Electronic Supplementary Material:

Designing a statistical procedure for monitoring global carbon dioxide emissions

Mikkel Bennedsen*

February 25, 2021

Contents

1 Introduction **2**

2 Additional statistical analysis on the GCB data **2**

2.1 Tests for stationarity 2

2.2 Autocorrelation and partial autocorrelation plots 3

2.3 QQ plots 3

3 Testing for misreporting in the historical data **3**

3.1 Testing for misreporting in the entire historical period 4

3.2 Testing for misreporting with unknown starting point 4

3.3 Counterfactual analyses 6

4 Simulation studies **7**

4.1 Empirical size of the test of H_0 8

4.2 Empirical power of the test of H_0 8

*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

This Electronic Supplementary Material file complements the paper "Designing a statistical procedure for monitoring global carbon dioxide emissions".

Section 2 contains results from a statistical analysis of the data studied in the main paper. Specifically, we perform stationarity tests (Section 2.1), provide plots of the autocorrelation and partial autocorrelations of the data (Section 2.2), and assess the Gaussianity of the key data series used in the main paper via QQ-plots (Section 2.3).

Section 3 contains analyses seeking to verify the assumption that the historical budget imbalance data sets do not contain systematic misreportings. Specifically, we test the hypothesis that the budget imbalance data studied in the main paper have zero mean. When analyzing the full data set, we find that we cannot reject the null of a zero mean in the data (Section 3.1). That is, we find no evidence of systematic biases in the full budget imbalance data sets. We then test whether possible under-reportings of fossil fuel emissions might have been initiated at an unknown time during the historical sample (Section 3.2). Again, we find no evidence of any under-reportings of these data in the historical sample. Lastly, to assess to what degree these tests would actually have power to detect under-reportings in the historical sample, we study a number of counterfactual situations, where fossil fuel emission are counterfactually constructed to be under-reported (Section 3.3). These counterfactual analyses show that the tests do have power to detect potential under-reportings in the historical sample. However, if this under-reporting is very small and/or happens very late in the historical period, it will, naturally, be difficult to detect.

2 Additional statistical analysis on the GCB data

2.1 Tests for stationarity

To test the hypothesis that the budget imbalance data are stationarity, we conduct the "KPSS" test of Kwiatkowski et al. (1992). The null hypothesis of this test is that the data are stationary, while the alternative is that the data contain a unit root. We also conduct the augmented Dickey-Fuller ("ADF") test (Dickey and Fuller, 1979), where the null is that the data contain a unit root and the alternative is that the data are stationary. The results of the tests are given in Table 1, 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. The tests are performed at a 5% significance level. Besides the data, the tests require 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 $0, 1, \ldots, 5$ lags for computing the long run variance and we assume that there is no deterministic trend under the null. As shown in Table 1, when running these tests on the budget imbalance data from the four data sets, we can never reject the null (of stationarity) when performing the KPSS test, but

we always reject the null (of a unit root) when performing the ADF test. This provides evidence that the budget imbalance data are historically well-described by a stationary process.

2.2 Autocorrelation and partial autocorrelation plots

Figure 1 shows the autocorrelation (left) and partial autocorrelation (right) calculated from the budget imbalance data. We see that there are some significant spikes for the shorter lags and that these gradually decay towards zero. This indicates that an AR(1) model could be adequate for the data. After fitting such an AR(1) model to the budget imbalance data, we can analyze the residuals from this model using similar methods. The results are in Figure 2, where we plot the autocorrelation (left) and partial autocorrelation (right) calculated from the AR(1) residuals of the budget imbalance data. We see that there does not appear to be any autocorrelation left in the data, which is evidence that the AR(1) model is adequate for the budget imbalance data studied here.

2.3 QQ plots

To examine the hypothesis that the AR(1) residuals from the budget imbalance data are Gaussian, Figure 3 contains a QQ-plot of these data. If the Gaussian assumption is adequate, the data points (blue) should lie on the 45 degree line (red). We see that the data are indeed clustered around the 45 degree line, thus corroborating the findings from the statistical tests for Gaussianity provided in Table 1 of the main paper. Taken together, the statistical tests of the main paper and the QQ-plots of this document provide evidence that the Gaussian distribution provides an adequate fit to the empirical distribution of the AR(1) residuals of the budget imbalance data studied here.

3 Testing for misreporting in the historical data

The main paper is proposing a test for systematic misreporting of the data in the *future*. However, a key assumption behind the procedure is that data are not systematically misreported in the *past*. This section examines the validity of this assumption.

Let $\{B_t^{\mathrm{IM}}\}_{t=1}^n$ be our historical budget imbalance data. We consider the following simple model for these data:

$$B_t^{\mathrm{IM}} = \mu_t + u_t, \quad t = 1, 2, \ldots, n,$$

where μ_t is a (possibly time-varying) mean parameter and u_t is a zero-mean process. In the main paper, we found that u_t is well-described by a stationary AR(1) process.

To test whether there are any systematic misreportings in the historical data, we proceed using two different methods. The first method (Section 3.1) assumes that μ_t is constant, i.e. $\mu_t = \mu \in \mathbb{R}$ for all $t = 1, 2, \ldots, n$, and then uses the full sample to test whether $\mu = 0$. The second method

(Section 3.2) acknowledges that misreporting might not have been going on through the full period and instead tests whether $\mu_t = 0$ for $t < \tau$ and $\mu_t = \mu \neq 0$ for $t \geq \tau$, where τ is an unknown break point.

3.1 Testing for misreporting in the entire historical period

In this section, we assume that the mean is not time varying, i.e. that $\mu_t = \mu \in \mathbb{R}$ for all $t = 1, 2, \ldots, n$. Now, to test whether there are systematic misreportings in the historical budget imbalance data, we are interested in testing the hypothesis

$$H_0 : \mu = 0, \quad \text{against} \quad H_1 : \mu \neq 0. \tag{3.1}$$

Alternatively, if we are only interested in the alternative that $\mu < 0$ (which would be the case when emissions are under-reported), we could also consider the test

$$H_0 : \mu = 0, \quad \text{against} \quad H_1 : \mu < 0. \tag{3.2}$$

The second test will have greater power to detect under-reportings. Table 2 contains the t-statistics from conducting these tests (see the caption of the table for the critical values for the tests in (3.1) and (3.2)). We provide two t-statistics for the null hypotheses: one calculated assuming no autocorrelation in the budget imbalance data ("t-stat") and one where we control for the autocorrelation in the data ("t-stat (corr)"). In no case can we reject the null hypotheses (3.1) and (3.2).[1]

3.2 Testing for misreporting with unknown starting point

We here entertain the possibility that the misreporting begins at an unknown time τ, where $1 \leq \tau \leq n$. To be precise, we are interested in testing the hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = 0, \quad \text{against} \quad H_1 : \mu_1 = \mu_2 = \ldots \mu_{\tau-1} = 0, \mu_\tau = \mu_{\tau+1} = \ldots \mu_n = \mu,$$

where $\mu \neq 0$ and τ is unknown. This is a classical "off-line" change-point analysis and the literature on this problem is voluminous. For our purposes, it is important to use a procedure which can take into account the autocorrelation in the term u_t. Here, we assume that u_t is an AR(1) process, which means that we can use the theory outlined in, e.g., Example 1 of Aue and Horváth (2012). Consider the process

$$Z_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nx \rfloor} B_t^{\mathrm{IM}}, \quad x \in [0, 1],$$

[1] Using the raw t-stat, it would be possible to reject the null in (3.1) at a 10% significance level for the GCB2019 data; however, after controlling for autocorrelation in the data ("t-stat (corr)"), this apparent significance disappears.

where $\lfloor y \rfloor$ denotes the greatest integer smaller than or equal to $y \in \mathbb{R}$. Under the null, we have (Aue and Horváth, 2012)

$$Z_n \Rightarrow \omega B, \quad n \to \infty,$$

where B is a Brownian motion on $[0,1]$ and ω^2 is the long-run variance of u_t. Here, "$\Rightarrow$" denotes weak convergence in the Skorohod space $D[0,1]$. Let $\hat{\omega}$ denote an estimate of ω, and consider the test statistic

$$M_n = \frac{1}{\hat{\omega}} \min_{k=1,2,\ldots,n} Z_n\left(\frac{k}{n}\right), \tag{3.3}$$

which adheres to

$$M_n \Rightarrow \inf_{x \in [0,1]} B(x), \quad n \to \infty.$$

The reason for considering the minimum of $Z_n\left(\frac{k}{n}\right)$ in Equation (3.3), instead of the maximum of $Z_n\left(\frac{k}{n}\right)$ or, say, the maximum of $\left|Z_n\left(\frac{k}{n}\right)\right|$, is that the relevant alternative hypothesis in our context is where the introduced mean is negative, i.e. $\mu < 0$. For this reason, we prefer only to reject the null when $Z_n\left(\frac{k}{n}\right)$ is sufficiently negative for some $k = 1, 2, \ldots, n$.

