Monitoring the carbon budget: Real-time verification of CO\textsubscript{2} emissions

Mikkel Bennedsen\textsuperscript{*}

April 9, 2019

Abstract

International agreements to reduce CO\textsubscript{2} emissions necessitate that nations will have to commit to drastic reductions in their emissions in the near future. However, CO\textsubscript{2} emissions are reported by individual nations and cannot be easily verified by the global community. This implies the need to develop reliable methods to verify reported CO\textsubscript{2} emissions. This paper puts forth a simple statistical method to do this. In particular, we propose to use Earth system data on CO\textsubscript{2} sinks, measured independently of reported emissions, to monitor, sequentially in real-time, whether the Earth system data are compatible with reported CO\textsubscript{2} emissions.

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\textsubscript{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. 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\textsubscript{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\textsubscript{2} emissions, nevertheless appear to be fulfilling their 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; Nature, 2018; Zhang et al., 2019). Indeed, the problem of verifying the reported CO\textsubscript{2} emissions was one of the key topics discussed at the recent COP24 meeting in Katowice, Poland (IPCC, 2018a).

\textsuperscript{*}Department of Economics and Business Economics and CREATEES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V, Denmark. E-mail: mbennedsen@econ.au.dk
These considerations highlight the increasing incentive to develop independent ways of verifying reported global CO$_2$ emissions (Peters et al., 2017). This paper proposes an easy-to-use statistical procedure for detecting systematic deviations between the CO$_2$ emissions that are reported and the CO$_2$ emissions that are actually emitted to the biosphere. It does so by exploiting the constraints on emissions imposed by other Earth system variables, variables which can be measured independently. In particular, we propose to utilize the concept of a balanced carbon budget: in every time period, the fluxes of CO$_2$ in the Earth system (atmospheric growth, terrestrial land sink, and oceanic sink) must equal global CO$_2$ emissions (Le Quéré et al., 2018). The underlying idea of the proposed statistical method, is to monitor, sequentially in time as new data arrive, the fluxes in the CO$_2$ sinks and then compare these with reported emissions. This allows us to propose a statistical tool that can detect if reported emissions are systematically different from actual emission. The tool relies on a monitoring/sequential testing framework (Page, 1954; Chu et al., 1996).

A Supplementary Material file, complementing this paper, is available online.¹

### Historical behaviour of the carbon budget

The carbon budget is a way to tally the fluxes of CO$_2$ in the carbon sources (emissions) and sinks (atmosphere, land, ocean). Formally, it is represented as (Le Quéré et al., 2018)

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

where $E_t^{ANT}$ is the anthropogenic CO$_2$ emissions at time $t$, while $G_t^{ATM}, S_t^{OCN}$, and $S_t^{LND}$ represent the flux to the sinks of CO$_2$, namely atmospheric growth, the flux to the oceans, and the flux to the terrestrial biosphere, respectively. See the Data section in Methods for further information on the data used in this paper. The term $B_t^{IM}$ is called the budget imbalance: it is implicitly defined so that Equation (1) holds. Figure 1 plots the time series of the variables from Equation (1) from $t = 1959$ to $t = 2017$.

The budget imbalance, $B_t^{IM}$, represents the shortcomings in our measurements of the emissions and Earth system variables in the carbon budget; for instance errors or systematic biases in the measurements. In the Methods (cf. also the Supplementary Material) we conduct a thorough statistical analysis of the time series properties of $B_t^{IM}$. We find that it is well-described by a zero-mean autoregressive process of order one. This implies that, on average, the carbon budget is balanced. It also implies that there is a slight serial dependence in the budget imbalance: if $B_t^{IM} > 0$ in period $t$, we expect that $B_{t+1}^{IM} > 0$ in period $t + 1$ (and vice versa for $B_t^{IM} < 0$), although it will revert towards the mean of zero as time goes.

