Mitigation and Adaptation Strategies for Global Change (2019) 24:1143–1168
https://doi.org/10.1007/s11027-019-09866-5

ORIGINAL ARTICLE

# Learning in greenhouse gas emission inventories in terms of uncertainty improvement over time

Jolanta Jarnicka[1] · Piotr Żebrowski[2]

Received: 31 August 2018 / Accepted: 26 February 2019 / Published online: 1 June 2019

## Abstract

This paper addresses the problem of learning in greenhouse gas (GHG) emission inventories understood as reductions in uncertainty, i.e., inaccuracy and/or imprecision, over time. We analyze the National Inventory Reports (NIRs) submitted annually to the United Nations Framework Convention on Climate Change. Each NIR contains data on the GHG emissions in a given country for a given year as well as revisions of past years' estimates. We arrange the revisions, i.e., estimates of historical emissions published in consecutive NIRs into a table, so that each column contains revised estimates of emissions for the same year, reflecting different realizations of uncertainty. We propose two variants of a two-step procedure to investigate the changes of uncertainty over time. In step 1, we assess changes in inaccuracy, which we consider constant within each revision, by either detrending the revisions using the smoothing spline fitted to the most recent revision (method 1) or by taking differences between the most recent revision and the previous ones (method 2). Step 2 estimates the imprecision by analyzing the columns of the data table. We assess learning by detecting and modeling a decreasing trend in inaccuracy and/or imprecision. We analyze carbon dioxide ($CO_2$) emission inventories for the European Union (EU-15) as a whole and its individual member countries. Our findings indicate that although there is still room for improvement, continued efforts to improve accounting methodology lead to a reduction of uncertainty of emission estimates reported in NIRs, which is of key importance for monitoring the realization of countries' emission reduction commitments.

**Keywords** Uncertainty · Inaccuracy · Imprecision · GHG emission inventory · Learning · Regression model

✉ Jolanta Jarnicka
jolanta.jarnicka@ibspan.waw.pl

[1] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

[2] International Institute for Applied Systems Analysis, Laxenburg, Austria

## 1 Introduction

Assessing the uncertainty of greenhouse gas (GHG) inventories is a complex problem that has been investigated for many years; however, no commonly accepted solution has been found. Low uncertainty of GHG emission inventories, namely, high accuracy and precision of emission estimates, is key to setting reduction targets for climate treaties (Jonas et al. 2010), monitoring treaty implementation (Bun et al. 2010), and establishing reliable emission trading schemes (Ermolieva et al. 2014).

According to the Guidelines for National Greenhouse Gas Inventories (cf. IPCC 2006, vol 1, Ch. 3), *accuracy* is an agreement between the true value and the average of repeated measured observations or estimates of a variable. Thus, *inaccuracy* (systematic error) is a result of failure to capture all relevant processes involved, because the available data are not representative of all real-world situations, or because of instrument error. *Precision*, in turn, is the agreement among repeated measurements or estimates of the same variable. High precision corresponds to a low random error.

Over time, as methods for accounting GHG emissions evolve (from the tier 1 and tier 2 approaches recommended in IPCC (2000, 2006) to the tier 3 approach considered in IPCC (2006), both the accuracy and precision of GHG inventories may change, undermining or improving the effectiveness of policies. The evolution of accounting methodology is particularly well reflected in the emission estimates published each year by the parties to the United Nations Framework Convention on Climate Change (UNFCCC) in the form of National Inventory Reports (NIRs). Each of these reports contains GHG emission data for a given year and revised estimates of past years' emissions. These estimates are considered to reflect the best available knowledge and are therefore treated as "true emissions." Yet, they are bound to change with the following year's revisions, as new data and knowledge about emission sources and processes become available to the institutions preparing the GHG inventories. The emergence of this new knowledge may allow the reporting institutions "to learn" how to prepare better quality GHG inventories. Here, we understand learning in a positive (not normative) sense as a detectable increase in the accuracy of revisions and/or an increase in the precision of initial estimates of new GHG emissions over time.

The problem of investigating learning is in line with the discussion on uncertainty assessment of NIRs considered, for example, in Nahorski and Jęda (2007), where the uncertainty of each reported revision was analyzed separately, and in Marland et al. (2009) and Hamal (2010), where changes in uncertainty over time were investigated. The concept of learning was also discussed in Żebrowski et al. (2015). Here, we especially build upon the work of Jarnicka and Nahorski (2015), and Jarnicka and Nahorski (2016), where models for evolution of uncertainty structure over time were developed and applied to $CO_2$ emission inventories submitted by parties to the UNFCCC in their NIRs; however, we distinguish between uncertainty related to reported revisions and uncertainty related to emissions, referring to them as inaccuracy and imprecision. This allows for learning to be considered in terms of reduction of inaccuracy and imprecision over time.

In this paper, we discuss methods of detecting and assessing learning in a set of consecutive NIRs. More specifically, we exclude estimates of carbon dioxide ($CO_2$) emissions from the land use, land use change, and forestry (LULUCF) sector, as the uncertainties of LULUCF emissions are large and may easily overshadow subtle trends in emission estimates. Detecting learning requires a two-stage analysis. First, information on inaccuracy and imprecision needs to be extracted from revisions of GHG inventories. We deal with this problem in Section 2, where we describe our main method of assessing uncertainty components (method 1), based on the detrending of consecutive revisions. Subtraction of the estimated trend extracts inaccuracy and the transformed emission estimates are thus used to evaluate imprecision. The method works on the assumption that detrending "cleans" the data of the information on the "real emission,"[1] leaving only the inventory uncertainty. To assess the quality of this "cleaning," we use an auxiliary method (method 2), which follows a similar analysis, but with the estimated trend being replaced by the most recent revision of historical emission estimates. We conclude Section 2 with a graphical illustration of methods 1 and 2. The second stage of our analysis—the detection of learning—is discussed in Section 3; there, we consider the question of detecting trends in changes in inaccuracy and imprecision over time and how to interpret those trends as learning and develop an algorithm to detect and assess learning (algorithm 1). Section 4 presents the results obtained by applying this procedure to the GHG emission inventories of the EU-15 and its individual member countries. Section 5 presents conclusions.

## 2 Data presentation and uncertainty assessment

The idea of investigating learning is strictly connected with the structure of the data. Each report contains inventory data on GHG emissions from a given year and revised estimates of emissions in past years, back to 1990; in other words, it contains a revised time series of historical emissions. The NIRs are submitted annually, providing revisions for the data from 2001 up to 2015.[2] We organize these data in a table, the rows of which consist of estimates published in consecutive NIRs, as presented in Table 1. The $j$-th row of Table 1 corresponds to revisions of estimates published in the year[3] $j$ and relating to emission years $n = 1990, \dots ,j$. The $E_j^n$ symbol denotes the inventory data for the year $n$, revised in the year $j$. The $n$th column of Table 1 contains the estimates of emissions for the year $n$, revised in years $j = 2001, \dots ,2015$.

We start by interpreting the data in such a way that the uncertainty can be extracted. Following Jarnicka and Nahorski (2015, 2016), we assume that each inventory data $E_j^n$ represents the "real emission" $RE_j^n$ (i.e., all emissions covered by the accounting scheme that would be reported if our knowledge of activity data and emission factors were perfect), distorted by uncertainty $U_j^n$. Accordingly, each revision $j$ (row of Table 1) is a time series (with time indexed by $n$), given by

$$E_j^n = RE_j^n + U_j^n, \quad n = 1990, \dots, j. \tag{1}$$

Uncertainty $U_j^n$ represents an interplay between the inaccuracy and the imprecision unique to each data point $E_j^n$. We observe that inaccuracy is associated with each revision, namely, an entire row of Table 1, rather than its single entries. Indeed, for each year $j$, $j = 2001, \dots , 2015$, the estimates $E_j^n, \ n = 1990, \dots, j$, published in that year, were calculated using the same accounting method (by this, we mean choices on adopting specific emission factor values and on ascribing activity data to subsectors, but still following the accounting schemes suggested

[1] We explain this notion in greater detail in Section 2.

[2] Calculation of the emission estimates, based on the measurements collected, takes approximately 2 years; thus, the most recent data reported in 2017 originate from the year 2015.

[3] To simplify the notation, we omit the delay in publishing the data and assume that the NIR containing the estimates of emissions for the year $j$ and the revised estimates of all previous years were published in the year $j$.

**Table 1** Indexing the data

$$\text{Revisions}\left\{\underbrace{\begin{array}{ccccccccccccc}
E^{1990}_{2001} & \cdots & E^{n-1}_{2001} & E^{n}_{2001} & E^{n+1}_{2001} & \cdots & E^{2001}_{2001} & & & & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & & & & \\
E^{1990}_{j-1} & \cdots & E^{n-1}_{j-1} & E^{n}_{j-1} & E^{n+1}_{j-1} & \cdots & E^{2001}_{j-1} & \cdots & & & & & \\
E^{1990}_{j} & \cdots & E^{n-1}_{j} & E^{n}_{j} & E^{n+1}_{j} & \cdots & E^{2001}_{j} & \cdots & E^{j}_{j} & & & & \\
E^{1990}_{j+1} & \cdots & E^{n-1}_{j+1} & E^{n}_{j+1} & E^{n+1}_{j+1} & \cdots & E^{2001}_{j+1} & \cdots & E^{j}_{j+1} & E^{j+1}_{j+1} & & & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & & \\
E^{1990}_{2015} & \cdots & E^{n-1}_{2015} & E^{n}_{2015} & E^{n+1}_{2015} & \cdots & E^{2001}_{2015} & \cdots & E^{j}_{2015} & E^{j+1}_{2015} & \cdots & & E^{2015}_{2015}
\end{array}}_{\text{Emissions}}\right.$$

by the UNFCCC) and thus have the same systematic error, that is, the same inaccuracy. However, inaccuracy differs across revisions (for instance, due to improved emission factors or minor changes in the classification of activity data, which occurs from revision to revision). The *evolution of inaccuracy* is described by the time series $U_j,\ \ j = 2001, \ldots, 2015$, where $U_j$ denotes the inaccuracy of the $j$th revision.

Imprecision, on the other hand, is an attribute of a set of repeated estimates of the same quantity. It is therefore associated with the columns of Table 1, where the $n$th column, $n =$ 1990, … , 2015, contains repeated estimates of emissions that occurred in the year $n$. The *changes in imprecision* of emission estimates are reflected by the time series $U^n,\ \ n = 1990,$ … , 2015, where $U^n$ is the estimate of imprecision based on $U^n_j, j = \max\{2001, n\}, \ldots, 2015$.

Table 1 carries entangled information on the "real emission," the inaccuracy of revisions, and the imprecision of emission estimates in the period covered by the inventory. To disentangle this information and detect learning, we proceed with the analysis summarized in Fig. 1.

