}_t}$, where $\hat{y}_{t|t-1}$ and $\hat{F}_t$ are formally defined as $y_{t|t-1} := \mathbb{E}(y_t|y_1, y_2, \ldots, y_{t-1})$ and $F_t := Var(y_t|y_1, y_2, \ldots, y_{t-1})$, respectively. These variables are directly obtained as output from the Kalman filter. The time series of v_t coming from the SADFMs are presented in Figure 5a). If the model is correctly specified, these prediction residuals have properties that can be verified. We focus particularly on the properties that the prediction residuals are serially uncorrelated and normally distributed. We report Durbin-Watson (DW) and Ljung-Box (LB) test statistics for serial correlation, skewness (skew) and kurtosis (kurt), and the Jarque-Bera test for normality. The results for the SADFM specifications can be summarized as follows. The LB test cannot reject the null of no serial correlation in $\hat{\epsilon}_t^y$ at a 5% significance level. Figures 5b) and c) present the sample autocorrelation and partial autocorrelation functions, respectively, for the prediction residuals and corroborate this finding. However, the DW test statistics indicate that there is still some positive autocorrelation left in the prediction residuals.[15] The reported normality diagnostics do not raise much concern. We have also investigated the temporal stability of the parameters of the SADFMs and have found no compelling evidence for time-varying parameters; see Appendices C and D for more details.

[15]Although Durbin-Watson statistics reported in Table 4 indicate positive autocorrelation in v_t and therefore in $\hat{\epsilon}_t^y$, Figure 5a) provides evidence that this is mostly due to observations in the 1960s. Indeed, if we redo the analysis with data starting in 1970 instead of 1961, the evidence for autocorrelation in v_t disappears completely (results not given here for brevity, but are available from the authors upon request).

Overall, we conclude that the SADFM describes the data well. The BIC selects the restricted version, SADFM-R, and since this is more convenient for out-of-sample analysis, it will be the preferred model in the following.

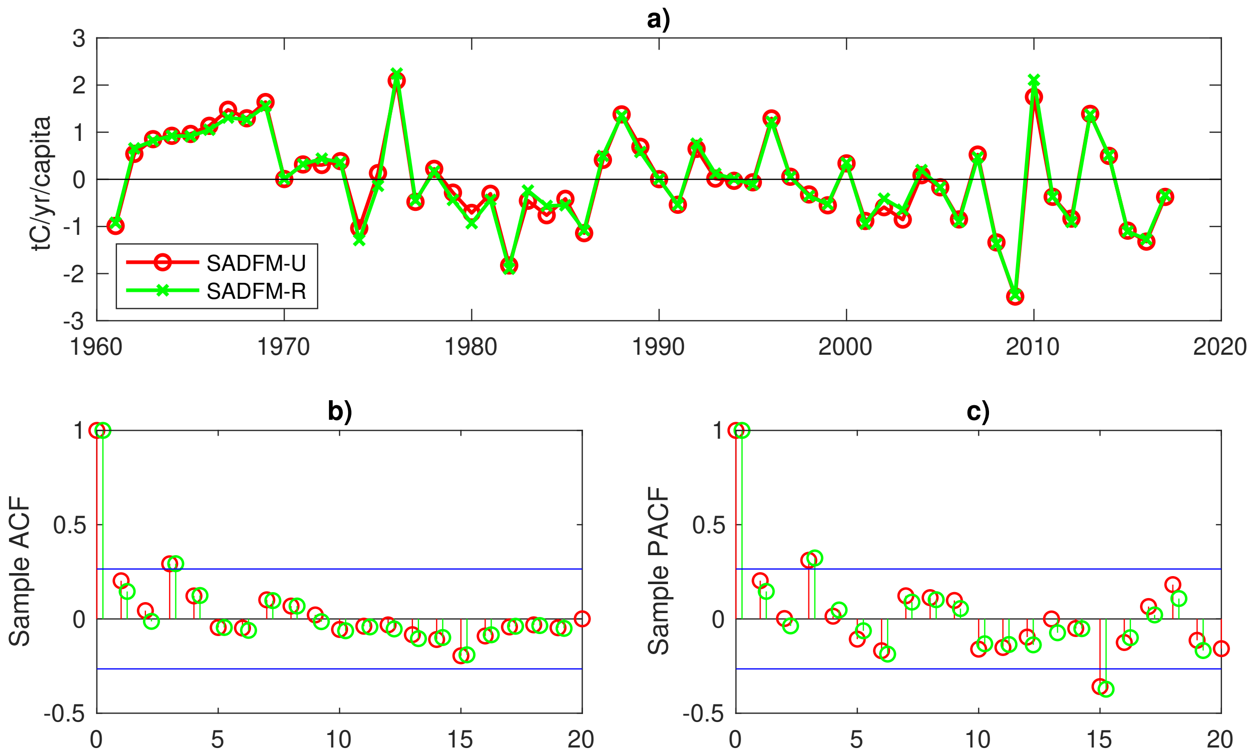

Figure 5: *a): Standardised one-step ahead prediction errors* v_t *from the SADFMs obtained from the Kalman filter. b) Sample autocorrelation of the standardised one-step ahead prediction errors with* $\pm 2/\sqrt{n}$ *confidence bands (blue lines), where* $n = 57$ *is the number of observations. c) Sample partial autocorrelation of the standardised one-step ahead prediction errors with* $\pm 2/\sqrt{n}$ *confidence bands.*

6 Out-of-sample Analysis

Dynamic factor models have been successfully applied to many macroeconomic forecasting problems (e.g. Stock and Watson, 2002; Bräuning and Koopman, 2014) and nowcasting problems (e.g. Giannone et al., 2008; Hindrayanto et al., 2016). Given the good in-sample fit of the SADFMs when applied to CO_2 emissions data, this section investigates their out-of-sample performance.

Since the sample period is short, containing only $T = 57$ observations, an out-of-sample experiment must necessarily be limited. Let $h \geq 0$ denote the forecast horizon in years; $h = 0$ corresponds to "nowcasting", while $h > 0$ corresponds to forecasting. We consider the in-sample period from $t = 1960$ to $t = 2001 - h$ and use an expanding window for estimation. That is, we estimate the models using the initial period from 1960 to $2001 - h$, containing $42 - h$ observations, and then nowcast ($h = 0$) or forecast ($h > 0$) per capita emissions growth in year 2001. We then extend the in-sample period by one year and estimate the models using data from 1960 to $2002 - h$ to nowcast/forecast per capita emissions growth in 2002, and so forth. For $h = 0$ we do not use the emissions data in the year we are nowcasting, but only the economic data. This setup ensures that the out-of-sample period covers 17 h-step ahead forecasts of per capita CO_2 emissions growth from 2001 to 2017, regardless of the value of h.

We consider two different loss metrics: root mean squared error (RMSE) and mean absolute error (MAE):

$$RMSE_h = \sqrt{\frac{1}{17}\sum_{t=2001}^{2017}(\hat{y}_{t|t-h} - y_t)^2},$$
$$MAE_h = \frac{1}{17}\sum_{t=2001}^{2017}|\hat{y}_{t|t-h} - y_t|,$$

where $\hat{y}_{t|t-h}$ denotes the forecast of y_t using information at time $t - h$ for any given model. The results for this exercise with $h = 0, 1, 2$ for the SADFM-R model are shown in Figure 6. The top row in the figure displays the raw losses $RMSE_h$ and MAE_h; the bottom row of the plot displays the corresponding losses as fractions of the corresponding losses from the constant growth model. The constant growth model simply predicts h-step ahead emissions to be the historical mean of y_t up to year $t - h$, that is $\hat{y}_{t|t-h} = \frac{1}{t-h-1961+1}\sum_{t=1961}^{t-h} y_t$. Numbers smaller than one indicate that the SADFM-R outperforms the constant growth model and vice versa for numbers greater than one. Figure 6 shows that for $h = 0$ and $h = 1$, the SADFM-R model outperforms the constant growth model. For $h = 2$, the advantage disappears.

