Biogeosciences
Discussions

The long-solved problem of the best-fit straight line: Application to isotopic mixing lines

Richard Wehr[1], Scott R. Saleska[1]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, 85721, U.S.A.

5   *Correspondence to*: Richard Wehr (rawehr@email.arizona.com)

Abstract. It has been almost 50 years since York published an exact and general solution for the best-fit straight line to independent points with normally distributed errors in both *x* and *y*. York's solution is highly cited in the geophysical literature but almost unknown outside of it, so that there has been no ebb in the tide of books and papers wrestling with the problem. Much of the post-1969 literature on straight-line fitting has sown confusion not merely by its content but by its very existence. The

10   optimal least-squares fit is already known; the problem is already solved. Here we introduce the non-specialist reader to York's solution, and demonstrate its application in the interesting case of the isotopic mixing line, an analytical tool widely used to determine the isotopic signature of trace gas sources for the study of biogeochemical cycles. The most commonly known linear regression methods—ordinary least squares regression (OLS), geometric mean regression (GMR), and orthogonal distance regression (ODR)—have each been recommended as the best method for fitting isotopic mixing lines. In fact, OLS, GMR, and

15   ODR are all special cases of York's solution that are valid only under particular measurement conditions, and those conditions do not hold in general for isotopic mixing lines. Using Monte Carlo simulations, we quantify the biases in OLS, GMR, and ODR under various conditions and show that York's general—and convenient—solution is always the least biased.

1 Introduction

A common analytical task in the physical sciences is to find the true straight-line relationship underlying independently
20   measured points with normally distributed measurement errors in both the ordinate *y* and abscissa *x*. The literature on this topic is profuse, but much of it was outdated even before it was written. Looking through it, one discovers a menagerie of least-squares regression methods, none strictly appropriate to the task at hand. What one is unlikely to find, unfortunately, is the general, exact, and convenient least-squares solution published in 1969 by geophysicist Derek York (York, 1969).

25   York was motivated by rubidium-strontium isochrons but his landmark solution was universal. Unfortunately, while it became standard in geophysics, York's solution has remained largely unknown to the broader scientific community: York's paper has been cited almost 2000 times within the geophysical literature and only about two dozen times in all the rest of the scientific literature combined. Meanwhile, the number of papers and chapters expounding on the subject as if the solution did not exist is staggering. Examples include widely consulted books like *Biometry* (Sokal and Rohlf, 1995) and *Numerical Recipes* (Press et al.,
30   2007) as well as articles in fields as diverse as: anthropology (Smith, 2009), water resources (Hirsch and Gilroy, 1984), clinical chemistry (Stöckl et al., 1998), marine biology (Laws and Archie, 1981), aerosol science (Leng et al., 2007), and astronomy (Feigelson and Babu, 1992).

One scientific problem in need of York's solution is the isotopic mixing line. In 1958, Charles Keeling (Keeling, 1958)
35   introduced the idea of using an isotopic mixing line to determine the stable isotopic signature of a trace gas source. If the isotopic

composition δ of a trace gas with a single source or sink is plotted versus the inverse of its mixing ratio in air c, the points describe a straight line whose y-intercept gives the desired source or sink signature. The "Keeling plot" became a key method in isotope ecology and biogeochemistry, and is regularly used, for example, to determine the isotopic composition of respired CO_2. Unaware of York's general least-squares solution, researchers concerned with Keeling plots have debated which of the more

5 widely known, special-case least-squares regression methods is best. Some studies have concluded in favor of ordinary least squares regression (OLS) (Zobitz, et al., 2006; Kayler, et al., 2010), others in favor of orthogonal distance regression (ODR) (Ogée, et al., 2003; Miller & Tans, 2003), and still others in favor of geometric mean regression (GMR) (Pataki, et al., 2003; Miller & Tans, 2003; Kayler, et al., 2010). The disagreement arises because in each of these special cases, the measurement conditions influence the fit line bias—but the debate is immaterial given the existence of York's general solution.

10

Here we introduce the reader to York's solution and its practical application, using the isotopic mixing line as an example. We then use Monte Carlo simulations to precisely quantify the biases inherent to York's solution and to the popular special-case regression methods under various common (and some uncommon) conditions. In Appendix A, we provide a short, fast, and easily implemented algorithm for computing York's best-fit slope and intercept, as well as their associated uncertainties.

15 **2 Background**

2.1 The taxonomy of straight-line fitting methods

The goal of straight-line fitting is to retrieve the true straight-line relationship between two variables, x and y, from their measured values, $\hat{x}$ and $\hat{y}$ (hats will denote measured values throughout this paper). Though it is common to neglect the error in $\hat{x}$, doing so can lead to a biased fit, as both $\hat{x}$ and $\hat{y}$ are measurements and therefore corrupted by error at some level. The nature

20 of that error might vary from point to point (a situation known as heteroscedasticity), and it might also be that the error in $\hat{x}_i$ is correlated with that in $\hat{y}_i$, where the subscript i specifies a measurement pair (i.e. a data point). Such correlation can arise, for example, if both $\hat{x}$ and $\hat{y}$ are derived from the same quantity or measured by the same apparatus, and can be described by a (Pearson's) correlation coefficient r that might also vary from point to point. Finally, it is conceivable that the errors in the various $\hat{x}_i$ might be correlated with one another, and similarly for the $\hat{y}_i$; that is, the points might not be independent.

25

If the points are independent of one another and their errors are normally distributed, then the problem can be treated by Least-Squares Estimation (LSE), which is equivalent to Maximum Likelihood Estimation (MLE) (Myung, 2003) in this situation. Most of the literature on straight-line fitting concerns LSE, as it is appropriate to the vast majority of straight-line fitting problems.

30 York's solution is the general LSE solution: his equations provide the best possible, unbiased estimates of the true intercept a and slope b in all cases where the points are independent and the errors are normally distributed. In contrast, the LSE methods commonly used and debated—namely OLS, ODR, and GMR—provide unbiased estimates only in very specific special cases. York (1966, 2004) considers those situations mathematically and shows that his equations reduce algebraically to OLS, GMR, etc. when appropriate.