### The carbon budget when CO$_2$ emissions are misreported

When CO$_2$ emissions are misreported, the budget imbalance will cease to be described by a stationary process: a so-called structural break will be introduced into $B_t^{IM}$. In fact, when CO$_2$ emissions are under-reported, arguably the most important case in practice, the structural break will imply

¹https://sites.google.com/site/mbennedsen/research-supplementary_material_arxiv_v23.pdf.
a negative budget imbalance (Methods). To illustrate, we consider a hypothetical scenario where countries agreed in the year 1999 to cut emission growth to 0%, i.e., to hold the 1999 level of emissions constant (in our data, $E_{1999}^{ANT} = 7.73$ GtC), while they actually continued emitting CO$_2$ according to the real historical levels. (The average growth rate of emissions from 1999 to 2017 has been approximately 2.0% per year.) In other words, we suppose that global CO$_2$ emissions were reported as constant from 1999 and onwards, while actual CO$_2$ emissions follow the historical path we have seen in the data. This setup thus mimics what would happen if reported CO$_2$ emissions begin to be systematically different from actual emissions.

Figure 2 illustrates the hypothetical scenario and compares it to the actual real-world data. Panel a) of Figure 2 shows the CO$_2$ emissions trajectories: The blue solid line is the actual CO$_2$ emissions trajectory, while the green dashed line is the hypothetical emissions trajectory, which we here take to be the reported emissions. Panel b) of the figure shows the corresponding budget imbalances: If CO$_2$ emissions were reported as actual values, the blue solid line would materialise, while if CO$_2$ emissions are misreported, we would observe the green dashed line when calculating the budget imbalance. The structural break introduced in $B_{t}^{IM}$ in $t = 2000$, indicated by the vertical black line, is apparent.

**Detecting misreportings of CO$_2$ emissions**

The foregoing discussion suggests a way to detect misreportings in CO$_2$ emissions: monitor the budget imbalance of the carbon budget and test whether a structural break has occurred. We
Figure 2: Hypothetical scenario where emissions are reported as constant starting in $t = 2000$ (black vertical line). 

a): CO$_2$ emission trajectories. Actual emissions (blue solid line) and hypothetically reported emissions (green dashed line).

b): Budget imbalance data. Actual budget imbalance data (blue solid line) and hypothetical budget imbalance data (green dashed line). The former is what has actually been observed historically, the latter is what would have materialised under the hypothetical misreporting.

c): Test statistic $Z_t$ and critical boundaries $C_t^\alpha$. Actual test statistic $Z_t^{act}$ (blue solid line) and hypothetical test statistic $Z_t^{hyp}$ (green dashed line). The former is what has actually been observed historically, the latter is what would have materialised under the hypothetical misreporting. The critical boundaries $C_t^\alpha$ are, from bottom to top, $\alpha = 5\%, 10\%$, and $32\%$. 


propose to do this sequentially in time: when new data arrive on the Earth system variables and
the reported emissions, we calculate the implied data point for the budget imbalance and conduct
a test for a structural break. Such tests abound in the statistical literature (Ploberger et al., 1989;
Andrews, 1993; Bai and Perron, 1998). However, since we are not sure when (or if) misreporting
will begin, we will have to keep testing every time new data arrive, continuing into the indefinite
future. This engenders a so-called multiple-testing problem: if we keep performing these tests
at a pre-specified significance level $\alpha$ (such as 10% or 32%), then, due to the sequential nature
of the procedure, we are likely to eventually find evidence of misreporting, even if there is none
(committing a so-called Type I error). (Robbins, 1970) Therefore, it is important that the critical
boundaries of the test are chosen in such a way that the probability of making a Type I error
is bounded by the chosen significance level, while the test still has power to detect potential
misreportings.