The results in Figure 6 indicate that the SADFM-R model is well-suited to nowcasting ($h = 0$) and forecasting one year ahead ($h = 1$). In relative terms, the SADFM-R improves on the constant growth model. It has 64% of the constant model's MAE and 67% of the RMSE when nowcasting and 20% of both measures when forecasting one year ahead. In the next two sections we study the forecasting and nowcasting performances of our proposed models: Section 6.1 conducts an extended forecast exercise with more benchmark models than just the constant growth model; Section 6.2 examines the nowcasts from the various models considered in this paper and also takes a closer look at more recent nowcasts and compares with those from the Global Carbon Project.

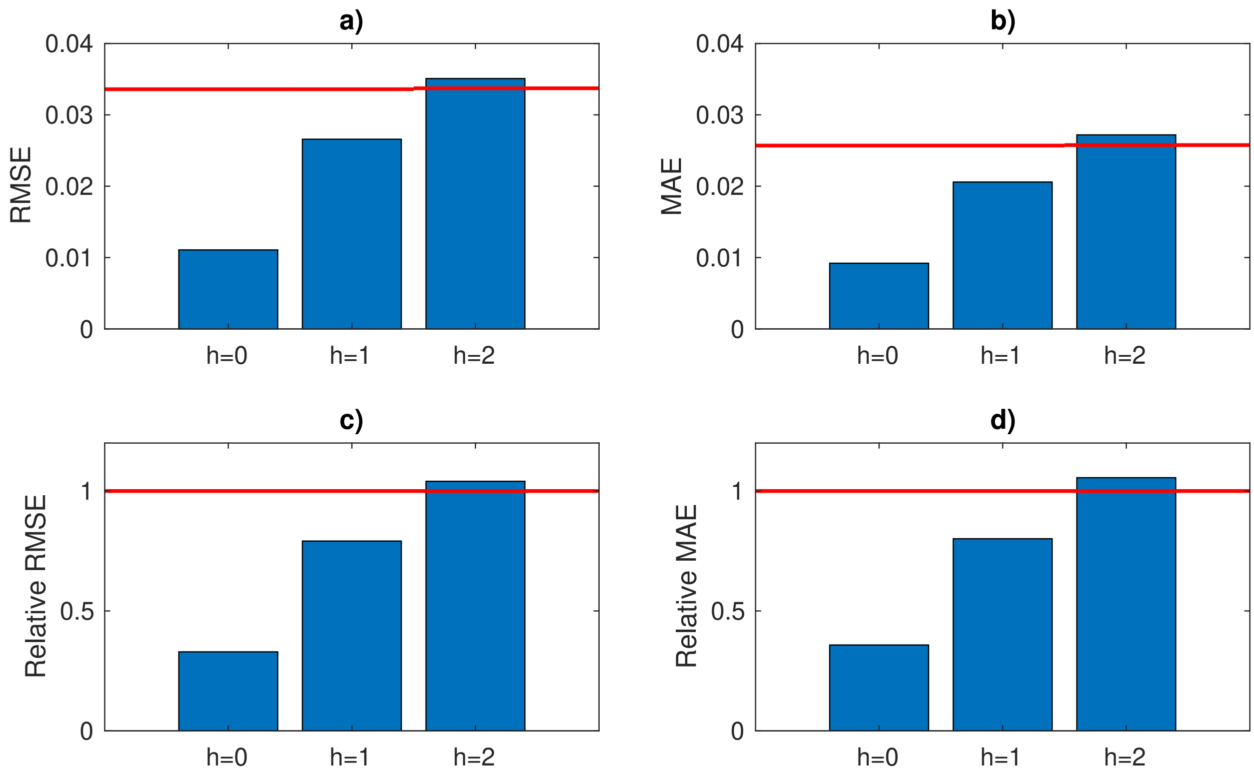

Figure 6: *Out-of-sample forecasting h years ahead for $h = 0, 1, 2$. $h = 0$ is "nowcasting". a): Root mean squared error (RMSE) of SADFM-R. b): Mean absolute error (MAE) of SADFM-R. c): RMSE of SADFM-R as fraction of the RMSE of the constant growth model. d): MAE of SADFM-R as fraction of the RMSE of the constant growth model. The red line shows the performance of the constant growth model.*

6.1 Forecasting

Accurate forecasting of the future growth in country-level CO_2 emissions is important, since such forecasts can be used to gauge whether a country is on track to keep their emissions targets, for instance. This section conducts a pseudo-out-of-sample study to assess how the models for forecasting CO_2 emissions growth proposed above compare with standard univariate and multivariate alternatives. The main benchmark is the constant growth model ("Cst"), which predicts the historical mean of y_t up until the time of the forecast. Other alternative models are:

- RW: A random walk model. This model predicts y_{t+h} using the current level of CO_2 growth, i.e. y_t.
- AR1: An autoregressive model of order one. This model assumes $y_t = a + \phi y_{t-1} + \epsilon_t$, where $a, \phi \in \mathbb{R}$ and ϵ_t is zero-mean white noise.
- MA1: A moving-average model of order one. This model assumes $y_t = a + \epsilon_t + \theta\epsilon_{t-1}$, where $a, \theta \in \mathbb{R}$ and ϵ_t is zero-mean white noise.
- ARMA(1,1): An autoregressive-moving-average model of order $(1, 1)$. This model assumes $y_t = a + \phi y_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$, where $a, \phi, \theta \in \mathbb{R}$ and ϵ_t is zero-mean white noise.

We also consider two VAR models, i.e.,

$$Y_t = a + \sum_{j=1}^{p} \Psi_j Y_{t-j} + \epsilon_t, \tag{6.1}$$

where $Y_t = (y_t, x_t^{(6)}, x_t^{(17)})'$, and $x_t^{(6)}, x_t^{(17)}$ are the IP index and the IP: Residential Utilities index, respectively.

- VAR: The 3-dimensional VAR of (6.1).
- SVAR: The structural counterpart of (6.1). That is, here we consider the model $B_1 Y_t = \Psi Y_{t-1} + \epsilon_t$, with B_1 similar, mutatis mutandis, to the B matrix considered in Section 3.

Inspired by the "direct" forecasting approach of Stock and Watson (2002), we also consider the predictive regression

$$y_t = a + b' F_{t-h} + \epsilon_t, \tag{6.2}$$

where F_t are principal components estimates of the economic factors. The numbers of factors, r, in F_t is determined by the p_2 criterion of Bai and Ng (2002). The parameters $(a, b')'$ can be estimated by OLS and the forecast of y_{t+h} is $\hat{y}_{t|t-h} = \hat{a} + \hat{b}' F_{t-h}$.

- PCA (all): This approach uses all r factors in the predictive regression (6.2).
- PCA ($|t| > 1.96$): Same as "PCA (all)" but instead of using all r factors in (6.2), only the factors that have significant t-statistics at a 5% level are included.

The results from the forecasting experiment for $h = 1$ are shown in Table 5. We again consider the two loss functions root mean squared error (RMSE) and mean absolute error (MAE). The loss numbers MAE and RMSE are fractions of the loss from the benchmark (constant growth) model. Thus, numbers less than one indicate that a particular model outperforms the benchmark. The factor-based models appear to be superior, improving on the benchmark by 11–21%. There is not much difference between the "direct" approaches and the "indirect" approaches. The large differences between the (S)VAR and the (SA)DFMs indicate that including the macroeconomic factors improves the forecasting accuracy for $x_t^{(i_1)}$, for $x_t^{(i_2)}$ and, indeed, for y_t.