It is straightforward to calculate the critical values and p-values of this test by simulation using an approach as that described in Section 3 of the main paper. We can also assess the test visually, by plotting $X_k = \hat{\omega}^{-1} Z_n(k/n)$ and seeing if, for some k, it crosses the critical values. Consider also the alternative test statistic,

$$Z_n^R(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nx \rfloor} B_{n+1-t}^{\mathrm{IM}}, \quad x \in [0,1],$$

which is similar to $Z_n(x)$ but "reversed", in the sense that the sum is constructed by summing budget imbalance terms in reverse, starting from the last term B_n^{IM} and ending with the first term B_1^{IM}. Under the null, the process Z_n^R will enjoy the same limiting behavior as $Z_n(x)$ and can therefore also be used to test the hypothesis considered here. However, using $Z_n^R(x)$ instead of $Z_n(x)$ might improve the power of the test if the τ is close to n, i.e. if the misreporting begins late in the period. Since this is the most likely case in our setting of the budget imbalance, our main test will be based on $Z_n^R(x)$ and the associated test statistic M_n^R which is given as M_n in Equation (3.3) above, but with Z^R in place of Z.

Figure 4 plots the processes $X_k = \hat{\omega}^{-1} Z_n(k/n)$ and $X_k^R = \hat{\omega}^{-1} Z_n^R(k/n)$ for the budget imbalance data in the four data sets GCB2017, GCB2018, GCB2019, and GCB2020, where $\omega = \sigma/(1-\phi)$ can be estimated using OLS estimates of ϕ and σ, as outlined in the main paper. If either process crosses the critical values (black lines) at some point, there is evidence against H_0 at the given confidence level. It is clear that for none of the data series can we reject the null. Indeed, the p-values connected to the main statistic based on M_n^R are 0.56, 1.00, 1.00, and 0.8028 for the four data sets, respectively. (These numbers are also reported in the left-most column of Table 4.)

It is worth providing some remarks about the differences between the test considered in this section and the test proposed in the main paper. Firstly, the test considered in this section is "off-line", in the sense that we are studying a historical data set of a fixed size, n, and running one hypothesis test. In contrast, the test considered in the main paper is "on-line", in the sense that we are conducting a new test every time a new (future) data point arrives. Secondly, in this section, we have studied the four vintages of the GCB data separately and conducted a separate test for each data set. Hence, there was no reason to be concerned with the data series having different statistical properties (such as different values for the coefficients in the AR(1) models describing u_t). In contrast, in the main paper, the focus is on constructing a test which is applicable to future data sets, which might plausibly have different statistical properties. Therefore, we were led to construct the test as "pivotal", in the sense that the objects entering the test statistic did not depend on the statistical properties of a given vintage of the data.

As subsections 3.1 and 3.2 have shown, there is no evidence of a non-zero mean in either the full data sample (Section 3.1) nor for a non-zero mean being introduced at any point during the sample (Section 3.2).[2] The following section uses counterfactual analyses to investigate what kind of misreportings we would have been able to detect, if they had been present in the historical sample.

3.3 Counterfactual analyses

Recall that the budget imbalance data are given by

$$B_t^{\mathrm{IM}} = E_t^{\mathrm{FF}} + E_t^{\mathrm{LUC}} - S_t^{\mathrm{LND}} - S_t^{\mathrm{OCN}}.$$

In this section, we analyse data on a *counterfactual budget imbalance* CB_t^{IM}, given by

$$CB_t^{\mathrm{IM}} = (1 - m_t) \cdot E_t^{\mathrm{FF}} + E_t^{\mathrm{LUC}} - S_t^{\mathrm{LND}} - S_t^{\mathrm{OCN}}, \tag{3.4}$$

where $m_t \in [0, 1]$ controls the amount of under-reporting of fossil fuel emission at time t. Note that if $m_t = 0$ for all t we have $CB_t^{\mathrm{IM}} = B_t^{\mathrm{IM}}$, which we analysed in the previous two sections.

Consider first the case, where we have misreporting during the entire sample, i.e. $m_t = m > 0$ for all t. In this case, we might detect the misreporting by testing the overall mean of the data CB_t^{IM} with an approach similar to the one in Section 3.1 above. In Table 3, we report the results from this when $m = 0.05$ and $m = 0.10$, i.e. in the cases where there is 5% and 10% under-reporting of fossil fuel emissions in the entire sample. These results should be compared with those in Table 2, where $m = 0$. As expected, and contrary to what was found for the actual data in Section 3.1 (Table 2), we see that the overall mean is always negative for the counterfactual data. Further, as

[2] In this section, we have testet for the alternative of a negative mean being introduced, i.e. $\mu < 0$, since this appears to be the most relevant alternative in the application of the main paper. Although not shown here, we also cannot reject the null in the case of a non-zero mean, i.e. $\mu \neq 0$, nor, indeed, in the case of a positive mean, i.e. $\mu > 0$.

seen from the t-statistics, we can always reject the null hypothesis in (3.2) at the 5% level when $m = 0.10$ (bottom panel), while this is the case for two out of the four data series when $m = 0.05$ (top panel).

Consider now the case where $m_t = 0$ for $t < \tau$ and $m_t = m > 0$ for $t \geq \tau$, i.e. the case where misreporting does not start until the period τ, after which fossil fuel emissions are under-reported by a fraction m. This setting can be analyzed with the approach considered in Section 3.2. Figure 5 plots the test statistic $Z_n^R(k/n)$ for the four counterfactual data sets when $\tau \in \{30, 45\}$ and $m \in \{0.05, 0.10\}$. Here, $\tau = 30$ corresponds to misreporting starting roughly halfway through the sample, while $\tau = 45$, which corresponds to misreporting starting roughly 75% through the sample. When $m = 0.05$, fossil fuel emissions are under-reported by 5% per year, starting in year τ, while when $m = 0.10$, the under-reporting is 10% per year. The p-values for the test based on M_n^R are given in Table 4.

As can be seen from Figure 5, and the p-values reported in Table 4, when $\tau = 30$ and $m = 0.10$, we can reject the null at a 5% level for all four data series; when $\tau = 45$ and $m = 0.10$, we can reject the null at a 10% level for the first data series and at a 5% level for the last data series. When the amount of under-reporting is smaller, i.e. $m = 0.05$, it of course becomes more difficult to detect. However, when $\tau = 30$, we can still reject the null at a 10% level for the first data series and at a 5% level for the last data series. When $m = 0.05$ and $\tau = 45$, we can only reject the null at a 10% level for the last data series.