First, we "clean" the data of information about the "real emission" to extract uncertainty. We perform that "cleaning" by operating on the rows of Table 1 and propose two variants of the "cleaning" procedure. The first variant is based on detrending the rows of Table 1. The second complementary method makes use of the most recent revision (the last row of Table 1) in place of the estimated trend, in order to assess the amount of information captured by the trend. We analyze the data thus transformed row-wise to extract the inaccuracy of consecutive revisions, reflected by the time series $U_j,\ \ j = 2001, \ldots, 2015$. Finally, once the inaccuracy of revisions is extracted from the data, we perform a column-wise estimation of the imprecision of emission estimates $U^n$, $n = 1990, \ldots, 2015$.

We start the above-mentioned analysis with estimating the "real emission" $RE^n_{2015,}$ by fitting the smoothing spline $Sp^n_{2015}$ to the most recent revision data $E^n_{2015}$, as presented in Nahorski and Jęda (2007). Residuals of this nonparametric approach are asymptotically normally distributed, with the mean value equal to zero and standard deviation $\sigma_{2015}$; we thus assume that the detrending of $E^n_{2015}$ with the smoothing spline $Sp^n_{2015}$ gives

$$d^n_{2015} = Sp^n_{2015} - E^n_{2015},\quad d^n_{2015} \sim N(0, \sigma_{2015}) \tag{2}$$

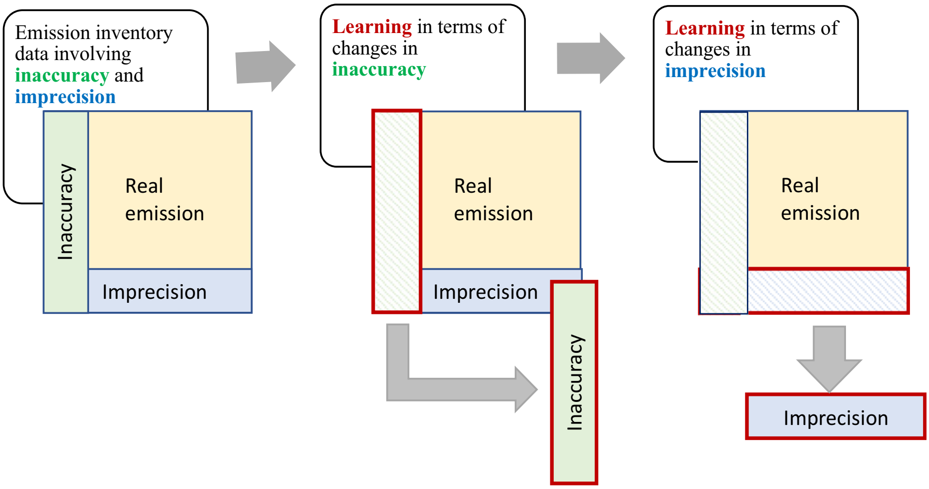

Emission inventory data involving inaccuracy and imprecision
Inaccuracy
Real emission
Imprecision
Learning in terms of changes in inaccuracy
Real emission
Imprecision
Inaccuracy
Learning in terms of changes in imprecision
Real emission
Imprecision

**Fig. 1** The idea of quantifying learning by means of the inaccuracy (changing from revision to revision) and the imprecision (changing in time as our knowledge about emission processes accumulates) of reported GHG emission estimates

Next, we detrend each of the earlier revision time series (1) using the smoothing spline $Sp^n_{2015}$, by subtracting them from this spline, and we assume that the differences obtained follow the same type of distribution

$$d^n_j = Sp^n_{2015} - E^n_j, \quad d^n_j \sim N(0, \sigma_j), \quad j = 2001, \dots, 2014 \tag{3}$$

Parameters $\sigma_j$,$j$ = 2001, … ,2015 can be estimated using the maximum likelihood estimators (e.g., Cowan 1998; Soong 2004), which leads us to the following *model*

$$d^n_j \sim N(0, \hat{\sigma}_j), \quad \text{where } \hat{\sigma}_j = \sqrt{\frac{1}{N_j} \sum_{n=1990}^{j} \left(d^n_j - m_j\right)^2}, \ j = 2001, \dots, 2015 \tag{4}$$

where $m_j$ denotes the mean value for the sample $d^{1990}_j, \dots, d^j_j$ and $N_j = j - 1990$.

Differences (2) and (3) correspond to the inaccuracy of revisions. Inaccuracy is understood as a systematic bias, i.e., the difference between the true value and the average of its repeated estimates. However, each revision consists of a series of different values (i.e., just one estimate for each year, starting in 1990), not repeated estimates of the same value. Hence, using the standard deviation is a suitable way of describing the inaccuracy of revisions.

If differences (2) and (3) are normally distributed, with the population mean value equal to zero and with $\sigma_j$ (different for each revision but equal for all estimates in this revision) as in model (4), then the detrending can be interpreted in terms of extracting inaccuracy. To estimate the inaccuracy of revisions, namely, the time series $U_{j,}$ $j$ = 2001, … ,2015 we normalize parameters $\hat{\sigma}_j$, $j$ = 2001, … ,2015, dividing them by the "real emission," assumed to be represented by the smoothing spline. This gives the following *relative inaccuracy estimates*

$$\hat{U}_j = \frac{\hat{\sigma}_j}{Sp^j_{2015}}, \quad j = 2001, \dots, 2015 \tag{5}$$

To assess the imprecision of emission estimates, i.e., $U^n$,n = 1990, … ,2015, we analyze the columns in the data table, the rows of which were detrended to assess the inaccuracy, i.e., we analyze columns, indexed by $n = 1990, \dots ,2015$, and having entries $d^n_j$, $j = \max\{2001, n\}$, … ,2015. Note that, although each column contains estimates of emissions for the same year, they are based on different activity data and different emission factors. Thus, they are realizations of different time series, and, in consequence, not readily comparable. To analyze them, we first bring them to the same units by means of standardization, consistent with model (4), where the population mean value was assumed to be zero. For each $j = 2001, \dots ,2015$, we divide difference $d^n_j$ by corresponding $\hat{\sigma}_j$, which gives columns of the form

$$e^n_j = \frac{d^n_j}{\hat{\sigma}_j}, \quad \text{indexed with time } n = 1990, \dots, j.$$

At this point, two problems arise. Firstly, the converted columns are not identically distributed. This means that we cannot use distribution parameters, as in model (4), but have to deal with sample characteristics instead. Secondly, samples $e^n_j$ are quite small and vary in size (the columns for $n = 1990, \dots ,2001$ are of equal size, and from then on, each column is one data point shorter than the previous one). This makes it difficult to compare standard deviations, which we use to estimate imprecision, as the sample standard deviation is sensitive to sample size. Hence, to compare them, a size correction is required. To calculate the size-corrected standard deviations, we first take the sample standard deviation given by

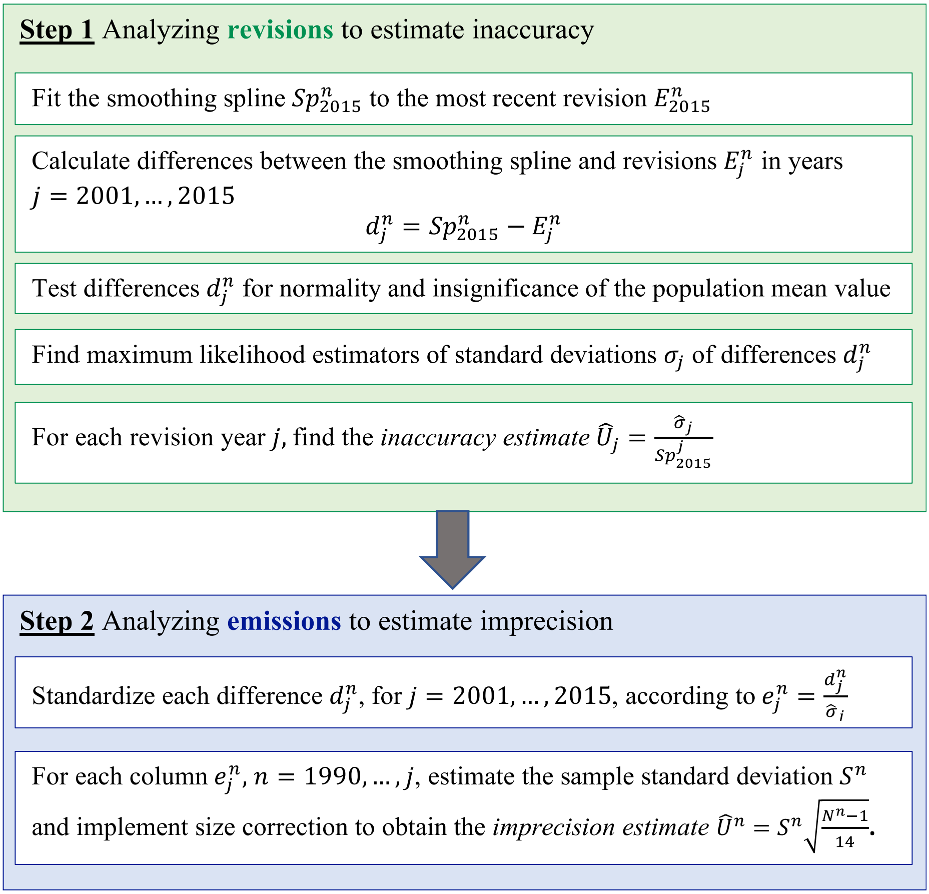
$$S^n = \sqrt{\frac{1}{N^n-1} \sum_{j=2001}^{2014} \left(e^n_j - m^n\right)^2},$$

where $(S^n)^2$ is the unbiased sample estimator of variance, $m^n$ denotes the sample mean value, and $N^n$ is the sample size. Then, we implement the size correction by multiplying $S^n$ by $\sqrt{\frac{N^n-1}{N^{1990}-1}} = \sqrt{\frac{N^n-1}{14}}$. This gives the following *imprecision estimates*

$$\hat{U}^n = \sqrt{\frac{1}{N^{1990}-1} \sum_{j=2001}^{2014} \left(e^n_j - m^n\right)^2}, \quad n = 1990, \dots, 2014 \tag{6}$$

The above discussion leads us to the two-step procedure aiming to estimate inaccuracy and imprecision. We will refer to it as method 1 and we present it graphically in Diagram 1.