The asterisks in Table 5 denote whether a particular model, for the specified loss function, is included in the *Model Confidence Set* (MCS) procedure of Hansen et al. (2011). For this procedure, we consider the significance levels $\alpha = 10\%$ and $\alpha = 25\%$, corresponding to $1-\alpha$ Model Confidence Sets, denoted by $\mathcal{M}_{1-\alpha}$. Models in the $\mathcal{M}_{90\%}$ are denoted by one asterisk, while models also in the $\mathcal{M}_{75\%}$ are denoted by two asterisks. We learn from Table 5 that only the direct PCA ($|t| > 1.96$) and SADFM-R are in $\mathcal{M}_{75\%}$ for both the RMSE and MAE loss metrics, while the DFM is in $\mathcal{M}_{75\%}$ for RMSE only. None of the other models are in $\mathcal{M}_{90\%}$.

We have also carried out the forecasting experiment for $h = 2$. These results are omitted for brevity but the overall conclusion is that the models have less forecasting power at this horizon. This finding resonates with the macroeconometric literature, where it is often reported that economic time series are difficult to forecast far into the future; see, for example, Giannone et al. (2008).

Finally, we forecast 2019 changes in U.S. CO_2 emissions. For this exercise, we updated the economic data set to include 2018 data. Using these, the 2019 forecasts for the series $x_t^{(6)}$ and $x_t^{(17)}$ are -0.16 and -0.27, respectively. These are studentized values; they correspond to an increase in IP of 1.9 percent and an increase in IP: Residential Utilities of 1.8 percent. Together with the estimated parameter vector $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = (-.00150, .0200, .0117)$, this results in a forecasted CO_2 emissions growth of -0.008, or about minus one percent.

Table 5: *Out-of-sample forecast exercise $h = 1$ period ahead. Boldface numbers indicate the models with lowest root mean squared error (RMSE) or mean absolute error (MAE). For RMSE and MAE, two stars indicate that the model is in the 75% Model Confidence Set, $\mathcal{M}_{75\%}$. RMSEs (MAEs) are fractions of the RMSE (MAE) of the main benchmarks. N is the Jarque-Bera test statistic for normality. DW is the Durbin-Watson test statistic for serial correlation. LB(20) is the Ljung-Box test statistic for serial correlation with 20 lags; this test has a 5% critical value of 31.4104. The abbreviations for the benchmark models are listed in the beginning of Section 6.1. DFM denotes the Dynamic Factor model; SADFM denotes the Structural Augmented Dynamic Factor Models, with "U" denoting the unrestricted SADFM and "R" the restricted SADFM.*

			Diagnostics (forecast errors)						
	RMSE	MAE	mean	std	skew	kurt	N	DW	LB(20)
Main benchmark									
Cst	1.00	1.00	0.02	0.03	0.59	3.53	3.86	2.35	13.48
Simple benchmarks									
RW	1.26	1.27	0.00	0.04	−1.11	4.06	14.11	2.79	19.31
AR1	1.04	1.07	0.01	0.03	−0.64	3.53	4.42	2.69	16.46
MA1	1.01	1.08	0.01	0.03	−0.39	3.04	1.39	2.75	14.81
ARMA(1,1)	1.04	1.10	0.01	0.03	−0.54	3.36	3.05	2.76	14.62
VAR methods									
VAR	1.05	1.09	0.00	0.04	−0.52	2.82	2.56	2.71	18.73
SVAR	1.14	1.17	0.00	0.04	−0.73	2.93	5.04	2.79	19.79
Direct methods									
PCA (all)	0.84	0.91	0.01	0.03	−0.12	2.61	0.49	2.01	14.42
PCA ($\|t\| > 1.96$)	0.82**	0.87**	0.02	0.02	−0.08	2.60	0.43	2.39	16.97
Indirect methods									
DFM	0.81**	0.85	0.02	0.02	0.24	2.64	0.85	2.10	18.46
SADFM-U	0.81	0.86	0.02	0.02	0.30	2.77	0.95	2.11	18.46
SADFM-R	**0.79****	**0.80****	0.02	0.02	0.40	2.91	1.48	2.05	18.92

6.2 Nowcasting

Estimating a variable of interest in period t while period t is still in progress is termed nowcasting in the macroeconomic literature. A common application in macroeconomics is nowcasting quarterly GDP growth, see, e.g., Giannone et al. (2008). In calculating annual CO_2 emissions, a country tallies consumption of energy carriers – use of coal, oil, gas, etc. – and calculates emissions implied by these statistics. Numbers on energy use in year t are generally not available until year $t+1$ (Le Quéré et al., 2015a, p. 55), introducing a lag in the reporting of country-level CO_2 emissions. Economic data used in this paper are generally published with a shorter lag than CO_2 emissions. We propose using our statistical framework to nowcast growth in CO_2 emissions in year t as soon as the economic data series $x_t^{(i_1)}$ and $x_t^{(i_2)}$ are available.

In Figure 6, we have already indicated the nowcasting performance of the SADFM-R compared to the constant growth model ($h = 0$). Table 6 reports the results for the various models considered in this paper, including a model with time-varying parameters (TVP) for the two IP variables; see Appendix D for a detailed discussion of this model. All models outperform the constant growth model. The SADFM-R, the SADFM-U, and the TVP models are in the 75% model confidence set, $\mathcal{M}_{75\%}$. None of the other models are in the $\mathcal{M}_{75\%}$ or the $\mathcal{M}_{90\%}$ sets.

In the annual research report on the global carbon cycle, published by the Global Carbon Project (GCP, see e.g., Le Quéré et al., 2018b), the authors collect and maintain annual data on sources and sinks of global CO_2 emissions. Because official CO_2 emissions are reported with a lag, the GCP paper published in year t contains a nowcast of year t emissions. These nowcasts are based on emissions estimates made by the U.S. Energy Information Administration (EIA) and on data on cement production from the United States Geological Survey (USGS), see, e.g., Le Quéré et al. (2018b) p. 2168.

The GCP has been supplying nowcasts of U.S. emissions since 2015, so that we only have four data points for comparison. We note that our (SA)DFMs use economic data from the entirety of a given year t. Consequently, the GCP nowcasts are potentially constructed using less data than the ones from the (SA)DFM, because the GCP (and EIA and USGS) might not have the full year t data set available for nowcasting. The nowcasting results should be interpreted in this light.

Figure 7 compares the (later) reported realized per capita emissions growth rates from 2015-2017 with the nowcasts made by the GCP, as published in Le Quéré et al. (2015b), Le Quéré et al. (2016), and Le Quéré et al. (2018a), and with those from the SADFM-R, SADFM-U, and TVP

Table 6: *Out-of-sample nowcast exercise ($h = 0$). Boldface numbers indicate the models with lowest root mean squared error (RMSE) or mean absolute error (MAE). For RMSE and MAE, two stars indicate that the model is in the* 75% *model confidence set,* $\mathcal{M}_{75\%}$. *RMSEs (MAEs) are fractions of the RMSE (MAE) of the constant growth model. N is the Jarque-Bera test statistic for normality. DW is the Durbin-Watson test statistic for serial correlation. LB(20) is the Ljung-Box test statistic for serial correlation with* 20 *lags; this test has a* 5% *critical value of* 31.4104. *DFM denotes the Dynamic Factor model; SADFM denotes the Structural Augmented Dynamic Factor Models, with "U" denoting the unrestricted SADFM and "R" the restricted SADFM. TVP refers to the time-varying parameter model* $y_t = (1, x_t^{(6)}, x_t^{(17)})\alpha_t^* + \gamma' z_t + \epsilon_t^y$, *where* α_t^* *is modelled by a random walk, see Equation* (D.1) *in Appendix D.*

			Diagnostics (nowcast errors)						
	RMSE	MAE	mean	std	skew	kurt	N	DW	LB(20)
Constant growth model									
Cst	1.00	1.00	0.02	0.03	0.59	3.53	3.93	2.35	13.48
Other models									
DFM	0.60	0.63	0.00	0.02	0.05	2.18	1.60	1.94	10.07
SADFM-U	**0.31****	**0.33****	−0.00	0.01	0.48	2.99	2.22	1.86	14.95
SADFM-R	0.33**	0.36**	−0.00	0.01	0.55	2.71	3.05	1.88	14.12
TVP	0.34**	0.34**	−0.00	0.01	1.14	3.26	12.53	1.77	12.58

models. We adjust the GCP nowcasts using the realized growth rates of population in the U.S. to arrive at nowcasts comparable to our per capita nowcasts.[16] The nowcasts produced by the SADFM-R and TVP models are very accurate for the three years 2015-2017. They appear to be closer to the eventual true value than the GCP nowcasts for 2015 and 2017, while for 2016, the GCP nowcast was better. The nowcasts from the SADFM-U model appear to be quite close to those from the GCP. These results come with the two caveats: we only have three reliable data points, and our models use *all* macroeconomic data from year t to make the nowcast for year t, while the GCP nowcasts are most likely using less data.

