Atmos. Meas. Tech. Discuss., 8, 4191–4218, 2015
www.atmos-meas-tech-discuss.net/8/4191/2015/
doi:10.5194/amtd-8-4191-2015

Atmospheric Measurement Techniques Discussions
Open Access

Non-parametric and least squares Langley plot methods

P. W. Kiedron[1] and J. J. Michalsky[2]

[1]Formerly with the Cooperative Institute for Research in Environmental Sciences, University of Colorado, and Earth System Research Laboratory, National Oceanic and Atmospheric Administration, USA
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, and Earth System Research Laboratory, National Oceanic and Atmospheric Administration, USA

Received: 25 February 2015 – Accepted: 25 March 2015 – Published: 27 April 2015

Correspondence to: J. J. Michalsky (joseph.michalsky@noaa.gov)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Langley plots are used to calibrate sun radiometers primarily for the measurement of the aerosol component of the atmosphere that attenuates (scatters and absorbs) incoming direct solar radiation. In principle, the calibration of a sun radiometer is a straightforward application of the Bouguer–Lambert–Beer law $V = V_0 e^{-\tau \cdot m}$, where a plot of ln($V$) voltage vs. m air mass yields a straight line with intercept ln(V_0). This ln(V_0) subsequently can be used to solve for τ for any measurement of V and calculation of m. This calibration works well on some high mountain sites, but the application of the Langley plot calibration technique is more complicated at other, more interesting, locales. This paper is concerned with ferreting out calibrations at difficult sites and examining and comparing a number of conventional and non-conventional methods for obtaining successful Langley plots. The eleven techniques discussed indicate that both least squares and various non-parametric techniques produce satisfactory calibrations with no significant differences among them when the time series of ln(V_0)'s are smoothed and interpolated with median and mean moving window filters.

1 Introduction

Langley plots are used to determine the instrumental constant V_0, i.e., to calibrate, sun radiometers from a series of measurements V_i at various air masses m_i. According to the Bouguer–Lambert–Beer (BLB) law, the optical depth τ is determined from pairs of points (V_i, m_i) that are fit to the linear equation

$$\ln(V) = \ln(V_0) - \tau \times m. \quad (1)$$

If τ = constant, the equation defines a straight line; the graph is called a Langley plot. When data are not perfect and contain outliers (τ is not always the same for all measurements when time t and air mass $m(t)$ change), the Langley plot is obtained after

AMTD
8, 4191–4218, 2015

removing the outliers. Thus, one can still obtain V_0 from Eq. (1). The derived instrumental constant V_0, if valid, is used to retrieve the optical depth $\tau(m)$ for any measured V and calculated m: $\tau(m) = -m^{-1} \ln(V/V_0)$. The main purpose of sun radiometer is to retrieve the optical depth of atmospheric constituents, mainly aerosols, but also O_3, CO_2, SO_2, H_2O, etc. from the direct beam's atmospheric transmittance V/V_0.

After multiplying the transmittance by the extraterrestrial solar flux one obtains the flux measured by the instrument. The flux at the site of the measurements can be used, e.g., to validate radiative transfer models (Mlawer et al., 2000). The Multi-Filter Rotating Shadowband Radiometer (MFRSR) (Harrison et al., 1994) and the Rotating Shadowband Spectroradiometer (RSS) (Harrison et al., 1999) measure the diffuse components of the flux, as well as, the direct. These instruments were calibrated via the Langley method and with standard lamps (Kiedron et al., 1999). Schmid and Wehrli (1995) concluded that when retrieving optical depth the calibration with Langleys is superior to the calibration based on standard light sources, however, the quality of calibration by Langleys depends on the site at which the instrument is located.

It is often overlooked, that the existence of the straight line in the data set does not imply that the actual τ is constant. The existence of a straight-line fit is the necessary condition, but not a sufficient one for the constancy of the optical depth. Shaw (1976) may have been the first to point this out. He observed that in cases when the optical depth of aerosols is a parabolic function of time the pairs (V_i, m_i) create a perfect straight line, but its intercept is not the actual V_0. Then the intercept is biased, and thus it cannot be used as a calibration constant. Several authors: Tanaka et al. (1986); Nieke et al. (1999); Harrison et al. (2003); and Campanelli et al. (2004) mention this problem more or less explicitly. More recently, Marenco (2007) devoted his paper to this phenomenon. Equation (1) implies that when the actual τ_A contain a varying component that is inversely proportional to the air mass

$$\tau_A = \tau + \varepsilon/m \tag{2}$$

AMTD
8, 4191–4218, 2015

the data align along the straight line with a slope τ, but the intercept is now $(\ln(V_0) - \varepsilon)$. The hyperbolic dependence $1/m$ produces a straight line.