Interpreting the results obtained when applying method 1 depends on the fulfillment of assumptions in model (4), in particular, on the normality of differences $d^n_j$. To verify normality, we use the Shapiro-Wilk test (considered the most reliable normality test) with significance level $\alpha = 0.05$ and confirm the results with the Lilliefors test (recommended for use in small samples). If the normality assumption is satisfied, we also test the differences in model (4) for the significance of the population mean value, using the two-tailed $t$ test with $\alpha = 0.05$. If the normality condition is not met, the $t$ test cannot be used, as we deal with small samples (see, e.g., Cowan 1998). We can apply its nonparametric version, i.e., the Mann-Whitney test, but need to take into account that it refers to the median, not the mean value. In fact, that test only



**Step 1** Analyzing **revisions** to estimate inaccuracy

Fit the smoothing spline $Sp^{n}_{2015}$ to the most recent revision $E^{n}_{2015}$

Calculate differences between the smoothing spline and revisions $E^{n}_{j}$ in years $j = 2001, \ldots, 2015$

$$d^{n}_{j} = Sp^{n}_{2015} - E^{n}_{j}$$

Test differences $d^{n}_{j}$ for normality and insignificance of the population mean value

Find maximum likelihood estimators of standard deviations $\sigma_j$ of differences $d^{n}_{j}$

For each revision year $j$, find the *inaccuracy estimate* $\widehat{U}_j = \frac{\hat{\sigma}_j}{Sp^{j}_{2015}}$

**Step 2** Analyzing **emissions** to estimate imprecision

Standardize each difference $d^{n}_{j}$, for $j = 2001, \ldots, 2015$, according to $e^{n}_{j} = \frac{d^{n}_{j}}{\hat{\sigma}_{i}}$

For each column $e^{n}_{j}$, $n = 1990, \ldots, j$, estimate the sample standard deviation $S^{n}$ and implement size correction to obtain the *imprecision estimate* $\widehat{U}^{n} = S^{n}\sqrt{\frac{N^{n}-1}{14}}$.

**Diagram 1** Illustrating method 1 for estimation of inaccuracy and imprecision of reported GHG emission inventories

provides some information on the mean value for normal-like distributions (in particular symmetric ones) when the mean and median are close to each other.

The assumption on the insignificance of the population mean value is of secondary importance and is needed only to formally confirm the way the standardization is performed. The assumption of normality, however, is of critical importance. If this assumption is satisfied, we can say that detrending "cleans" the data sufficiently, removing all the information on the "real emission," so that we are left only with information on inaccuracy. If normality condition is not met, this may indicate that the estimation of the "real emission" was not good enough (most likely due to substantial approximation errors), which makes detrending less effective. This may affect the inaccuracy assessment and lead to different results in the learning investigation.

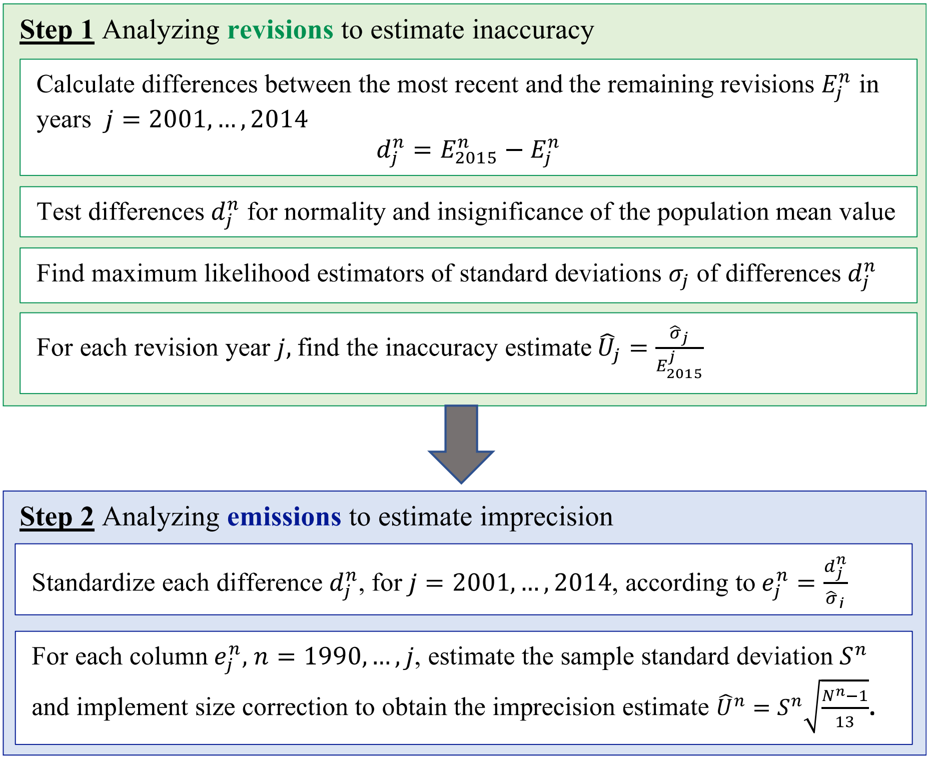
On the other hand, normality of analyzed differences does not guarantee that detrending with spline removes only the information on the "real emission" while leaving the information on uncertainty intact. As a nonparametric approach, the smoothing spline gives asymptotically normally distributed residuals that are likely to pass normality tests (not only in the case of the difference between the smoothing spline and the most recent revision, but also for most of the remaining differences). However, the smoothing spline, fitted to the most recent revision $E^{n}_{2015}$, may grasp not only the "real emission," but also a part of the uncertainty. Thus, we cannot be certain if the detrended data fully reflect the uncertainty. To test this in practice, we consider a modified version of method 1, with no extra estimate of the "real emission," and called method 2 (depicted in Diagram 2).

In method 2, we assume that the most recent revision is the best approximation of the "real emission," and we consider differences between the most recent and the remaining revisions

$$d_j^n = E_{2015}^n - E_j^n, \quad \text{for} \quad j = 2001, \dots, 2014.$$

Note that, there is one row of data less to be analyzed in method 2, compared with method 1, as for every $n$, $E_{2015}^n - E_{2015}^n = 0$. Moreover, as opposed to (2), the difference $d_{2014}^n = E_{2015}^n - E_{2014}^n$ does not represent residuals in a nonparametric regression approach. We can therefore expect that the normality condition may not be met (not only for this difference but for other differences too). This should result in a different behavior of these differences, compared with the approach based on the smoothing spline, but we have to check whether it helps in the learning investigation.

According to the above interpretation, verification of normality provides two types of information. If the normality condition is met, we can assume that differences (both in method 1 and method 2) consist only of inaccuracy (which needs to be estimated), but we must be aware that this information may be incomplete. On the other hand, the lack of normality means that part of



**Step 1** Analyzing **revisions** to estimate inaccuracy

Calculate differences between the most recent and the remaining revisions $E_j^n$ in years $j = 2001, \dots, 2014$

$$d_j^n = E_{2015}^n - E_j^n$$

Test differences $d_j^n$ for normality and insignificance of the population mean value

Find maximum likelihood estimators of standard deviations $\sigma_j$ of differences $d_j^n$

For each revision year $j$, find the inaccuracy estimate $\widehat{U}_j = \frac{\hat{\sigma}_j}{E_{2015}^j}$

**Step 2** Analyzing **emissions** to estimate imprecision

Standardize each difference $d_j^n$, for $j = 2001, \dots, 2014$, according to $e_j^n = \frac{d_j^n}{\hat{\sigma}_i}$

For each column $e_j^n$, $n = 1990, \dots, j$, estimate the sample standard deviation $S^n$ and implement size correction to obtain the imprecision estimate $\widehat{U}^n = S^n\sqrt{\frac{N^n - 1}{13}}$.

**Diagram 2** Illustrating method 2 for estimation of inaccuracy and imprecision of reported GHG emission inventories the "real emission" has been left over in the analyzed differences, which may affect the behavior of inaccuracy (and therefore also imprecision), and make it difficult to capture learning.

Note that the interpretation of inaccuracy estimates (5) obtained with method 1 is similar to that for the inaccuracy estimates calculated with method 2, as in both cases, the relative estimates are calculated with respect to the "real emission" represented either by the smoothing spline $\mathrm{Sp}_{2015}^{n}$ or by the most recent revision $E_{2015}^{n}$. The relative imprecision estimates calculated in the second step of methods 1 and 2 are based on the results obtained in the first step—thus, they are also relative to the "real emission."

## 3 Investigating learning

To detect and assess learning, if present, in inaccuracy and imprecision, we analyze the time series of their estimates $\hat{U}_{j,} \quad j = 2001, \ldots, 2015$ and $\hat{U}^{n}, n = 1990, \ldots, 2015$, obtained using method 1 or method 2 (presented in Section 2).

We assume that learning refers to improvement in the certainty and precision of emission inventories over time, that is, to an observed reduction in uncertainty. We distinguish between learning in the inaccuracy of revisions and learning in the imprecision of emission estimates; however, we may not be able to fully disentangle the two.

We check the aforementioned time series of inaccuracy and imprecision estimates for a trend, namely, the presence of a trend and then its monotonic behavior. In both cases, learning corresponds to the trend decreasing over time (the downward trend), where time is understood as a year of revision in the case of inaccuracy, and as a year in which emissions occurred, in the case of imprecision. This trend can be modeled by a regression curve taking positive values, being decreasing, and approaching zero asymptotically. We can expect some residual uncertainty always to be present. In that case, the trend will stabilize around some level above zero, which in principle can be modeled within the framework proposed here. However, assumptions on asymptotic behavior are of low practical importance, as we work with short samples. For simplicity, we assume that the trend decreases to zero. In addition, we require the curve modeling the trend to be concave up. This is a mild technical assumption, facilitating the use of regression models to assess learning, as we want to avoid the situation where the curve modeling the trend crosses the horizontal axis and takes on negative values.

Examples of changes in uncertainty over time where learning can be observed are depicted in Fig. 2.

Figure 3 illustrates uncertainty structure, where no learning can be detected due to (a) strong random oscillations instead of a clear trend, (b) an upward instead of a downward trend in

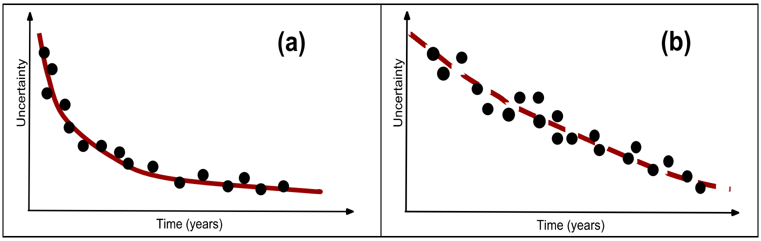


**Fig. 2** Examples of learning in uncertainty

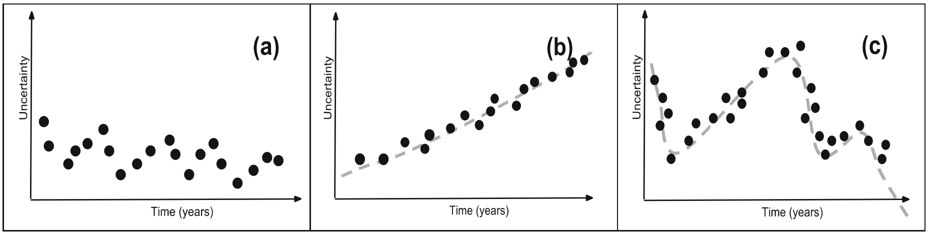


**Fig. 3** Examples where no learning is detected

uncertainty, or (c) polynomial-like non-monotone behavior (where the curve fitted crosses the horizontal axis at some point, leading to a negative uncertainty).