4 Simulation studies

This section will investigate the finite sample properties of the monitoring procedure proposed in the main paper when it is applied to simulated data. We set up the simulation study such as to mimic the setting of the Global Carbon Budget data studied in Section 2 of the main paper. To be precise, for $n \geq 1$, we consider the following AR(1) data generating process (DGP)

$$u_t^n = \phi_n u_{t-1}^n + \sigma_n \epsilon_t^n, \qquad t = 1, 2, \ldots, n,$$

where $\phi_n \in (0, 1)$, $\sigma_n > 0$, and $\epsilon_t^n \sim N(0, 1)$ is an iid sequence for $t = 1, 2, \ldots, n$. To capture the fact that in the GCP data ϵ_t^n appears to be correlated with ϵ_t^m for $m \neq n$, cf. the bottom plot of Figure 1 in the main paper, we let $Corr(\epsilon_t^n, \epsilon_t^{n+1}) = \rho$ for $t = 1, 2, \ldots, n$, with $\rho = 0.90$ (the results are very insensitive to the choice of ρ). The process u_t^n is initialized at its stationary distribution, i.e. $u_0^n \sim N\left(0, \sigma_n^2/(1 - \phi_n^2)\right)$.

We consider the following four specifications for the DGP of u_t^n:

- DGP1: $\phi_n = 0.35$, $\sigma_n = 0.72$ for all n.
- DGP2: $\phi_n = \frac{0.35}{2}$, $\sigma_n = 0.72$ for all n.
- DGP3: $\phi_n = 0.35$, $\sigma_n = \frac{0.72}{2}$ for all n.

- DGP4: $\phi_n \sim \mathrm{U}(0,1)$, $\sigma_n \sim \mathrm{Exp}(0.72)$ for all n.[3]

DGP1 is designed to mimic the budget imbalance data from GCB2020, cf. Table 1 of the main paper. DGP2 and DGP3 are similar to DGP1, but with lower values for the autoregressive parameter ϕ and standard deviation of the error term σ, respectively. Thus, the results from DGP2 will allow us to gauge the effect of a less autocorrelated budget imbalance, while those from DGP3 will allow us to gauge the effect of a less volatile budget imbalance. Lastly, DGP4 is designed to examine the effects of letting the parameters of the DGP vary (here, independently) from vintage to vintage.

Using a specific DGP, we simulate u_t^n, for $n = K+1, K+2, \ldots, K+T$ and $t = 1, 2, \ldots, n$, where K is the length of the initial period and T is the length of the monitoring period. In the study, we consider $K, T \in \{30, 60, 120\}$. For $n = K+1, K+2, \ldots, K+T$, the observations available for the procedure are $\{y_t^n\}_{t=1}^n$ with

$$y_t^n = u_t^n + \xi_t^n, \qquad t = 1, 2, \ldots, n,$$

where $\xi_t^n = 0$ for $t < \tau$ and non-zero otherwise. Using the procedure outlined in Section 3 of the main paper, the null hypotheses in Equation (3.3) of that section are sequentially tested for $n = K+1, K+2, \ldots, K+T$. All simulation studies presented below follow this procedure for 10 000 Monte Carlo replications.

4.1 Empirical size of the test of H_0

To see how the proposed monitoring procedure performs under the null, i.e. when CO_2 emissions are faithfully reported, we here set $\tau = \infty$ such that $\{y_t^n\}_{t=1}^n = \{u_t^n\}_{t=1}^n$ for all n. In other words, we set $\xi_t^n = 0$ for all t and n. Here, where the null hypothesis H_0 is true, any rejection of this null will be an instance of a "false positive" (Type I error). By construction, the probability of a Type I error is asymptotically (as $K \to \infty$) equal to the significance level α.

Table 5 presents the rejection rates of H_0 in this case where H_0 is true. From the table, we see that the test is slightly over-sized when K is small ($K = 30$). For moderate values of K, the empirical size of the test is very close to the nominal size. In particular, for the case $K = 60$, which is most relevant for our application, the empirical size of the test is satisfactory. These comments hold for all four DGPs considered.

4.2 Empirical power of the test of H_0

To see how the proposed monitoring procedure performs when the null is false, i.e. when there is under-reporting of CO_2 emissions, we here set $\tau = K+1$ meaning a structural break is introduced

[3] We write $X \sim \mathrm{U}(a,b)$ for a random variable which is uniformly distributed on the interval $[a,b]$ with $a < b$; we write $X \sim \mathrm{Exp}(\lambda)$ for a random variable which is exponentially distributed with parameter $\lambda > 0$. Note that in the latter case, we have $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda^2$.

immediately after the initial period. The structural break process is defined as the difference between "reported CO_2 emissions" ($E_t^{\text{FF},*}$) and "actual CO_2 emissions" (E_t^{FF}), i.e. for all n set

$$\xi_t^n = \xi_t = E_t^{\text{FF},*} - E_t^{\text{FF}}, \qquad t = K+1, K+2, \ldots, K+T.$$

Note that, in practice, only $E_t^{\text{FF},*}$ would be observed by the researcher and both E_t^{FF} and ξ_t would be unobserved. In the simulation study, we consider two specifications for the paths of the emissions processes:

- Exponential Abatement:

$$E_t^{\text{FF},*} = E_o(1-g)^{t-K}, \quad \text{and} \quad E_t^{\text{FF}} = (1-m)E_t^{\text{FF},*} + mE_o,$$

  where $E_o = 9.9456$, $g = 0.0692$, $m \in [0,1]$, and $t = K+1, K+2, \ldots, K+T$.

- Linear Abatement:

$$E_t^{\text{FF},*} = \max(E_o - a(t-K), 0), \quad \text{and} \quad E_t^{\text{FF}} = (1-m)E_t^{\text{FF},*} + mE_o,$$

  where $E_o = 9.9456$, $a = 0.4931$, $m \in [0,1]$, and $t = K+1, K+2, \ldots, K+T$.

Here, the initial amount of reported emissions is $E_o = 9.9456$, which correspond to the level of fossil fuel emissions in 2019 according to the GCB2020 data set, i.e. $E_{2019}^{\text{FF}} = 9.9456$ GtC/yr. The decay rate (g) in the Exponential Abatement case and the slope (a) in the Linear Abatement case are set such that reported emissions after 11 periods are equal to $\frac{1}{2}E_{2010}^{\text{FF}}$, where $E_{2010}^{\text{FF}} = 9.0430$ GtC/yr. These numbers are chosen to mimic a CO_2 abatement situation consistent with the Paris Agreement; recall from the introduction of the main paper, that to achieve the Paris objectives, CO_2 emissions should be cut in approximately half, as compared to 2010 levels, by 2030.

In both abatement settings, the form of the actual emissions, E_t^{FF}, is set to $E_t^{\text{FF}} = (1-m)E_t^{\text{FF},*} + mE_o$. Here, we can interpret the *misreporting parameter* m to be the fraction of the emissions that are not abated, while a fraction $(1-m)$ of the emissions are abated according to the Paris objectives. This situation would obtain if all countries report their fossil fuel emissions in a way which is consistent with the Paris objectives, but a number of countries, representing a fraction m of emissions in 2019, actually keep their emissions constantly equal to their 2019 levels (i.e. they are under-reporting their emissions). Figure 6 illustrates the paths of $E_t^{\text{FF},*}$ and E_t^{FF} for different values of m and the two abatement schemes. Which of the two schemes are most relevant in practice is an empirical question, which might be answered when (and if) the international community begins CO_2 abatement in earnest.

As we shall see, the empirical properties of the test are broadly similar in these two cases, indicating that it is the *magnitude*, and not the *form*, of the abatement which matters for detecting potential systematic under-reporting of CO_2 emissions.

We report the power of the monitoring test of H_0, i.e. the rejection rate of H_0 when H_1 is true, calculated over 10 000 Monte Carlo replications. We also report the (conditional) "Average

detection time", i.e. the time from the structural break occurs until the statistical test rejects H_0, conditional on a detection being made. Hence, the "Average detection time" is only an accurate description of the actual average detection time when the power of the test is close to one. Figure 7 reports the results for the Exponential Abatement scheme and Figure 8 shows the results for the Linear Abatement scheme. We report the results for $K = 60$, $T = 30$; the results for other settings are similar.