35

Adding confusion to the literature is the fact that OLS, ODR, and GMR are each known by another name: OLS is called 'the regression of y on x', ODR is called 'major axis regression', and GMR is called 'reduced major axis regression'. Worse still, ODR has also been called 'least normal squares' and 'the line of closest fit', while GMR has also been called 'the line of organic

correlation', 'the unique solution', and 'the equivalence line' (Hirsch and Gilroy, 1984). The methods are also often categorized, with methods that consider error only in $\hat{y}$ (e.g. OLS) being called 'Model I' regressions, and those that consider error in both $\hat{x}$ and $\hat{y}$ (e.g. ODR, GMR) being called 'Model II' regressions (Sokal and Rohlf, 1995).

5   OLS is by far the most widely known fitting method. The OLS fit line is unbiased only when there is negligible error in $\hat{x}$ and when the error variance for the $\hat{y}_i$ does not vary with i. In this case, the problem reduces to minimizing the sum of the squares of the vertical distances of the points from the fit line. A variant known as weighted least-squares allows the error variance for the $\hat{y}_i$ to vary with i.

10 The ODR fit line (Pearson, 1901) is unbiased only when the error variances for the $\hat{x}_i$ and the $\hat{y}_i$ are equal and independent of i, and when the error in $\hat{x}_i$ is uncorrelated with that in $\hat{y}_i$. (The technique Pearson invented to arrive at the ODR fit line is called principal components analysis, which is widely used in its own right.) In this case, the problem reduces to minimizing the sum of the squares of the perpendicular distances of the points from the fit line. However, the condition of equal error variances is almost never satisfied in reality. The excessively restrictive nature of ODR is highlighted by the fact that the ODR result is not

15 invariant under a change of scale. In other words, if one scales the y-axis by a factor of 10, the fit line slope does not scale by a factor of 10. This flaw led to the development of GMR (Kermack and Haldane, 1950), which effectively performs ODR on transformed coordinates: the $\hat{x}_i$ divided by the standard deviation of the $\hat{x}_i$, and the $\hat{y}_i$ divided by the standard deviation of the $\hat{y}_i$ (not by the standard deviations of their errors). However, GMR is also highly restrictive, as its fit line is unbiased only in the peculiar circumstance that the variance of the $\hat{x}$ error divided by the variance of the $\hat{x}_i$ is equal to the variance of the $\hat{y}$ error

20 divided by the variance of the $\hat{y}_i$.

Neither OLS, ODR, nor GMR uses estimates of the actual measurement uncertainty in its determination of the best-fit line. A superior fitting method, called *fitexy*, is provided as an algorithm in Press et al. (2007) and its earlier editions. Press et al. (2007) started (unknowingly) from a similar point as did York (1969), seeking to minimize a χ^2 function that is identical to S in

25 Eq. (2) of York (1969) if the correlation coefficient there is set to zero. However, rather than taking advantage of York's (1966, 1969) algebra, Press et al. (2007) treat the problem as one of nonlinear minimization. This approach leads them to a more complicated (and slower) algorithm than one based on York's solution (such as that provided here in Appendix A). Moreover, *fitexy* is less general than York's solution in that it does not allow the errors in $\hat{x}_i$ and $\hat{y}_i$ to be correlated.

30 If the errors in any or all of the $\hat{x}_i$ and $\hat{y}_i$ are known to be distributed non-normally, then strictly speaking, LSE does not apply and one should retreat to the even more general formalism of MLE. However, the bias introduced into York's solution by non-normal error distributions is not always a problem, as we show in Section 4.1. The bias will vary on a case-by-case basis, but if the correct distribution is known, then the significance of the bias can be estimated fairly simply by Monte Carlo simulation, as we do here.

35

If LSE must be rejected, MLE may or may not be tractable. MLE requires that the correct error distributions be written analytically, and that a useable expression be derived for the likelihood function L appropriate to those distributions. We say 'useable' because simply applying the definition of L to the case of a straight line with errors in both $\hat{x}$ and $\hat{y}$ yields an expression that requires knowledge of the true values x_i and is therefore useless in and of itself. Expressing the likelihood

function in a form that does not include the true values x_i was essentially York's first step in deriving his solution (see Eqs. (7) through (14) of York (1966), in which $S \propto -\ln(L)$).

2.2 The isotopic mixing line problem: Keeling and Miller/Tans plots

A Keeling plot is a scatterplot of the stable isotopic composition δ of a trace gas (typically CO_2) versus the inverse of its mixing

5   ratio c in air (Fig. 1). The isotopic composition δ is defined as a relative deviation from a reference material: $\delta = (R - R_{ref})/R_{ref}$, where R is the ratio of the rare isotope (e.g. ^{13}C) abundance to the most common isotope (e.g. ^{12}C) abundance, and R_{ref} is that ratio for the reference material. The standard units for δ are ‰ (parts per thousand), while those for c are ppm (parts per million). It is easy to show that in the case of a single source or sink changing the trace gas concentration in the atmosphere, the y-intercept of a straight line fit to a Keeling plot gives the isotopic signature (i.e. the δ) of that source or sink. To wit, if the

10   subscripts t, 0, and s refer to the atmosphere at some time t, the atmosphere at time 0, and the source, respectively, then:

$$c_t = c_0 + c_s$$
$$\delta_t c_t = \delta_0 c_0 + \delta_s c_s = \delta_0 c_0 + \delta_s(c_t - c_0)$$
$$\Rightarrow \delta_t = \delta_s + (\delta_0 c_0 - \delta_s c_0) \times (1/c_t)$$

Multiplying the Keeling fit line equation $\delta = a + b(1/c)$ through by c gives an alternate straight-line equation, $\delta c = b + ac$, which can be fit to a plot of δc versus c, called a Miller/Tans plot (Miller & Tans, 2003). In that case, it is the slope of the line that gives

15   the source or sink signature. (Ironically, Keeling reported the inverse relationship between δ and c in 1958 but no "Keeling plots" appear in any of his early papers, nor does he discuss straight line fitting *per se*; he chose instead to show curved fit lines to plots of δ versus c.)

In many ecosystems, the source/sink signatures of interest differ from one another by just 1 ‰ or less (Bowling et al., 2014), so

20   that Keeling or Miller/Tans plot fit biases of 0.1 ‰ can be important.