In the Methods section we propose such a monitoring procedure, the theoretical validity of which
is proved in the Supplementary Material. The goal is to test whether CO$_2$ emissions are reported
without systematic bias (the null hypothesis, $H_0$), against the alternative that CO$_2$ emissions are
systematically under-reported (the alternative hypothesis, $H_1$). Our test is based on monitoring
a test statistic, $Z_t$, which is the cumulated sum of budget imbalances since monitoring started.
For instance, if we begin monitoring at time period $t_0$, then, at $t \geq t_0$, the test statistic takes
the value $Z_t = \sum_{\tau=t_0}^{t} B_{\tau}^{IM}$. If, at some time $t$, the test statistic $Z_t$ is less than some critical
value $C_t^{\alpha}$, then we reject $H_0$ in favour of $H_1$ at an $\alpha$-significance level, where $\alpha$ is e.g. 10% or
32%. Under the null hypothesis, $Z_t$ will be a cumulative sum of a stationary process and will thus
have increasing variance over time. To control the size of the test, it is therefore necessary that
the critical boundary $C_t^{\alpha}$ is expanding over time. In the Methods, we explain how to calculate
the critical values $C_t^{\alpha}$ for a pre-specified significance level $\alpha$. Conversely, under the alternative
hypothesis, $Z_t$ will include the cumulative sum of a negative structural break process, representing
the amount of under-reporting, which will tend to push $Z_t$ downwards. Hence we reject the null
in favour of the alternative when $Z_t$ is sufficiently negative, i.e. when $Z_t < C_t^{\alpha}$.

Figure 2, Panel c), illustrates the procedure, when applied to the example discussed above,
where a hypothetical agreement is made in the year 1999 and global CO$_2$ emissions are misreported
starting in $t = 2000$. The statistic $Z_t^{act}$ (blue solid line) is the sequential test statistic which has
materialised in actuality since year 2000. The statistic $Z_t^{hyp}$ (green dashed line) is the hypothetical
sequential test statistic that would have materialised if global CO$_2$ emissions had been reported
constant since 1999 but had actually grown at historical levels. The black lines denote the critical
boundaries, $C_t^{\alpha}$, at different significance levels $\alpha$, as indicated in the caption of the figure. When
the test statistic $Z_t$ crosses the critical boundary $C_t^{\alpha}$, there is evidence of misreporting at an $\alpha$
significance level. In the hypothetical case studied here, we would have evidence at a 32% level that
CO$_2$ emissions are misreported in year 2006 and at 10% and 5% in 2008 and 2010, respectively.
Estimating mean detection time through simulations

To investigate how the proposed monitoring scheme will perform in practice, 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. Our initial data set is the realised budget imbalance data from 1959 to 2017, i.e. the data studied above. We then simulate 10 000 instances of possible future paths of the budget imbalance from 2018 to 2100. Further details are given in the Methods. The Supplementary Material contains more extensive simulation studies.

As mentioned above, to achieve the Paris objectives, CO$_2$ emissions should be cut in approximately half, as compared to 2010 levels, by 2030. Since $E_{2010}^{ANT} = 10.44$ GtC and $E_{2017}^{ANT} = 11.26$ GtC, this means that CO$_2$ emissions should shrink by 5.74% each year from 2018 onwards. This is our baseline scenario: in the simulations to come, we suppose that CO$_2$ emissions are reported as falling 5.74% each year from 2018 to 2030. We then consider three different scenarios for actual emissions: emissions shrinking by 5.74% each year (no misreporting); emissions falling 3% each year (2.74% under-reporting per year); and constant emissions (5.74% under-reporting per year). The resulting paths of the emissions are shown in the left panel of Figure 3, cf. a), c), and d). The blue solid line corresponds to the “actual” emissions, while the green dashed line is the “reported emissions”. The right panel of Figure 3 presents 100 example paths of the simulated test statistic $Z_t$, cyan solid lines, obtained from the simulated budget imbalances and the CO$_2$ emission scenarios described above, cf. b), d), and f). See Methods for further information. The black lines indicate the critical values, $C^\alpha_t$, for $\alpha = 5\%$, $10\%$, and $32\%$. In the first scenario, cf. a) and b), emissions are reported truthfully. Consequently, when $Z_t$ crosses the critical boundary, it will result in 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%, 4.72%, and 17.27% for $\alpha = 5\%$, $10\%$, and $32\%$, respectively (asymptotically, the false positive rates will equal the significance level $\alpha$; however, in finite samples the false positive rate might be different from the nominal significance level). When emissions are reported 2.74% below actual emissions, cf. c) and d), a structural break is introduced into the budget imbalance and, consequently, all the paths of the test statistic $Z_t$ will cross the critical boundaries eventually. The mean crossing times are 9.30, 8.51, and 6.83 years for $\alpha = 5\%$, $10\%$, and $32\%$, respectively. Similarly, when emissions are reported 5.74% below actual emissions, cf. e) and f), the paths of $Z_t$ will cross the critical boundaries even faster. Here, the mean crossing times are 5.24, 4.84, and 3.96 years for $\alpha = 5\%$, $10\%$, and $32\%$, respectively.