From Figure 7 and Figure 8 we see, as expected, that the power of the test is quite low when the amount of under-reporting (captured by the parameter m) is small. Conversely, when the amount of under-reporting is moderate-to-high, the power of the test is close to one, meaning that the under-reporting of CO_2 emissions is (practically) always detected during the monitoring period $\{K+1, K+2, \ldots, K+T\}$. Compared to DGP1, the power and the average detection time are slightly superior in the case of DGP2 and very superior in the case of DGP3. This indicates that lowering the autocorrelation of the budget imbalance data will result in slightly better properties of the test under the alternative, while lowering the standard deviation will greatly improve the properties of the test under the alternative. DGP4 exhibits the best power properties and detection times of all the DGPs studied here; however, this finding is of limited interest, since DGP4 is not a realistic representation of the budget imbalance data (that is, while we expect that ϕ and σ might change from one vintage to the next, we do not expect the values of these parameters to be wholly independent of each other from year to year).

For DGP1, which is the most relevant DGP in practice, since here the parameters are set to those obtained by the most recent data set GCB2020, we see that the average detection time is quite large for the smaller amounts of under-reportings, i.e. for small values of m. Indeed, in both the Exponential Abatement and Linear Abatement cases, it is only for $m \geq 0.10$ that the power (i.e. probability of detecting under-reporting inside the monitoring period) becomes close to unity. For the Exponential Abatement case (Figure 7) , $m = 0.20$ results in an average detection time between 7 years ($\alpha = 32\%$) and 12 years ($\alpha = 5\%$); for $m = 0.30$, the average detection time is between 5 years ($\alpha = 32\%$) and 10 years ($\alpha = 5\%$); and for $m \geq 0.35$ the average detection time is smaller than 5 years (for $\alpha = 32\%$). Similar, but marginally larger, numbers are found in the case of Linear Abatement (Figure 8).

References

Aue, A. and L. Horváth (2012). Structural breaks in time series. *Journal of Time Series Analysis 34*(1), 1–16.

Dickey, D. A. and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association 74*(366a), 427–431.

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.

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

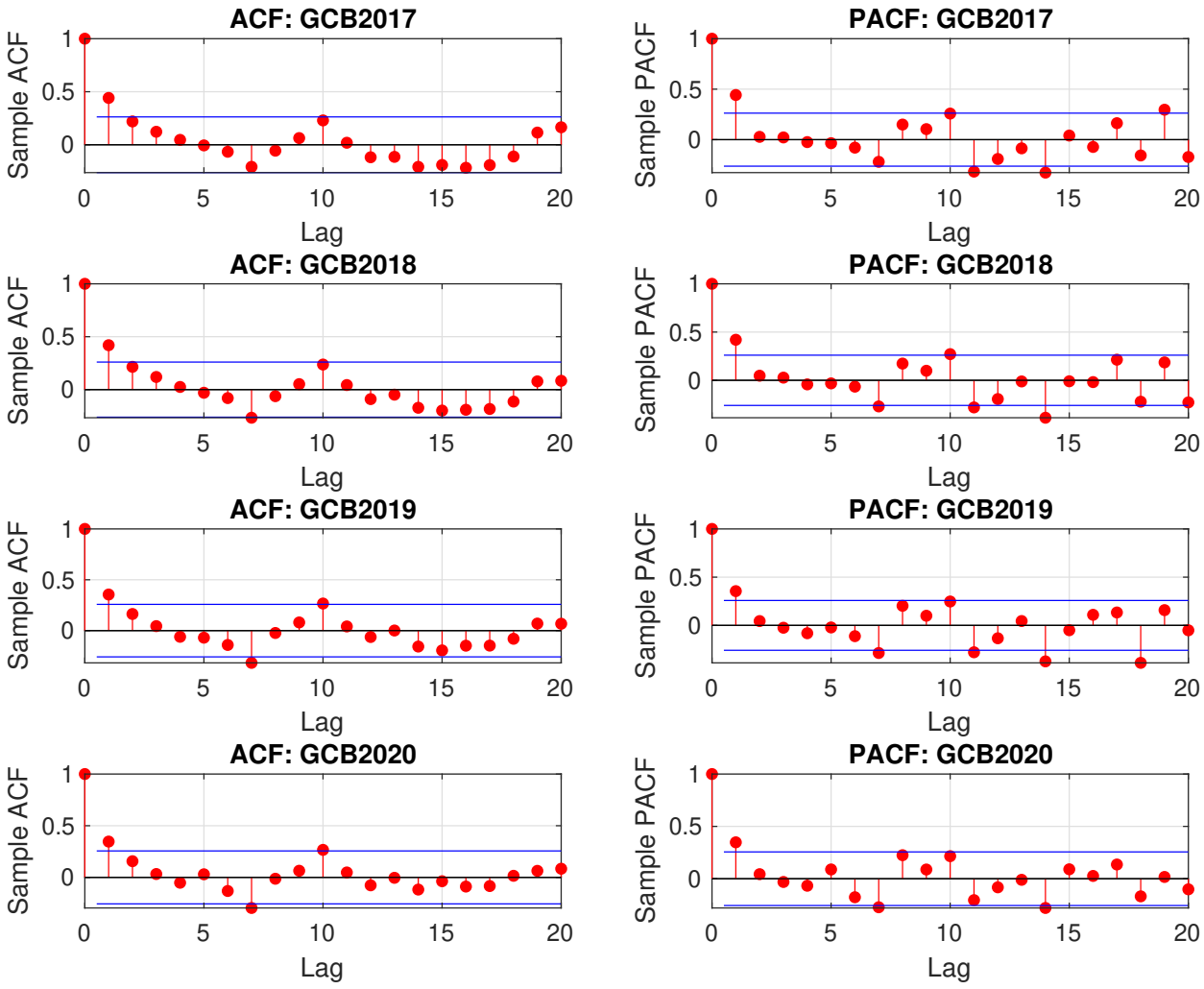

Figure 1: *Autocorrelation (ACF) and partial autocorrelation (PACF) plots of the budget imbalance data.*