3 Methods

3.1 Applying York's solution to isotopic mixing lines

An algorithm that solves the York equations (see Appendix A) requires as input, for each data point i: the abscissa $\hat{x}_i$, the ordinate $\hat{y}_i$, the standard deviations of the errors for $\hat{x}_i$ and $\hat{y}_i$, and a correlation coefficient r_i describing the correlation between

25   the error in $\hat{x}_i$ and that in $\hat{y}_i$. This set of information is the same set required for any accurate straight-line fit—if a method does not ask for some of these parameters, then it is implicitly assuming values for them. Solution of the York equations does require one additional piece of information, though: an initial guess slope. This guess can be very rough and still sufficient, as the solution can be iterated (using the previous best-fit slope as the new guess slope) to an arbitrary degree of accuracy. In practice, ten iterations are almost always sufficient. If desired, a good initial guess slope can be obtained from an OLS fit.

30

The accuracy of the best-fit line will depend on the accuracy of the error and correlation estimates, but almost any reasonable estimates will be better than the estimates implicit in OLS, GMR, and ODR. (And good error estimates are to be sought anyway, as a measurement is only meaningful to the extent that its uncertainty has been quantified.) The Keeling plot is an interesting

application partly because, if we take the errors in $\hat{c}$ to be normally distributed, then the errors in the Keeling abscissa $\hat{x} = 1/\hat{c}$ are not normally distributed—though the errors in $\hat{c}$ tend to be small relative to $\hat{c}$ itself, so that the resulting bias in the fit line tends to be negligible (see Section 4.1). Miller/Tans plots are also interesting, for a different reason: though we take the errors in $\hat{\delta}_i$ and $\hat{c}_i$ to be uncorrelated (which is usually the case in practice) the errors in the Miller/Tans abscissa $\hat{x}_i = \hat{c}_i$ and ordinate

5   $\hat{y}_i = \widehat{(\delta c)}_i = \hat{\delta}_i \hat{c}_i$ are always correlated because both contain $\hat{c}_i$. Moreover, the errors in the Keeling abscissa and Miller/Tans ordinate are both (slightly) heteroscedastic.

Regarding correlation between the errors in $\hat{\delta}_i$ and $\hat{c}_i$, it is perhaps worth mentioning that the high-frequency (e.g. 1 Hz) random noise in modern spectroscopic measurements of δ and c is generally correlated due to the fact that certain causes of random

10   spectroscopic noise (e.g. laser frequency instability) necessarily affect retrievals of both δ and c. The correlation may nonetheless be negligible for data averaged over a minute or longer.

The standard deviations of the errors in $\widehat{\delta c}$ and $1/\hat{c}$, as well as the Miller/Tans correlation coefficient, are given by Eqs. (1), (2), and (6), below. For a series of i measurement pairs $\{\hat{c}_i, \hat{\delta}_i\}$, let the standard deviations of the normally distributed measurement

15   errors be $\{\varepsilon_i, \eta_i\}$. Then the standard deviation of the measurement errors in the Miller/Tans ordinate $\widehat{(\delta c)}_i$ is given by (again, assuming no correlation between the errors in $\hat{\delta}_i$ and $\hat{c}_i$):

$$\phi_i = \sqrt{\varepsilon_i^2 \delta_i^2 + \eta_i^2 c_i^2} \approx \sqrt{\varepsilon_i^2 \hat{\delta}_i^2 + \eta_i^2 \hat{c}_i^2}. \tag{1}$$

20   Meanwhile the standard deviation of the (not really normally distributed) measurement errors in the Keeling abscissa $(1/\hat{c})_i$ is given by:

$$\theta_i = \varepsilon_i / c_i^2 \approx \varepsilon_i / \hat{c}_i^2. \tag{2}$$

25   Eqs. (1) and (2) are standard expressions for the propagation of uncertainty through the operations of multiplication and division. For a Keeling fit, the correlation coefficient r_i should be set to zero for all i (unless there is correlation between the measurement errors in δ_i and c_i). For a Miller/Tans fit, r_i is defined as:

$$r_i = \langle (\delta c)_i' c_i' \rangle / \phi_i \varepsilon_i, \tag{3}$$

30

where the angle brackets denote expectation value and where we have decomposed the measured values as $\hat{c}_i = c_i + c_i'$, $\hat{\delta}_i = \delta_i + \delta_i'$, and $\widehat{(\delta c)}_i = (\delta c)_i + (\delta c)_i'$, primes denoting measurement errors and unaccented variables denoting true values. Note that the expectation value here is not an average over the data points, which would involve variations in the true δ and c. It is rather the expectation value for a given data point, i.e. what the value for the data point would be if it were an average of many

35   measurements with the true δ and c held constant.

It follows immediately from the above decompositions that:

$$(\delta c)_i' = \delta_i' c_i + c_i' \delta_i + c_i' \delta_i',\tag{4}$$

and so:

$$
\begin{aligned}
\langle (\delta c)_i' c_i' \rangle &= \langle c_i' \delta_i' c_i + (c_i')^2 \delta_i + (c_i')^2 \delta_i' \rangle \\
&= \delta_i \langle (c_i')^2 \rangle \\
&= \delta_i \varepsilon_i^2 \\
&\approx \hat{\delta}_i \varepsilon_i^2
\end{aligned}
\tag{5}
$$

5

where the second line follows because none of the variables is correlated with another. Thus the correlation coefficient is given by:

$$r_i = \hat{\delta}_i \varepsilon_i / \phi_i.\tag{6}$$

10

So for the Keeling plot, we have $\hat{x}_i = 1/\hat{c}_i$ with uncertainty θ_i, $\hat{y}_i = \hat{\delta}_i$ with uncertainty η_i, and $r_i = 0$. For the Miller/Tans plot, we have $\hat{x}_i = \hat{c}_i$ with uncertainty ε_i, $\hat{y}_i = \hat{\delta}_i \hat{c}_i$ with uncertainty ϕ_i, and $r_i = \hat{\delta}_i \varepsilon_i / \phi_i$.

The approximations in Eqs. (1), (2), and (5) are usually excellent. Despite the precision afforded our Monte Carlo simulations by
15 using 2.5×10^7 data points, we detect bias due to these approximations only under extreme circumstances (see Section 4.1).

On a modest 2009-model notebook computer, using the Igor Pro programming language (WaveMetrics, Inc.), 5000 five-iteration York fits to a 5000-point mixing line took 215 seconds (compare OLS at 14 seconds and *fitexy* at 1410 seconds), while 100,000 such fits to a 20-point mixing line took just 25 seconds (compare OLS at 66 seconds and *fitexy* at 150 seconds). For small
20 numbers of fits, all of these methods are effectively instantaneous. It is perhaps ironic that for most real-world Keeling or Miller/Tans plots, involving fewer than 100 points, solving the York equations is actually faster than OLS.