We have not been able to obtain an official estimate of U.S. CO_2 emissions for the year 2018, since this number has not been reported on the UNFCCC web site at the time of writing.[17] This illustrates why nowcasting is a useful exercise: at the time of writing towards the end of 2019, the official U.S. estimate of CO_2 emissions in $t = 2018$ is still not available. Figure 7 shows the GCP nowcast from 2018, published in Le Quéré et al. (2018b), and the nowcasts from our models. For this exercise, we updated the economic data set to include 2018 data. To convert the GCP nowcast into a per capita emissions growth rate, we adjust by the realized growth rate in U.S. population (0.6% in 2018). The GCP nowcasts a 2018 growth of 1.9% in per capita emissions, while the SADFM-R, SADFM-U, and TVP models nowcast 2.3%, 2.07%, and 4.1%, respectively.

Finally, we focus on the nowcast of 2019. We use the data set from the Federal Reserve Bank of St. Louis through September 2019.[18] From this data set, we use the two monthly data series on IP and IP: Residential Utilities. We average the values from January to September (9 observations with monthly frequency), to estimate annual growth in the two annual series $x_t^{(6)}$ and $x_t^{(17)}$ in our models. These estimates are (in studentized values) -0.4095 and -1.3698, respectively. This corresponds to 0.9 percent growth in IP and a 1.8 percent contraction in IP: Residential Utilities. With these, we can nowcast 2019 emissions growth from the SADFM-R and the TVP models. The results are shown in Figure 7: With the estimated parameter vector $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = (-.00150, .0200, .0117)$, the SADFM-R model nowcasts that 2019 U.S. CO_2 emissions per capita will decline by approximately 2.6% in 2019; the TVP model nowcasts a decline of 2.2%.[19]

Nowcasting with the SADFM-U model when the economic data set is not available for the

[16] If the GCP nowcast of emissions growth in year t is e_t^*, we report $y_t^* = e_t^* - p_t$, where p_t is the population growth in the U.S. in year t.

[17] See e.g. `https://di.unfccc.int/time_series`, accessed on November 18, 2019.

[18] `https://research.stlouisfed.org/econ/mccracken/fred-databases/`, downloaded on November 18, 2019.

[19] See Appendix A for the details of the annualization and stationarity transformations.

entire year is more cumbersome, since this model relies on the (annual) factors f_t which must be estimated from the monthly data series available at the time of nowcasting. For this reason, we refrain from nowcasting 2019 using this model. This highlights an advantage of the SADFM-R (and TVP), compared to SADFM-U, since the former easily incorporates monthly data. A "mixed frequency SADFM-U" that is capable of incorporating monthly economic data and factors for forecasting annual CO_2 emissions, would be a worthwhile pursuit. We leave the development of such mixed frequency SADFMs for further research.

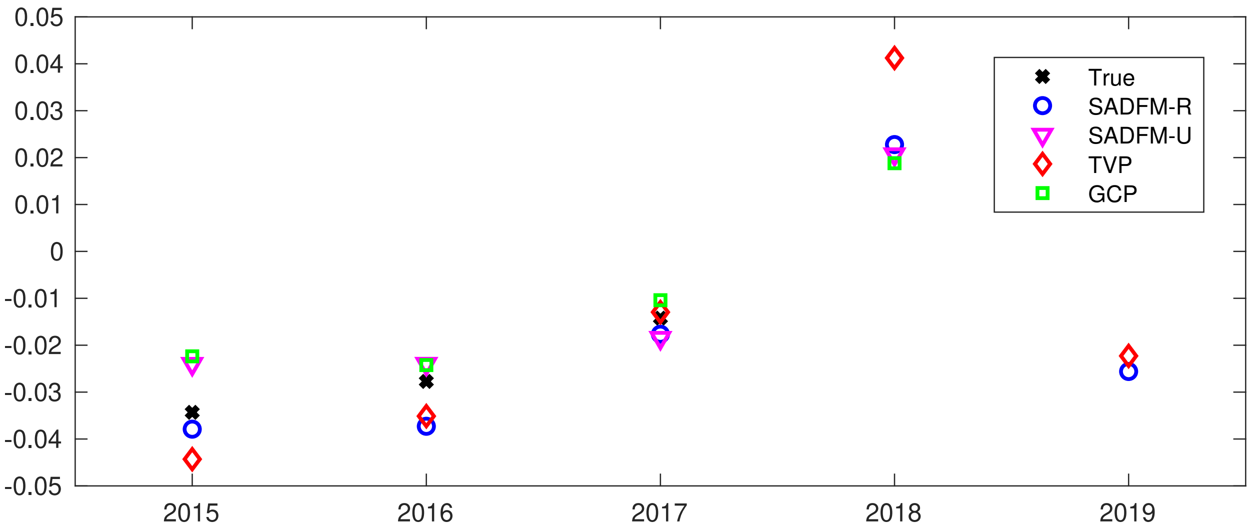

Figure 7: *Nowcasting U.S. CO_2 emissions per capita growth. Black crosses: True growth rates. Blue circles: Nowcasts from SADFM-R. Magenta triangles: Nowcasts from SADFM-U. Red diamond: Nowcasts from TVP model (see Appendix C). Green square: Nowcasts from the Global Carbon Project. At the time of writing, official U.S. CO_2 emissions for* 2018 *and* 2019 *were not available.*

7 Discussion and Conclusion

We have proposed a structural augmented dynamic factor model (SADFM) for U.S. CO_2 emissions depending on macroeconomic activity. Emissions are best explained by contemporaneous industrial production (IP), both in manufacturing and in residential utilities. In order to forecast IP, we employed a dynamic factor structure using a large set of annual time series on U.S. macroeconomic variables. The model has good in-sample and out-of-sample properties. Using 2018 data (except emissions for 2018, which at the time of writing this paper were not yet available in the UNFCCC inventories), our model forecasts 2019 emissions to decrease by approximately 1%. Using data up to September 2019, the model nowcasts a decrease of 2.6% (2.2% for the model with time-varying parameters). We recommend to use industrial production indices that cover manufacturing and residential utilities instead of the commonly employed GDP time series in explaining U.S. CO_2 emissions.

A limitation of our model is that it cannot capture changes in the energy mix, due to its focus on emissions as dependent variable. Emissions are, in essence, a linear combination of consumption quantities of the different energy carriers, and including explanatory variables on their relative consumption quantities would lead to definitional overlap on both sides of our model equation. As an alternative, one could replace CO_2 emissions on the left-hand side with consumption quantities, for instance $y_t = (COAL_t, OIL_t, GAS_t)'$, with self-evident notation, and connect these to macroeconomic variables using model equation (3.1). This could capture technological changes, for example towards renewable energy, at the cost of losing the focus on emissions. We leave this avenue for future research.

Other future research questions may include applying this model to other countries and macroeconomic areas such as the European Union, considering other dependent variables such as CO_2 emissions normalized by total energy use instead of population, allowing for non-linear dependencies of emissions on predictor variables (Auffhammer and Steinhauser, 2012), and using data of mixed frequency to further improve forecasting and nowcasting performance.

References

Arrow, K., B. Bolin, R. Costanza, P. Dasgupta, C. Folke, C. Holling, B. Jansson, S. Levin, K. Maler, C. Perrings, and P. D. (1995). Economic growth, carrying capacity and the environment. *Science 268*, 520–521.