Both examples presented in Fig. 2 illustrate learning, although the one depicted in Fig. 2b, illustrates it at a much slower rate. This shows that we can also assess the rate of learning based on the model fitted and on its goodness of fit. Thus, having estimated inaccuracy and imprecision, we first check them for a downward trend (detecting learning) and then assess that learning (if detected).

### 3.1 Detecting trends in uncertainty

To test uncertainty estimates for a downward trend, we first perform the Bartels test[4] for randomness (Bartels 1982), testing the null hypothesis $H_0$: *randomness* against the left-sided alternative hypothesis $H_1$: *trend*. This nonparametric rank test is very sensitive in trend detection, showing evidence of a trend even if it is very weak. It does not, however, distinguish between a downward and an upward trend. To check this, the Cox-Stuart test[5] (Cox and Stuart 1995) can be used, with null hypothesis $H_0$: *randomness* against the left-sided alternative hypothesis $H_1$: *downward trend.*

Both the above tests are quite easy to perform and work well for small samples (as in the analysis considered here) but as nonparametric ones they may, in some cases, be insufficiently powerful. Their combination is therefore important, allowing us to confirm the presence of the trend detected by the Bartels test (slightly oversensitive and therefore ideal for initial analysis) and, at the same time, to apply the Cox-Stuart test (less powerful) only to those data where the trend is present. To perform the aforementioned tests, we take the most common significance level $\alpha = 0.05$, (e.g., Cowan 1998; Brandt 2014), as it works well in most cases. Setting $\alpha$ at 0.05 means that there is 5% chance of rejecting the null hypothesis when it is true (a type I error). By reducing $\alpha$ (e.g., to 0.01), we reduce the chance of a type I error but increase the chance of not rejecting $H_0$ when the alternative hypothesis is true (a type II error). Thus, 5% seems to be a good balance between these two issues.

[4] The Bartels test is the nonparametric version of von Neumann's ratio test for randomness. It ranks the observations from the smallest to the largest and tests the ratio of the sequential variance calculated from consecutive ranks to the variance based on deviations of ranks from the mean. For values far from the test statistic (two-sided test), there is evidence for non-randomness. In the left-sided test (used in our analysis), randomness is tested against trend, while in the right-sided against regular oscillations.

[5] The Cox-Stuart sign test is based on the binomial distribution. Its test statistic is the number of positive slopes between points that are separated by about half of the observations. The null hypothesis on randomness can be interpreted in terms of positive and negative slopes being equally likely. Both two-sided or one-sided alternative hypotheses can be considered. The left-sided alternative hypothesis, (considered here for the analysis) indicates that negative slopes are more likely than positive ones, which corresponds to a downward trend.

### 3.2 Assessing learning

If a downward trend in uncertainty is present, we can model it by fitting a regression curve. Since the linear regression cannot be used (a straight line does not satisfy the model requirements as it crosses the horizontal axis at some point) and we want to keep the analysis as simple as possible, we consider nonlinear regression models that can be transformed into a standard linear regression (e.g., Myers 1990; Hocking 2013). This allows us to use coefficients of determination $R^2$ to compare the results.

We focus on the following models:

– *exponential model*

$$Y = e^{at+b}, a < 0, \qquad (M1)$$

which can be log-transformed into $Y' = at + b$, taking $Y' = \ln(Y)$,

– *power model*

$$Y = e^{a\ ln(t)+b},\ \ a < 0, \qquad (M2)$$

which can be transformed into $Y' = at' + b$, by $Y' = \ln(Y)$ and $t' = \ln(t)$.

Variable $Y$ represents uncertainty (inaccuracy or imprecision), while $t$ corresponds to time (in years). Thus, both take only positive values and can be log-transformed. If $a < 0$, both curves are decreasing to zero, but the first one at a much faster rate. The difference between their shapes can be observed in Fig. 2, where panel (a) illustrates model (M1), while panel (b) corresponds to model (M2).

Because of that difference, we distinguish between *strong learning* (learning at a faster rate) and *weak learning* (learning at a slower rate). We say that there is a strong learning in uncertainty when the observed downward trend can be modeled using (M1) with a reasonably good fit. If model (M2) is fitted instead, we call it weak learning (or learning at a slower rate).

We select the model based on its goodness of fit, measured by $R^2$, which indicates how much of the relationship between variables $Y$ and $t$ (uncertainty and time, respectively) is explained by the model used (e.g., Soong 2004; Ryan 2008). For instance, the value of $R^2 < 0.5$ indicates that less than 50% of the relationship between variables is explained (and in such a case, the model most likely fails to satisfy the assumptions required, e.g., on the normality of residuals).

In this paper, we will consider such explanatory capabilities of the model as being insufficient and will use a cutoff value for $R^2$ equal to 0.5. This choice of the cutoff value is arbitrary, as there are no strict rules regarding the threshold, although it is often assumed that it should equal at least 60–70%. In some areas, low values of $R^2$ (around 30%) are considered sufficient. Taking a cutoff value at 50% seems to be reasonable here.

The values of $R^2 < 0.5$ for model (M1) will be interpreted as no evidence of a strong learning. In such cases, model (M2) will be used, but if $R^2$ for this model is again smaller than 0.5, we will say that even a weak learning could not be detected.

The method for detecting and assessing learning is described by the following algorithm (depicted in Table 2).

**Table 2** Algorithm to detect and assess learning

```
Algorithm 1. Detecting and assessing learning.
Input: $u$ – time series of uncertainty estimates
Output: NoLearning, StrongLearning or WeakLearning
1. Detecting learning
    Test $u$ for a downward trend
    if no downward trend is detected
        then return NoLearning              ▷ It stops if no downward trend is detected
2. Assessing learning                   ▷ We suspect learning at this point
    Model the downward trend in $u$ using (M1) and (M2)
    Validate the model and check its goodness of fit $R^2$
    if model (M1) provides $R^2 > 0.5$
        then return StrongLearning
    if model (M2) provides $R^2 > 0.5$
        then return WeakLearning
    return NoLearning
```

According to Algorithm 1, the exponential model is preferred over the power model, which is consistent with the interpretation given above. If fitting the exponential model gives $R^2 > 0.5$, this is equivalent to a strong learning, in which case the power model is not considered. We use the power model, if fitting the exponential model gives $R^2 < 0.5$. This means that the criterion for the choice of model (M1) or (M2) is, in fact, the cutoff value and that the values of $R^2$ obtained as the results should be compared independently for each model.

## 4 Learning in the EU-15 emission inventories

The method of detecting learning discussed in previous sections is generic and can be applied to any set of consecutive GHG inventories or their parts (specific sectors). Here, we demonstrate that potential, by applying the method to analyze the estimates of total $CO_2$ emissions excluding LULUCF sector, submitted annually to the UNFCCC in the form of the NIRs[6] produced by each of the EU-15 member countries, along with the emission estimates for the entire EU-15.[7] The emission estimates analyzed cover the period from 1990 to 2015, published in the years 2001–2015.

### 4.1 Analyzing the EU-15 emission inventories

We start by estimating the "real emission" in two ways. In method 1, the "real emission" is estimated by the smoothing spline $Sp^n_{2015}$ fitted to the most recent revision $E^n_{2015}$ (see Fig. 4). Method 2 works on the assumption that the most recent revision involves the best knowledge

[6] Available at http://unfccc.int/national_reports/annex_i_ghg_inventories/national_inventories_submissions/items/8812.php .

[7] EU reports are the aggregate of GHG emission inventories of all member countries. Originally, these were EU-15 countries, but after expansion of European Union these reports contain also emissions of new member states. However, for comparison, the EU-15 data are included in reports of expanded EU.

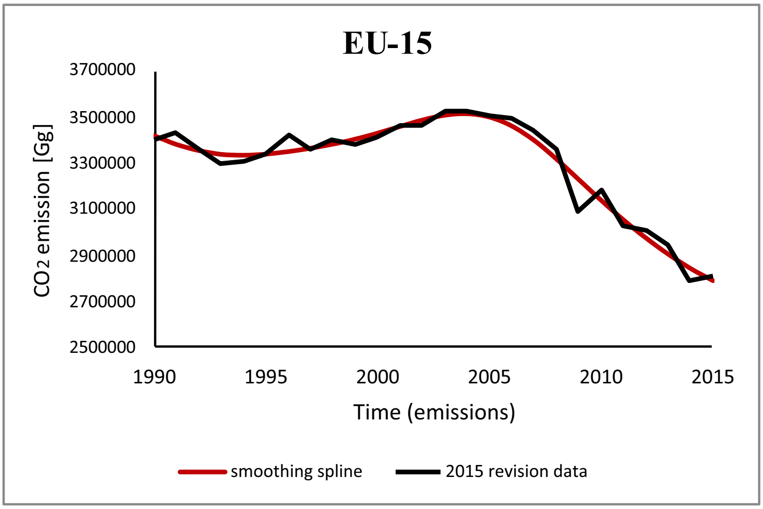


**Fig. 4** Estimating the "real emission" in method 1, using the smoothing spline fitted to the most recent revision (considered the "real emission" in method 2)

on the "real emission" and can be considered its best approximation. Thus, no extra estimate for $E^{n}_{2015}$ is used.

We calculate differences between the "real emission" and revisions, using both the aforementioned methods. We find differences between the smoothing spline and consecutive revisions (depicted in Fig. 5a), as well as between the most recent and earlier revisions (Fig. 5b). As discussed in Section 2, these approaches are based on a different interpretation of uncertainty extraction. By estimating the "real emission" with the smoothing spline and finding the differences (see Diagram 1), we detrend consecutive revision data series. When the most recent revision is considered to be the "real emission" (as presented in Diagram 2), the differences, illustrating changes between the most recent and earlier revisions, do not actually detrend the data. This means that these differences remove a different amount of information regarding the "real emission," which results in each behaving completely differently.

The detrended differences oscillate randomly around zero. However, if we compare them, we can observe some regularities, as if they were following the same pattern (see Fig. 5a). The differences calculated according to the second method show rather chaotic behavior (Fig. 5b),

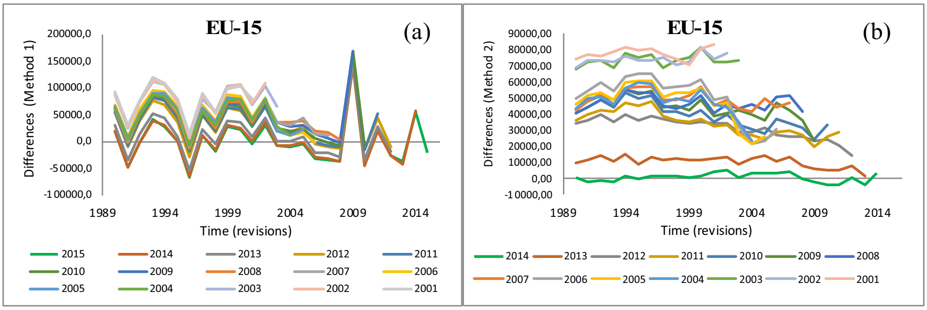


**Fig. 5** Illustrating differences (in [Gg]) **a** between the smoothing spline and consecutive revisions (method 1) and **b** between the most recent and earlier revisions (method 2)

but we can also observe groupings of differences with similar behavior, for example, those related to the most initial or most recent revisions.