3.2 Monte Carlo simulations

We tested the York, OLS, ODR, GMR, and *fitexy* methods using simulated measurements of an isotopic mixing line typical of CO_2 respiration in a forest. In this way, the true line is known and the fit bias can be assessed. Our true mixing line had a source
25 isotopic signature of exactly -25 ‰, a background c of 380 ppm, and a background δ of -9 ‰. (A Keeling plot of this line has a slope of 6080 and a y-intercept of -25.) We simulated measurements of this line for a variety of mixing ratio ranges and measurement precisions. For each set of conditions, we generated 5000 'measured' lines, each comprising 5000 points spread evenly over the mixing ratio range Δc, with normally distributed, uncorrelated errors added to c (with standard deviation ε) and to δ (with standard deviation η). The same values of ε and η were used for all points. We then expressed each line as a Keeling
30 plot and as a Miller/Tans plot, and fit each plot by each of the methods.

Real-world mixing line plots are not likely to contain 5000 points each, but using a large number of points per plot can be important when precisely quantifying fit bias in an ensemble of lines. A demonstration of this point is provided in Appendix B for the interested reader. Uninterested readers will be content to know that using more points per plot does not add any new bias

6

to the results—although it might clarify existing bias from an inappropriate fit model, which is what is going on in the discussion under "Sample size effect" in Kayler et al. (2010).

We performed tests for Δc = 1, 5, 10, and 50 ppm, and for various combinations of ε (ppm) and η (‰), including some corresponding to common instrumentation: $\varepsilon = 0.01$, $\eta = 0.01$ (for the best existing laser spectrometer); $\varepsilon = 0.2$, $\eta = 0.3$ (for a popular commercial laser spectrometer); $\varepsilon = 0.05$, $\eta = 0.05$ (for some pairings of an infrared gas analyzer (IRGA) and an isotope ratio mass spectrometer (IRMS)); and $\varepsilon = 0.15$, $\eta = 0.01$ (for other IRGA/IRMS pairings). The example Keeling plot in Fig. 1 corresponds to the latter case. Because Kayler et al. (2010) were concerned with biases at very large Δc and ε in studies of leaf respiration and the like, we also conducted simulations for $\Delta c \geq 100$ ppm with $\eta = 0.2$ ‰ as in Kayler et al. (2010) and with $\varepsilon = 1, 5,$ and 20 ppm.

4 Results and discussion

4.1 Comparison of fitting methods

Isotopic source signatures retrieved from our Keeling plots for $\Delta c \leq 50$ ppm are reported in Table 1. The main cells contain three numbers: the upper is the York result (in bold), while the middle is the OLS result, and the lower is the GMR result. We have not tabulated the *fitexy* or ODR results; the *fitexy* results are discussed below, and it is enough to say that the ODR results were no better than the GMR results (ODR being a seriously flawed precursor to GMR, as explained in Section 2.1). The units of ‰ here (and throughout this paper) are simply the units of δ, and not an indication of relative error in the results. Numbers in parentheses represent the standard error in the last digit, calculated from the distribution of retrieved values rather than by York's equations, although the two results did agree closely (see Section 4.2). We have omitted our Miller/Tans results from Table 1 because there were no significant differences between the Miller/Tans and Keeling results for the York and OLS methods, and because the GMR results were very poor for both line types.

Indeed, as reported by Zobitz et al. (2006), GMR produces highly biased fit lines (that is, the retrieved source signature falls much more than 3 standard errors from -25 ‰). For CO_2 ranges less than 50 ppm, the GMR bias is only negligible in practice (that is, less than about 0.1 ‰) when the measurement uncertainty in δ is extraordinarily low (0.01 ‰). OLS does better than GMR in most of the tested circumstances because the relative error in the c measurement is usually much less than the relative error in the δ measurement, but OLS still involves important levels of bias when ε is large or Δc is small.

The York equations, on the other hand, produce unbiased Keeling and Miller/Tans fit lines for all conditions in the table. Because the emergence of high-frequency isotopic measurements is starting to raise the issue, we show in detail how some OLS- and York-retrieved isotopic source signatures compare at the lowest Δc in Figure 2, where the error bars represent ±2σ, i.e. twice the standard error in the mean of 5000 fits.

Our retrieved isotopic source signatures for $\Delta c \geq 100$ ppm are reported in Table 2, where we have tabulated both the Keeling and the Miller/Tans results. Again, the York method is by far the best, although the York fit lines do exhibit small but detectible biases for some sets of conditions here. The York Miller/Tans fits are biased by at most -0.020 ‰, while bias in the York

Keeling fits is worse, reaching -0.204 ‰ when ε = 20 ppm (an exceptionally high value) and Δc = 100 ppm. This bias is still an order of magnitude less than the bias from any of the other methods under those conditions.

The York Miller/Tans bias is due to the approximations made in Eqs. (1) and (5). We have confirmed this by comparing
5 simulations with and without the approximations (a luxury not available with real data). The Keeling bias is due partly to the approximation in Eq. (2), but mostly to the non-normal error distribution in $1/\hat{c}$ (see Section 3.1): a distribution that is asymmetric, with a non-zero mean. In Table 1, where ε (maximum 0.15 ppm) is always a small fraction of c (380 ppm), the skew of the error in $1/\hat{c}$ is small and its effect on the fit negligible; however, in Table 2, where ε/c can be as large as 5 %, the skew becomes relatively large (the uncertainty on one side of a point is about 10 % larger than on the other side) and its effect on the
10 fit becomes detectible in our simulations. The bias induced in the fit by the non-normal error distribution should increase as ε increases and as Δc decreases, which are the trends we observe. The preceding explanation is confirmed by the fact that when we alter our simulations by adding normally distributed measurement error directly to $1/c$ rather than to c (and giving the correct information to the York fitting algorithm), we find that the York Keeling fits are completely unbiased (results not tabulated). Luckily, the York fit biases we report in Table 2 are very small considering the measurement uncertainties necessary to induce
15 them, and are unlikely to be the limiting source of error in any experiment.