Auffhammer, M. and R. Steinhauser (2012). Forecasting the path of U.S. CO_2 emissions using state-level information. *The Review of Economics and Statistics 94*(1), 172–185.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Blanco, G., R. Gerlagh, S. Suh, J. Barrett, H. C. de Coninck, C. D. Morejon, R. Mathur, N. Nakicenovic, A. O. Ahenkorah, J. Pan, et al. (2014). Drivers, trends and mitigation. In O. Edenhofer, Y. Pichs-Madruga, E. Sokona, et al. (Eds.), *Climate Change 2014: Mitigation of Climate Change, Contribution of*

Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

Boden, T. A., G. Marland, and R. J. Andres (2018). Global, regional, and national fossil-fuel CO_2 emissions. Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, Tenn., USA, available at: https://cdiac.ess-dive.lbl.gov/trends/emis/overview.html, last access: 28 June 2017.

Bosetti, V., C. Carraro, M. Galeotti, E. Massetti, and M. Tavoni (2006). WITCH - A world induced technical change hybrid model. *The Energy Journal*, 13–37.

Bräuning, F. and S. J. Koopman (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting 30*(3), 572–584.

Brock, W. A. and M. S. Taylor (2005). Economic growth and the environment: a review of theory and empirics. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth, Chapter 28*, pp. 1749–1821. Elsevier.

Calvin, K., P. Patel, L. Clarke, G. Asrar, B. Bond-Lamberty, R. Y. Cui, A. Di Vittorio, K. Dorheim, J. Edmonds, C. Hartin, et al. (2019). GCAM v5.1: Representing the linkages between energy, water, land, climate, and economic systems. *Geoscientific Model Development (Online) 12*(PNNL-SA-137098).

Doornik, J. (2009). Autometrics. In J. Castle and N. Shephard (Eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 88–121. Oxford University Press.

Doornik, J. A. and D. F. Hendry (2015). Statistical model selection with 'big data'. *Cogent Economics and Finance*, DOI: 10.1080/23322039.2015.1045216.

Doornik, J. A. and D. F. Hendry (2018). *PcGive, Empirical Econometric Modelling* (15 ed.). London: Timberlake Consultants Ltd.

Doz, C., D. Giannone, and L. Reichlin (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics 164*(1), 188–205.

Doz, C., D. Giannone, and L. Reichlin (2012). A quasi—maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics 94*(4), 1014–1024.

Durbin, J. and S. J. Koopman (2012). *Time series analysis by state space methods.* Number 38. Oxford University Press.

Durbin, J. and G. S. Watson (1971). Testing for serial correlation in least squares regression. *Biometrika 58*(1), 1 – 19.

Elliott, G., C. Granger, and A. Timmermann (2006). *Handbook of Economic Forecasting Vol. 1.* Elsevier.

Elliott, G. and A. Timmermann (2013). *Handbook of Economic Forecasting Vol. 2A and 2B*. Elsevier.

Elliott, G. and A. Timmermann (2016). *Economic Forecasting*. Princeton University Press.

Fujimori, S., T. Masui, and Y. Matsuoka (2017). AIM/CGE v2.0 model formula. In S. Fujimori, M. Kainuma, and T. Masui (Eds.), *Post-2020 Climate Action*, pp. 201–303. Springer.

Gambhir, A., I. Butnar, P.-H. Li, P. Smith, and N. Strachan (2019). A review of criticisms of integrated assessment models and proposed approaches to address these, through the lens of BECCS. *Energies 12*(9), 1747.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Grossman, G. M. and A. B. Krueger (1991). Environmental impacts of a north american free trade agreement. Technical report, National Bureau of Economic Research Working Paper 3914.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Hillebrand, E. and S. Koopman (Eds.) (2016). *Dynamic Factor Models, Advances in Econometrics Vol. 35*. Emerald Group Publishing.

Hindrayanto, I., S. J. Koopman, and J. de Winter (2016). Forecasting and nowcasting economic growth in the euro area using factor models. *International Journal of Forecasting 32*(4), 1284–1305.

Jarque, C. M. and A. K. Bera (1987). A test for normality of observations and regression residuals. *International Statistical Review 2*, 163–172.

Jungbacker, B. and S. J. Koopman (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal 18*(2), C1–C21.

Le Quéré, C., R. M. Andrew, and J. G. e. Canadell (2016). Global Carbon Budget 2016. *Earth System Science Data 8*(2), 605–649.

Le Quéré, C., R. M. Andrew, and P. e. Friedlingstein (2018a). Global Carbon Budget 2017. *Earth System Science Data 10*(1), 405–448.

Le Quéré, C., R. M. Andrew, and P. e. Friedlingstein (2018b). Global Carbon Budget 2018. *Earth System Science Data 10*(4), 2141–2194.

Le Quéré, C., R. Moriarty, and R. M. e. Andrew (2015a). Global Carbon Budget 2014. *Earth System Science Data 7*(1), 47–85.

Le Quéré, C., R. Moriarty, and R. M. e. Andrew (2015b). Global Carbon Budget 2015. *Earth System Science Data 7*(2), 349–396.

Lee, C.-C. (2005). Energy consumption and gdp in developing countries: A cointegrated panel analysis. *Energy Economics 27*(3), 415 – 427.

Ljung, G. M. and G. E. P. Box (1978). On a measure of lack of fit in time series models. *Biometrika 65*(2), 297–303.

Luderer, G., M. Leimbach, N. Bauer, E. Kriegler, L. Baumstark, C. Bertram, A. Giannousakis, J. Hilaire, D. Klein, A. Levesque, et al. (2015). Description of the REMIND model (version 1.6). Technical report, SSRN Working Paper 2697070.

Marland, G. and R. M. Rotty (1984). Carbon dioxide emissions from fossil fuels: a procedure for estimation and results for 1950–1982. *Tellus B 36B*(4), 232–261.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Messner, S. and M. Strubegger (1995). User's guide for MESSAGE III. Technical report, Institute for Applied Systems Analysis, Laxenburg, Austria, WP-95-069.

Millimet, D. L., J. A. List, and T. Stengos (2003). The environmental Kuznets curve: Real progress or misspecified models? *Review of Economics and Statistics 85*(4), 1038–1047.

Nordhaus, W. and P. Sztorc (2013). DICE2013R: Introduction and user's manual. Technical report, Yale University.

Oh, W. and K. Lee (2004). Causal relationship between energy consumption and GDP revisited: the case of Korea 1970–1999. *Energy Economics 26*(1), 51 – 59.

Ozturk, I. (2010). A literature survey on energy–growth nexus. *Energy Policy 38*(1), 340 – 349.

Raupach, M. R., G. Marland, P. Ciais, C. Le Quéré, J. G. Canadell, G. Klepper, and C. B. Field (2007). Global and regional drivers of accelerating CO2 emissions. *Proceedings of the National Academy of Sciences 104*(24), 10288–10293.

Schmalensee, R., T. M. Stoker, and R. A. Judson (1998). World carbon dioxide emissions: 1950–2050. *The Review of Economics and Statistics 80*(1), 15–27.

Stehfest, E., D. van Vuuren, L. Bouwman, and T. Kram (2014). *Integrated assessment of global environmental change with IMAGE 3.0: Model description and policy applications.* Netherlands Environmental Assessment Agency (PBL).

Stern, D. I. (1993). Energy and economic growth in the USA: A multivariate approach. *Energy Economics 15*(2), 137 – 150.

Stern, D. I. (2000). A multivariate cointegration analysis of the role of energy in the US macroeconomy. *Energy Economics 22*(2), 267 – 283.

Stern, D. I. (2017). The environmental Kuznets curve after 25 years. *Journal of Bioeconomics 19*(1), 7–28.

Stock, J. and M. Watson (2010). *Dynamic Factor Models*. Oxford: Oxford University Press.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H., M. W. Watson, J. B. Taylor, and H. Uhlig (2016). *Chapter 8 - Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics*, Volume 2, pp. 415–525. Elsevier.

