DEPARTMENT OF ECONOMICS
AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

Designing a sequential testing procedure for verifying global CO2 emissions

Mikkel Bennedsen

CREATES Research Paper 2020-01

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Designing a sequential testing procedure for verifying global CO_2 emissions*

Mikkel Bennedsen†

January 14, 2020

Abstract

Following the Paris Agreement, most countries have agreed to reduce their CO_2 emissions according to individually set Nationally Determined Contributions (NDCs). However, national CO_2 emissions are reported by individual countries, and cannot be directly measured or verified by third parties. This engenders a potential misreporting problem, where nations that are not living up to their Paris commitments could, by under-reporting emissions, nevertheless appear to be fulfilling their NDC targets. This paper uses the theory of sequential testing to design a statistical CO_2 monitoring procedure, that can detect systematic misreportings of CO_2 emissions. The data series that we monitor is the so-called carbon budget imbalance, which is a time series derived from reported CO_2 emissions and independently measured Earth system data. We show that, when emissions are truthfully reported, the budget imbalance constitutes a stationary process, while, if emissions become systematically misreported, a structural break occurs. Our proposed procedure monitors the budget imbalance data and sequentially tests the null that the budget imbalance is stationary; rejection of the null provides evidence for systematic misreportings of CO_2 emissions. By constructing the procedure appropriately, detection time can be made sufficiently fast to help inform the 5 yearly global "stocktake" of the Paris Agreement.

Keywords: CO_2 emissions; Paris agreement; Global Carbon Budget; sequential testing.

JEL Codes: Q54. C12.

*The author would like to thank Eric Hillebrand, Siem Jan Koopman, and participants in the session on "Climate change mitigation, impacts, and adaptation" at the European Geoscience Union (EGU) General Assembly, Vienna, 2019, for many helpful comments regarding this manuscript. Financial support from the Independent Research Fund Denmark for the project "Econometric Modeling of Climate Change" is acknowledged.

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

1 Introduction

The Paris Agreement of 2015 instituted a transnational commitment to limit global temperature rise to between 1.5 and 2.0 degrees centigrade above pre-industrial levels (UNFCCC, 2015). It is widely accepted that to achieve this goal, substantial reductions of anthropogenic CO_2 emissions are needed (Millar et al., 2017; Tokarska and Gillett, 2018). Indeed, the recent IPCC report (IPCC, 2018b) states that to stay below 1.5°C, emissions should be reduced by almost half by 2030 (from 2010 levels) with a level close to zero in 2050 (Sanderson et al., 2016; Tanaka and O'Neill, 2018; Luderer et al., 2018).

Reducing emissions substantially requires all nations to work towards this goal, particularly the nations that are currently emitting the most (Larkin et al., 2018). The Paris Agreement therefore requires signing parties to deliver mandatory emissions reports, which are to be assessed during 5 yearly "stocktakes" of the global emissions status. Unfortunately, since data on CO_2 emissions are *reported* by the nations themselves, instead of being *measured* by the global community, this could create incentives for individual nations to misreport emissions (Peters et al., 2017). In this way, nations that are not living up to their Paris commitments could, by misreporting their CO_2 emissions, nevertheless appear to be fulfilling their Nationally Determined Contribution (NDC) targets. This is especially worrisome, as some countries have notoriously opaque emissions reporting and verification practices (Guan et al., 2012; Duflo et al., 2013; Transparency International, 2013; Ghanem and Zhang, 2014; Korsbakken et al., 2016; Nature, 2018; Zhang et al., 2019). Indeed, the problem of verifying the reported CO_2 emissions was one of the key topics discussed at the recent COP24 meeting in Katowice, Poland (IPCC, 2018a).

The aim of this paper is to suggest a statistically rigorous procedure to verify global anthropogenic CO_2 emissions, a problem which has not received much research attention in the climate literature (Peters et al., 2017). By employing the theory of sequential testing (e.g., Page, 1954), or "monitoring" (e.g., Chu et al., 1996), the procedure uses available climate data, collected independently of emissions data, to verify reported anthropogenic CO_2 emissions sequentially in time as new data become available. To do this, we exploit the idea of a balanced carbon budget (Friedlingstein et al., 2019): the amount of CO_2 that is emitted to the atmosphere must equal the amount of CO_2 absorbed in the three carbon sinks, namely the atmosphere, the terrestrial biosphere, and the oceans. This insight gives rise to the carbon budget equation

$$E_t^{ANT} = G_t^{ATM} + S_t^{OCN} + S_t^{LND} + B_t^{IM}, \tag{1.1}$$

where E_t^{ANT} is year-t anthropogenic CO_2 emissions, and $G_t^{ATM}, S_t^{OCN}, S_t^{LND}$ denote the year-t uptake of CO_2 in the atmosphere, the oceanic carbon sink, and the terrestrial ("land") carbon sink, respectively. In theory, because the carbon system is closed, the total CO_2 flux to the three carbon sinks must be equal to the amount of anthropogenic emissions, which implies that $B_t^{IM} = 0$ in (1.1). Indeed, originally, the budget equation was simply stated as $E_t^{ANT} = G_t^{ATM} + S_t^{OCN} + S_t^{LND}$ (e.g., Le Quéré et al., 2016). However, due to measurement errors in the various data sources making up the carbon budget, i.e. $E_t^{ANT}, G_t^{ATM}, S_t^{OCN}$, and S_t^{LND}, this equation will not hold in practice, when the measurements of the various terms are inserted. For this reason Le Quéré et al. (2018)

introduced the residual term B_t^{IM}, dubbed the *budget imbalance*, into the carbon budget equation, making the equation (1.1) balanced at all times.

When emissions are truthfully reported, we do not expect the time series of the budget imbalance B_t^{IM} to contain any large systematic biases (Le Quéré et al., 2018). Below we perform a statistical analysis of the observed budget imbalance B_t^{IM} from $t = 1959$ to $t = 2018$, and do indeed find that these data are historically well-described by a zero-mean stationary process. In contrast, when emissions are *not* truthfully reported, we show that the budget imbalance B_t^{IM} will undergo a *structural break*: some non-stationary process will be introduced into the data. These insights allow us to cast the problem of verifying reported emissions data as a sequential testing problem, and thus draw on well-known results in this field. Our procedure monitors the residuals of the global carbon budget, i.e. B_t^{IM}, through time, and decide whether or not a structural break has occurred. In effect, we sequentially test the null hypothesis that reported emissions E_t^{ANT} are compatible with the independently measured Earth system data G_t^{ATM}, S_t^{OCN}, and S_t^{LND}. If, for some time period t, this null hypothesis can be rejected, we conclude that there is evidence for E_t^{ANT} being systematically misreported.

Using simulations, we illustrate the use of the theoretical results proposed in the paper and investigate their finite sample performance. We find that, under realistic conditions, our monitoring scheme is able to detect misreporting quickly and with high probability, while having controlled size. Indeed, the simulations indicate that the empirical size of our proposed test is slightly below the nominal level when the null of no misreporting is true, i.e. when CO_2 emissions are reported truthfully. The empirical (power) properties of the test when the alternative is true, i.e. when CO_2 emissions are misreported, depend on the magnitude of misreporting. We find that when the magnitude of misreporting is very small, misreporting can be difficult to detect in practice. For moderately larger magnitudes of misreporting, however, the mean detection time of the method is on the order of 5 years, which is the frequency at which the Paris "stocktakes" take place. Consequently, the method proposed in this paper can potentially help the global community in future efforts of verifying reported CO_2 emissions.

The rest of the paper is structured as follows. Section 2 is theoretical, containing the proposed monitoring procedure and its theoretical justification, and a reader only interested in the practicalities and implementation of the testing procedure can skip this section. Section 3 briefly reviews the data that we work with and reports the results of a statistical analysis of the residual from the carbon budget, i.e. the budget imbalance B_t^{IM}. Section 4 illustrates the practical use of the sequential testing procedure on simulated and real data. An Appendix, containing a proof of the main theoretical result, as well as a detailed statistical analysis of the budget imbalance data, is given at the end. An online Supplementary Material file contains additional details and simulation results.[1] A MATLAB software package is available online, with which the methods of the paper can be easily implemented and adapted (Bennedsen, 2020).

[1]The Supplementary Material file is available at https://sites.google.com/site/mbennedsen/research/monitoring_gcb_v36_supplementary.pdf.