We also tested *fitexy* using our Monte Carlo simulations. As expected, the *fitexy* results were always identical to the York results when fitting to Keeling plots but were in error when fitting to Miller/Tans plots, because the latter plots involved correlation between the x and y errors. For example, with the fairly large error variances ε = 0.2 ppm and η = 0.3 ‰, the *fitexy* Miller/Tans
20 slopes were biased by -4.259 ‰ for Δc = 1 ppm and -0.027 ‰ for Δc = 10 ppm.

4.2 Estimating the fit errors and the goodness of fit

York et al. (2004) provide compact equations not only for the slope and intercept, but for their standard errors as well. They further show that these error estimates are algebraically identical to those of the more general error formulation of MLE. Note, however, that while the York equations for the slope and intercept are exact (if the measurement uncertainties are normally
25 distributed), the York/MLE estimates of the errors in the slope and intercept are approximations (Titterington and Halliday, 1973). In Table 3, we compare York's error estimates against the standard deviations of the Keeling plot intercepts retrieved from our Monte Carlo simulations. York's error estimates agree extremely well with the Monte Carlo results except when the measurement error variances are so large as to approach the total variances in x and y (i.e. when $\varepsilon/\Delta c$ and $\eta/\Delta\delta$ approach 1). Under those conditions, the agreement is nonetheless within a factor of 2.
30

The errors estimated by *fitexy* (not tabulated) were slightly higher than those estimated by the York equations, tending to be farther from the Monte Carlo results for small $\varepsilon/\Delta c$ (when the line is well-constrained) and closer to the Monte Carlo results for large $\varepsilon/\Delta c$ (likely by coincidence). The York and *fitexy* estimated errors were of the same order of magnitude, however, and their disagreement may relate to a subtle point raised in York et al. (2004) concerning whether the errors are estimated using the
35 original data points or the adjusted data points.

The standard error is a measure of precision; it does not speak to how well the straight-line model represents the data—a concept known as goodness of fit. York et al. (2004) note that the goodness of fit of the York solution is estimated by the quantity $S/(R-2)$, where R is the number of points in the fit and S is given by:

5
$$S = \sum W_i (Y_i - bX_i - a)^2. \tag{7}$$

We denote the goodness of fit by G, i.e. $G = S/(R\text{-}2)$. G is essentially comparing the deviations of the points from the fit line to the assigned measurement error standard deviations. If the variables x and y are in fact related by a straight line, and if the assigned measurement errors are correct (and normally distributed), then G will equal 1. A value of G significantly different

10   from 1 indicates the failure of one or both of those assumptions, where we suggest that significance be defined relative to the standard error in G, which depends only on the number of points per fit, R, and is given by:

$$\sigma_G = \sqrt{2/(R-2)}. \tag{8}$$

15   We have tested this equation by Monte Carlo simulation and found it to be correct to within the simulation precision of roughly 1 part in 200.

The right-hand column of Table 3 gives the mean value of G for each 5000-line ensemble described in the table. With $R = 5000$, our σ_G equals 0.02, giving a standard error in our mean G of 3×10^{-4}. Even at this level of precision, G indicates that our fits are

20   almost always good, as expected given the nature of our simulations. However, under those conditions for which the non-normal distribution of $1/c$ is a source of detectible bias, i.e. when $\varepsilon = 20$ ppm, we see the expected deviation in G, which reaches a low of 0.986. In practice, without many fit lines of many points each, such a small drop in G would be undetectable, as would the associated slope and intercept biases.

5 Conclusion

25   We have shown that the general least-squares solution for the best-fit straight line, published by Derek York in 1969, is the least biased estimator of the isotopic signature of a trace gas source or sink from a Keeling or a Miller/Tans plot, regardless of the measurement conditions. In contrast, the popular regression methods considered in the literature are unbiased only under particular, often unrealistic conditions. The isotopic mixing line illustrates the virtue and convenience of York's solution, which is applicable to line fitting problems in many scientific disciplines. We have provided a short, fast pseudo-code algorithm for

30   computing York's solution, and derived simple equations for the required measurement error and correlation parameters in the case of a Keeling or Miller/Tans plot. Being both accurate and convenient, York's solution supersedes all other least-squares straight-line fit methods.

Appendix A: Computer pseudo-code algorithm

Here we provide an algorithm in pseudo-code for computing the slope and intercept of the best-fit straight line according to Eqs.

35   (13) of York et al. (2004). The data consist of R data points indexed by the subscript i. The inputs are: the abscissa $X_i = \hat{x}_i$; the ordinates $Y_i = \hat{y}_i$; their measurement error standard deviations $\sigma_{X,i}$ and $\sigma_{Y,i}$; the correlation coefficient between those errors, r_i;

and an initial guess slope b_0. The outputs are: the intercept a and slope b, their standard errors σ_a and σ_b, the goodness of fit G, and its standard error σ_G. The algorithm iterates until the slope converges to within the tolerance T. The fist block computes the slope and intercept, after which the second block computes the fit errors and the goodness of fit.

$$b = b_0$$

5   While $b_{\text{diff}} > T$ do

Begin loop

$$b_{\text{old}} = b$$
$$\bar{X} = 0 \ \& \ \bar{Y} = 0 \ \& \ W_{\text{sum}} = 0$$

For $i = 1, R$ step 1 do

Begin loop

$$\omega_{X,i} = 1/\sigma_{X,i}^2 \ \& \ \omega_{Y,i} = 1/\sigma_{Y,i}^2$$
$$\alpha_i = \sqrt{\omega_{X,i}\omega_{Y,i}}$$
$$W_i = \alpha_i^2/(b^2\omega_{Y,i}) + \omega_{X,i} - 2br_i\alpha_i$$
$$\bar{X} = \bar{X} + W_iX_i \ \& \ \bar{Y} = \bar{Y} + W_iY_i$$
$$W_{\text{sum}} = W_{\text{sum}} + W_i$$

End loop

$$\bar{X} = \bar{X}/W_{\text{sum}} \ \& \ \bar{Y} = \bar{Y}/W_{\text{sum}}$$
$$Q_1 = 0 \ \& \ Q_2 = 0$$

10 For $i = 1, R$ step 1 do

Begin loop

$$U_i = X_i - \bar{X} \ \& \ V_i = Y_i - \bar{Y}$$
$$\beta_i = W_i\big[(U_i/\omega_{Y,i}) + (bV_i/\omega_{X,i}) - (bU_i + V_i)\,r_i/\alpha_i\big]$$
$$Q_1 = Q_1 + W_i\beta_iV_i \ \& \ Q_2 = Q_2 + W_i\beta_iU_i$$