This suggests that the detrended differences have been “cleaned” sufficiently, while those based on the most recent revision may still involve some information on the “real emission.” To verify this, we carry out normality tests (the Shapiro-Wilk and the Lilliefors test), with $\alpha = 0.05$, and (if possible) $t$ tests to verify the insignificance of the population mean value. The tests conducted show that in most cases, no statistical evidence can be found against the null hypothesis on the normality of the detrended differences. The tests fail in the case of the most initial revisions, which can partly be explained by the small sample sizes. In all cases where normality condition is met, we also conduct the two-tailed $t$ tests, which show that in most cases, the true population mean is statistically insignificant and can be assumed to be zero.

Checking normality for differences based on the most recent revision shows, in turn, that in most cases, the differences cannot be considered to be normally distributed. This translates into a different behavior and properties of differences calculated by method 1 and method 2.

**Corollary 1** According to the above discussion, we can conclude that

- By detrending the revisions, we managed to remove all the information on the “real emission,” leaving only the inaccuracy.
- By subtracting the most recent revision, we “cleaned” the data only partially; some information on the “real emission” is still present.

We find $\hat{\sigma}_j$ and use them to evaluate changes in inaccuracy over time, as described in Diagrams 1 and 2 (for methods 1 and 2, respectively) and apply algorithm 1 (depicted in Table 2) to check them for learning. First, we analyze the inaccuracy estimates obtained using method 1. The Bartels test for randomness, with null hypothesis $H_0$: *randomness* against the left-sided alternative hypothesis $H_1$: *trend*, performed taking $\alpha = 0.05$, detects a trend in inaccuracy (as $p$ value $= 0.0028 < \alpha$, we reject the null hypothesis on randomness). To check if it is a downward trend, we use the Cox-Stuart test, with $H_0$: *randomness* against $H_1$: *downward trend.* As $p$ value $= 0.77 > \alpha$, we reject $H_1$ on a downward trend. However, to explain the results obtained by applying the Bartels test, we also use the right-sided Cox-Stuart test, with the alternative hypothesis on an upward trend. This time $p$ value $= 0.007 < \alpha$, which shows evidence for an upward trend in inaccuracy. Therefore, *no learning in inaccuracy* is detected.

Now, we consider the columns of the data table, with the rows detrended in the first step of the analysis. First, we standardize the differences, dividing them by corresponding $\hat{\sigma}_j$. Then, we find estimates of imprecision, using formula (6). According to algorithm 1, we test these estimates for a downward trend, applying both aforementioned tests for randomness. The Bartels test gives $p$ value $= 6.5 \times 10^{-8} < \alpha$, thus we accept the alternative hypothesis on the presence of a trend. The Cox-Stuart test shows evidence of a downward trend ($p$ value $= 0.000024 < \alpha$, thus we accept $H_1$ on a downward trend). Once learning in imprecision is detected, we can assess it by fitting model (M1) or (M2). Model (M1) provides a good fit (see Table 3), with a determination coefficient $R^2 = 0.69$. Thus, we can observe *strong learning in imprecision*.

The results are depicted in Table 3 both for inaccuracy and imprecision. The relative inaccuracy estimates are presented in Fig. 6a. The relative imprecision estimates, along with the model fitted, are depicted in Fig. 6b.

**Table 3** Investigating learning in EU-15 $CO_2$ emission inventories (method 1)

| Tests for randomness vs trend | | | model $Y=e^{at+b}$ | | | | |
|---|---|---|---|---|---|---|---|
| Inaccuracy | Bartels test<br>$p=0.0028$<br>Trend | Cox-Stuart test<br>$p=0.007$<br>Upward trend | No learning in inaccuracy detected | | | | |
| Imprecision | Bartels test | Cox-Stuart test | Significance tests | | | Resid. | Fit |
| | $p=6.5\times10^{-8}$ | $p=0.000024$ | $b$ | 280.2 | $p=5.6\times10^{-8}$ | SE = 0.7 | $R^2$ |
| | | | $a$ | −0.14 | $p=5.3\times10^{-7}$ | Norm. (S-W) | 0.69 |
| | Trend | Downward trend | F-test | | $p=5.3\times10^{-7}$ | $p=0.34$ | |
| | | | Strong learning in imprecision | | | | |

Similarly, we estimate changes in inaccuracy, evaluated using method 2. We start with tests for randomness, taking $\alpha=0.05$. Both the Bartels test and the Cox-Stuart test show that there is no trend in inaccuracy (see Table 4); therefore, we can say that *no learning in inaccuracy* can be observed.

We then convert columns in the data table and estimate changes in imprecision, following the procedure described in Diagram 2. As described in Algorithm 1, we check the estimates obtained for a downward trend. The Bartels test with null hypothesis $H_0$: *randomness* against the left-sided alternative hypothesis $H_1$: *trend* gives $p$ value $=9.3\times10^{-9}<\alpha$. This means that we reject the null hypothesis and accept $H_1$ on the presence of a trend. We then use the Cox-Stuart test with the left-sided alternative hypothesis on a downward trend. Since $p$ value $=0.000021<\alpha$, we clearly accept the alternative hypothesis on a downward trend. Therefore, learning in imprecision is detected. To assess it, we fit the exponential model, which provides $R^2=0.47$. Thus, we use the power model instead. This gives $R^2=0.79$ (the results of its validation are presented in Table 4), and hence we can say that *weak learning in imprecision* is observed. The results of learning investigation using method 2 are presented in Fig. 7.

The analysis carried out according to algorithm 1 with both methods 1 and 2 showing that there is no learning in inaccuracy. Method 1 enabled a weak upward trend to be detected. Using method 2, we could observe random inaccuracy behavior over time. As the differences in method 2 were non-normally distributed, it can be concluded that the inaccuracy has not been sufficiently extracted. Both methods allowed us to capture learning in imprecision, but method 1 resulted in detecting learning at a faster rate, while method 2 detected learning at a slower rate. This can be explained by a worse "cleaning" of the data when using method 2.

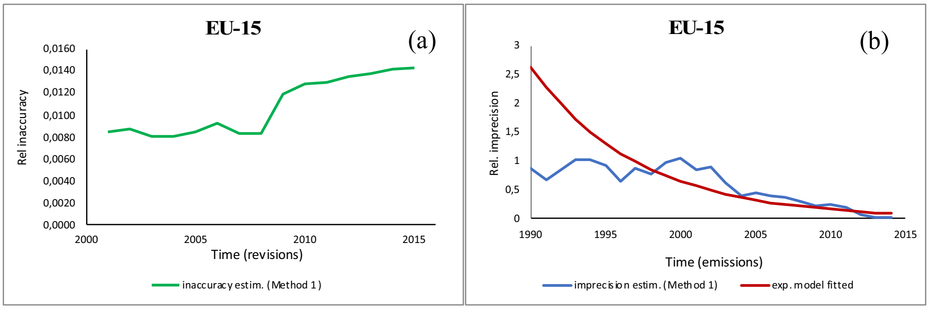


**Fig. 6** Investigating learning in EU-15 emission inventories (method 1). **a** No learning in (relative) inaccuracy. **b** Strong learning in (relative) imprecision

**Table 4** Investigating learning in EU-15 $CO_2$ emission inventory (method 2)

| Tests for randomness vs trend | | | model $Y = e^{a\ ln(t) + b}$ | | | | |
|---|---|---|---|---|---|---|---|
| Inaccuracy | Bartels test<br>$p = 0.312$<br>Randomness | Cox-Stuart test<br>$p = 0.773$<br>Randomness | No learning in inaccuracy detected | | | | |
| Imprecision | Bartels test | Cox-Stuart test | Significance tests | | | Resid. | Fit |
| | $p = 9.3 \times 10^{-9}$ | $p = 0.000021$ | $b$ | 1654.8 | $p = 7.0 \times 10^{-9}$ | $SE = 0.4$ | $R^2$ |
| | | | $a$ | − 217.5 | $p = 6.9 \times 10^{-9}$ | Norm. (S-W) | 0.79 |
| | Trend | Downward trend | F-test | | $p = 6.9 \times 10^{-9}$ | $p = 0.21$ | |
| | | | Weak learning in imprecision | | | | |

**Corollary 2** We can observe that

- There is no learning in inaccuracy (none of the approaches used allowed us to capture it).
- We have not lost any information on uncertainty due to detrending, while extracting uncertainty with method 2 was insufficient
- There is strong learning in imprecision (even insufficient extraction of uncertainty allowed us to capture it, although at a slower rate).

## 4.2 Learning assessment for the EU-15 member countries

The data on GHG emissions in the EU Inventory Reports checked for possible learning in Section 4.1, are obtained by adding those reported by member countries. Analysis of the NIR data for each of the EU-15 member countries should explain and confirm the previous results. Firstly, some countries are expected to follow the same scheme, where strong learning in imprecision is captured by applying method 1, and only weak learning in imprecision is captured by applying method 2. This refers to countries with high emissions reported (as their contribution to the data is significant), and those with particularly strong learning in imprecision detected using method 1. Secondly, there are likely to be countries showing no learning at all (which may have slightly weakened the downward trend in imprecision observed for the EU-15). Of interest to us are any results in between, far from these extreme cases, and whether or not any similarities between neighboring countries can be observed.

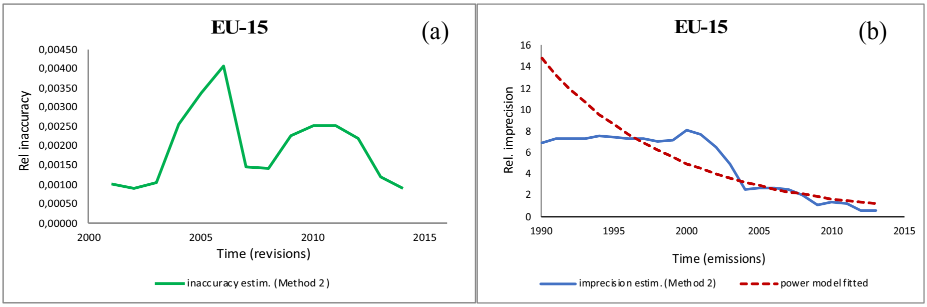


**Fig. 7** Investigating learning in EU-15 emission inventories (method 2). **a** No learning in inaccuracy. **b** Weak learning in imprecision

We conduct the analysis, using both method 1 and method 2, and applying algorithm 1 to detect and assess learning, as in Section 4.1, and compare the results obtained for various countries.