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

UNFCCC (2018). https://di.unfccc.int/time_series.

Wagner, M. (2008). The carbon Kuznets curve: a cloudy picture emitted by bad econometrics? *Resource and Energy Economics 30*(3), 388–408.

Wagner, M. (2015). The environmental Kuznets curve, cointegration and nonlinearity. *Journal of Applied Econometrics 30*(6), 948–967.

Zhang, X.-P. and X.-M. Cheng (2009). Energy consumption, carbon emissions, and economic growth in China. *Ecological Economics 68*(10), 2706 – 2712.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67*(2), 301–320.

APPENDICES

A Economic data

In Section 2 we consider the $T \times N$ matrix X of economic data. In our study, we have $N = 226$ time series of length $T = 57$ years. The $N_{fed} = 126$ first data series are taken from the "FRED-MD" database, compiled and maintained by the Federal Reserve Bank of St. Louis.[20] Detailed information on these data can be found in the Online Appendix of McCracken and Ng (2016).[21] In Table 7 we supply similar information for the additional economic data series considered in this paper, namely $N_{agri} = 18$ variables that represent U.S. agricultural production[22]; $N_{trade} = 76$ variables that represent U.S. foreign trade[23]; $N_{cement} = 2$ time series that represent U.S. cement production[24]; and $N_{transport} = 4$ represent U.S. transport[25]. Appendix B explains how we handle the remaining missing values in the time series.

The $N_{agri} + N_{trade} = 18 + 76 = 94$ data series obtained from the World Bank are, like CO_2 emissions, recorded at an annual frequency. The $N_{fed} + N_{cement} + N_{transport} = 126 + 2 + 4 = 132$ data series obtained from the St. Louis Fed are originally all recorded at a monthly frequency. Since the emissions data used in this paper are available at an annual frequency, we aggregate these economic data to an annual frequency, arriving at the $T = 57$ observations for each time series. The method of aggregation might vary from time series to time series and is based on the nature of the series in question: for data in levels, we take the arithmetic mean over the 12 observations from January to December in year t to arrive at the year t observation for the data series; for data in monthly growth rates, we sum over the 12 observations from January to December in year t to arrive at the year t growth rate; for interest rates we take the geometric mean over the 12 observations from January to December in year t to arrive at the year t interest rate.

We then follow the procedure described in the online appendix of McCracken and Ng (2016) to transform the economic time series to a set of stationary variables. The exact method used

[20] https://research.stlouisfed.org/econ/mccracken/fred-databases/, downloaded on May 7, 2019.

[21] https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/Appendix_Tables_Update.pdf, accessed on November 22, 2019.

[22] Collected from the World Bank website, https://data.worldbank.org, downloaded on September 19, 2019.

[23] Collected from the World Bank website, https://data.worldbank.org, downloaded on September 19, 2019.

[24] Collected from the website of the U.S. Federal Reserve Bank, https://fred.stlouisfed.org, downloaded on October 10, 2019.

[25] Collected from the website of the U.S. Federal Reserve Bank, https://fred.stlouisfed.org, downloaded on October 10, 2019.

to transform a particular data series is based on the statistical properties of the series. If the series is positive, we take the natural logarithm. After this possible transformation, we examine the integration properties of the data to check whether they need to be differenced before they can be considered stationary. To be precise, we test for a unit root using the Augmented Dickey Fuller (ADF) test; if the null of a unit root is rejected, we stop; if the null is not rejected, we take first differences of the time series and perform another unit root test; we proceed until the null is rejected. In Table 7 we give the transformation code ("tcode") denoting which transformation has been applied to a particular time series. For one time series, transport services as percent of service imports, we differenced the series even though the null of a unit root was rejected by the ADF test; this decision was taken on the basis of graphical inspection of the time series. The transformation code for this time series is supplemented with an asterisk.

B The treatment of missing values

The data matrix X constructed using the procedure described in the preceding section contains a number of missing values. Of the $N \cdot T = 226 \cdot 57 = 12882$ data points, $M = 492$ (3.8%) are missing. There are only very few missing values in the $N_{fed} = 126$ first economic data series based on the FRED-MD data set ($M_{fed} = 24$); hence, most of the missing values come from the remaining 100 data series. Most of the missing values are located in the beginning of the data set.

Before the analyses described in the paper are performed, we impute the missing values using the iterative factor-based approach suggested in Stock and Watson (2002) as implemented by McCracken and Ng (2016).[26]

[26] We use the MATLAB code available at https://research.stlouisfed.org/econ/mccracken/fred-databases/, downloaded on February 1, 2019.

Table 7: The column "tcode" denotes the data transformation for a series x following McCracken and Ng (2016): (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $\log x_t$; (5) $\Delta \log x_t$; (6) $\Delta^2 \log x_t$; (7) $\Delta(x_t/x_{t-1} - 1)$. The transformation code of one series was set manually by the authors; this is denoted by an asterisk.

Group 9: Agriculture

	id	tcode	Indicator name	description
1	127	5	TX.VAL.AGRI.ZS.UN	Agricultural raw materials exports (% of merchandise exports)
2	128	6	SP.RUR.TOTL.ZS	Rural population (% of total population)
3	129	2	SP.RUR.TOTL.ZG	Rural population growth (annual %)
4	130	6	SP.RUR.TOTL	Rural population
5	131	5	AG.YLD.CREL.KG	Cereal yield (kg per hectare)
6	132	5	AG.SRF.TOTL.K2	Surface area (sq. km)
7	133	5	AG.PRD.LVSK.XD	Livestock production index (2004-2006 = 100)
8	134	5	AG.PRD.FOOD.XD	Food production index (2004-2006 = 100)
9	135	5	AG.PRD.CROP.XD	Crop production index (2004-2006 = 100)
10	136	5	AG.PRD.CREL.MT	Cereal production (metric tons)
11	137	5	AG.LND.TRAC.ZS	Agricultural machinery, tractors per 100 sq. km of arable land
12	138	5	AG.LND.TOTL.K2	Land area (sq. km)
13	139	5	AG.LND.CROP.ZS	Permanent cropland (% of land area)
14	140	5	AG.LND.CREL.HA	Land under cereal production (hectares)
15	141	5	AG.LND.ARBL.ZS	Arable land (% of land area)
16	142	5	AG.LND.ARBL.HA.PC	Arable land (hectares per person)
17	143	5	AG.LND.ARBL.HA	Arable land (hectares)
18	144	5	AG.AGR.TRAC.NO	Agricultural machinery, tractors

C Temporal stability of the estimation results

We study the stability of the SADFM-R results over time. We split the sample period in two parts: 1961 to 1988 and 1990 to 2017. In order for both periods to have the same number of observations (28), the observation for 1989 is discarded and the dummy variable in 1990 is removed. We then estimate the parameters for SADFM-R on the sub-samples and focus on the parameters in the equation for y_t, that is $(\alpha, \beta_1, \beta_2)'$. We adopt the same maximum likelihood estimation strategy as for the full sample. Figure 8 reports the results for the parameters α, β_1 and β_2. While it is plausible that the effect of the various economic processes on CO_2 emissions growth changes over time, we find only changes in the intercept across the two sub-periods. The estimates of β_1 and β_2 for the two sub-periods are very similar and statistically not different from each other, indicating parameter stability. The estimates for the intercept α appear to be larger in the second period when compared to the first period, providing tentative evidence that the base level of emission growth has increased slightly over time.