End loop

$$b = Q_1/Q_2$$
$$b_{\text{diff}} = |b - b_{\text{old}}|$$

End loop

$$a = \bar{Y} - b\bar{X}$$

$$\bar{x} = 0$$

15 For $i = 1, R$ step 1 do

Begin loop

$$x_i = \bar{X} + \beta_i$$
$$\bar{x} = \bar{x} + W_ix_i$$

End loop

$$\bar{x} = \bar{x}/W_{\text{sum}}$$
$$\sigma_b = 0 \ \& \ G = 0$$

For $i = 1, R$ step 1 do

Begin loop

$$u_i = x_i - \bar{x}$$

$$\sigma_b = \sigma_b + W_i u_i^2$$

$$G = G + W_i(Y_i - bX_i - a)^2$$

End loop

$$\sigma_b = \sqrt{1/\sigma_b}$$

$$\sigma_a = \sqrt{\bar{x}^2 \sigma_b^2 + 1/W_{\text{sum}}}$$

$$G = G/(R-2)$$

$$\sigma_G = \sqrt{2/(R-2)}$$

Appendix B: Biased estimators and the sample size effect

Why simulate 5000 points per plot when real plots are more likely to contain 20 or 50 points? Consider the simple picture in
5   Figure 3. Here a true line with slope 1 and y-intercept 0 passes through the points (2,2) and (6,6). Measurements at those two
points contain errors in y only, and the errors are supposed to be either +1 or -1. The two measured lines shown have errors that
are symmetric in the sense that they are equal in magnitude and opposite in sign, and yet these measured lines are not symmetric
about the true line in the sense of being reflections of one another in the true line. Such symmetry would arise only if there were
equal error in both x and y. Nonetheless, the measured slopes (1/2 and 3/2) and y-intercepts (2 and -2) are indeed symmetric
10   about the true values. Now imagine reflecting this picture about $y = x$ (or rotate the page 90˚), so that the line is measured with
errors of ±1 in x only. In this case, the measured slopes (2/3 and 2) and y-intercepts (4/3 and -4) are highly asymmetric about the
true values although the fit lines are just as good as before. Instead, the x-intercepts and inverse slopes have become the
symmetric parameters. All this follows from the asymmetries inherent to the definitions of the slope and intercept.

15   Translating the simple picture of Figure 3 into practical reality, one finds that when there is normally distributed measurement
error only in y, straight line fits can be characterized by their slopes and y-intercepts, which are normally distributed parameters
that can be averaged over an ensemble of fits to give an unbiased estimate of the slope and y-intercept of the mean fit line (by
which we mean the line that would be obtained by combining all of the measured data from the whole ensemble on one plot and
doing a single fit to that). However, when error is present also (or only) in x, the slopes and y-intercepts of an ensemble of fit
20   lines are not normally distributed about the mean-line values. Instead the distributions are skewed, with means and modes that
differ from one another and from the mean-line values. In the general case, with uncertainty in both x and y that might vary from
point to point, there are no standard fit line parameters that can be averaged to give the corresponding parameters of the mean fit
line.

25   Fortunately, as the CO_2 range or the number of points per fit line becomes large, the deviations of the fit lines from the true line
become small and the skew in the parameter distributions becomes insignificant. We exploited this fact in our Monte Carlo
simulations by using 5000 points per line, which enabled us to simply average our fit slopes and intercepts without introducing
significant bias even for $\Delta c = 1$ ppm. Real-world Keeling plots often contain only 10 or 20 points, but usually span more than 10
ppm, and so the skew remains small. Moreover, real-world Keeling plots are not found in ensembles of 5,000, and so the
30   statistical error in the mean slope and intercept would swamp any skew-related bias even for $\Delta c = 1$ ppm.

11

One might be worried about increasing the number of points per plot because it has been reported that fit bias appears to increase with the number of points per plot (Kayler et al., 2010). Actually, increasing the number of points per plot merely clarifies the bias associated with a poor fitting method. Consider Figure 4, where we plot distributions of y-intercepts retrieved from Keeling

5 plots with $\Delta c = 1$, $\varepsilon = 0.15$, and $\eta = 0.01$. These conditions are extreme in that $\varepsilon/\Delta c \approx \eta/\Delta\delta \approx 0.2$ and so the measurement error spread in each dimension is roughly equal to the extent of the true line. Effectively similar conditions ($\Delta c = 40$, $\varepsilon = 5$) were used to produce Table 3 of Kayler et al. (2010). In our figure, the thick curve is the distribution with just 2 points per line, in which case there is no distinction between fit methods because "fitting" consists merely of drawing a line exactly between the two measured points. As the number of points per line increases, they form a distribution in Cartesian space that each fit method

10 interprets differently (according to its assumptions about the measurement error distributions and correlation), so that the intercept distributions obtained from different fit methods diverge. The thin solid curves in Figure 4 show the distributions retrieved from the York equations as the number of points per plot increases to 5 and then 20. The dashed curves show the corresponding distributions retrieved from OLS. By 20 points, the distributions are nearly Gaussian, and the York distribution is nearly centered on the true mean while the OLS distribution is nearly centered on the value corresponding to its true fit bias

15 under these uncertainty conditions. Note that the skew in Figure 4 is not due to the subtly skewed noise distribution in the abscissa: the figure is not visibly changed when normally distributed noise is added directly to $1/c$ rather than to c.

Author contributions

R. Wehr designed and performed the study, and prepared the manuscript with contributions from S. R. Saleska.

Competing interests

20 The authors declare that they have no conflict of interest.

Acknowledgments

References

25 Bowling, D. R., Ballantyne, A. P., Miller, J. B., Burns, S. P., Conway, T. J., Menzer, O., Stephens, B. B., and Vaughn, B. H.: Ecological processes dominate the ^{13}C land disequilibrium in a Rocky Mountain subalpine forest. Global Biogeochem. Cy., 28 (4), 352–370, doi:10.1002/2013GB004686, 2014.

30 Feigelson, E.D. and Babu, G.J.: Linear regression in astronomy. II. Astrophys. J., 397, 55-67, 1992.

Hirsch, R.M., and Gilroy, E.J.: Methods of fitting a straight line to data: Examples in water resources. Water Resour. Bull., 20 (5), 705-711, 1984.