The time series of $\ln V_0$'s over several days are used to weed out cases when ε is not zero and estimate the true calibration. An individual Langley plot cannot identify the value of ε. The data (V_i, m_i) do not contain information on the presence of a nonzero ε.

The process of removing outliers from the Langley plot may actually facilitate selecting a straight line from the data that will contain a spurious value of ε. Different methods of removing outliers may cause an inadvertent selection of a different value of ε, some larger and some smaller (positive or negative).

We will call a Langley plot with a nonzero ε an *anomalous* Langley plot. The results from the anomalous Langley plots can be weeded out only (a) when a time series, an ensemble of $\{V_0\}$, is analyzed, or (b) if additional measurements are performed. Method (a) is the most common procedure, as usually there are no other available independent measurements. One example of an additional procedure (b) that may help in identifying anomalous Langleys is measuring the flux from the solar aureole (Tanaka et al., 1986).

The experimental perspective on Langley plots is as follows: mountain tops like Mauna Loa in Hawaii or Izaña on Tenerife provide environments where the constancy of the optical depth is frequently assured, and, thus, the Langley plots do not pose difficulties in their analyses; these results are rarely controversial. In most places where sun-radiometers are deployed, periods of stable atmospheres are much less common, and they are frequently interrupted by cloud passages, changes in atmospheric conditions like varying humidity that promulgate aerosol size changes, and by aerosol plume incursions. These events create outliers, non-linearity, and anomalous Langley plots. Our paper is directed to the practitioners of Langley calibrations at the sites that are difficult rather than the easy ones like mountaintops.

This paper is not concerned with the identification of anomalous Langley plots. Also, we will not deal with issues of ambiguities related to the definition of the air mass when various air constituents, aerosols, in particular, are present and higher order effects like

AMTD
8, 4191–4218, 2015

atmospheric refraction's dependence on wavelength impact air mass. In other words, we presume that the BLB as given by Eq. (1) is valid. Our comparisons use real data from real sun radiometers that contain data departing from the BLB law chiefly because air masses of Rayleigh scattering and aerosols are different, and the latter is usually known only approximately. The main objective of the paper is to analyze the efficacy of non-parametric and least squares methods of straight line fitting to identify Langley plots useful for calibration.

Non-parametric statistics is the branch of statistics where no a priori assumptions on the statistical distribution of variables are made (Kendall, 1938). Non-parametric does not mean that there are no parameters, but that the involved parameters have no assigned statistical properties a priori. For instance, a histogram is used as a substitute for the probability distribution function rather than, for instance, a fitted Gaussian to the histogram. For Langley plot processing the non-parametric methods are particularly appropriate. We do not know the statistical distribution of outliers in a Langley plot even if we understand their physical origins. We do not know the statistics of V_0's in a time series of them, but we still want to get the best estimate of the calibration constant of a sun radiometer.

For simplicity of notation in the rest of the paper Eq. (1) is replaced with a linear equation: $y = \alpha + \beta x$ where $y = \ln(V), x = m$, $\alpha = \ln(V_0)$ and $\beta = -\tau$.

The organization of the paper is as follows. In Sect. 2 we define a Langley plot. In Sects. 3 to 7, eleven methods of finding a Langley plot are described: in Sect. 3, two least squares methods; in Sect. 4, the so-called objective algorithm method; in Sect. 5, four non-parametric regressions methods following Thiel (1950) and Siegal (1982); in Sect. 6, a non-parametric method of identifying outliers and a modified Siegal (1982) method with sequential removal of outliers; and in Sect. 7, a method of analyzing histograms of slopes and intercepts. In Sect. 8 we describe the set of data used in the comparisons for all methods. Section 9 presents results of these analyses and comparisons. The final section summarizes the paper.