Firstly, *no learning in inaccuracy* is detected for any of them, when using method 1 (see Fig. 8). To confirm the lack of learning in inaccuracy, and to make sure that the results obtained are not consequences of possible exaggerated "cleaning" of the data by detrending, and thus also removing part of the information on inaccuracy (as discussed at the end of Section 2), we also use method 2. The changes in imprecision are also analyzed using both methods.

The results of learning investigation allow for division of the countries analyzed into six groups.

### 4.2.1 Group I: no learning in inaccuracy, strong learning in imprecision

There are three countries whose data on $CO_2$ emission inventories follow the scheme observed for the EU-15 (Fig. 9). This applies to the data reported by Germany, Netherlands, and the UK

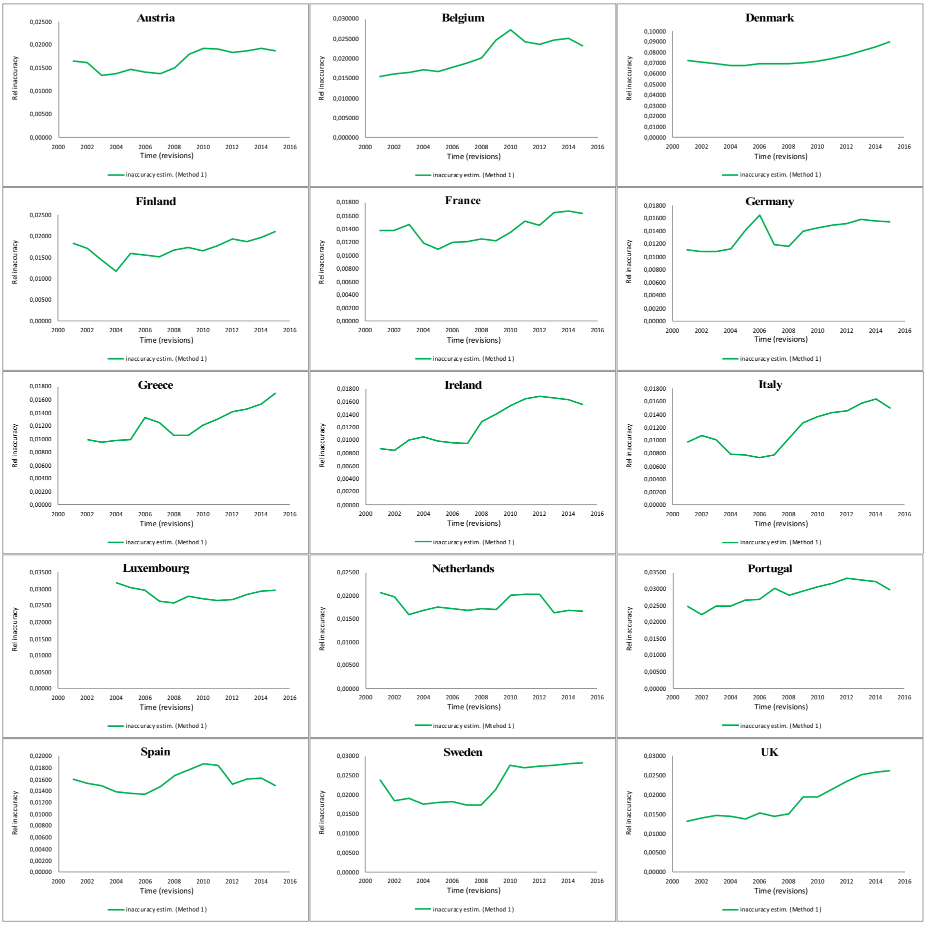


**Fig. 8** No learning in inaccuracy detected in $CO_2$ emission inventories for the EU-15 member countries, using method 1

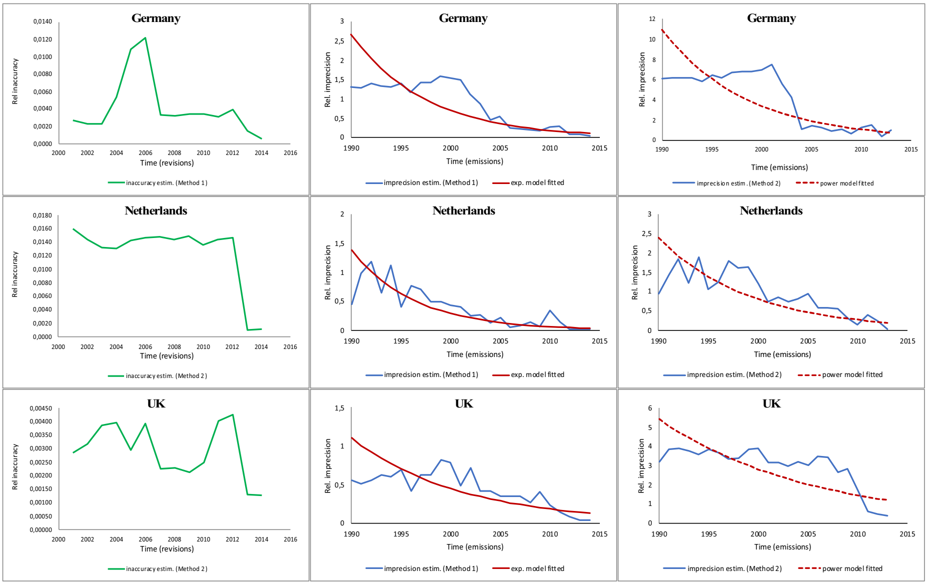


**Fig. 9** Illustrating learning investigation in $CO_2$ emission inventories for Germany, Netherlands, and the UK. No learning in inaccuracy detected when using method 2, strong learning in imprecision due to method 1, and weak learning in imprecision using method 2

for which there is evidence of *strong learning in imprecision* but *no learning in inaccuracy* is detected.

When analyzing inaccuracy estimated with method 2, the Bartels test indicated the presence of a trend, but that result was not confirmed in further analysis. As with the EU-15, learning at a faster rate was captured using method 1, with the fit of the exponential model $R^2 = 0.79$ for Germany, $R^2 = 0.74$ for Netherlands, and $R^2 = 0.59$ for the UK. A weak learning was captured, using method 2, where the fit of the power model, used to illustrate changes in imprecision for those countries, was equal to $R^2 = 0.73$, $R^2 = 0.62$, and $R^2 = 0.52$, for Germany, Netherlands, and the UK respectively.

Given that the $CO_2$ emissions for those countries are quite high compared with other countries, they have a large impact on the results obtained by the entire EU-15. This impact is also due to the fact that similar statistical properties of the differences analyzed can be observed. The detrended differences turned out to be mostly normally distributed with the population mean value zero, while those obtained based on the most recent revision, as for the EU-15, were mostly non-normal. This can be interpreted, as before, in terms of sufficiently and insufficiently “cleaned” revision data series.

### 4.3 Group II: weak learning in inaccuracy, strong learning in imprecision

In the case of two countries Austria and Finland, we managed to capture strong *learning in imprecision* and *weak learning in inaccuracy* (Fig. 10).

By investigating learning with method 1, we managed to observe strong learning in imprecision. Tests for randomness showed the presence of a downward trend in imprecision, and the exponential model fitted to this trend gave $R^2 = 0.77$ for Austria and $R^2 = 0.84$ for Finland.

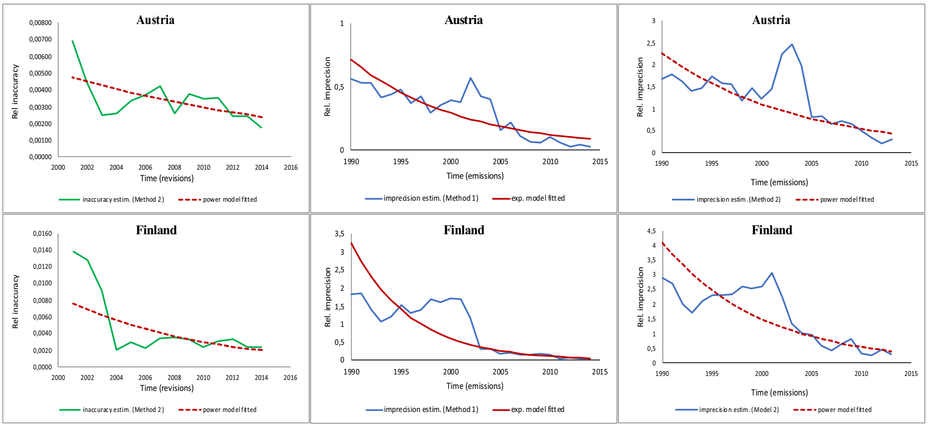


**Fig. 10** Illustrating learning investigation in $CO_2$ emission inventories for Austria and Finland. Weak learning in inaccuracy detected using method 2, strong learning in imprecision due to method 1, and weak learning in imprecision using method 2

Method 2, in turn, allowed learning to be captured both in inaccuracy and imprecision, although both at a slower rate. Tests for randomness showed evidence of a downward trend in inaccuracy. To assess this, the power model was used, giving a fairly poor fit with $R^2$, slightly over 50%, namely, $R^2 = 0.58$ for Austria and $R^2 = 0.59$ for Finland. This, however, enabled us to consider it a weak learning in inaccuracy. The analysis of changes in imprecision also indicated a weak learning, with the fit of the power model $R^2 = 0.61$ for Austria and $R^2 = 0.75$ for Finland. Such results may eventually indicate a strong learning in imprecision, as a weak learning was captured despite the insufficiently "cleaned" data. As we did not detect learning in inaccuracy in the case of detrending, the learning can be considered so weak that the sufficient "cleaning" of the data (by detrending) makes capturing it impossible.

### 4.3.1 Group III: weak learning in inaccuracy, strong learning in imprecision (detected only when using method 1)

The results of the investigation of emission inventory data for Ireland (Fig. 11) partly follow the scheme observed for Austria and Finland. *Strong learning* was detected *in imprecision*, thanks to method 1. The fit of the exponential model used in that case was quite good ($R^2 =$ 0.63). We also captured *weak learning in inaccuracy*, with a fairly good fit of the power model ($R^2 = 0.61$), using method 2, but in this case, no learning in imprecision was detected.