2 Sequential monitoring procedure

Let Y_t denote a time series of observations of a stochastic process, given by

$$Y_t = \begin{cases} u_t & t = 1, 2, \ldots, \tau - 1, \\ u_t + \epsilon_t^* & t = \tau, \tau + 1, \ldots, \end{cases} \tag{2.1}$$

where u_t is a stationary stochastic process and ϵ_t^* is a possibly non-stationary process. We assume that Y_t is observed over an initial time period $t = 1, \ldots, K$, where it is known that $\epsilon_t^* = 0$. The initial period is used to estimate the long run variance of u_t and then the monitoring algorithm is initiated from time $t = K + 1$. At some later (unknown) time $\tau \geq K + 1$, a structural break occurs and the process ϵ_t^* is introduced into the observations. If no structural break occurs, set $\tau = \infty$.

Technically, we make the following assumption on the process u_t.

Assumption 2.1. *The process u_t is a zero-mean stationary process satisfying a functional central limit theorem of the form*

$$(\lambda \mapsto \sum_{i=1}^{\lfloor \lambda K \rfloor} u_i) \Rightarrow (\lambda \mapsto \omega_u B(\lambda)), \quad \lambda > 0,$$

where ω_u^2 is the long run variance of u_t and $B(\cdot)$ is a Brownian motion. Here, "$\Rightarrow$" means convergence in distribution.

Assumption 2.1 is satisfied for a wide variety of stochastic processes u_t. For instance, u_t can be a stationary martingale difference sequence or u_t can be a "short memory" linear process. An AR(1) process with autoregressive parameter $|\phi| < 1$ falls into this latter category. More general and high-level conditions, such as mixing conditions, are also sufficient. We refer the interested reader to Davidson (2006) for a comprehensive list of sufficient conditions for Assumption 2.1 to hold.

Our goal is to design a monitoring scheme, which, at each time period t, conducts a statistical test for whether $t \geq \tau$, i.e. for whether a structural break has occurred. Monitoring can be done over an indefinte time horizon or over a fixed time horizon. Let $\Lambda > 0$ be a constant denoting the length of the monitoring period, compared to the length of the initial period K.[2] That is, we monitor over the period $\mathbb{F}_{K,\Lambda}$, where

$$\mathbb{F}_{K,\Lambda} := \{K + 1, K + 2, \ldots, K(\Lambda + 1)\}.$$

We are interested in testing the null hypothesis

$$H_0 : \tau > K(\Lambda + 1),$$

i.e. that there is no structural break in the monitoring period $\mathbb{F}_{K,\Lambda}$, against the alternative

$$H_1 : \tau \in \mathbb{F}_{K,\Lambda},$$

[2] It is straight forward to adapt our methods to allow for an open-ended monitoring period, i.e. to $\Lambda = \infty$, cf. Remark 2.1.

i.e. that there is a structural break in the monitoring period.

Our proposed scheme is based on the following statistic

$$\tilde{Z}_t := \frac{1}{\sqrt{K\hat{\omega}_K^2}} \sum_{i=K+1}^{t} Y_i, \qquad t \geq K+1, \tag{2.2}$$

where $\hat{\omega}_K^2$ is an estimate of the long run variance of u_t using the first K observations $\{Y_i\}_{i=1}^K$, which are equal to $\{u_i\}_{i=1}^K$ by assumption. In other words, the test statistic is the cumulated sum of the observations after monitoring has started, scaled appropriately.

Before we state the limit theory as it relates to our setup, we need to introduce the concept of a boundary function. Specifically, we consider functions h satisfying the following assumption.

Assumption 2.2. *The function h, defined on $(1,\infty)$, is such that $h(\lambda) > 0$ for all $\lambda > 1$.*

Remark 2.1. To allow for $\Lambda = \infty$, it is necessary to assume some more stringent assumption on the boundary function h, see, e.g., the assumptions imposed in Chu et al. (1996).

The main theoretical result which will allow us to construct monitoring procedures is as follows. The proof relies on results in Davidson (2006) and Leisch et al. (2000). The details are given in Appendix A.

Theorem 2.1. *Let $\tilde{Z}_t$ be given by (2.2), where u_t and the boundary function h satisfy Assumptions 2.1 and 2.2, respectively, and let $\hat{\omega}_K^2$ be a consistent estimator of ω_u^2 as $K \to \infty$. Suppose H_0 is true and let $\Lambda > 0$. Then,*

$$\lim_{K\to\infty} P\left(|\tilde{Z}_t| \geq h(t/K), \text{ for some } t \in \mathbb{F}_{K,\Lambda}\right) = P\big(|B(\lambda)| \geq h(1+\lambda), \text{ for some } \lambda \in (0,\Lambda]\big),$$

where $B(\cdot)$ is a Brownian motion.

Theorem 2.1 shows that we can use the statistic (2.2) to sequentially test H_0 against H_1. To be precise, let $\alpha \in (0, 1/2]$ denote the desired nominal significance level of the test, and choose the function h such that

$$\alpha = P\big(|B(\lambda)| \geq h(1+\lambda), \text{ for some } \lambda \in (0,\Lambda]\big),$$

where B is a Brownian motion. Now, by Theorem 2.1, for $t = K+1, K+2, \ldots, K(\Lambda+1)$, the sequential rule

"reject H_0 in favor of H_1 if $|\tilde{Z}_t| \geq h(t/K)$",

will, asymptotically, result in a test of H_0 with size α.

Conversely, if $\epsilon_t^* \neq 0$ for $t \geq \tau$, a structural break will be introduced in Y_t. As we will see below, cf. also Remark 3.1, realistic specifications for ϵ_t^* will quickly result in large values of the test statistic $|\tilde{Z}_t|$, thus leading to rejection of the null hypothesis. In other words, after settling on the length of the monitoring period (which can be infinite, cf. Remark 2.1) and a nominal significance level, it is only necessary to track the value of $|\tilde{Z}_t|$ and in each period compare it to

the value of the boundary $h(t/K)$. If the test statistic exceeds the boundary at any time in the monitoring period, there is statistical evidence for a structural break in Y_t at significance level α. The behavior of the test under H_1, i.e. when there is misreporting, will depend on the exact form of ϵ_t^* and on the chosen boundary function.

As we will see below, the situation most relevant to our application is when ϵ_t^* is a negative process. This implies that it is only necessary to test if $\tilde{Z}_t$ is sufficiently negative, as compared to testing whether $|\tilde{Z}_t|$ is sufficiently positive. In particular, the test may gain some power against the alternative of a negative test statistic, compared to the double-sided alternative. A straight forward consequence of Theorem 2.1 is the following one-sided version of the testing procedure.

Corollary 2.1. *Suppose the setup of Theorem 2.1. Then,*

$$\begin{aligned}\lim_{K\to\infty} & P\left(\tilde{Z}_t < -h(t/K), \text{ for some } t \in \mathbb{F}_{K,\Lambda}\right) \\ &= P\big(B(\lambda) < -h(1+\lambda), \text{ for some } \lambda \in (0,\Lambda]\big) \\ &= \frac{1}{2}P\big(|B(\lambda)| > h(1+\lambda), \text{ for some } \lambda \in (0,\Lambda]\big).\end{aligned}$$

Proof. The second line follows from similar arguments as those of Theorem 2.1. The third line follows since the law of $B(t)$ is symmetric around 0. □

2.1 The boundary function

The researcher has considerable freedom when choosing the boundary function, as long as it satisfies Assumption 2.2. The choice of function can thus be tailored to the objective the researcher has. The Supplementary Material contains an extensive discussion on boundary functions, including simulation studies motivating our preferred choice of boundary function, which is as follows:[3]

$$h(\lambda) = c \cdot \sqrt{\lambda(\lambda-1)\left(1+\log\left(\frac{\lambda}{\lambda-1}\right)\right)}, \quad \lambda > 1. \tag{2.3}$$

The constant $c = c_{\Lambda,\alpha} > 0$ depends on the length of the monitoring period, i.e., on Λ, and on the nominal significance level, α, of the test of H_0. Appendix B contains details on how to calculate the constant c given Λ and α in the context of the one-sided test of Corollary 2.1.[4] We also supply a computer program that can do these calculations automatically (Bennedsen, 2020).

Remark 2.2. From (2.2) and (2.3), we see that we can base our test on the more simple test statistic

$$Z_t := \sum_{i=K+1}^{t} Y_i, \qquad t \geq K+1,$$

in place of $\tilde{Z}_t$, provided we use the alternative boundary function (or "critical value function"):

$$C_t^\alpha := \sqrt{K\hat{\omega}_K^2} \cdot h(t/K), \quad t \geq K+1, \tag{2.4}$$

[3] For a given $t = K+1, K+2, \ldots$, the corresponding value for λ is $\lambda = \lambda_t = \frac{t}{K}$.

[4] The Supplementary Material contains more extensive details on this procedure, including implementation of the two-sided test as well as implementation in the case of alternative boundary functions h.

where $h(\cdot)$ is given by (2.3). This is the approach taken in what follows. Consequently we will use the simpler test statistic Z_t and the appropriately modified boundary function (2.4) in the remainder of the paper.