AMTD
8, 4191–4218, 2015

2 Our definition of a Langley plot

For any set of points $P = \{(x_i, y_i) : i = 0, \ldots, n-1\}$ and for any nonnegative δ find a subset $L \subset P$ for which a line $y = \alpha + \beta x$ can be defined such that one of the following metrics

$$\Delta = \sqrt{\frac{1}{k}\sum_{j=0}^{k-1} r_j^2} \quad \text{or} \quad \frac{1}{k}\sum_{j=0}^{k-1} |r_j| \quad \text{or} \quad \max_j |r_j| \tag{3}$$

that measures the magnitude of residuals, is smaller than δ ($\Delta < \delta$) on the set L, where the $r_j = y_j - \alpha - \beta x_j$ are residuals and k is the number of points in the subset L. We refer to the points of the subset L as δ-collinear. The subset L is not unique, or it may not exist when δ is too small, ignoring subsets consisting of two points only. Therefore, we add a requirement that L should be the largest subset with this property of residuals. In other words, a Langley plot is the most numerous δ-collinear subset of set P. The size of the set L, i.e., the number of points that "actually" define the line is important in judging the viability of the resulting Langley plot, i.e., the subset L. The quotation marks around "actually" are justified for non-parametric methods, because they do not identify outliers explicitly.

This problem is related to the pattern recognition problem. The human eye and mind are able to solve the problem in a qualitative way very quickly by identifying points that are approximately collinear. The human eye and mind can perform this task regardless of the plot orientation. The result is rotationally invariant, i.e., neither of the axes x or y is treated preferentially. Furthermore, the human eye and mind can almost instantaneously identify a data set P that has no potential of containing any subset L of a significant size and rejects this case as not providing a viable Langley plot. Mathematically this problem reduces to the straight line fit and to a method of identifying and removing outliers. Any one of the criteria (Eq. 3) can be used to define the quality of the fit.

AMTD
8, 4191–4218, 2015

Some researchers (Augustine et al., 2003) seemingly avoid the issue of removing outliers altogether by selecting clear sky days based on the method of Long and Ackerman (2000) for which measurements from collocated broadband shaded and unshaded pyranometers are required. This approach, while effective, misses many Langley plots from partially clear days, so it does not fit the scope of this paper. Furthermore, many sites with sun radiometers do not have collocated pyranometers.

3 LSF with Sequential Removal of Outliers (SRO)

This is the most straightforward and, probably, the most commonly used method in existence. The LSF is applied to the set of points P and the largest residual (negative or positive) is removed. Then the root-mean-square (rms) of residuals is calculated. The process is repeated until either $\text{rms} \leq \text{rms}_{\max}$ or the number of remaining points $k = k_{\min}$, with $\text{rms}_{\max}$ and $k_{\min}$ chosen based on experience. Usually, most of the outliers are negative (e.g., cloud passages), but there are less frequent cases when the atmosphere has periods of stability at larger τ than when it is not stable during a day. For this case the outliers at smaller τ are positive. For this reason the method must allow removal of the positive outliers. On our data set we note that we obtained good results (meaning that both the number of false and missed identifications of Langley plots were small) when about every 5th outlier that was removed was a positive one. However, we do not claim the value of five is a general rule.

Usually measurements with sun radiometers are performed at equal time steps. This means that the values of x, the air masses, are not evenly distributed. For one minute intervals Δx at $x = 2$ might be 10 times smaller than Δx at $x = 6$ at mid latitudes, where the air mass x is the Rayleigh air mass (Kasten, 1965). Some researchers recognized the bias introduced by the uneven distribution of x on α due to the larger number of points at low air masses. For example, Forgan (2000) performed Langley plots on $(y/x, 1/x)$ for his sun photometric studies. In the Dobson ozone spectrophotometer community Langley plots for the ozone extraterrestrial constant (ETC) are performed

AMTD
8, 4191–4218, 2015

in coordinates $(y/x, 1/x)$ to give a smaller weight to points that are more sparse at large air masses (Dobson and Normand, 1958). Note, however, that for the Brewer UV ozone spectrophotometers the standard equation $y = \alpha + \beta x$ is used (Redondas, 2005; Ito et al., 2014).

The change of variable from x to $1/x$ leads to the linear equation $y/x = \beta + \alpha(1/x)$. This approach is equivalent to applying weights $w = 1/x^2$ to the LSF of the original equation $y = \alpha + \beta x$. Herman et al. (1981) considered applying other weighting methods.

A sequential removal of outliers can be applied and the method may have the same terminating criterion in terms of $\text{rms} < \text{rms}_{\text{max}}$, however, the residuals must be calculated for the equation $y = \alpha + \beta x$.

We label these two methods $\text{LSF}_{\text{SRO}} - x$ and $\text{LSF}_{\text{SRO}} - 1/x$, where the subscript SRO stands for "sequential removal of outliers".