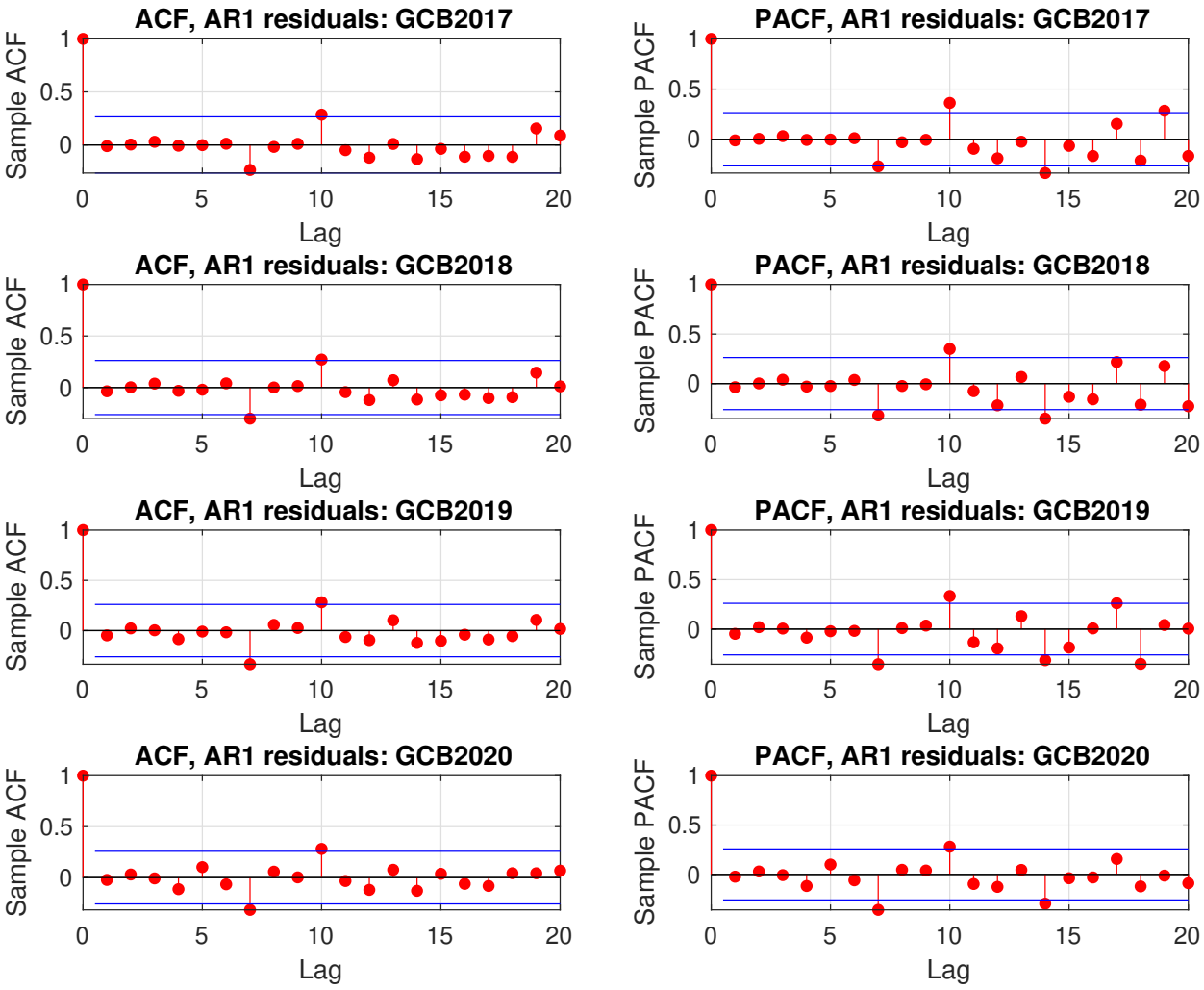

Figure 2: *Autocorrelation (ACF) and partial autocorrelation (PACF) plots of the AR(1) residuals from the budget imbalance data.*

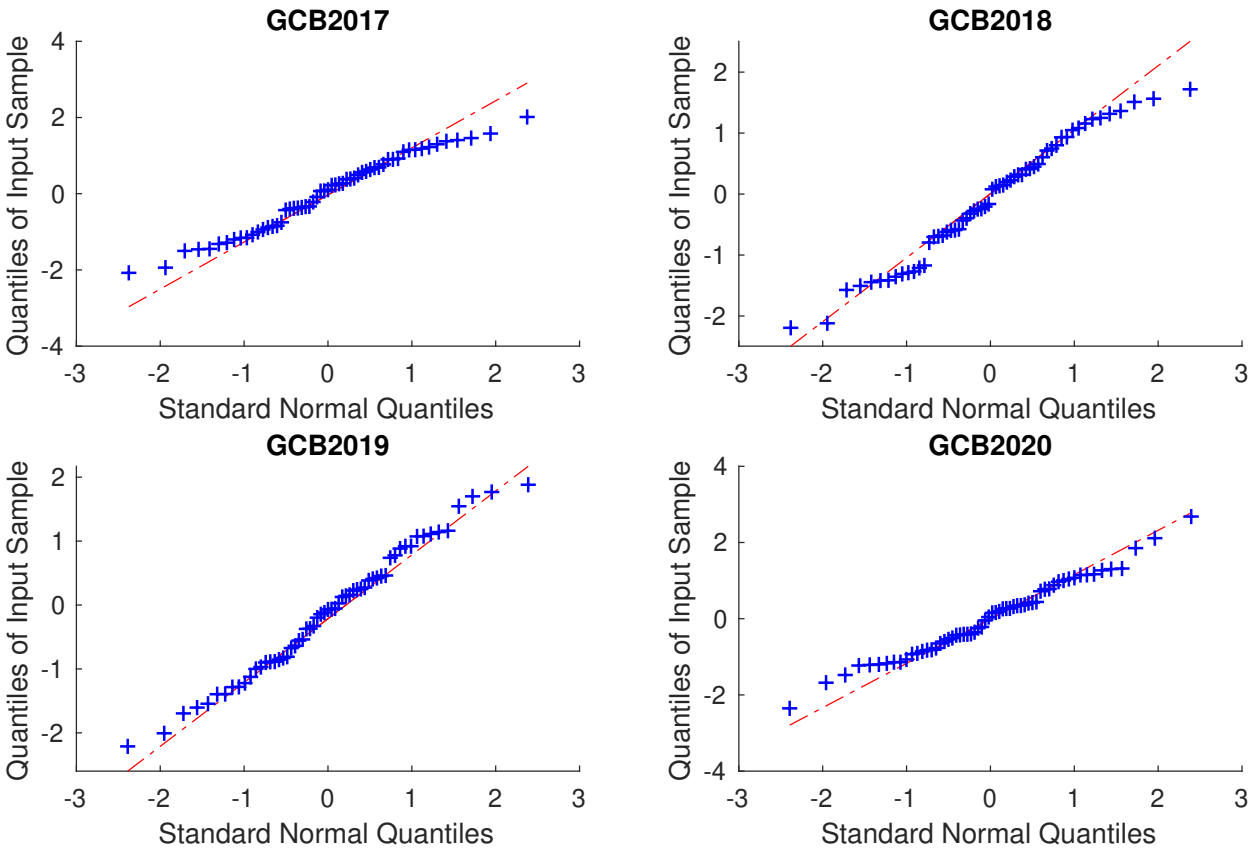

Figure 3: *QQ-plot for the AR(1) residuals.*

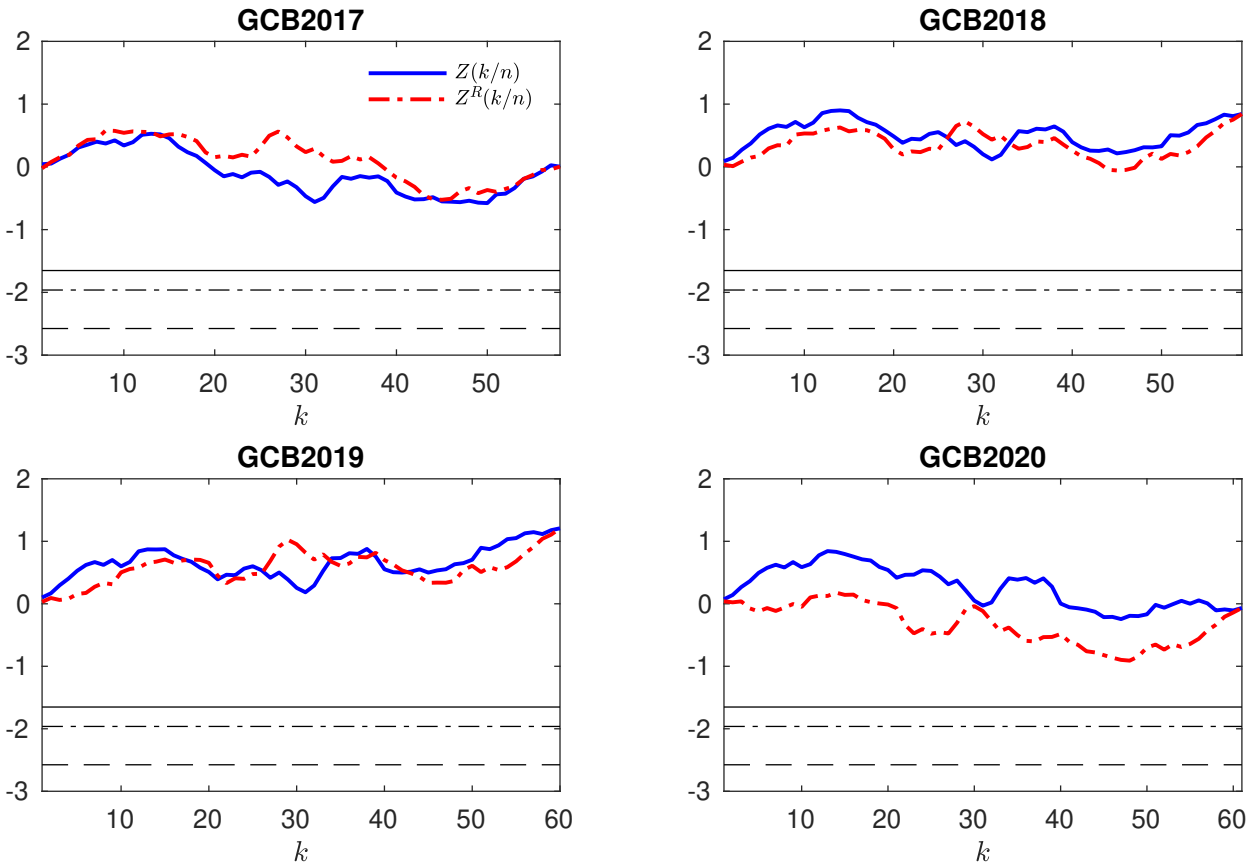

Figure 4: *Paths of $Z_n(k/n)$ (solid blue) and $Z_n^R(k/n)$ (dashed red), along with the critical values for the M_n and M_n^R statistics (black) for the significance levels* 10%, 5%, *and* 1%. *Data: Historical budget imbalance B_t^{IM}, as presented in Section 2 of the main paper. The p-values associated to the test based on M_n^R, through the $Z_n^R(k/n)$ processes shown here, are given in the left-most column of Table 4.*