Monitoring the future carbon budget

If we wish to monitor the carbon budget starting now, that is, starting with 2018 data, Table 1 presents the critical values, $C^\alpha_t$, for the test proposed in this paper. These are the critical boundaries which were used in the simulation experiment, cf. Figure 3. To monitor the future carbon budget, we proceed as follows. Every year $t=2018,2019,\ldots$, when new data arrive, the monitoring statistic $Z_t$ is updated and compared to the critical values given in Table 1. That is, we calculate the cumulative sum of the budget imbalances through time, $Z_t = \sum_{\tau=2018}^{t} B_{\tau}$, and
Figure 3: Illustration of simulation study. Left panels show hypothetical future CO₂ emissions trajectories from 2018 to 2030: actual emissions (blue solid line) and reported emissions (green dashed line). Right panels show 100 simulated paths of the test statistic Zₜ (cyan lines) and critical boundaries Cₜ(α) (black lines) for α = 5%, 10%, and 32%, from 2018 to 2030. a)+b): Both actual and reported emissions are shrinking by 5.74% each year. c)+d): Actual emissions are shrinking 3% each year, while emissions are reported to shrink by 5.74% each year. e)+f): Actual emissions are shrinking 0% each year, while emissions are reported to shrink by 5.74% each year.
compare this to the appropriate critical value $C_{t}^{\alpha}$ 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 $\alpha$. In other words, if $Z_{t} < C_{t}^{\alpha}$ for some $t = 2018, 2019, \ldots$, 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 details given in Methods and the Supplementary Material. To facilitate easy application of our methods, we supply a simple computer program that can do this automatically (Bennedsen, 2019).

Table 1: Critical values $C_{t}^{\alpha}$ for the test of misreporting in CO$_2$ emissions. The values in the table have been calculated using $B_{t}^{LM}$, $t = 1959, \ldots, 2017$, as input.

|     | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 | 2026 | 2027 |
|-----|------|------|------|------|------|------|------|------|------|------|
| $\alpha = 5\%$ | -4.17 | -5.54 | -6.53 | -7.34 | -8.04 | -8.67 | -9.24 | -9.77 | -10.26 | -10.73 |
| $\alpha = 10\%$ | -3.69 | -4.90 | -5.78 | -6.49 | -7.11 | -7.67 | -8.17 | -8.64 | -9.08 | -9.49 |
| $\alpha = 32\%$ | -2.68 | -3.55 | -4.19 | -4.71 | -5.16 | -5.56 | -5.92 | -6.26 | -6.58 | -6.88 |

**Discussion**

Verifying and monitoring reported CO$_2$ emissions will soon become important to ensure that predetermined goals for CO$_2$ emissions, such as those following the Paris Agreement, are met. This paper has proposed a statistical method to test whether or not CO$_2$ emissions are systematically under-reported. The method is easy to use: simply compute the cumulated sum of the budget imbalances since monitoring started and compare with the appropriate critical values in Table 1. If monitoring starts at a later time than 2018, the critical values should be re-computed; the computations are relatively easy to make and we provide a computer program that can perform these automatically (Bennedsen, 2019).