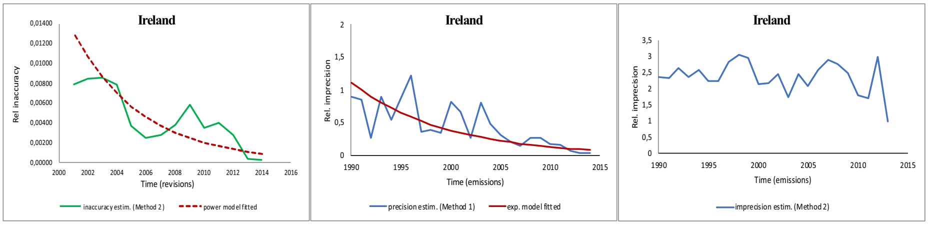


**Fig. 11** Illustrating learning investigation in $CO_2$ emission inventories for Ireland: weak learning in inaccuracy was detected using method 2 and strong learning in imprecision using method 1

Comparing the results obtained for Ireland with those for Austria and Finland, we can see a good fit of the exponential model, used to illustrate the changes in imprecision over time. This translates into strong learning in imprecision. In the case of Ireland, the fit is slightly worse, with $R^2 = 0.63$, which may indicate that learning in imprecision is slightly less pronounced and becomes undetectable after extracting inaccuracy with method 2. Thus, leaving some information on the "real emission" (in method 2) enables a weak learning in inaccuracy to be captured, at the price, however, of not detecting learning in imprecision.

The case of Ireland illustrates the discussion in Section 2, confirming that using different approaches may, in some cases, be crucial.

### 4.3.2 Group IV: no learning in inaccuracy, weak learning in imprecision

In the case of Italy, Portugal, and Spain, only *weak learning in imprecision* was detected (Fig. 12), when using method 1. The power model fitted gave $R^2 = 0.67$ for Italy, $R^2 = 0.69$ for Portugal, and $R^2 = 0.61$ for Spain. By using method 2, we were unable to detect learning either in inaccuracy or in imprecision. Following the interpretation used in the case of Ireland, this can be explained by a really weak learning in imprecision. In the case of "noisy" revision data, where some information on the "real emission" is left over in the differences analyzed (method 2), it becomes undetectable.

At the same time, analysis of changes in inaccuracy over time with method 2 confirmed that there is *no learning in inaccuracy*. The behavior of inaccuracy estimates was, however, different than under method 1 (see Figs. 8 and 12).

In the first case, we observed an upward trend in inaccuracy. Method 2 showed, in turn, that changes in inaccuracy are random (as confirmed by tests for randomness). For Spain, the

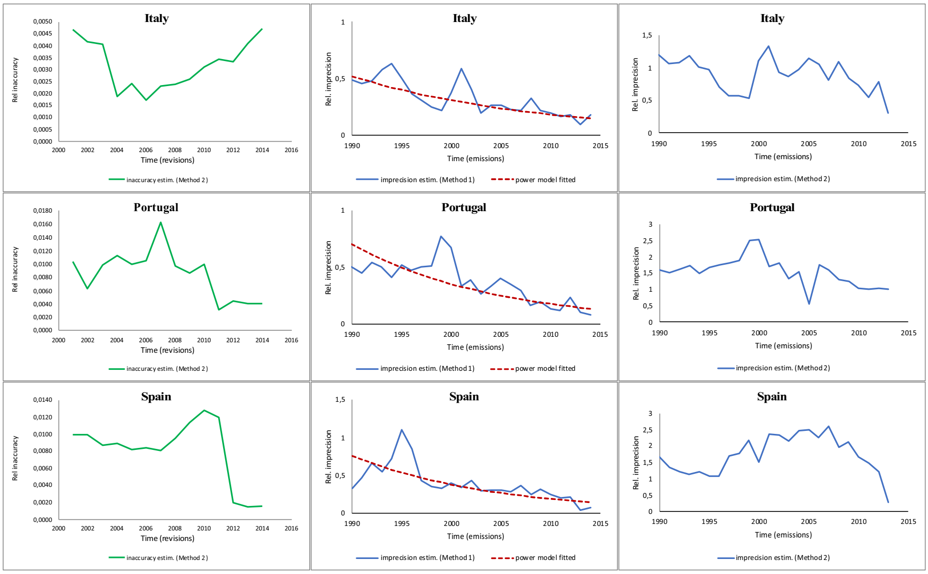


**Fig. 12** Illustrating learning investigation in $CO_2$ emission inventories for Italy, Portugal, and Spain: weak learning in imprecision was detected using method 1 and no learning in inaccuracy using both method 1 and method 2

Bartels test indicated the presence of a trend, but further analysis did not confirm this result (the Cox-Stuart test showed the evidence for the randomness of the data).

### 4.3.3 Group V: weak learning in inaccuracy, no learning in imprecision

The analysis of emission inventories for Denmark and Sweden allowed only *weak learning in inaccuracy* to be captured using method 2 (see Fig. 13).

The power model fitted provided $R^2 = 0.59$ and $R^2 = 0.67$ for Denmark and Sweden, respectively. Neither method 1 nor method 2 enabled learning in imprecision to be captured. Tests for randomness showed no presence of a trend in imprecision, indicating random changes in imprecision over time and hence no *learning in imprecision*.

It is easy to observe the similarity in the behavior of the estimated uncertainty over time with respect to the changes both in inaccuracy and imprecision. We should stress that the data analyzed, both for Denmark and Sweden, seem to be chaotic and random. This was already noticeable when the differences were being analyzed. The detrended differences were mainly non-normally distributed, which means that detrending did not sufficiently "clean" the data. The same could be observed for differences based on the most recent revision. Thus, due to the nature of the data for Denmark and Sweden, we were, in fact, unable to sufficiently extract the uncertainty.

### 4.3.4 Group VI: no learning in inaccuracy, no learning in imprecision

We close the classification with four countries, where *no learning* was detected when using both method 1 and method 2: Belgium, France, Greece, and Luxembourg. The tests for randomness conducted in the case of those countries showed no presence of a trend, either in inaccuracy or in imprecision, indicating the randomness of the data analyzed (see Fig. 14).

It should be noted that as two of these countries (i.e., Greece and Luxembourg) started their official reporting to the UNFCCC later (Greece since 2002, and Luxembourg since 2004), the samples analyzed in those cases were slightly shorter. However, this did not affect the results obtained. It is worth mentioning that for each of these four countries, as in the case of the data for Denmark and Sweden, the random and chaotic behavior could be observed. Only some of

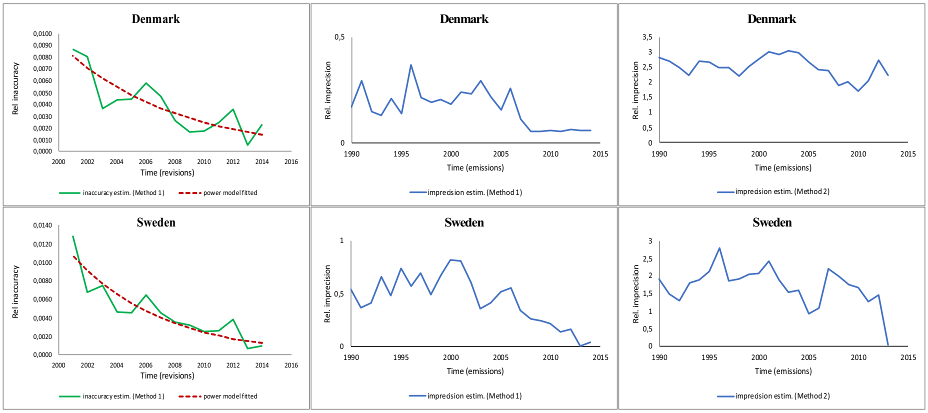


**Fig. 13** Illustrating learning investigation in $CO_2$ emission inventories for Denmark and Sweden: weak learning in inaccuracy was detected using method 2 and no learning in imprecision using both method 1 and method 2

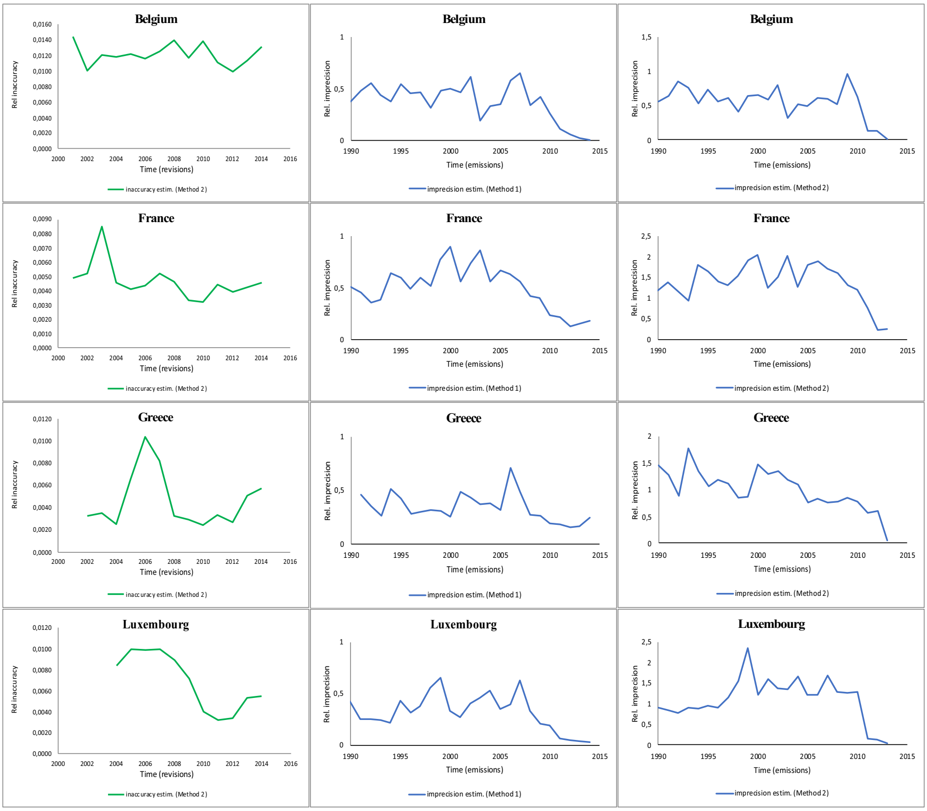


**Fig. 14** Illustrating learning investigation in $CO_2$ emission inventories for Belgium, France, Greece, and Luxembourg: no learning in inaccuracy or in imprecision detected

the detrended differences turned out to be normally distributed, which, as in the previous case, confirms the random nature of the emission inventories for these countries.

**Corollary 3** Summarizing the results, we can observe that

- Only for three countries (Austria, Finland, and Ireland) we managed to capture learning both in imprecision and in inaccuracy (the latter one at a slower rate).
- Only three countries (Germany, Netherlands, and the UK) followed the scheme observed for the entire EU-15, with learning in imprecision.
- For 9 of the 15 countries considered, the $CO_2$ emission inventories showed random changes in inaccuracy and imprecision rather than learning.
- In most cases, we managed to detect only weak learning, either in inaccuracy (Denmark and Sweden) or in imprecision (Italy, Portugal, and Spain), or we detected no learning at all (Belgium, France, Greece, and Luxembourg).

The results of learning investigation based on algorithm 1, with the use of both method 1 and method 2 are summarized in Table 5. Countries are sorted alphabetically, but we also indicate the group to which the given country belongs.