Remark 2.3. The critical value function (2.4) requires an estimate of the long run variance ω_K^2, obtained from the first K observations of the data $\{Y_t\}_{t=1}^K$. We will see below that a good model for $\{Y_t\}_{t=1}^K$ in our application is an AR(1) process, which has long run variance given by

$$\omega_{AR(1)}^2 = \frac{\sigma_\epsilon^2}{(1-\phi)^2},$$

and can thus be straightforwardly estimated given σ_ϵ^2 and ϕ. These parameters can in turn be estimated from a (ordinary least squares) regression of Y_{t+1} on Y_t for $t = 1, 2, \ldots, K-1$. See Section 3.1 for an example of this, where we obtain $\widehat{\sigma_\epsilon^2} = 0.52$ and $\hat{\phi} = 0.38$ and thus $\hat{\omega}_K^2 = \hat{\omega}_{AR(1)}^2 = 1.37$. In our empirical applications, when computing the critical values of the monitoring test through (2.4), we use this AR(1)-based approach to estimate the long run variance ω_K^2. Note, however, that the methods proposed above apply equally well when ω_K^2 is estimated using other (consistent) means. For instance, if the parametric AR(1) assumption on the dynamics of Y_t is believed to be too strong, ω_K^2 can be estimated non-parametrically. This possibility is explored in the Supplementary Material.

3 Data

The global carbon budget is a physical accounting identity, describing that the amount of anthropogenically emitted CO_2 in a given time period, must equal the total flux of CO_2 in the atmosphere, the oceans, and the terrestrial biosphere (the so-called "carbon sinks"). The global carbon budget is thus given by the equation (Friedlingstein et al., 2019)

$$E_t^{FF} + E_t^{LUC} = G_t^{ATM} + S_t^{OCN} + S_t^{LND} + B_t^{IM}, \tag{3.1}$$

where E_t^{FF} is CO_2 emissions from fossil fuel burning, cement production, and gas flaring; E_t^{LUC} is CO_2 emissions from land-use change; G_t^{ATM} is growth of atmospheric CO_2 concentration; S_t^{OCN} is the flux of CO_2 from the atmosphere to the oceans; and S_t^{LND} is the flux of CO_2 from the atmosphere to the terrestrial biosphere. We use the data set provided by The Global Carbon Project (Friedlingstein et al., 2019).[5] The fossil fuel emissions data E_t^{FF} are from Boden et al. (2018), while the land-use change data, E_t^{LUC} are averages over the model-based estimates of Hansis et al. (2015) and Houghton and Nassikas (2017), updated as in Friedlingstein et al. (2019). The growth rate in atmospheric CO_2 data, G_t^{ATM}, is from Dlugokencky and Tans (2018), while the sink data, S_t^{OCN} and S_t^{LND}, are averages over several independent model-based estimates, constructed as explained in Friedlingstein et al. (2019). We define the total amount of anthropogenic CO_2 emissions, E_t^{ANT}, as the sum of fossil fuel emissions, cement production, and gas flaring and emissions from land-use change. That is, $E_t^{ANT} := E_t^{FF} + E_t^{LUC}$.

[5]The data are available at `http://www.globalcarbonproject.org/` and were downloaded on December 22, 2019.

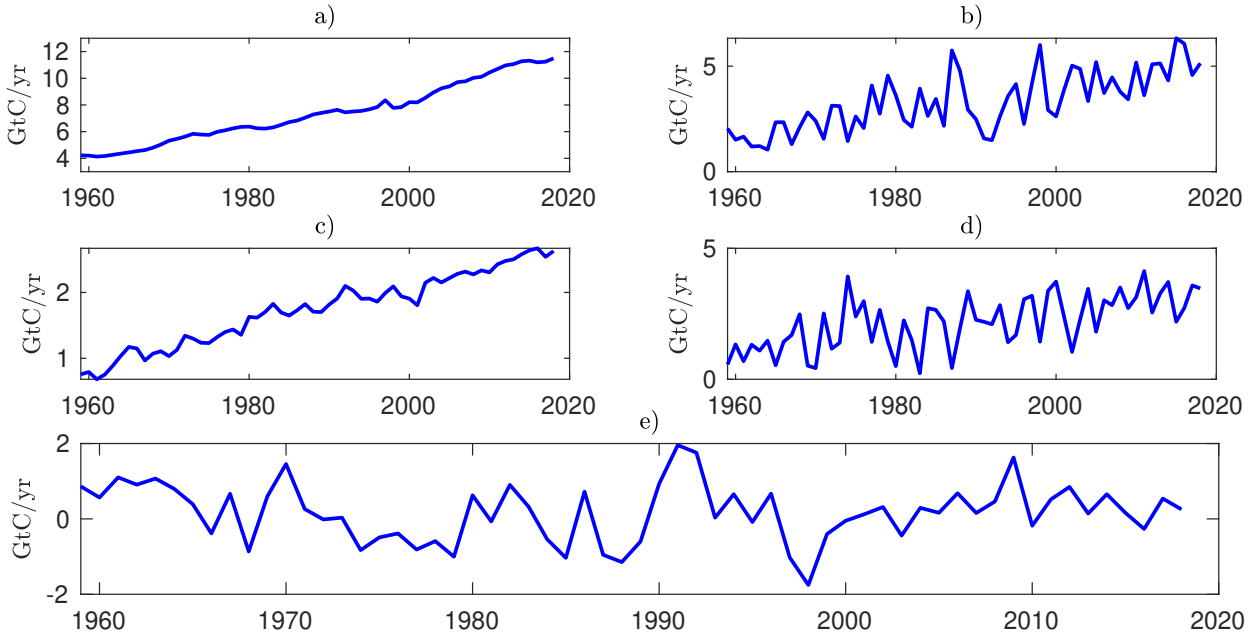

Figure 1: *Time series data from the carbon budget, Equation* (3.1)*, from* 1959 *to* 2018*. a): Total anthropogenic emissions,* $E_t^{ANT} := E_t^{FF} + E_t^{LUC}$*. b): Atmospheric growth,* G_t^{ATM}*. c): Ocean sink flux,* S_t^{OCN}*. d): Terrestrial sink flux,* S_t^{LND}*. e): Budget imbalance,* B_t^{IM}*.*

The quantity

$$B_t^{IM} = E_t^{ANT} - G_t^{ATM} - S_t^{OCN} - S_t^{LND}. \tag{3.2}$$

is the so-called budget imbalance. It is implicitly defined so as to balance the carbon budget equation (3.1). In principle, the budget should be balanced at all times, so that $B_t^{IM} = 0$, but due to measurement errors in the sources and sinks of CO_2, the budget imbalance will in general be non-zero. The next section presents the results of a statistical analysis of the time series properties of the budget imbalance B_t^{IM}.

All data are given in gigatonnes of carbon (GtC) and are recorded at a yearly frequency, beginning in 1959 and ending in 2018, resulting in 60 observations for each term in (3.1). Figure 1 plots the time series of the variables from Equation (3.1) from $t = 1959$ to $t = 2018$.

3.1 The budget imbalance and its statistical properties

Figure 1e) plots the budget imbalance data B_t^{IM}, $t = 1959, \ldots, 2018$. In Appendix C, an in-depth statistical analysis of the time series properties of these data are conducted, the main points of which we briefly review here.

Firstly, there is statistical evidence that the data are stationary with zero mean and positive autocorrelation. Secondly, in the autoregressive moving average (ARMA) class of models, the Bayesian Information Criterion (BIC) of Schwarz (1978) indicates that the best fitting parametric model is a zero-mean autoregressive process of order one (AR(1)). Subsequent tests on the residuals

from this model confirm that it provides a good fit to the data. That is, a reasonable statistical model of the historical budget imbalance time series data is

$$B_t^{IM} = \phi B_{t-1}^{IM} + \epsilon_t,$$

where ϵ_t is an iid noise sequence and $\phi \in \mathbb{R}$. Our estimate (obtained by an ordinary least squares regression) of the autoregressive parameter is $\hat{\phi} = 0.38$ and for the variance of ϵ_t, $\widehat{Var(\epsilon_t)} = 0.52$. See Appendix C for further details. As mentioned above, the sequential testing framework of this paper does not hinge on this parametric assumption. Indeed, the methods proposed below are valid under significantly weaker (e.g., non-parametric) assumptions as well. See Remark 2.3 and the Supplementary Material for details.