The time it takes to detect potential misreporting depends on the magnitude of the misreporting. Using simulations, cf. Figure 3, we estimate that it on average takes 3.96 years before we can reject the null of no misreporting at a 32% confidence level, when emissions are misreported by 5.74% every year. When the magnitude of misreporting is 2.74% per year, the average detection time increases to 6.83 years. These numbers are on the order of the 5 years that are between the Paris stocktakes, indicating that the methods proposed here can help inform the global community whether countries are reporting their emissions truthfully. Still, reducing detection time even further would be very beneficial. As described in the Supplementary Material, the statistical approach proposed here has been tuned so as to favour quick detection; for this reason we believe that it is difficult to reduce detection time further by statistical methods. However, if we can get more precise estimates of the Earth system variables, it will be possible to reduce detection time. In effect, what we would to like have are smaller measurement errors and thus a smaller budget.
imbalance. We replicated the simulation study from above using a budget imbalance which was half the size of the one we actually observe, i.e. we used \( \bar{B}_t^{IM} = 0.5 \cdot B_t^{IM}, \ t = 1959, \ldots, 2017, \) in the analysis. This resulted in mean detection times of 2.51 and 4.17 years in the two scenarios considered above, at a 32% level (see Methods). In other words, more precisely measured Earth system variables will lead to a smaller budget imbalance, which will in turn allow us to detect potential misreportings of CO\(_2\) emissions much more quickly. This highlights the importance of getting more precise measurements of Earth system variables going forward (Peters et al., 2017).

**Methods**

**Data**

Recall the carbon budget equation (1):

\[
E_t^{ANT} = G_t^{ATM} + S_t^O + S_t^L + B_t^{IM}.
\]

Here \( E_t^{ANT} = E_t^{FF} + E_t^{LUC} \), 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; \( S_t^O \) is the flux of CO\(_2\) from the atmosphere to the oceans; and \( S_t^L \) is the flux of CO\(_2\) from the atmosphere to the terrestrial biosphere. We use the data set provided by The Global Carbon Project (Le Quéré et al., 2018).\(^2\) The fossil fuel emissions data \( E_t^{FF} \) are from Boden et al. (2017), 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 Le Quéré et al. (2018). The growth rate in atmospheric CO\(_2\) data, \( G_t^{ATM} \), is from Dlugokencky and Tans (2018), while the sink data, \( S_t^O \) and \( S_t^L \), are averages over several independent model-based estimates, constructed as explained in Le Quéré et al. (2018). All data are given in gigatonnes of carbon (GtC) and are recorded at a yearly frequency, beginning in 1959 and ending in 2017, resulting in 59 observations for each term in (1).

**Statistical analysis of the budget imbalance**

We subject the budget imbalance time series \( B_t^{IM} \) to a number of statistical tests. We report the most important findings; details are given in Section 1 of the Supplementary Material.

The mean of the time series is not significantly different from zero, indicating that the carbon budget is balanced on average. When conducting the “KPSS” test of stationarity, we can not reject the null that the data is stationary (Kwiatkowski et al., 1992). The stationarity hypothesis is further strengthened by simple visual inspection of the data. The Durbin-Watson (Durbin and Watson, 1971) and Ljung-Box (Ljung and Box, 1978) tests indicate that the budget imbalance contains (positive) serial autocorrelation.

Given that the budget imbalance appears to be a stationary process, we employ the so-called Box-Jenkins approach to modelling \( B_t^{IM} \) (Box and Jenkins, 1970). That is, we fit an autoregressive moving average (ARMA) model to the data. After inspecting the autocorrelations and partial

\(^2\)The data are available at [http://www.globalcarbonproject.org/](http://www.globalcarbonproject.org/) and were downloaded on 8 January 2019.
autocorrelations of the data, we conclude that an autoregressive process of order one is a good model for the data. This choice is corroborated by automatic selection of ARMA lag lengths using information criteria (Akaike, 1974; Schwarz, 1978). We estimate the autoregressive parameter as \( \hat{\phi} = 0.43 \) and the variance of the error term as \( \hat{\sigma}^2 = 0.53 \).