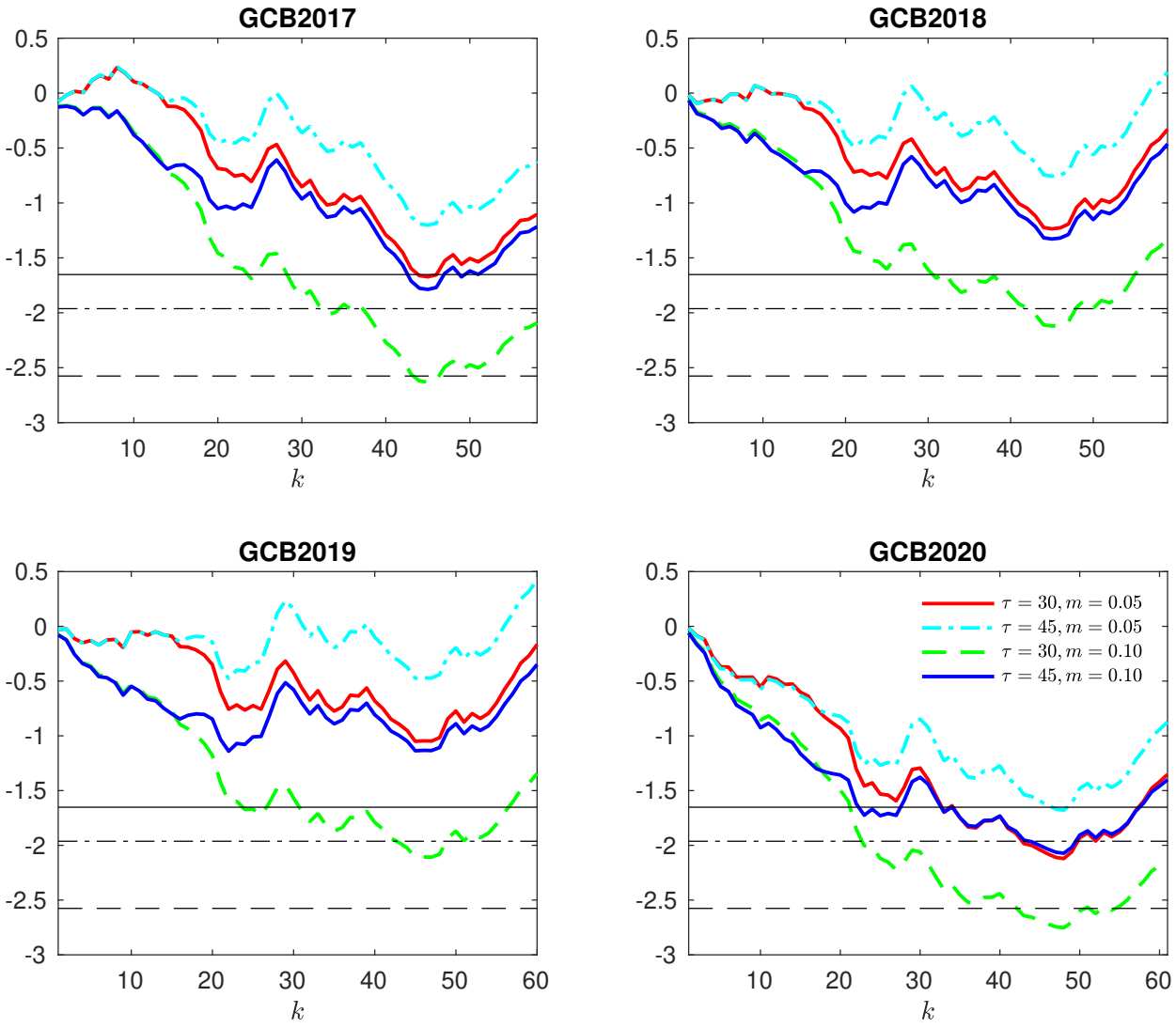

Figure 5: *Paths of* $Z_n^R(k/n)$*, along with the critical values for the* M_n^R *statistic (black) for the significance levels* 10%*,* 5%*, and* 1%*. Data: Counterfactual budget imbalance* CB_t^{IM} (3.4) *with* $\tau \in \{30, 45\}$ *and* $m \in \{0.05, 0.10\}$*. The p-values corresponding to the* M_n^R *statistics are given in Table 4.*

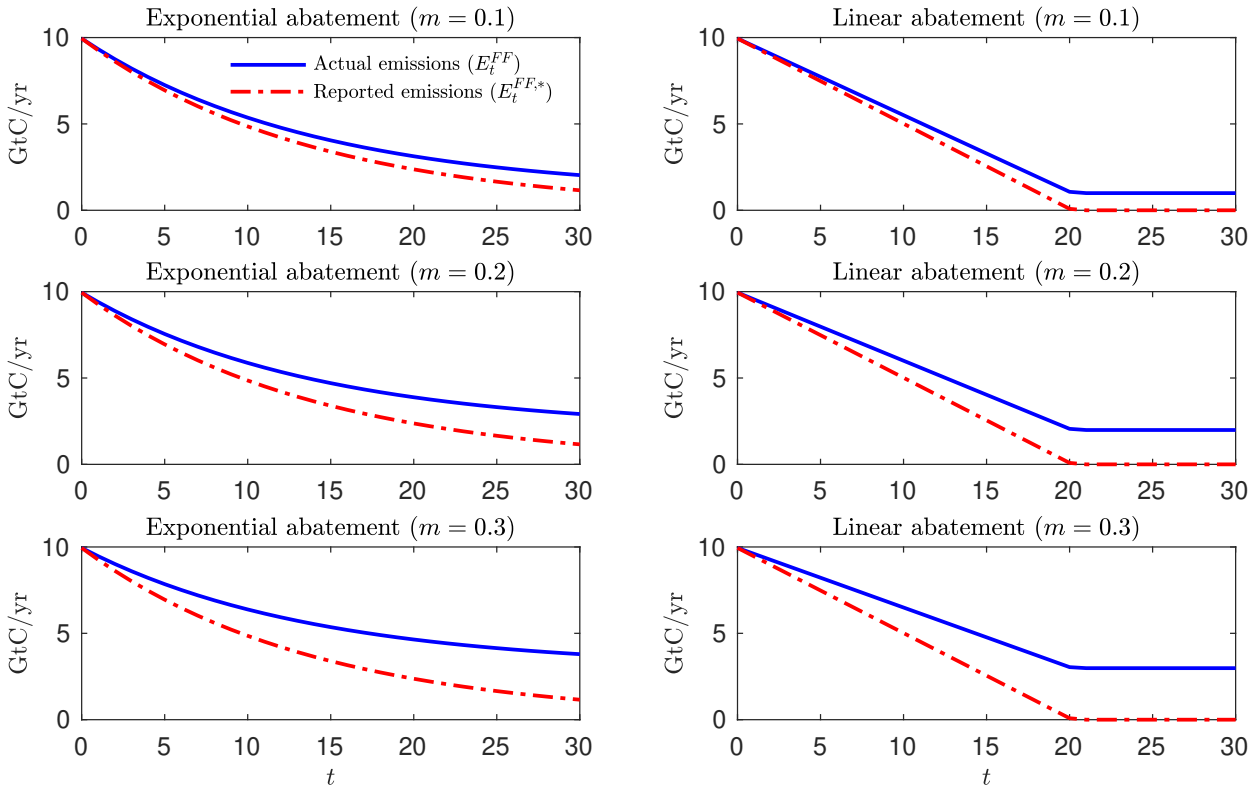

Figure 6: *Example paths of actual emissions ($E_t^{FF,*}$) and reported emissions (E_t^{FF}), for various values of misreporting parameter m, used in the simulation studies of power properties.*