**Table 5** Summary of learning results for EU-15 countries

| Country | Learning in inaccuracy | | Learning in imprecision | | Group |
|---|---|---|---|---|---|
| | Method 1 | Method 2 | Method 1 | Method 2 | |
| Austria | – | Weak | Strong | Weak | II |
| Belgium | – | – | – | – | VI |
| Denmark | – | Weak | – | – | V |
| Finland | – | Weak | Strong | Weak | II |
| France | – | – | – | – | VI |
| Germany | – | – | Strong | Weak | I |
| Greece | – | – | – | – | VI |
| Ireland | – | Weak | Strong | – | III |
| Italy | – | – | Weak | – | IV |
| Luxembourg | – | – | – | – | VI |
| Netherlands | – | – | Strong | Weak | I |
| Portugal | – | – | Weak | – | IV |
| Spain | – | – | Weak | – | IV |
| Sweden | – | Weak | – | – | V |
| UK | – | – | Strong | Weak | I |

## 5 Conclusions and policy recommendations

The practice of revising GHG inventories provides a unique opportunity to conduct a diagnostic analysis of the quality of emission estimates, in terms of both their accuracy and precision. The volume of data collected over the last 15 years has just become sufficient to allow for the application of statistical methods to detect a reduction of uncertainty (i.e., learning) in accounting-based estimates of national GHG emissions published in NIRs. We emphasize that further collection of new data (both new emission estimates and revision of the old ones) is recommended, as longer data samples increase the confidence in the results obtained in Section 4.

In general, method 1 appears to be better at detecting learning in imprecision compared with method 2. For the EU-15, with method 1, we were able to find evidence of strong learning in imprecision, while with method 2, we captured only weak learning. This conclusion is strengthened by an observation (cf. Table 5) that whenever method 1 detects a strong learning in imprecision, method 2 indicates only a weak learning (for countries from groups I and II); Ireland (group III) is an exception here, as a weak learning in inaccuracy instead of imprecision was detected by method 2. Moreover, whenever method 1 detects weak learning in imprecision, method 2 fails to find any evidence of learning (for countries in group IV). Method 2, however, occasionally allows detection of weak learning in inaccuracy, when method 1 fails to find evidence of learning in inaccuracy (for countries in groups II, III, and V). Yet, this comes at the price of a generally worse performance in detecting learning in imprecision (only weak learning was detected for countries in group II, and no learning for groups III and V).

A closer look at the fulfillment of normality assumption sheds some light on this apparent difference in performance between the two methods discussed. In most cases, the differences between the emission estimates and the trend used in method 1 are normally distributed. Thus, detrending removes all the information about the “real emission,” but potentially also some information on inaccuracy. This may render inconspicuous trends in relative inaccuracy that are virtually undetectable using method 1. On the other hand, the differences between the most recent revision and the older ones, as analyzed in method 2, are in general not normally distributed. This means that some information on the "real emissions" was still left over in the data transformed, interfering with the estimation of inaccuracy and thus affecting the assessment of imprecision. This may be the reason why method 2 detects only weak learning (if any). However, this insufficient "cleaning" of the data from information on the "real emissions" may, in some cases, retain some information on inaccuracy (while being removed by method 1), making method 2 more suitable for detecting feeble trends in inaccuracy. To summarize, method 1 may have a slight tendency to underestimate learning in inaccuracy, while method 2 may be more pessimistic in assessing learning in imprecision.

We should note that there is no central agency providing independent inventorying of GHG emissions for the whole EU-15, and the NIRs for the EU-15 are simply obtained as the aggregated NIRs of its member countries. Thus, any learning which we were able to detect in emission data for the EU-15 is due to improvements in GHG inventorying at the national level. This aggregation, however, has a smoothing effect on the evolution of inaccuracy and imprecision for the EU-15 (Figs. 6 and 7) compared with individual member countries (Figs. 8, 9, 10, 11, 12). The reduced variability helps with detection of learning and in drawing stronger conclusions about the satisfactory performance of the methods proposed.

The results presented in this paper have several practical consequences for policy. First, the analysis carried out both for the entire EU-15 (Section 4.1), and for its individual member countries (Section 4.2) shows that there is still much room for further reductions in the uncertainty of emission inventories reported to the UNFCCC. Evidence of a slow increase in accuracy is feeble at best, while many countries also fail to improve the precision of their emission estimates in a noticeable way.

We were unable to detect learning in inaccuracy in the emission estimates of the EU-15 as a whole, which is generally consistent with our findings for individual member countries. Only in several cases of relatively small emitters did method 2 capture weak learning (as presented in Table 5). This apparent general lack of improvement in accuracy of inventories (both for the entire EU-15 and on the national level) is likely to be explained by the fact that all emission estimates (both new and revised) are based on the same accounting schemes suggested by the UNFCCC in IPCC (2000), and later in IPCC (2006). However, the result of introducing new accounting guidelines in IPCC (2006) is noticeable in the formation of peaks in the differences between the smoothing spline and the most recent revision (Fig. 5a), as well as in inaccuracy estimates for the EU-15 (Figs. 6a and 7a) and for most countries analyzed (see Figs. 8, 9, 10, 11, 12, 13, 14). This observation suggests that subsequent updates of GHG emissions accounting guidelines have the potential to reduce the inaccuracy of emission estimates.

An improvement in the precision of the EU-15 emission estimates was detected by both methods proposed. We ascribe this effect to learning in imprecision detected for individual countries, mainly by the big emitters: Germany and the UK (strong learning), and possibly Italy and Spain (weak learning). A possible explanation of this improved precision is the availability of better knowledge about emission processes and emission factors. Further efforts to improve this knowledge are recommended, as they have been proven to reduce the inaccuracy of GHG estimates in the past.

Methods 1 and 2 presented here offer alternative ways of assessing uncertainty suggested in the reporting guidelines, namely, the tier 1 or tier 2 approach, and later also the tier 3 approach IPCC (2000, 2006). These different approaches used in uncertainty assessments published in NIRs make it difficult not only to compare uncertainty for various countries (using different approaches), but often also to track changes in uncertainty over time for a given country (which used different approaches in consecutive years). Step 1 of the proposed methods (cf. Diagrams 1 and 2), together with the evaluation of inaccuracy changes over time, can be useful in such cases (similar analysis was carried out in Jarnicka and Nahorski (2016) where the parametric model was considered, although the results were compared with official assessments only in a few available cases). Moreover, uncertainty estimates published in NIRs are not revised (except for emissions in the base year, usually 1990). This limits the insights into the evolution of uncertainty that could be collected from NIRs. The method proposed here offers a way of building a more complete picture of the evolution of uncertainty.

We conclude with a recommendation for continuation and expansion of the practice of annual revisions of GHG emission estimates published in consecutive NIRs. With the help of methods proposed here, these revisions allow monitoring of improvements in the quality of national GHG inventories and can possibly identify countries (or sectors) for which uncertainty of emission estimates are still not satisfactory. Reducing uncertainty of national GHG inventories is of key importance for monitoring whether countries have achieved their emission reduction commitments and for setting future reductions targets that are likely to ensure the desired results.

## References

Bartels R (1982) The rank version of von Neumann's ratio test for randomness. J Am Stat Assoc 77(377):40–46

Brandt S (2014) Data analysis: statistical and computational methods for scientists and engineers, 4th edn. Springer, New York

Bun A, Hamal K, Jonas M, Lesiv M (2010) Verification of compliance with GHG emission targets: annex B countries. Clim Chang 103(1–2):215–225. https://doi.org/10.1007/s10584-010-9906-6

Cowan G (1998) Statistical data analysis. Clarendon Press, Oxford

Cox DR, Stuart A (1995) Some quick tests for trend in location and dispersion. Biometrika 42(1/2):80–95

Ermolieva T, Ermoliev J, Jonas M, Obersteiner M, Wagner F, Winiwarter W (2014) Uncertainty, cost-effectiveness and environmental safety of robust carbon trading: integrated approach. Clim Chang 124(3): 663–646. https://doi.org/10.1007/s10584-013-0824-2

Hamal K (2010) Reporting GHG emissions: change in uncertainty and its relevance for detection of emission changes. Interim Report IR-10-003. IIASA, Laxenburg

Hocking RR (2013) Methods and applications of linear models: regression and the analysis of variance. In: Wiley series in probability and statistics, 3rd edn. John Wiley & Sons, Inc., Hoboken

IPCC (2000) Good practice guidance and uncertainty management in national greenhouse inventories, http://www.ipccnggip.iges.or.jp/public/gp/english/. Accessed 28 May 2019

IPCC (2006) Guidelines for national greenhouse gas inventories, http://www.ipcc-nggip.iges.or.jp/public/2006gl/Accessed 13 Nov 2018

Jarnicka J, Nahorski Z (2015) A method for estimating time evolution of precision and accuracy of greenhouse gases inventories from revised reports. Proc. 4th Intl Workshop on Uncertainty in Atmospheric Emissions, Kraków, Poland, 2015, pp. 97–102, available at http://www.ibspan.waw.pl/unws2015/images/publications/4thWorkshopProceedings.pdf. Accessed 28 May 2019

Jarnicka J, Nahorski Z (2016) Estimation of temporal uncertainty structure of GHG inventories for selected EU countries. In: Ganzha M, Maciaszek L, Paprzycki M (eds) Proceedings of the 2016 FedCSiS Conference ACSIS, vol 8. IEEE, pp 459–465. https://doi.org/10.15439/2016F318

Jonas M, Gusti M, Jęda W, Nahorski Z, Nilsson S (2010) Comparison of preparatory signal analysis techniques for consideration in the (post-)Kyoto policy process. Clim Chang 103(1–2):175–213. https://doi.org/10.1007/s10584-010-9914-6

Marland G, Hamal K, Jonas M (2009) How uncertain are estimates of CO2 emissions? J Ind Ecol 13:4–7. https://doi.org/10.1111/j.1530-9290.2009.00108.x

Myers RH (1990) Classical and modern regression with applications, 2nd edn. Duxbury Press, Belmont

Nahorski Z, Jęda W (2007) Processing national $CO_2$ inventory emission data and their total uncertainty estimates. Water Air Soil Pollut Focus 7:513–527. https://doi.org/10.1007/s11267-006-9114-6

Ryan TP (2008) Modern regression methods, 2nd edn. Wiley Series in Probability and Statistics, John Wiley & Sons, New York

Soong TT (2004) Fundamentals of probability and statistics for engineers. John Wiley & Sons, New York

Żebrowski P, Jonas M, Rovenskaya E (2015) Assessing the improvement of greenhouse gases inventories: can we capture diagnostic learning? Proc. 4th Intl Workshop on Uncertainty in Atmospheric Emissions, Kraków, Poland, 2015, pp. 90–96, available at: http://www.ibspan.waw.pl/unws2015/images/publications/4thWorkshopProceedings.pdf. Accessed 28 May 2019