4 The Objective Algorithm of Harrison and Michalsky

We describe some aspects of the Objective Algorithm (OA) because (a) its development is an excellent example of how a mathematical method was stimulated by the human eye-and-mind approach, (b) it is based on physical phenomena that are responsible for the curve shape and the outliers, and (c) it is basically a non-parametric method despite the fact that LSF is used for the final filtering.

When Harrison and Michalsky (1994) developed the OA they tested it by comparing a set of cases from 384 days using 500 nm channel data where α and β were obtained by the eye-and-mind method of Michalsky who disqualified the non-viable cases and identified the ones that, after the removal of outliers, produced Langley plots. Then he performed the LSF on the retained points. The OA did not try to produce "an artificial intelligence" emulating Michalsky's approach. Instead it identified several physical phenomena (like cloud passages, overcast skies, curvature in the plot, etc.) that were responsible for outliers and the non-linearity of the Langley plots. This justifies the term

AMTD
8, 4191–4218, 2015

"objective" in the method's name. The method applies consecutive filters: each meant to deal with outliers produced by one of the identified physical phenomena that produced them. The last filter is LSF that shaves off outliers larger than 1.5 of the SD of all residuals of points that survived the previous filtering.

The successful Langley plot for the OA is the one for which rms ≤ 0.006 (of retained residuals) and the $k/n \geq 1/3$. The value of $\mathrm{rms}_{\max} = 0.006$ was chosen to maximize the agreement with the eye-and-mind Michalsky method using 143 cases of successful Langley plots. The value of $\mathrm{rms}_{\max}$ is valid for the wavelength $\lambda = 500$ nm. For other wavelengths the $\mathrm{rms}_{\max}$ will be different because aerosols' impact on outliers is wavelength dependent.

5 Non-Parametric Fits (NPF)

LSFs use means of $\{x_i\}$, $\{y_i\}$, $\{x_i^2\}$ and $\{y_i, x_i\}$. The so-called breakdown point of the mean is 0 %, i.e., a single outlier can significantly change the value of a mean. On the other hand, the breakdown point of a median is 50 %. Theil (1950) opened the field of the so-called non-parametric regression fits that are based on medians and, thus, are much more robust.

The set P produces an $n \times n$ matrix of all possible slopes $\{b_{i,j}\}$, where $b_{i,j} = (y_i - y_j)/(x_i - x_j)$. The matrix is symmetric with a diagonal that has indeterminate values. Its upper or lower triangles each have $n(n-1)/2$ points. They are used to calculate the slope:

$$\beta = \underset{i<j}{\mathrm{med}}\{b_{i,j}\} \tag{4}$$

From the slope β the intercept is obtained also as a median:

$$\alpha = \underset{i}{\mathrm{med}}\{y_i - \beta x_i\} \tag{5}$$

AMTD
8, 4191–4218, 2015

Theil's (1950) algorithm robustness is 29.3 %, which means that when outliers exceed 29.3 % of all points the performance of the algorithm is not guaranteed; at this level it reaches its breakdown point. The increase of robustness to 50 % was achieved by Siegel (1982) with his method of "repeated medians" that uses two medians in Eq. (4) rather than one: one median along the rows of the matrix $\{b_{i,j}\}$ and then the median of this column of medians

$$\beta = \underset{i}{\text{med}} \left\{ \underset{j}{\text{med}} \left\{ b_{\text{i, j}} \right\} \right\} \tag{6}$$

where all $n(n-1)$ values of the matrix $\{b_{i,j}\}$ are used.

The Langley plot is chiefly concerned with obtaining the intercept and, unlike Theil's focus, the slope is secondary. Instead of obtaining the slope first, one can obtain the intercept first. From the matrix $\{a_{i,j}\}$ of intercepts: $a_{i,j} = (y_j x_i - y_i x_j)/(x_i - x_j)$ one gets the intercept α with Eqs. (4) or (6) and then gets the slope from

$$\beta = \underset{i}{\text{med}} \left\{ \frac{y_i - \alpha}{x_i} \right\} \tag{7}$$

The weighted median methods (Jaeckel, 1972) can also be applied. The uncertainty of slopes $b_{i,j}$ stems from the measurement errors of y_i values. The uncertainty is inversely proportional to $|x_i - x_j|$ if uncertainties for y_i are the same. When $|x_i - x_j|$ is small, the measurement errors have inversely proportional $1/|x_i - x_j|$ larger impact on the slope. For the intercepts the weights for $a_{i,j}$ are proportional to $\left|x_i - x_j\right| /(x_i^2 + x_j^2)^{1/2}$. The exact formulas for using weighted medians for Theil (1950) methods can be found in Birknes and Dodge (1993). The weighted medians, one would expect, should offer an advantage when x_i are not uniformly distributed, which is exactly the case for air masses. However, our simulations with weighted medians did not confirm this expectation. In fact, in our experience the weighted median methods introduced unacceptable biases in α and β.