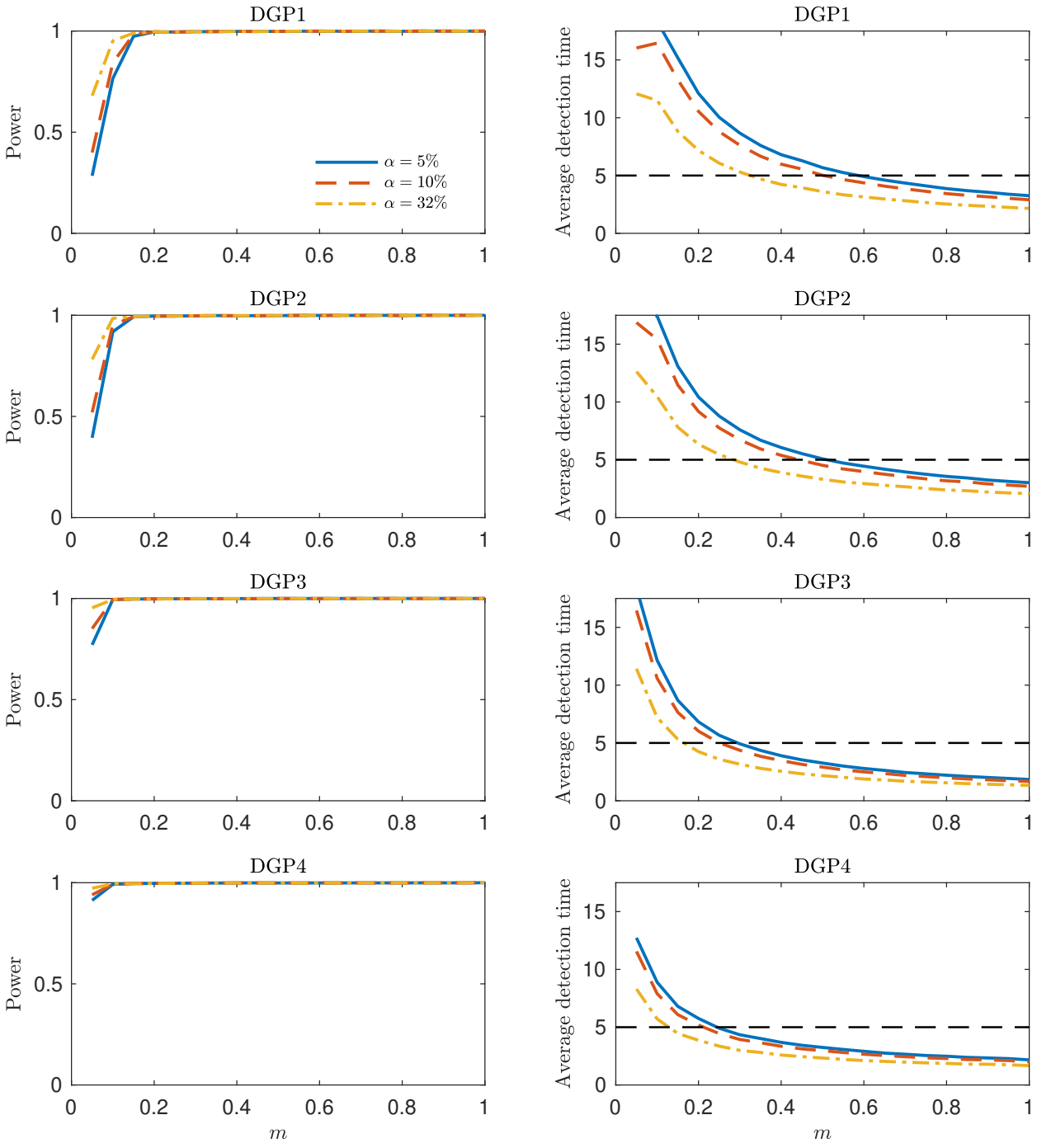

Figure 7: *Exponential abatement: Detection rate (power) and average detection time (conditional on a detection having been made) as a function of misreporting parameter* m*, calculated from* 10 000 *Monte Carlo replications. We consider significance levels* $\alpha = 5\%$ *(solid line),* 10% *(dashed line),* 32% *(dashed and dotted line). The black dashed horizontal line denotes* 5 *years.*