3.2 The budget imbalance when emissions are misreported

Suppose that from some time point τ, anthropogenic CO_2 emissions are misreported as the amount $E_t^{ANT,*}$, while the true value emitted to the biosphere is $E_t^{ANT} \neq E_t^{ANT,*}$. Then, for $t \geq \tau$, the observed budget imbalance data become

$$\begin{aligned} B_t^{IM,*} &= E_t^{ANT,*} - G_t^{ATM} - S_t^{OCN} - S_t^{LND} \\ &= u_t + \epsilon_t^*, \end{aligned}$$

where

$$u_t = E_t^{ANT} - G_t^{ATM} - S_t^{OCN} - S_t^{LND},$$

is the budget imbalance under the true emission path E_t^{ANT}, while

$$\epsilon_t^* = E_t^{ANT,*} - E_t^{ANT},$$

denotes the amount of misreporting in CO_2 emissions at time $t \geq \tau$.

In this case, the budget imbalance will take the form

$$B_t^{IM,*} = \begin{cases} u_t & t < \tau, \\ u_t + \epsilon_t^* & t \geq \tau. \end{cases} \tag{3.3}$$

Equation (3.3) shows that at the (unknown) time $t = \tau$, the budget imbalance time series will undergo a *structural break*: it will go from being a zero-mean stationary process to being the sum of this process and the term ϵ_t^*. The upshot is that the reported budget imbalance data $Y_t = B_t^{IM,*}$ can be expected to conform with the theoretical monitoring framework presented in Section 2 cf. Equations (2.1) and (3.3). This shows that we can use the monitoring methodology of Section 2 to verify reported CO_2 emissions. Section 4 gives the practical details.

Lastly, note that the arguably most important case is when emissions are *under*-reported, i.e., when $E_t^{ANT,*} < E_t^{ANT}$. This implies that the structural break term $\epsilon_t^* = E_t^{ANT,*} - E_t^{ANT}$ will be a *negative* process. This insight motivates our use of the one-sided testing procedure as given in Corollary 2.1.

Remark 3.1. The properties of the process ϵ_t^* will decide the effect of the structural break on the observations. A setup which we find particularly useful and reasonably realistic is the following. For $t \geq \tau$, assume that true emissions grow at a constant rate $g \in (-1, \infty)$, i.e.,

$$E_t^{ANT} = (1+g)^{t-\tau+1} E_{\tau-1}^{ANT}, \qquad t = \tau, \tau+1, \ldots,$$

while (mis)reported emissions grow at the rate $m \in (-1, \infty)$, relative to the value for emissions before misreporting began:

$$E_t^{ANT,*} = (1+m)^{t-\tau+1} E_{\tau-1}^{ANT}, \qquad t = \tau, \tau+1, \ldots.$$

For instance, if actual emissions grow at a rate of 1% per year while emissions are reported fall 2% per year, then $g = 0.01$ and $m = -0.02$. Such a situation could follow due to some agreement, such as the Paris Agreement, where emissions are agreed (and therefore reported) to fall, i.e. that $m < 0$. If the reported value differs from the actual value such that $g > m$, i.e. that actual emissions are above reported emissions, then

$$\epsilon_t^* = E_t^{ANT,*} - E_t^{ANT} = \left[(1+m)^{t-\tau+1} - (1+g)^{t-\tau+1}\right] E_{\tau-1}^{ANT} < 0, \quad t = \tau, \tau+1, \ldots. \tag{3.4}$$

4 Implementation and numerical investigations

As shown in the previous section, the budget imbalance data $B_t^{IM,*}$, implied by the reported emissions data $E_t^{ANT,*}$ and the independent data on the Earth system variables G_t^{ATM}, S_t^{OCN}, and S_t^{LND}, through Equation (3.2), are likely to conform to the setup of Section 2. Therefore, to sequentially verify whether reported CO_2 emissions are compatible with the Earth system data, we can use the monitoring theory presented above. In practice, Remark 2.2 tells us that we should monitor the cumulated sum of the budget imbalance data since the monitoring period started, i.e.

$$Z_t = \sum_{i=K+1}^{t} B_i^{IM,*}, \qquad t \geq K+1,$$

and, at each time point t, check whether Z_t has crossed the critical boundary C_t^α of Equation (2.4). As mentioned above, we are especially interested in the alternative hypothesis that CO_2 emissions are under-reported; hence we reject the null when Z_t becomes smaller than C_t^α. That is, if for some t it is the case that $Z_t < C_t^\alpha$, then the null hypothesis that CO_2 emissions are accurately reported can be rejected at an α significance level. Details on how to calculate C_t^α are given in Appendix B, and a computer program written in the MATLAB programming language, which performs the calculation of the critical values C_t^α automatically, is supplied online (Bennedsen, 2020).

Section 4.1 illustrates the use of this method in a simulation study, which also serves to estimate the mean detection time of the test under various assumptions on the amount of misreporting being conducted (i.e. on ϵ_t^*). Section 4.2 uses all currently available data to set up the critical values for a monitoring procedure, which can be used in practice to verify global CO_2 emissions going forward.

4.1 Estimating mean detection time through simulations

The goal of this section is to investigate how the proposed monitoring scheme will perform under realistic conditions going forward. We simulate future paths of the budget imbalance and compare the results from the test when emissions are reported correctly and when they are misreported.

To be precise, we use the historical budget imbalance data B_t^{IM}, $t = 1959, \ldots, 2018$, as initial data (implying $K = 60$) and then simulate 10 000 different future paths of B_t^{IM}, $t = 2019, \ldots, 2078$. That is, we set $B_t^{IM} = u_t + \epsilon_t^*$, where u_t is simulated as a stationary autoregressive process of order one and $\epsilon_t^* = E_t^{ANT,*} - E_t^{ANT}$, where $E_t^{ANT,*}$ is the reported emissions and E_t^{ANT} is the actual emissions. The autoregressive parameter for u_t is set to $\phi = 0.38$, while the variance of error term of u_t is set to $\sigma^2 = 0.52$, which are the parameters we estimated from the initial budget imbalance data in Section 3.1.

To achieve the Paris objectives, CO_2 emissions should be cut in approximately half, as compared to 2010 levels, by 2030 (Sanderson et al., 2016; Luderer et al., 2018). Since $E_{2010}^{ANT} = 10.44$ GtC and $E_{2018}^{ANT} = 11.4905$ GtC, this means that CO_2 emissions should decrease by 6.37% each year from 2019 onwards. This is our baseline scenario: in the simulations to come, we suppose that reported CO_2 emissions, $E_t^{ANT,*}$, are decreasing with 6.37% each year for $t = 2019, 2020, \ldots$. We consider a range of different scenarios for actual emissions, E_t^{ANT}, where each scenario will imply a different behavior for the structural break process $\epsilon_t^* = E_t^{ANT,*} - E_t^{ANT}$. We use the setup discussed in Remark 3.1; since we assume that emissions are reported to decrease by 6.37% per year, this entails that $m = -0.0637$. The parameter g denotes the growth rate of actual emissions, E_t^{ANT}, for $t = 2019, 2020, \ldots$. We consider $g \in \{-0.0637, -0.0537, -0.0437, -0.0337, -0.0237, -0.0137, 0\}$. Here, $g = -0.0637$ corresponds to accurately reported emissions (no misreporting) and $g = 0$ corresponds to constant emissions (6.37% under-reporting per year). The remaining values of g each correspond to a certain magnitude of yearly misreporting; for instance, $g = -0.0537$ corresponds to 1% under-reporting, $g = -0.0437$ corresponds to 2% under-reporting, and so forth.

The resulting paths of the emissions, for the cases $g = -0.0637$ (no misreporting) and $g = -0.0437$ (2% under-reporting per year), are shown in Figures 2a) and 2c), respectively. The solid blue line corresponds to the "actual emissions", i.e. E_t^{ANT}, while the dashed green line is the "reported emissions", i.e. $E_t^{ANT,*}$. Figures 2b) and 2d) present 100 example paths of the simulated test statistic $Z_t = \sum_i B_i^{IM} = \sum_i u_i + \sum_i \epsilon_i^*$ (solid cyan lines), obtained from the simulated budget imbalances and the CO_2 emission scenarios just described. The black lines indicate the one-sided critical values, C_t^α, for $\alpha = 5\%, 10\%$, and 32%, calculated using Corollary 2.1 cf. also Remark 2.2. We set $\Lambda = 1$, which correspond to monitoring CO_2 emissions for $K \cdot \Lambda = 60$ years, i.e. from $t = 2019$ to $t = 2078$. The null hypothesis is rejected if $Z_t < C_t^\alpha$ for some $t = 2019, 2020, \ldots, 2078$.