**Detecting a structural break in the budget imbalance data**

Suppose that from some time point \( \tau \), anthropogenic CO\(_2\) emissions are misreported as \( E_{t}^{\text{ANT,*}} \), while the true value emitted to the biosphere is \( E_{t}^{\text{ANT}} \neq E_{t}^{\text{ANT,*}} \). Then, for \( t \geq \tau \), the observed budget imbalance data become

\[
B_{t}^{\text{IM,*}} = E_{t}^{\text{ANT,*}} - G_{t}^{\text{ATM}} - S_{t}^{\text{OCEAN}} - S_{t}^{\text{LAND}} = u_{t} + \epsilon_{t}^{*},
\]

where \( u_{t} = E_{t}^{\text{ANT}} - G_{t}^{\text{ATM}} - S_{t}^{\text{OCEAN}} - S_{t}^{\text{LAND}} \) is the budget imbalance under the true emission path \( E_{t}^{\text{ANT}} \), while \( \epsilon_{t}^{*} = E_{t}^{\text{ANT,*}} - E_{t}^{\text{ANT}} \) denotes the amount of misreporting in CO\(_2\) emissions at time \( t \geq \tau \). As discussed above, \( u_{t} \) is a stationary process. The process \( \epsilon_{t}^{*} \) describes the structural break introduced in the budget imbalance when emissions are misreported, i.e., when \( E_{t}^{\text{ANT,*}} \neq E_{t}^{\text{ANT}} \). Note that when emissions are under-reported, i.e. when \( E_{t}^{\text{ANT,*}} < E_{t}^{\text{ANT}} \), then \( \epsilon_{t}^{*} < 0 \), which is why the structural break will result in negative values of the observed budget imbalance data.

When monitoring starts at time \( t_{0} \), the test statistic \( Z_{t} \) introduced in the paper takes the form

\[
Z_{t} = \sum_{\tau=t_{0}}^{t} B_{\tau}^{\text{IM,*}} = \sum_{\tau=t_{0}}^{t} u_{\tau} + \sum_{\tau=t_{0}}^{t} \epsilon_{\tau}^{*}.
\]

If \( \epsilon_{\tau}^{*} = 0 \) for all \( \tau \), i.e. if there is no misreporting, we can derive critical boundaries \( C_{\alpha}^{t} \), such that \( Z_{t} \) will, asymptotically, stay above these boundaries with probability \( 1 - \alpha \). The boundaries we suggest are of the form

\[
C_{t}^{\alpha} = c_{\alpha,\Lambda} \cdot \sqrt{K} \cdot \hat{\omega}_{u}^{2} \cdot \sqrt{\lambda_{t}(\lambda_{t} - 1)} \left(1 + \log \left(\frac{\lambda_{t}}{\lambda_{t} - 1}\right)\right),
\]

where \( \lambda_{t} := \frac{t - t_{0} + K + 1}{K} \) is the number of observations in the initial data set, and \( \hat{\omega}_{u}^{2} \) is the estimate of the long run variance of the initial budget imbalance data (\( K = 59 \) and \( \hat{\omega}_{u}^{2} = 1.65 \) for the data studied in this paper). The theoretical validity of the test is proved in the Supplementary Material.

The constant \( c_{\alpha,\Lambda} \) depends on the nominal significance level \( \alpha \) and on a parameter \( \Lambda > 0 \), which controls how long the monitoring period is. The monitoring period is assumed to be of length \( K \cdot \Lambda \). In the paper, we set \( \Lambda = 1 \), corresponding to monitoring from 2018 to 2076, but a different value can be used if the length of the monitoring period is known \textit{a priori}. In practice, the test, and the results of the paper, is insensitive to the choice of \( \Lambda \) and we obtain roughly the same results with, e.g., \( \Lambda = 1 \) and \( \Lambda = 100 \) (Supplementary Material, Section 4.1.1). The Supplementary Material contains more information on boundary functions and explains how to calculate \( c_{\alpha,\Lambda} \). We also provide a computer program that can calculate \( c_{\alpha,\Lambda} \) and \( C_{t}^{\alpha} \) automatically (Bennedsen, 2019).