Group 10: Trade

	id	tcode	Indicator name	description
1	145	5	TX.VAL.TRVL.ZS.WT	Travel services (% of commercial service exports)
2	146	6	TX.VAL.SERV.CD.WT	Commercial service exports (current USD)
3	147	5	TX.VAL.OTHR.ZS.WT	Computer, communications and other services (% of commercial service exports)
4	148	5	TX.VAL.MRCH.WL.CD	Merch. exports by the reporting economy (current USD)
5	149	5	TX.VAL.MRCH.RS.ZS	Merch. exports by the reporting economy, residual (% of total merch. exports)
6	150	5	TX.VAL.MRCH.R6.ZS	Merch. exports to low- and middle-income in Sub-Saharan Africa (% of total merch. exports)
7	151	5	TX.VAL.MRCH.R5.ZS	Merch. exports to low- and middle-income in South Asia (% of total merch. exports)
8	152	5	TX.VAL.MRCH.R4.ZS	Merch. exports to low- and middle-income in Middle East & North Africa (% of total merch. exports)
9	153	5	TX.VAL.MRCH.R3.ZS	Merch. exports to low- and middle-income in Latin America & the Caribbean (% of total merch. exports)
10	154	5	TX.VAL.MRCH.R2.ZS	Merch. exports to low- and middle-income in Europe & Central Asia (% of total merch. exports)
11	155	5	TX.VAL.MRCH.R1.ZS	Merch. exports to low- and middle-income in East Asia & Pacific (% of total merch. exports)
12	156	5	TX.VAL.MRCH.OR.ZS	Merch. exports to low- and middle-income outside region (% of total merch. exports)
13	157	5	TX.VAL.MRCH.HI.ZS	Merch. exports to high-income (% of total merch. exports)
14	158	5	TX.VAL.MRCH.CD.WT	Merch. exports (current USD)
15	159	5	TX.VAL.MRCH.AL.ZS	Merch. exports to economies in the Arab World (% of total merch. exports)
16	160	5	TX.VAL.MMTL.ZS.UN	Ores and metals exports (% of merchandise exports)
17	161	5	TX.VAL.MANF.ZS.UN	Manufactures exports (% of merch. exports)
18	162	5	TX.VAL.INSF.ZS.WT	Insurance and financial services (% of commercial service exports)
19	163	5	TX.VAL.FUEL.ZS.UN	Fuel exports (% of merchandise exports)
20	164	5	TX.VAL.FOOD.ZS.UN	Food exports (% of merchandise exports)
21	165	5	TX.VAL.AGRI.ZS.UN	Agricultural raw materials exports (% of merchandise exports)
22	166	5	TM.VAL.TRVL.ZS.WT	Travel services (% of commercial service imports)
23	167	6	TM.VAL.SERV.CD.WT	Commercial service imports (current USD)
24	168	5	TM.VAL.OTHR.ZS.WT	Computer, communications and other services (% of commercial service imports)
25	169	5	TM.VAL.MRCH.WL.CD	Merch. imports by the reporting economy (current USD)
26	170	5	TM.VAL.MRCH.RS.ZS	Merch. imports by the reporting economy, residual (% of total merch. imports)
27	171	5	TM.VAL.MRCH.R6.ZS	Merch. imports from low- and middle-income in Sub-Saharan Africa (% of total merch. imports)
28	172	5	TM.VAL.MRCH.R5.ZS	Merch. imports from low- and middle-income in South Asia (% of total merch. imports)
29	173	5	TM.VAL.MRCH.R4.ZS	Merch. imports from low- and middle-income in Middle East & North Africa (% of total merch. imports)
30	174	5	TM.VAL.MRCH.R3.ZS	Merch. imports from low- and middle-income in Latin America & the Caribbean (% of total merch. imports)
31	175	5	TM.VAL.MRCH.R2.ZS	Merch. imports from low- and middle-income in Europe & Central Asia (% of total merch. imports)
32	176	5	TM.VAL.MRCH.R1.ZS	Merch. imports from low- and middle-income in East Asia & Pacific (% of total merch. imports)
33	177	5	TM.VAL.MRCH.OR.ZS	Merch. imports from low- and middle-income outside region (% of total merch. imports)
34	178	5	TM.VAL.MRCH.HI.ZS	Merch. imports from high-income economies (% of total merch. imports)
35	179	5	TM.VAL.MRCH.CD.WT	Merch. imports (current USD)
36	180	5	TM.VAL.MRCH.AL.ZS	Merch. imports from economies in the Arab World (% of total merch. imports)
37	181	5	TM.VAL.MANF.ZS.UN	Manufactures imports (% of merchandise imports)
38	182	5	TM.VAL.INSF.ZS.WT	Insurance and financial services (% of commercial service imports)

Group 10: Trade (contd.)

	id	tcode	Indicator name	description
39	183	5	TM.VAL.INSF.ZS.WT	Fuel imports (% of merchandise imports)
40	184	5	TG.VAL.TOTL.GD.ZS	Merchandise trade (% of GDP)
41	185	5	NY.EXP.CAPM.KN	Exports as a capacity to import (constant LCU)
42	186	5	NE.TRD.GNFS.ZS	Trade (% of GDP)
43	187	2	NE.RSB.GNFS.ZS	External balance on goods and services (% of GDP)
44	188	2	NE.RSB.GNFS.CD	External balance on goods and services (current USD)
45	189	5	NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
46	190	1	NE.IMP.GNFS.KD.ZG	Imports of goods and services (annual % growth)
47	191	5	NE.IMP.GNFS.KD	Imports of goods and services (constant 2010 USD)
48	192	5	NE.IMP.GNFS.CD	Imports of goods and services (current USD)
49	193	5	NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
50	194	1	NE.EXP.GNFS.KD.ZG	Exports of goods and services (annual % growth)
51	195	5	NE.EXP.GNFS.KD	Exports of goods and services (constant 2010 USD)
52	196	5	NE.EXP.GNFS.CD	Exports of goods and services (current USD)
53	197	5	MS.MIL.XPRT.KD	Arms exports (SIPRI trend indicator values)
54	198	5	MS.MIL.MPRT.KD	Arms imports (SIPRI trend indicator values)
55	199	5	GC.TAX.IMPT.ZS	Customs and other import duties (% of tax revenue)
56	200	5	EG.IMP.CONS.ZS	Energy imports, net (% of energy use)
57	201	5	BX.GSR.TRVL.ZS	Travel services (% of service exports, BoP)
58	202	5	BX.GSR.TOTL.CD	Exports of goods, services and primary income (BoP, current USD)
59	203	6	BX.GSR.NFSV.CD	Service exports (BoP, current USD)
60	204	5	BX.GSR.MRCH.CD	Goods exports (BoP, current USD)
61	205	5	BX.GSR.INSF.ZS	Insurance and financial services (% of service exports, BoP)
62	206	5	BX.GSR.GNFS.CD	Exports of goods and services (BoP, current USD)
63	207	5	BX.GSR.CMCP.ZS	Communications, computer, etc. (% of service exports, BoP)
64	208	5	BX.GSR.CCIS.ZS	ICT service exports (% of service exports, BoP)
65	209	5	BX.GSR.CCIS.CD	ICT service exports (BoP, current USD)
66	210	2	BN.GSR.MRCH.CD	Net trade in goods (BoP, current USD)
67	211	2	BN.GSR.GNFS.CD	Net trade in goods and services (BoP, current USD)
68	212	5	BM.GSR.TRVL.ZS	Travel services (% of service imports, BoP)
69	213	5^*	BM.GSR.TRAN.ZS	Transport services (% of service imports, BoP)
70	214	5	BM.GSR.TOTL.CD	Imports of goods, services and primary income (BoP, current USD)
71	215	6	BM.GSR.NFSV.CD	Service imports (BoP, current USD)
72	216	5	BM.GSR.MRCH.CD	Goods imports (BoP, current USD)
73	217	5	BM.GSR.INSF.ZS	Insurance and financial services (% of service imports, BoP)
74	218	5	BM.GSR.GNFS.CD	Imports of goods and services (BoP, current USD)
75	219	5	BM.GSR.CMCP.ZS	Communications, computer, etc. (% of service imports, BoP)
76	220	5	BG.GSR.NFSV.GD.ZS	Trade in services (% of GDP)