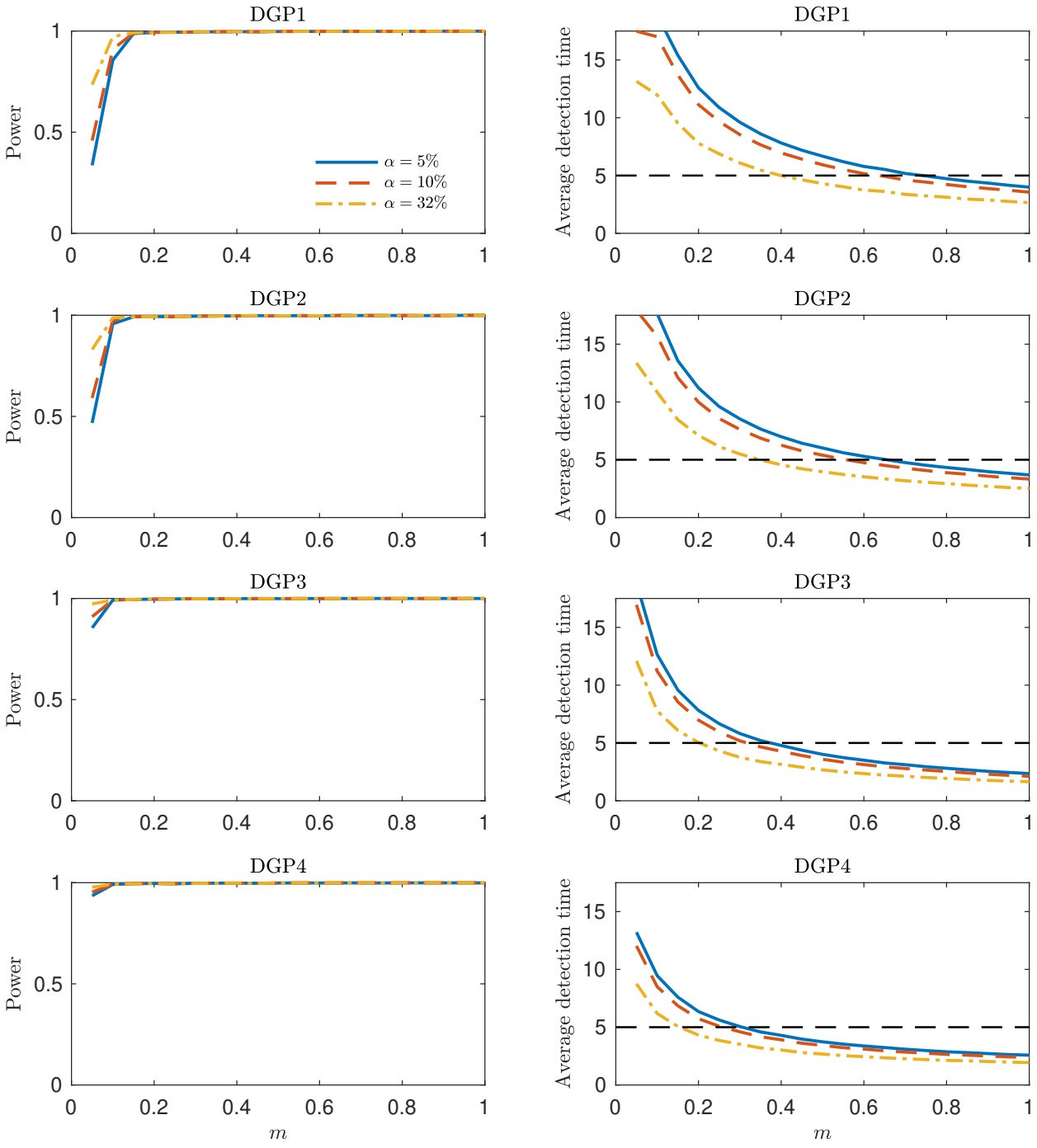

Figure 8: *Linear abatement: Detection rate (power) and average detection time (conditional on a detection having been made) as a function of misreporting parameter* m*, calculated from* 10 000 *Monte Carlo replications. We consider significance levels* $\alpha = 5\%$ *(solid line),* 10% *(dashed line),* 32% *(dashed and dotted line). The black dashed horizontal line denotes* 5 *years.*

Table 1: *Testing for stationarity of the data using the KPSS test (Kwiatkowski et al., 1992) and for a unit root using the Augmented Dickey-Fuller test (ADF, Dickey and Fuller, 1979). Numbers equal to one indicate that the test rejects the null, while numbers equal to zero indicate a failure to reject the null. The tests are performed at a 5% significance level. In the KPSS test, the null is that the data are stationarity and the alternative is that the data contain a unit root. In the ADF test, the null is that the data contain a unit root and the alternative is that the data are stationary. The table reports the results from using* $0, 1, \ldots, 5$ *lags for computing the long run variance of the data using the estimator proposed in Newey and West (1987).*

	KPSS						ADF					
Number of lags:	0	1	2	3	4	5	0	1	2	3	4	5
Budget imbalance												
GCB2017	0	0	0	0	0	0	1	1	1	1	1	1
GCB2018	0	0	0	0	0	0	1	1	1	1	1	1
GCB2019	0	0	0	0	0	0	1	1	1	1	1	1
GCB2020	0	0	0	0	0	0	1	1	1	1	1	1
AR(1) *residuals*												
GCB2017	0	0	0	0	0	0	1	1	1	1	1	1
GCB2018	0	0	0	0	0	0	1	1	1	1	1	1
GCB2019	0	0	0	0	0	0	1	1	1	1	1	1
GCB2020	0	0	0	0	0	0	1	1	1	1	1	1

Table 2: *Testing the null of a zero-mean in the entire historical data set* B_t^{IM}*. "t-stat" is the t-statistic for the null hypotheis that the mean is equal to zero assuming that the data are iid. "t-stat (corr)" is the corresponding t-statistic, corrected for the autocorrelation in the data (assuming an AR(1) structure). The two-sided critical values for the the alternative of a non-zero mean* (3.1) *are* 2.58, 1.96, *and* 1.64 *for tests at the* 1%, 5%, *and* 10% *significance levels, respectively. The one-sided critical values for the alternative of a negative mean* (3.2) *are* -2.33, -1.64, *and* -1.28 *for tests at the* 1%, 5%, *and* 10% *significance levels, respectively.*

	n	Mean	Std	t-stat	t-stat (corr)
Historical budget imbalance data					
GCB2017	58	0.00	0.86	0.01	0.00
GCB2018	59	0.14	0.80	1.31	0.84
GCB2019	60	0.17	0.77	1.74	1.22
GCB2020	61	-0.01	0.77	-0.10	-0.07

Table 3: *Testing the null of a zero-mean in the entire counterfactual data set* CB_t^{IM}. *"t-stat" is the t-statistic for the null hypotheis that the mean is equal to zero assuming that the data are iid. "t-stat (corr)" is the corresponding t-statistic, corrected for the autocorrelation in the data (assuming an AR(1) structure). The two-sided critical values for the the alternative of a non-zero mean* (3.1) *are* 2.58, 1.96, *and* 1.64 *for tests at the* 1%, 5%, *and* 10% *significance levels, respectively. The one-sided critical values for the alternative of a negative mean* (3.2) *are* −2.33, −1.64, *and* −1.28 *for tests at the* 1%, 5%, *and* 10% *significance levels, respectively.*

	n	Mean	Std	t-stat	t-stat (corr)
5% *under-reporting*					
GCB2017	58	−0.30	0.85	−2.64	−1.67
GCB2018	59	−0.16	0.81	−1.53	−0.98
GCB2019	60	−0.13	0.78	−1.30	−0.90
GCB2020	61	−0.31	0.80	−3.08	−2.07
10% *under-reporting*					
GCB2017	57	−0.59	0.86	−5.23	−3.29
GCB2018	58	−0.46	0.83	−4.23	−2.63
GCB2019	59	−0.43	0.80	−4.19	−2.80
GCB2020	60	−0.62	0.84	−5.78	−3.67

Table 4: *p-values associated to the one-sided test of the presence of a non-zero mean at an unknown point in the budget imbalance data using the test statistic* M_n^R *as explained in the test. The amount of under-reporting is denoted by* m *and* τ *is the time at which the non-zero mean (under-reporting) is introduced into the data. The first column with* $\tau = \infty$ *and* $m = 0$ *represents no misreporting, i.e. the actual budget imbalance data* B_t^{IM}. *These p-values are associated to the paths of* $Z_n^R(k/n)$ *in Figure 4 (red lines). The paths of the* $Z_n^R(k/n)$ *processes associated with the last four columns, i.e. the counterfactual data* CB_t^{IM} (3.4), *where under-reporting is present, are shown in Figure 5.*

	$\tau = \infty, m = 0$	$\tau = 30, m = 0.05$	$\tau = 45, m = 0.05$	$\tau = 30, m = 0.10$	$\tau = 45, m = 0.10$
GCB2017	0.5974	0.0956	0.2298	0.0085	0.0746
GCB2018	0.9495	0.2174	0.4486	0.0343	0.1856
GCB2019	1.0000	0.2941	0.6289	0.0350	0.2541
GCB2020	0.3625	0.0342	0.0939	0.0059	0.0386

Table 5: *Empirical size of monitoring scheme, i.e. rejection rates of H_0 over* 10 000 *Monte Carlo replications. The number K denotes the length of the "initial period", while T denotes the length of the "monitoring period". See the text for further details on the simulation setup*

	DGP1			DGP2			DGP3			DGP4		
Nominal size:	5%	10%	32%	5%	10%	32%	5%	10%	32%	5%	10%	32%
$K = 30, T = 30$	0.08	0.14	0.36	0.08	0.14	0.35	0.08	0.14	0.36	0.09	0.15	0.37
$K = 30, T = 60$	0.09	0.15	0.37	0.09	0.14	0.36	0.09	0.15	0.37	0.10	0.16	0.39
$K = 30, T = 120$	0.09	0.15	0.37	0.08	0.14	0.36	0.09	0.15	0.37	0.12	0.18	0.40
$K = 60, T = 30$	0.07	0.12	0.34	0.07	0.12	0.34	0.07	0.12	0.34	0.08	0.13	0.34
$K = 60, T = 60$	0.07	0.12	0.34	0.06	0.12	0.34	0.07	0.12	0.34	0.07	0.13	0.35
$K = 60, T = 120$	0.07	0.12	0.34	0.07	0.12	0.34	0.07	0.12	0.34	0.08	0.14	0.36
$K = 120, T = 30$	0.06	0.11	0.32	0.06	0.11	0.33	0.06	0.11	0.32	0.06	0.11	0.34
$K = 120, T = 60$	0.06	0.11	0.33	0.05	0.10	0.32	0.06	0.11	0.33	0.06	0.11	0.33
$K = 120, T = 120$	0.06	0.11	0.33	0.06	0.11	0.33	0.06	0.11	0.33	0.06	0.12	0.34