In the first scenario, cf. a) and b) of Figure 2, emissions are reported truthfully. Consequently, when Z_t crosses the critical boundary, it will result in a Type I error (a "false positive"): we will reject the null of no misreporting even in this case when emissions are truthfully reported. The false positive rates in the simulation experiment are 2.24%, 4.95%, and 18.79% for $\alpha = 5\%, 10\%$, and 32%, respectively, implying that the test is slightly under-sized in this setup. When emissions are under-reported, i.e. when $g > m$, such as in the case shown in c) and d) of Figure 2, the null is false. Hence, failure to reject the null will here result in a Type II error (a "false negative").

When $g = -0.0537$, i.e. when there is under-reporting of 1% per year, the Type II error rates are 1.36%, 6.92%, and 12.54% for $\alpha = 5\%, 10\%$, and 32%, respectively. When $g > -0.0537$, i.e. when the under-reporting is 2% or larger, the Type II error rates are all zero, meaning that misreporting is always detected before the end of the monitoring period.

Figure 3a) reports the estimated mean detection times as a function of the amount of misreporting, $m - g \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.0637\}$. Figure 3b) reports the associated standard deviations of the time until detection. Here, the false-negatives (Type II errors) have been removed for $m - g = 0.01$. (Recall that there were no false-negatives for $m - g \geq 0.02$.) From the figures, we see that when the magnitude of under-reporting is $m - g = 0.01$, the mean detection time is very large (15 to 25 years, depending on the significance level) and so is the standard deviation of the detection times. This indicates that if the magnitude of misreporting is very small, the misreporting can be difficult to detect in practice. Conversely, when $m - g = 0.02$, both mean detection time and the standard deviation of the detection time drop dramatically: the mean detection time is in this case between 7.5 and 11.5 years, depending on the significance level. For $m - g \geq 3\%$, the mean detection time is on the order of 5 years, indicated by the horizontal black line in Figure 3a), which is the time between the Paris "stocktakes" of the global emissions status. In these cases, the standard deviation of the detection time is also low, between 0.9 and 2 years, indicating that detection of misreporting is not only fast on average, but also reliable.

The Supplementary Material contains additional simulation results. There it is also shown that the test is correctly sized under H_0 when K is large.

4.2 Monitoring the future carbon budget

If we wish to monitor the carbon budget starting now, that is, starting when the 2019 data come in, Table 1 presents the critical values, C_t^α, for the test proposed in this paper. These are the critical boundaries which were used in the simulation experiment of Section 4.1, cf. Figure 2. To monitor the future carbon budget, we proceed as follows. Every year $t = 2019, 2020, \ldots$, when new data arrive, we calculate the budget imbalance using Equation (3.2), update the monitoring statistic Z_t, and compare it to the critical values given in Table 1. That is, we calculate the cumulative sum of the budget imbalances through time, $Z_t = \sum_{i=2019}^{t} B_i^{IM}$, and compare this to the appropriate critical value C_t^α in the table. If in some year t, the test statistic is below the corresponding critical value, i.e. if $Z_t < C_t^\alpha$, we reject the null of no misreporting against the alternative that under-reporting is taking place, at the given significance level α. In other words, if $Z_t < C_t^\alpha$ for some $t = 2019, 2020, \ldots$, we can conclude that there is statistical evidence that CO_2 emissions are being systematically under-reported.

In practice, it might be preferable to defer monitoring until more hard and fast commitments are made and misreporting becomes a serious issue to contend with. In this case, the critical values of Table 1 should be updated accordingly. This is easily done using the methods of Section 2. To facilitate easy application of our methods, we supply a simple computer program that can do this automatically (Bennedsen, 2020).

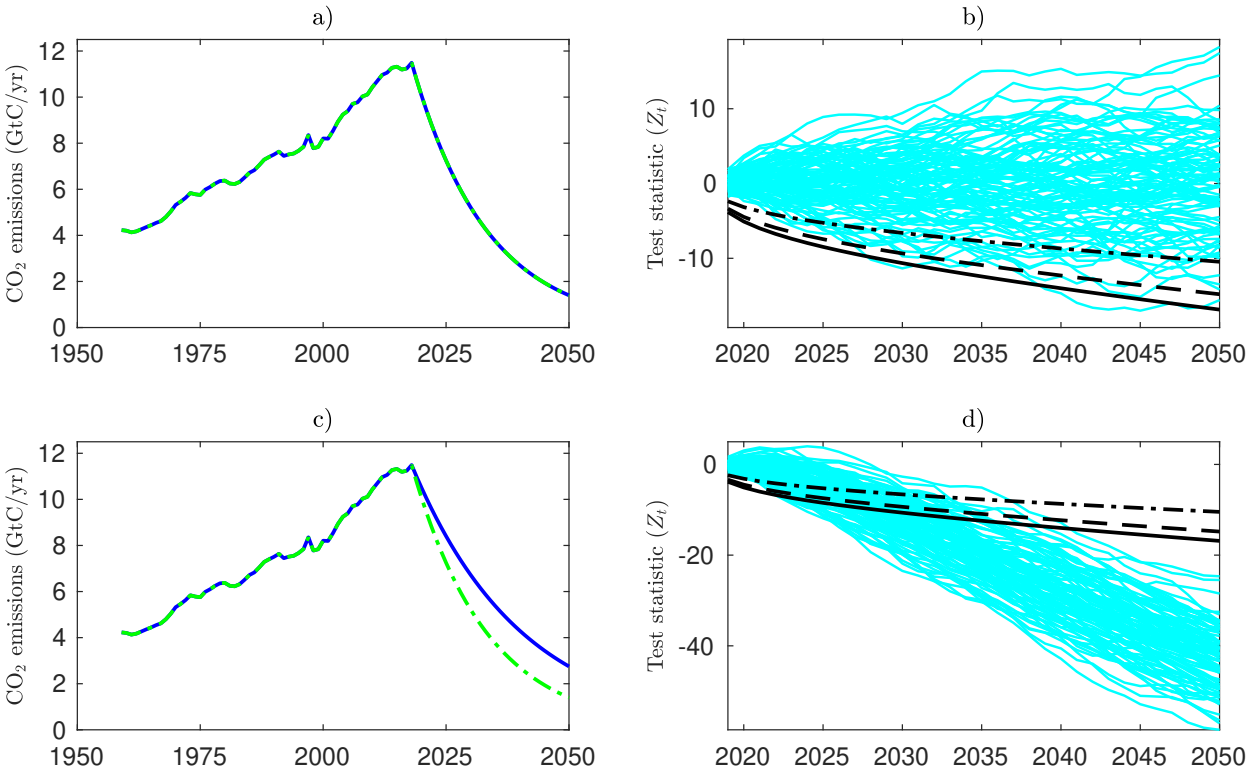

Figure 2: *Illustration of simulation study. From* 1959 *to* 2018 *real data are used; from* 2019 *to* 2050 *different emissions scenarios are considered and the budget imbalance is simulated using an AR(1) process, as explained in the text. Left panels show realized* CO_2 *emissions trajectories from* 1959 *to* 2018 *and hypothetical future* CO_2 *emissions trajectories from* 2019 *to* 2050*: actual emissions (blue solid line) and reported emissions (green dashed line). Right panels show* 100 *simulated paths of the test statistic* Z_t *(cyan lines) and critical boundaries* C_t^α *(black lines) for* $\alpha = 5\%$, 10%, *and* 32%, *from* 2019 *to* 2050*. a)+b): Both actual and reported emissions are decreasing by* 6.37% *each year (no misreporting). c)+d): Actual emissions are decreasing* 4.37% *each year, while emissions are reported to decrease by* 6.37% *each year (under-reporting of* 2% *per year).*