Group 11: Cement and transport

	id	tcode	Indicator name	description
1	221	5	IPN32732T9S	Industrial Production: Durable Goods: Concrete and product, Index 2012 = 100
2	222	5	IPN32731S	Industrial Production: Durable Goods: Cement, Index 2012 = 100
3	223	5	CES4300000001	All Employees: Transportation and Warehousing
4	224	5	CPITRNSL	Consumer Price Index: Transportation in U.S. City Average, All Urban Consumers
5	225	5	IPG3364T9S	Industrial Production: Durable manufacturing: Aerospace and miscellaneous transportation equipment
6	226	5	IPN32411AS	Industrial Production: Nondurable Goods: Aviation fuel and kerosene

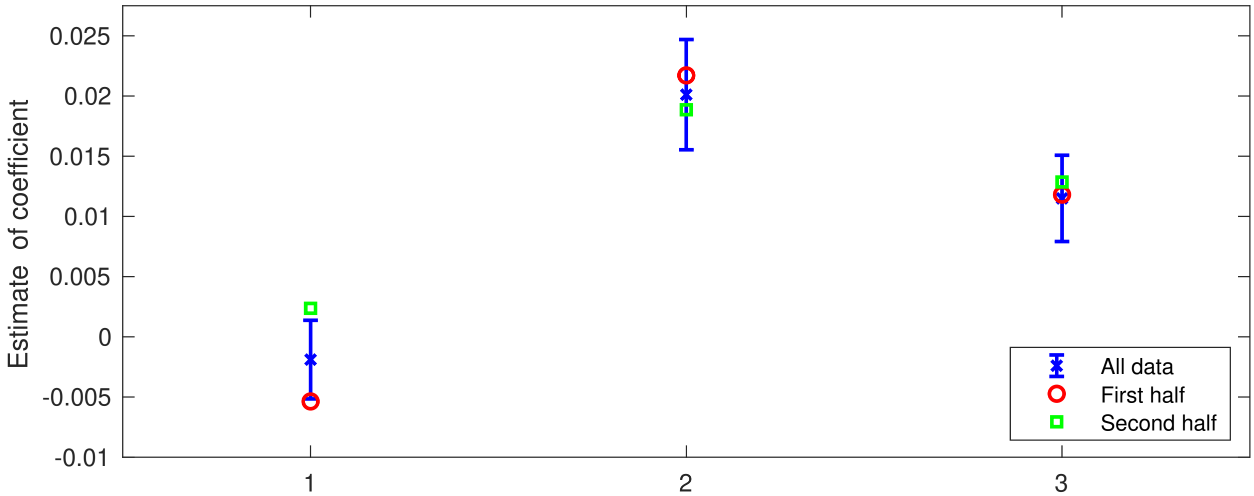

Figure 8: *Estimation of SADFM-R when the data are split in half ($T_1 = T_2 = 28$ observations in each half). The first coefficient is the intercept, the second coefficient relates to $x_t^{(6)}$, and the third coefficient relates to $x_t^{(17)}$. The estimates using the full data set, along with 95% confidence bands, are given as blue crosses; the estimates from using only the first half of the data are given in red circles; the estimates from using only the second half of the data are given in green squares.*

D Time-varying parameters for observation equation

Finally, as part of our in-sample analysis, we investigate the possibility of time-varying parameters in the observation equation for our SADFM-R specification, which is

$$y_t = \alpha + \beta_1 x_t^{(6)} + \beta_2 x_t^{(17)} + \gamma' z_t + \epsilon_t^y,$$

for $t = 1, \ldots, T$. Due to technological changes in the fuel mix, the β_1 and β_2 coefficients may change over time. Also, technology changes that are not associated with economic variables but are still affecting CO_2 emissions may lead to changes in the intercept α. To investigate these possible effects in our SADFM-R model, we replace α, β_1 and β_2 by the corresponding elements in the time-varying parameter vector $\alpha_t^* = (\alpha_t\,,\, \beta_{1t}\,,\, \beta_{2t})'$. The time-varying parameter (TVP) version of the observation equation of the SADFM-R model is then given by the system of equations

$$y_t = (1, x_t^{(6)}, x_t^{(17)})\alpha_t^* + \gamma' z_t + \epsilon_t^y, \qquad \alpha_{t+1}^* = \alpha_t^* + \kappa_t, \tag{D.1}$$

where κ_t is a 3×1 disturbance vector with mean zero and diagonal variance matrix Σ_κ. We let the time-varying parameters follow independent random walk processes. The resulting TVP model can be viewed as a specific linear state space model with initial moment conditions $\mathrm{E}(\alpha_1) = 0$ and

$\text{Var}(\alpha_1) = d \cdot I_3$ where d is typically a large positive value, say $d = 10^7$; see the discussion in Durbin and Koopman (2012, Chapters 4 and 5).

The estimation results for the time-varying parameters in the TVP model are presented in Figure 9. The overall fit of the TVP model appears to be highly satisfactory. The right-hand plot of Figure 9 presents the estimated time-varying paths for the three parameters in α_t^*. We find clear evidence of time-varying behavior in some of the elements of α_t^*. The intercept α_t is upward trending. This confirms the earlier finding that the estimated intercept in the second half of the sample is larger than the one in the first half. The estimated time-varying coefficient for the IP index $x_t^{(6)}$ reduces to a constant in the full sample. The estimated time-varying coefficient for the residential utility production index $x_t^{(17)}$ is varying somewhat over time. Its time-varying path appears to oscillate. The in-sample evidence for time-varying parameter behavior in the SADFM-R specification is therefore somewhat weak. Overall we may conclude that there is some evidence of a time-varying intercept α_t but not much evidence for time-varying β_{1t} and β_{2t} coefficients. Apart from the increasing general level of emissions captured by α_t, technology developments do not seem to have changed the economy-emissions relationship noticeably in our in-sample analysis.

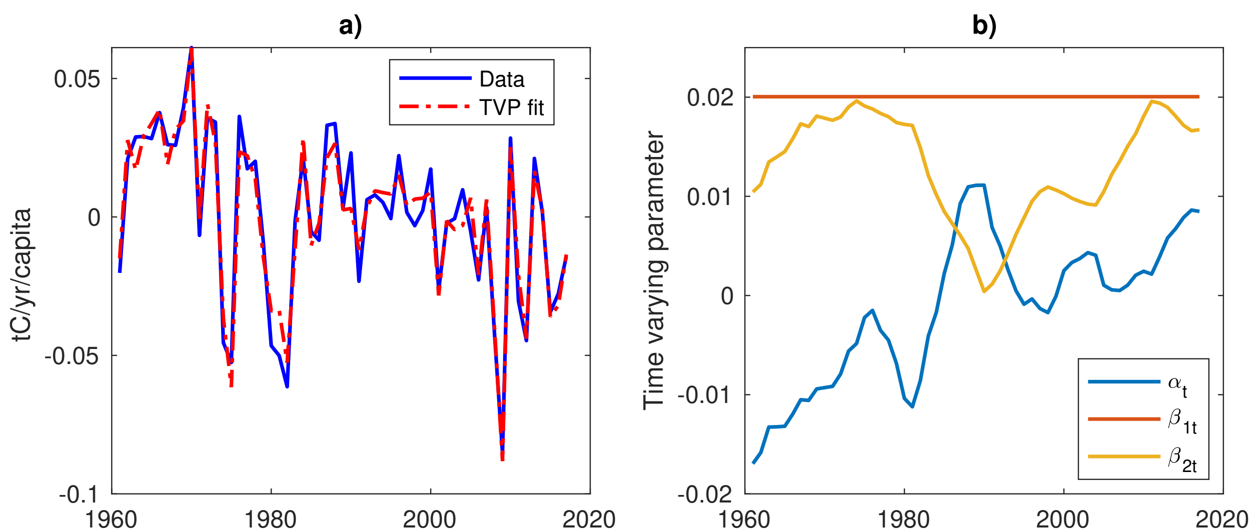

Figure 9: *TVP model. a): Fit of model. b): Smoothed estimates of α_t and β_t.*