Kayler, Z.E., Ganio, L., Hauck, M., Pypker, T.G., Sulzman, E.W., Mix, A.C., and Bond, B.J.: Bias and uncertainty of $\delta^{13}CO_2$ isotopic mixing models. Oecologia, 163, 227-234, doi:10.1007/s00442-009-1531-6, 2010.

5 Keeling, C.D.: The concentration and isotopic abundances of atmospheric carbon dioxide in rural areas. Geochim. Cosmochim. Ac., 13, 322-334, 1958.

Kermack, K.A. and Haldane, J.B.S.: Organic correlation and allometry. Biometrika, 37 (1/2), 30-41, 1950.

10 Laws, E.A. and Archie, J.W.: Appropriate use of regression analysis in marine biology. Mar. Biol., 65, 13-16, 1981.

Leng, L., Zhang, T., Kleinmann, L., and Zhu, W.: Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. J. Phys. Conf. Ser., 78, 012084, doi:10.1088/1742-6596/78/1/012084, 2007.

15 Miller, J.B. and Tans, P.P.: Calculating isotopic fractionation from atmospheric measurements at various scales. Tellus B , 55B, 207-214, 2003.

Myung, J.: Tutorial on maximum likelihood estimation. J. Math. Psychol., 47, 90-100, doi:10.1016/S0022-2496(02)00028-7, 2003.

20

Ogée, J., Peylin, P., Ciais, P., Bariac, T., Brunet, Y., Berbigier, P., Roche, C., Richard, P., Bardoux, G., and Bonnefond, J.-M.: Partitioning net ecosystem carbon exchange into net assimilation and respiration using 13CO2 measurements: a cost-effective sampling strategy. Global Biogeochem. Cy., 17 (2), 1070, doi:10.1029/2002GB001995, 2003.

25 Pataki, D.E., Ehleringer, J.R., Flanagan, L.B., Yakir, D., Bowling, D.R., Still, C.J., Buchmann, N., Kaplan, J.O., and Berry, J.A.: The application and interpretation of Keeling plots in terrestrial carbon cycle research. Global Biogeochem. Cy., 17 (1), 1022, doi:10.1029/2001GB001850, 2003.

Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag., 2(6), 559–572, 1901.

30

Press, W. H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P.: Numerical recipes: the art of scientific computing, 3rd Ed., Cambridge University Press, New York, 2007.

Smith, R.J.: Use and misuse of the reduced major axis for line-fitting. Am. J. Phys. Anthropol., 140, 476-486, 35 doi:10.1002/ajpa.21090, 2009.

Sokal, R.R. and Rohlf, F.J.: Biometry: the principles and practice of statistics in biological research, 3rd Ed., W.H. Freeman and Co., New York, 1995.

Stöckl, D., Katy, D., and Thienpont, L.M.: Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? Clin. Chem., 44(11), 2340-2346, 1998.

Titterington, D.M. and Halliday, A.N.: On the fitting of parallel isochrons and the method of maximum likelihood. Chem. Geol.,
5   26, 183-195, 1979.

York, D.: Least-squares fitting of a straight line. Can. J. Phys., 44 (5), 1079-1086, 1966.

York, D.: Least squares fitting of a straight line with correlated errors. Earth Planet. Sc. Lett., 5, 320-324, 1969.
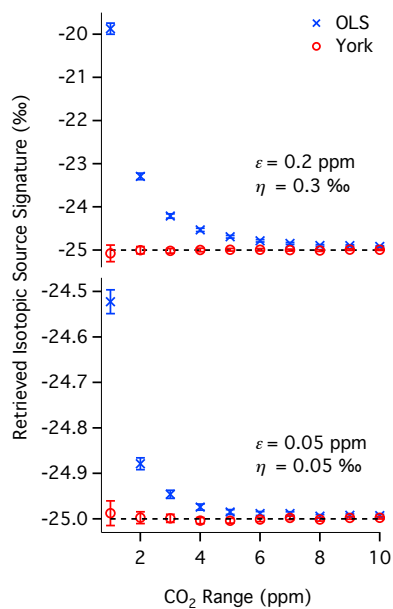10

York, D., Evensen, N.M., Martínez, M.L., and Delgado, J.D.: Unified equations for the slope, intercept, and standard errors of the best straight line. Am. J. Phys., 72 (3), 367-375, doi:10.1119/1.1632486, 2004.

Zobitz, J.M., Keener, J.P., Schnyder, H., and Bowling, D.R: Sensitivity analysis and quantification of uncertainty for isotopic
15 mixing relationships in carbon cycle research. Agr. Forest Meteorol., 136 (1-2), 56-75, doi:10.1016/j.agrformet.2006.01.003, 2006.

Figures, Tables, and Captions
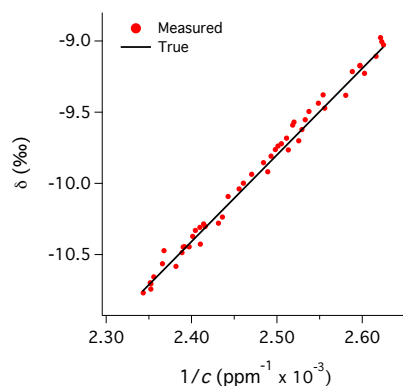
Figure 1: A Keeling plot for a typical measurement scenario, showing 50 simulated measurements of a true line, spread over a CO_2
20 **range of Δc = 50 ppm, with measurement error standard deviations of ε = 0.05 ppm and η = 0.05 ‰.**