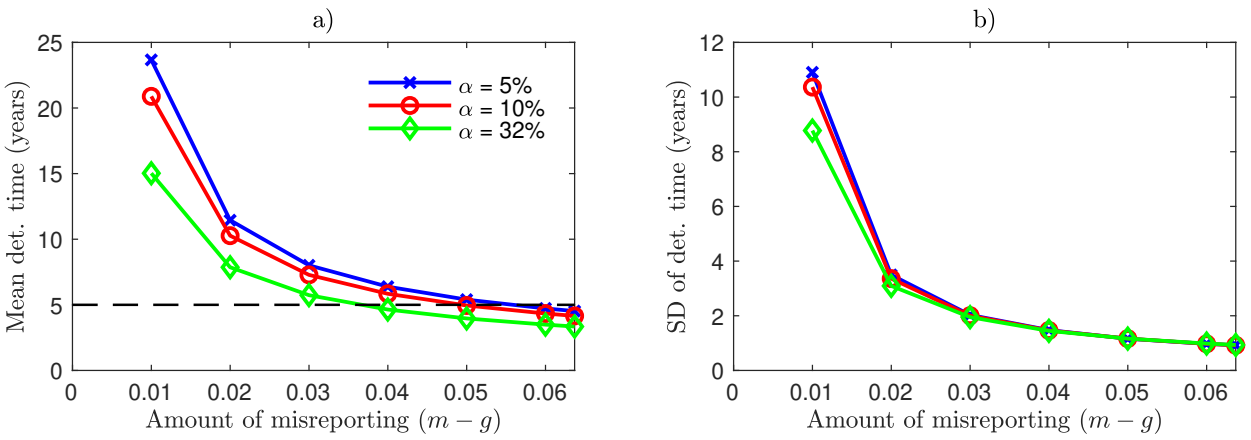

Figure 3: *Estimating the mean and standard deviation of the detection time using simulations. From* 1959 *to* 2018 *real data are used; from* 2019 *to* 2078 *different emissions scenarios are considered, which results in different magnitudes of misreportings,* $m - g$ *(x-axis), as explained in the text. a): Mean detection time as function of the amount of misreporting. The horizontal dashed black line denotes* 5 *years. b): Standard deviation of the detection time as function of the amount of misreporting. Three different significance levels are considered:* $\alpha = 5\%$ *(blue line, crosses),* $\alpha = 10\%$ *(red line, circles), and* $\alpha = 32\%$ *(green line, diamonds).*

Table 1: *Critical values* C_t^α *for the test of misreporting in* CO_2 *emissions. The values in the table have been calculated using* B_t^{IM}, $t = 1959, \ldots, 2018$, *as input.*

	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
$\alpha = 5\%$	−3.81	−5.06	−5.97	−6.71	−7.34	−7.91	−8.44	−8.92	−9.37	−9.80
$\alpha = 10\%$	−3.37	−4.47	−5.28	−5.93	−6.49	−7.00	−7.46	−7.89	−8.29	−8.66
$\alpha = 32\%$	−2.44	−3.24	−3.82	−4.30	−4.71	−5.07	−5.41	−5.72	−6.01	−6.28

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control 19*(6), 716–723.

Bennedsen, M. (2020). Software package for "Designing a sequential testing procedure for verifying global CO_2 emissions", available at https://sites.google.com/site/mbennedsen/research/Monitoring_GCB_Code.zip.

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: http://cdiac.ornl.gov/trends/emis/overview_2014.html.

Chu, C.-S. J., M. Stinchcombe, and H. White (1996). Monitoring structural change. *Econometrica 64*(5), 1045–1065.

Davidson, J. (2006). *Asymptotic Methods and Functional Central Limit Theorems*, Volume 1 of *Palgrave Handbook of Econometrics*, Chapter 5. Palgrave Macmillan UK.

Dlugokencky, E. and P. Tans (2018). Trends in atmospheric carbon dioxide. National Oceanic & Atmospheric

Administration, Earth System Research Laboratory (NOAA/ESRL), available at: http://www.esrl.noaa.gov/gmd/ccgg/trends/global.html.

Duflo, E., M. Greenstone, R. Pande, and N. Ryan (2013). Truth-telling by third-party auditors and the response of the polluting firms: Experimental evidence from India. *Quarterly Journal of Economics 128*, 1499–1545.

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

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *50*(4), 987–1007.

Friedlingstein, P., M. W. Jones, M. O'Sullivan, R. M. Andrew, J. Hauck, G. P. Peters, W. Peters, J. Pongratz, S. Sitch, C. Le Quéré, D. C. E. Bakker, J. G. Canadell, P. Ciais, R. B. Jackson, P. Anthoni, L. Barbero, A. Bastos, V. Bastrikov, M. Becker, L. Bopp, E. Buitenhuis, N. Chandra, F. Chevallier, L. P. Chini, K. I. Currie, R. A. Feely, M. Gehlen, D. Gilfillan, T. Gkritzalis, D. S. Goll, N. Gruber, S. Gutekunst, I. Harris, V. Haverd, R. A. Houghton, G. Hurtt, T. Ilyina, A. K. Jain, E. Joetzjer, J. O. Kaplan, E. Kato, K. Klein Goldewijk, J. I. Korsbakken, P. Landschützer, S. K. Lauvset, N. Lefèvre, A. Lenton, S. Lienert, D. Lombardozzi, G. Marland, P. C. McGuire, J. R. Melton, N. Metzl, D. R. Munro, J. E. M. S. Nabel, S.-I. Nakaoka, C. Neill, A. M. Omar, T. Ono, A. Peregon, D. Pierrot, B. Poulter, G. Rehder, L. Resplandy, E. Robertson, C. Rödenbeck, R. Séférian, J. Schwinger, N. Smith, P. P. Tans, H. Tian, B. Tilbrook, F. N. Tubiello, G. R. van der Werf, A. J. Wiltshire, and S. Zaehle (2019). Global Carbon Budget 2019. *Earth System Science Data 11*(4), 1783–1838.

Ghanem, D. and J. Zhang (2014). 'Effortless Perfection': Do Chinese cities manipulate air pollution data? *Journal of Environmental Economics and Management 68*(2), 203–225.

Guan, D., Z. Liu, S. Lindner, and K. Hubacek (2012). The gigatonne gap in China carbon dioxide inventories. *Nature Climate Change 2*, 672–675.

Hansis, E., S. J. Davis, and J. Pongratz (2015). Relevance of methodological choices for accounting of land use change carbon fluxes. *Global Biogeochem. Cy. 29*, 1230 – 1246.

Houghton, R. A. and A. A. Nassikas (2017). Global and regional fluxes of carbon from land use and land cover change 1850-2015. *Global Biogeochem. Cy. 31*, 456 – 472.

IPCC (2018a). Modalities, procedures and guidelines for the transparency framework for action and support referred to in Article 13 of the Paris Agreement. Technical report, Intergovernmental Panel on Climate Change, https://unfccc.int/sites/default/files/resource/APA-SBSTA-SBI.2018.Informal.2.Add_.6_1.pdf.

IPCC (2018b). Special Report on Global Warming of 1.5°C. Technical report, Intergovernmental Panel on Climate Change.

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

Korsbakken, J. I., G. P. Peters, and R. M. Andrew (2016, 3). Uncertainties around reductions in China's coal use and CO2 emissions. *Nature Climate Change 6*, 687–690.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics 54*(1), 159–178.

Larkin, A., J. Kuriakose, M. Sharmina, and K. Anderson (2018, 07). What if negative emission technologies fail at scale? Implications of the Paris Agreement for big emitting nations. *Climate Policy 18*(6), 690–714.

Le Quéré, C., R. M. Andrew, J. G. Canadell, S. Sitch, J. I. Korsbakken, G. P. Peters, A. C. Manning, T. A. Boden, P. P. Tans, R. A. Houghton, R. F. Keeling, S. Alin, O. D. Andrews, P. Anthoni, L. Barbero, L. Bopp, F. Chevallier, L. P. Chini, P. Ciais, K. Currie, C. Delire, S. C. Doney, P. Friedlingstein, T. Gkritzalis, I. Harris, J. Hauck, V. Haverd, M. Hoppema, K. Klein Goldewijk, A. K. Jain, E. Kato, A. Körtzinger, P. Landschützer, N. Lefèvre, A. Lenton, S. Lienert, D. Lombardozzi, J. R. Melton, N. Metzl, F. Millero, P. M. S. Monteiro, D. R. Munro, J. E. M. S. Nabel, S. Nakaoka, K. O'Brien, A. Olsen, A. M. Omar, T. Ono, D. Pierrot, B. Poulter, C. Rödenbeck, J. Salisbury, U. Schuster, J. Schwinger, R. Séférian, I. Skjelvan, B. D. Stocker, A. J. Sutton, T. Takahashi, H. Tian, B. Tilbrook, I. T. van der Laan-Luijkx, G. R. van der Werf, N. Viovy, A. P. Walker, A. J. Wiltshire, and S. Zaehle (2016). Global Carbon Budget 2016. *Earth System Science Data 8*(2), 605–649.