10
Additional details from the simulations of the future budget imbalance

In the simulation experiment, we use the historical budget imbalance data $B_{t}^{IM}$, $t = 1959, \ldots, 2017$, as initial data. Then we simulate 10,000 different future paths of $B_{t}^{IM}$, $t = 2018, \ldots, 2100$. We do this using the results from above, i.e., for $t \geq 2018$ we set $B_{t}^{IM} = u_{t} + \epsilon_{t}^{*}$, where $u_{t}$ is simulated as an 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 cyan lines in the right panel of Figure 3 are simulated instances of $Z_{t} = \sum_{\tau} B_{t}^{IM} = \sum_{\tau} u_{\tau} + \sum_{\tau} \epsilon_{\tau}^{*}$. The autoregressive parameter for $u_{t}$ was set to $\phi = 0.43$, while the variance of error term of $u_{t}$ was set to $\sigma^{2} = 0.53$, which are the parameters we estimated from the initial budget imbalance data.

In the Discussion section, we report the results from a simulation study where the magnitude of the budget imbalance is halved. This experiment is done in the same way as just discussed, the only difference being that the initial data used is $0.5 \cdot B_{t}^{IM}$, $t = 1959, \ldots, 2017$, which results in $\hat{\phi} = 0.43$ and $\hat{\sigma}^{2} = 0.13$ when estimating the autoregressive model for the data. We use these estimates as parameters when simulating the autoregressive process $u_{t}$.

References

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

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* 61(4), 821–856.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* 66(1), 47–78.

Bennedsen, M. (2019). Software package for “Monitoring the carbon budget: Real-time verification of 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 (2017). 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, last access: 28 June 2017.

Box, G. E. P. and G. M. Jenkins (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

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

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, last access: 9 March 2018.

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.

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.

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.

Le Quéré, C., R. M. Andrew, P. Friedlingstein, S. Sitch, J. Hauck, J. Pongratz, P. A. Pickers, J. I. Korsbakken, G. P. Peters, J. G. Canadell, A. Arneth, V. K. Arora, L. Barbero, A. Bastos, L. Bopp, F. Chevallier, L. P. Chini, P. Ciais, S. C. Doney, T. Gkritzalis, D. S. Goll, I. Harris, V. Haverd, F. M. Hoffman, M. Hoppema, R. A. Houghton, G. Hurr, T. Ilyina, A. K. Jain, T. Johannessen, C. D. Jones, E. Kato, R. F. Keeling, K. K. Goldewijk, P. Landschützer, N. Lefèvre, S. Lienert, Z. Liu, D. Lombardozzi, N. Metzl, D. R. Munro, J. E. M. S. Nabel, S.-I. Nakaoka, C. Neill, A. Olsen, T. Ono, P. Patra, A. Peregon, W. Peters, P. Peylin, B. Pfeil, D. Pierrot, B. Poulter, G. Rehder, L. Resplandy, E. Robertson, M. Rocher, C. Rödenbeck, U. Schuster, J. Schwinger, R. Séférian, I. Skjelvan, T. Steinhoff, A. Sutton, P. P. Tans, H. Tian, B. Tilbrook, F. N. Tubiello, I. T. van der Laan-Luijkx, G. R. van der Werf, N. Viovy, A. P. Walker, A. J. Wiltshire, R. Wright, S. Zaehle, and B. Zheng (2018). Global Carbon Budget 2018. *Earth System Science Data* 10(4), 2141–2194.

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. Fricco, S. Fujimori, P. Havlik, 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, E. (2018, December 5). Rules for a safe climate. Editorial.

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 CO₂ emissions. *Nature Climate Change* 7(12), 848–850.

Ploberger, W., W. Krämer, and K. Kontrus (1989). A new test for structural stability in the linear regression model. *Journal of Econometrics* 40(2), 307–318.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 41(5), 1397–1409.

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, [https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf](https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf).

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.