Figure 2: Mean isotopic signatures retrieved from ensembles of 5000 simulated Keeling plots (each containing 5000 points) using the York and OLS methods, for CO_2 ranges from 1 to 10 ppm and a true isotopic signature of -25 ‰. The standard deviations of the random noise added to the simulated measurements of c and δ are inset. Error bars show two standard errors. Note the differing scales for the ordinate.
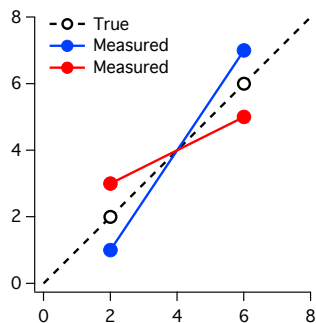
Figure 3: Two lines, each defined by measurements of the true points (2,2) and (6,6) that are in error by +1 or -1 in the ordinate only.
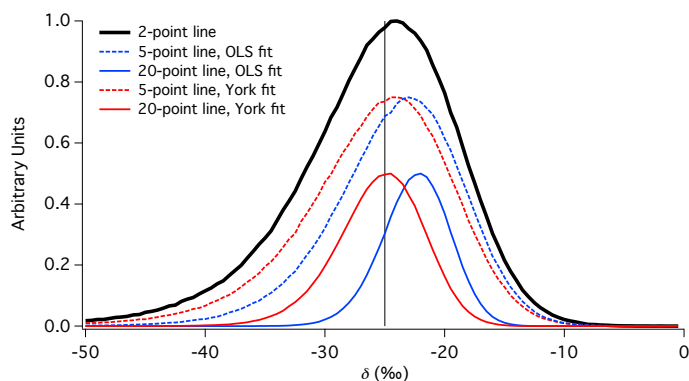
Figure 4: Distributions of the *y*-intercepts retrieved from Keeling plots with a true intercept of -25 ‰ (indicated by the vertical line). The black curve was obtained from 10^6 two-point plots. The red curves were obtained using the York equations on 10^6 five-point or twenty-point plots. The blue curves were obtained using OLS on 10^6 five-point or twenty-point plots. The heights of the curves have been adjusted independently for presentation and convey no meaning.

Table 1. Bias in retrieved isotopic source signatures from Keeling line fits for $\Delta c \leq 50$ ppm, expressed as differences from the true value of -25, in units of ‰. Each cell contains, from top to bottom: the York, the OLS, and the GMR result. 1-σ uncertainties in the final digit(s) are given in parentheses, based on the standard deviation of the ensemble of 5000 fits.

ε (ppm)	η (‰)	CO$_2$ Range (ppm)			
		1	5	10	50
0.01	0.01	**0.005(3)**	**0.000(1)**	**0.000(1)**	**0.000(0)**
		0.024(3)	0.001(1)	0.000(1)	0.000(0)
		-4.700(3)	-0.214(1)	-0.054(1)	0.002(0)
0.01	0.15	**-0.066(40)**	**-0.006(8)**	**0.007(4)**	**0.000(1)**
		-0.047(40)	-0.005(8)	0.008(4)	0.000(1)
		-181.95(3)	-26.592(6)	-9.400(4)	-0.477(1)
0.05	0.05	**0.010(14)**	**0.001(3)**	**0.004(2)**	**0.000(1)**
		0.475(13)	0.020(3)	0.009(2)	0.000(1)
		-50.840(10)	-4.781(3)	-1.343(2)	-0.065(1)
0.15	0.01	**0.002(4)**	**0.000(1)**	**0.000(1)**	**0.000(0)**
		3.398(3)	0.171(1)	0.043(1)	0.002(0)
		-2.385(3)	-0.128(1)	-0.032(1)	-0.001(0)
0.15	0.15	**-0.072(44)**	**-0.005(9)**	**0.001(4)**	**0.000(1)**
		3.339(35)	0.165(8)	0.043(4)	0.002(1)
		-160.00(3)	-26.673(6)	-9.657(4)	-0.575(1)
0.2	0.3	**0.108(97)**	**-0.028(16)**	**-0.006(8)**	**0.002(2)**
		5.251(65)	0.272(16)	0.070(8)	0.005(2)
		-303.09(89)	-64.388(12)	-26.100(6)	-2.185(2)

5

Table 2. Bias in retrieved isotopic source signatures from Keeling and Miller/Tans line fits for $\Delta c \geq 100$ ppm and $\eta = 0.2$ ‰, expressed as differences from the true value of -25, in units of ‰. Each cell contains, from top to bottom: the York, the OLS, and the GMR result. 1-σ uncertainties in the final digit are given in parentheses, based on the standard deviation of the ensemble of 5000 fits.

Model	ε (ppm)	CO$_2$ Range (ppm)			
		100	500	1000	5000
Keeling	1	**0.001(1)**	**0.000(1)**	**0.000(1)**	**0.000(0)**
		0.018(1)	0.001(1)	0.000(1)	0.000(0)
		-0.296(1)	-0.032(1)	-0.017(1)	-0.008(0)
	5	**-0.011(1)**	**-0.002(1)**	**-0.001(1)**	**0.000(0)**
		0.427(1)	0.019(1)	0.006(1)	0.001(0)
		-0.084(1)	-0.022(1)	-0.014(1)	-0.008(0)
	20	**-0.204(2)**	**-0.075(3)**	**-0.032(1)**	**-0.003(0)**
		4.741(2)	0.289(3)	0.092(1)	0.011(0)
		2.386(2)	0.127(3)	0.037(1)	-0.001(0)
Miller/Tans	1	**0.001(1)**	**0.000(1)**	**0.000(1)**	**0.000(0)**
		0.018(1)	0.000(1)	0.000(1)	0.000(0)
		-0.165(1)	-0.016(1)	-0.008(1)	0.004(0)
	5	**0.001(1)**	**0.003(1)**	**0.002(1)**	**0.001(0)**
		0.413(1)	0.012(1)	0.002(1)	0.000(0)
		0.122(1)	-0.006(1)	-0.006(1)	-0.004(0)
	20	**-0.020(2)**	**-0.008(1)**	**0.001(1)**	**0.002(0)**
		4.603(2)	0.192(1)	0.037(1)	0.001(0)
		3.399(2)	0.135(1)	0.022(1)	-0.004(0)

5

Table 3. Monte Carlo (MC) and York estimates of the error in the isotopic source signature retrieved from an individual Keeling plot, in units of ‰. The mean goodness of fit G over each 5000-fit ensemble is also shown.

Δc	ε (ppm)	η (‰)	MC	York	G
	0.01	0.01	0.190	0.186	1.000
	0.01	0.15	2.80	2.79	1.000
1	0.15	0.01	0.241	0.202	1.000
	0.2	0.3	6.84	4.60	0.999
	0.01	0.01	0.0189	0.0189	1.000
	0.01	0.15	0.283	0.283	1.000
10	0.15	0.01	0.0224	0.0221	1.000
	0.2	0.3	0.574	0.565	1.000
	1	0.2	0.0425	0.0426	1.000
100	20	0.2	0.153	0.147	0.986
	1	0.2	0.00797	0.00795	1.000
1000	20	0.2	0.0132	0.0131	0.996