Le Quéré, C., R. M. Andrew, P. Friedlingstein, S. Sitch, J. Pongratz, A. C. Manning, J. I. Korsbakken, G. P. Peters, J. G. Canadell, R. B. Jackson, T. A. Boden, P. P. Tans, O. D. Andrews, V. K. Arora, D. C. E. Bakker, L. Barbero, M. Becker, R. A. Betts, L. Bopp, F. Chevallier, L. P. Chini, P. Ciais, C. E. Cosca, J. Cross, K. Currie, T. Gasser, I. Harris, J. Hauck, V. Haverd, R. A. Houghton, C. W. Hunt, G. Hurtt, T. Ilyina, A. K. Jain, E. Kato, M. Kautz, R. F. Keeling, K. Klein Goldewijk, A. Körtzinger, P. Landschützer, N. Lefèvre, A. Lenton, S. Lienert, I. Lima, D. Lombardozzi, N. Metzl, F. Millero, P. M. S. Monteiro, D. R. Munro, J. E. M. S. Nabel, S.-I. Nakaoka, Y. Nojiri, X. A. Padin, A. Peregon, B. Pfeil, D. Pierrot, B. Poulter, G. Rehder, J. Reimer, C. Rödenbeck, J. Schwinger, R. Séférian, I. Skjelvan, B. D. Stocker, H. Tian, B. Tilbrook, F. N. Tubiello, I. T. van der Laan-Luijkx, G. R. van der Werf, S. van Heuven, N. Viovy, N. Vuichard, A. P. Walker, A. J. Watson, A. J. Wiltshire, S. Zaehle, and D. Zhu (2018). Global carbon budget 2017. *Earth System Science Data 10*(1), 405 – 448.

Leisch, F., K. Hornik, and C.-M. Kuan (2000). Monitoring structural changes with the generalized fluctuation test. *Econometric Theory 16*, 835–854.

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., Z. Vrontisi, C. Bertram, O. Y. Edelenbosch, R. C. Pietzcker, J. Rogelj, H. S. De Boer, L. Drouet, J. Emmerling, O. Fricko, S. Fujimori, P. Havlík, G. Iyer, K. Keramidas, A. Kitous, M. Pehl, V. Krey, K. Riahi, B. Saveyn, M. Tavoni, D. P. Van Vuuren, and E. Kriegler (2018). Residual fossil CO2 emissions in 1.5–2°C pathways. *Nature Climate Change 8*(7), 626–633.

Millar, R. J., J. S. Fuglestvedt, P. Friedlingstein, J. Rogelj, M. J. Grubb, H. D. Matthews, R. B. Skeie, P. M. Forster, D. J. Frame, and M. R. Allen (2017, 09). Emission budgets and pathways consistent with limiting warming to 1.5°C. *Nature Geoscience 10*, 741 EP –.

Nature (2018, December 5). Rules for a safe climate. Editorial.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix. *Econometrica 55*(394), 703–708.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika 41*(1), 100–115.

Peters, G. P., C. Le Quéré, R. M. Andrew, J. G. Canadell, P. Friedlingstein, T. Ilyina, R. B. Jackson, F. Joos, J. I. Korsbakken, G. A. McKinley, S. Sitch, and P. Tans (2017). Towards real-time verification of CO_2 emissions. *Nature Climate Change 7*(12), 848–850.

Sanderson, B. M., B. C. O'Neill, and C. Tebaldi (2016). What would it take to achieve the Paris temperature targets? *Geophysical Research Letters 43*(13), 7133–7142.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Tanaka, K. and B. C. O'Neill (2018). The Paris Agreement zero-emissions goal is not always consistent with the 1.5°C and 2°C temperature targets. *Nature Climate Change 8*(4), 319–324.

Tokarska, K. B. and N. P. Gillett (2018). Cumulative carbon emissions budgets consistent with 1.5 °C global warming. *Nature Climate Change 8*(4), 296–299.

Transparency International (Ed.) (2013). *Global Corruption Report: Climate Change*. London: Routledge.

UNFCCC (2015). Adoption of the Paris agreement.

Zhang, D., Q. Zhang, S. Qi, J. Huang, V. J. Karplus, and X. Zhang (2019). Integrity of firms' emissions reporting in China's early carbon markets. *Nature Climate Change 9*(2), 164–169.

A Proof of Theorem 2.1

Proof of Theorem 2.1. Let $t \geq K+1$ and suppose that H_0 holds, i.e., that $\epsilon_t^* = 0$ for all t. Note first, that we can write (2.2)

$$\tilde{Z}_t = \frac{1}{\sqrt{K\hat{\omega}_K^2}} \left(\sum_{i=1}^{t} u_i - \sum_{i=1}^{K} u_i \right).$$

Since $\hat{\omega}_K^2 \xrightarrow{P} \omega^2 > 0$, where $\xrightarrow{P}$ denotes convergence in probability, it holds, by Assumption 2.1 and Theorem 5.1. in Davidson (2006), that for $\tilde{\lambda} \in [1, \Lambda+1]$,

$$\left(\tilde{\lambda} \mapsto \tilde{Z}_{[K\tilde{\lambda}]} \right) \Rightarrow \left(\tilde{\lambda} \mapsto \tilde{B}(\tilde{\lambda}) - \tilde{B}(1) \right),$$

as $n \to \infty$, where $[x]$ denotes the integer part of $x \in \mathbb{R}$ and the convergence takes place in the Skorohod space $D[0,1]$, see, e.g., Davidson (2006). The process $\tilde{B}$ is a standard Brownian motion on $[1, \Lambda+1]$. Letting $\lambda := \tilde{\lambda} - 1$ and $B(\lambda) := \tilde{B}(\lambda+1) - \tilde{B}(1)$, it is clear that B is a Brownian motion on $[0, \Lambda]$. The result follows, using Assumption 2.2, from arguments similar to those in Theorem 2.1 of Leisch et al. (2000). □

B Calculation of the critical boundary

This section briefly explains how to calculate the critical values C_t^α of Equation (2.4) for use in the one-sided test of Corollary 2.1. Further details, including information on the two-sided test and alternative boundary functions, can be found in the Supplementary Material. A MATLAB software package that calculates C_t^α automatically is supplied online (Bennedsen, 2020).

Recall from Equation (2.4) that $C_t^\alpha = \sqrt{K\hat{\omega}_K^2}\cdot h(t/K)$, where the function h is given in Equation (2.3). Given an estimate of the long run variance ω^2, it therefore only remains to determine the constant $c = c_{\Lambda,\alpha}$ in (2.3). Let α be the chosen significance level of the test and let $f(\lambda) = c^{-1}h(\lambda)$, i.e. f is the function h without the normalizing constant c. When the monitoring period is finite, i.e. when $\Lambda < \infty$, we determine c by simulation as follows. A large number M of Brownian motions (here $M = 100\ 000$) are simulated on $[0, \Lambda]$. For each of these Brownian motions, $B(\lambda)$, we construct the path of $G(\lambda) = B(\lambda)/f(1+\lambda)$, and record the minimum of $G(\lambda)$ on $[0, \Delta]$. The α quantile of the M recorded minima is now the simulated value of $c = c_{\Lambda,\alpha}$ we were seeking (cf. also Leisch et al., 2000, Section 4). If the monitoring period is indefinite, i.e. when $\Lambda = \infty$, we approximate c using the same procedure, but setting Λ to a very large number. (E.g. $\Lambda = 100$, which corresponds to monitoring for $\Lambda \times K = 6\ 000$ years in our setup.)

C The budget imbalance and its statistical properties

This section contains further details of the analysis of the budget imbalance data B_t^{IM}, $t = 1959, ..., 2018$, the outcome of which was briefly reviewed in Section 3.1. The top row of Table 2 presents some descriptive statistics regarding these data. We see that the mean of the time series is not significantly different from zero, indicating that the carbon budget is balanced on average. Further, the Durbin-Watson ($DW = 1.21$) and Ljung-Box ($Q = 35.27$) test statistics indicate that the budget imbalance contains (positive) serial autocorrelation. (The caption of Table 2 contains additional information regarding the DW and Q statistics.) This, together with the visual impression of Figure 1e), provides a first indication of the budget imbalance being well-described by a stationary process with some positive correlation structure.

To further test the stationarity hypothesis, we conduct the "KPSS" test of Kwiatkowski et al. (1992). The null hypothesis of this test is that the data are (trend) stationary, while the alternative is that the data contain a unit root. The results of the test are given in Table 3, where a number equal to one indicates that we reject the null and a number equal to zero indicates that we fail to reject the null. Besides the data, the KPSS test requires two inputs: the number of lags used to calculate the long run variance of the data (using the estimator proposed in Newey and West, 1987) and whether or not the data contain a deterministic trend under the null. The table reports the results from using $0, 1, \ldots, 10$ lags for computing the long run variance and for both the cases of the presence of a deterministic trend. The results of the KPSS test provide additional evidence that the budget imbalance constitutes a stationary process: in most setups, we can not reject the null. The cases where the null is rejected are where we include a deterministic trend under the null and include a small number of lags when estimating the long run variance. It is well-known, that when there is (positive) autocorrelation in the process, this setup can result in over-rejections of the null, especially when the number of data points is small, which is the case here (Kwiatkowski et al., 1992, Section 5). What is more, a deterministic trend in the budget imbalance is very unlikely *a priori* and the most relevant case for our purpose is therefore the test where such a trend is excluded. In this case, we can not reject the null regardless of the number of lags used in the calculation of the long run variance.

Inspecting the empirical autocorrelation and partial autocorrelation functions of the data B_t^{IM} (not given here for brevity but see the Supplementary Material), provides evidence that an autoregressive process of order one (AR1) is an adequate statistical model for B_t^{IM} for $t = 1959, \ldots, 2018$. Likewise, the Bayesian Information Criterion (Schwarz, 1978) selects an AR(1) model from the class of autoregressive moving average (ARMA) models. Although the Akaike Information Criterion (Akaike, 1974) prefers a more complicated model (ARMA$(5, 5)$), we conclude that a reasonable model of the budget imbalance data is an autoregressive process of order one.[6]

After fitting an AR(1) model to the data, we subject the residuals of this fit to the same analysis as conducted on the budget imbalance data in the beginning of this section. The results are shown in Figure 4 and Tables 2 and 3. After fitting this model, there is practically no autocorrelation left in the residuals ($DW = 2.00$, $Q = 20.39$) and the KPSS test can not reject the null of stationarity in any of the setups considered here. Summing up, the diagnostics confirm that our chosen AR(1) model is a good model for the historical budget imbalance data. The estimate (obtained by an ordinary least squares regression) of the autoregressive parameter is $\hat{\phi} = 0.38$ and for the variance of ϵ_t, $\widehat{Var(\epsilon_t)} = 0.52$.

Lastly, we note that if, in spite of the findings of this section, the researcher does not want to specify a parametric structure for the budget imbalance data, the methods proposed in this paper are valid under significantly weaker assumptions as well. See Remark 2.3 and the Supplementary Material for further details.

Table 2: *Descriptive statistics and diagnostics of the budget imbalance data from Equation (3.2) in the paper. N is the test-statistic from the Jarque-Bera test (Jarque and Bera, 1987): the null hypothesis that the data comes from a Gaussian distribution can be rejected if N is larger than the 95% critical value of 5.99. DW is the Durbin-Watson test statistic (Durbin and Watson, 1971): If $DW < 2$ there is evidence of positive serial correlation in the data; if $DW > 2$ there is evidence of negative serial correlation in the data; data without serial correlation will have $DW = 2$. Q is the Ljung-Box Q test statistic (Ljung and Box, 1978) for presence of autocorrelation. The critical value for the Q-test is 31.41; hence if $Q > 31.41$ then the null of no autocorrelation can be rejected. The autoregressive parameter of the $AR(1)$ process is estimated as $\hat{\phi} = 0.38$.*

	Descriptive statistics					Diagnostics		
Data	Num. Obs.	Mean	Std. Dev.	Skewness	Kurtosis	N	DW	Q
Budget imbalance	60	0.17	0.77	-0.070	2.81	0.14	1.21	35.27
AR(1) residuals	59	-0.099	0.72	0.045	2.31	1.20	2.00	20.39

[6]Further, a test for conditional heteroskedasticity using the ARCH test of Engle (1982) shows no signs heteroskedasticity in the data. The specifics of this test are not included for brevity but they are available from the author upon request.

Table 3: *Testing for stationarity of the data using the KPSS test (Kwiatkowski et al., 1992). Numbers equal to one indicate that the test rejects the null of (trend) stationarity against the alternative of a unit root in the data, while numbers equal to zero indicate a failure to reject the null. The table reports the results from using* $0, 1, \ldots, 10$ *lags for computing the long run variance and for both the cases of the presence of a deterministic trend under the null.*

	Number of lags										
	0	1	2	3	4	5	6	7	8	9	10
Budget imbalance											
No Trend	0	0	0	0	0	0	0	0	0	0	0
Trend	1	1	0	0	0	0	0	0	0	0	0
AR(1) *residuals*											
No Trend	0	0	0	0	0	0	0	0	0	0	0
Trend	0	0	0	0	0	0	0	0	0	0	0

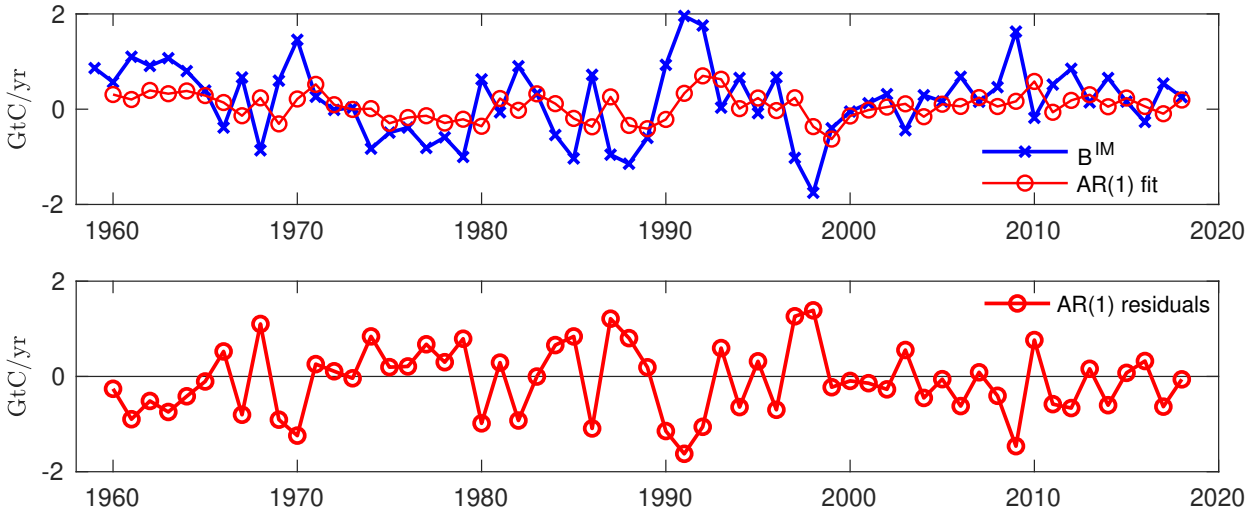

Figure 4: *Top: Budget imbalance data and AR*(1) *fit. Bottom: Residuals from AR(1) fit.*

Research Papers 2020

2019-09: Debopam Bhattacharya, Pascaline Dupas and Shin Kanaya: Demand and Welfare Analysis in Discrete Choice Models with Social Interactions

2019-10: Martin Møller Andreasen, Kasper Jørgensen and Andrew Meldrum: Bond Risk Premiums at the Zero Lower Bound

2019-11: Martin Møller Andrasen: Explaining Bond Return Predictability in an Estimated New Keynesian Model

2019-12: Vanessa Berenguer-Rico, Søren Johansen and Bent Nielsen: Uniform Consistency of Marked and Weighted Empirical Distributions of Residuals

2019-13: Daniel Borup and Erik Christian Montes Schütte: In search of a job: Forecasting employment growth using Google Trends

2019-14: Kim Christensen, Charlotte Christiansen and Anders M. Posselt: The Economic Value of VIX ETPs

2019-15: Vanessa Berenguer-Rico, Søren Johansen and Bent Nielsen: Models where the Least Trimmed Squares and Least Median of Squares estimators are maximum likelihood

2019-16: Kristoffer Pons Bertelsen: Comparing Tests for Identification of Bubbles

2019-17: Dakyung Seong, Jin Seo Cho and Timo Teräsvirta: Comprehensive Testing of Linearity against the Smooth Transition Autoregressive Model

2019-18: Changli He, Jian Kang, Timo Teräsvirta and Shuhua Zhang: Long monthly temperature series and the Vector Seasonal Shifting Mean and Covariance Autoregressive model

2019-19: Changli He, Jian Kang, Timo Teräsvirta and Shuhua Zhang: Comparing long monthly Chinese and selected European temperature series using the Vector Seasonal Shifting Mean and Covariance Autoregressive model

2019-20: Malene Kallestrup-Lamb, Søren Kjærgaard and Carsten P. T. Rosenskjold: Insight into Stagnating Life Expectancy: Analysing Cause of Death Patterns across Socio-economic Groups

2019-21: Mikkel Bennedsen, Eric Hillebrand and Siem Jan Koopman: Modeling, Forecasting, and Nowcasting U.S. CO2 Emissions Using Many Macroeconomic Predictors

2019-22: Anne G. Balter, Malene Kallestrup-Lamb and Jesper Rangvid: The move towards riskier pensions: The importance of mortality

2019-23: Duván Humberto Cataño, Carlos Vladimir Rodríguez-Caballero and Daniel Peña: Wavelet Estimation for Dynamic Factor Models with Time-Varying Loadings

2020-01: Mikkel Bennedsen: Designing a sequential testing procedure for verifying global CO2 emissions