AMTD
8, 4191–4218, 2015

The LSF methods have several advantages over NPF methods. They perform better when data have no outliers, but have uniformly distributed noise among y values, particularly when the noise is Gaussian. The solutions for α and β in LSF methods are closed-form formulas derived from smooth analytic functions. The NPF methods are inherently discrete. They depend on medians. The "granulation effect", in a small data sample may lead to errors because of discontinuities. The residuals from the LSF methods are unbiased (sum of residuals equals zero), while this is not guaranteed for the methods based on the discreet processes. Theil (1950) and Siegal (1982) NPF methods do not identify outliers explicitly. α and β are generated for any set P, but the outliers must be identified to get the value of the metrics and reach the decision whether these particular α and β define a Langley plot or not. The last problem we solved by using the following method of outlier identification.

For a given α and β we calculate residuals r_i and sort them ($r_j \le r_{j+1}$), where index j points to the index i_j of the unsorted sequence. Then we calculate the sequence $\mathrm{rms}_j \le \mathrm{rms}_{j+1}$. The rms_j is calculated on residuals from the smallest to the $(j-1)$th. Each index i_j for which $\mathrm{rms}_j > \mathrm{rms}_{\max}$ points to an outlier. If the number of outliers is smaller than $2n/3$, α and β define a Langley plot. The remaining residuals may be de-biased by performing the LSF on the remaining points. This will reduce the rms of residuals, but also slightly change α and β.

We chose to use rms (the first criterion in Eq. 3) because the data set for OA, which is used in comparisons, is based on rms metrics. The method of identifying outliers in the set P, when α and β are given, we label OSM (outlier sorting method). In Sect. 9 we also apply the OSM to the results obtained from the OA method.

We described four NPF methods of finding a Langley plot. We label them as $\mathrm{T_{OSM}}$-β, $\mathrm{T_{OSM}}$-α, $\mathrm{S_{OSM}}$-β and $\mathrm{S_{OSM}}$-α, where T and S stand for Theil and Siegel, respectively, and α and β stands for the "intercept first" and "slope first" methods and OSM for the outlier sorting method with a final residual de-biasing LSF. The de-biasing on average increases the intercept α. At most ($\mathrm{S_{OSM}}$-β method) α increases by 0.0028 (0.28 % in terms of V_0).

AMTD
8, 4191–4218, 2015

6 Identifying outliers from the dispersions of slopes

Neither the Theil (1950) nor Siegel (1982) methods identify outliers explicitly. They produce the slope and intercept directly for any set P. In this section we describe a non-parametric method that identifies outliers without calculating the values of residuals.

For each row i of matrix $\{b_{i,j}\}$ we calculate d_i which is the measure of dispersion among the points of the ith row. d_i can be the SD of the row or its median absolute deviation (MAD). We used the latter. The "largest" outlier is the one for which d_i is largest. Then we remove row i and column i from the matrix $\{b_{i,j}\}$ and calculate new values d_i and find the one that is the largest, and so on. The largest dispersion D_m, where m is the index of the iteration process, forms a descending sequence with a decreasing steepness. Large drops in the sequence indicate a removal of a significantly "large" outlier. The largeness or smallness of outliers should be understood as their values of dispersion, though it may correlate very well with the value of residuals from the straight line $y = \alpha + \beta x$.

One can analyze the sequence of $\{D_m\}$. Once it flattens, this indicates that points that remain approximate a straight line and the process of outlier removal can be stopped, but if $\{D_m\}$ remains strongly decreasing it implies that there is no "collinear" subset in the data. We did not explore the potential of finding a criterion for stopping the iteration process from features and behavior of the $\{D_m\}$ sequence. Instead we used the remaining points to calculate rms and stopped the process when rms became smaller than $\mathrm{rms}_{\max}$.

We applied the same method to the matrix of intercepts $\{a_{i,j}\}$ and obtained similar results, i.e., the sequences of removal of "large" outliers for both $\{a_{i,j}\}$ and $\{b_{i,j}\}$ matrices were the same, but not identical when only "small" ones were left.

This approach of outlier identification and removal led us to a modified Siegel (1982) method. At each stage when a row and a column are removed we calculate new α_m and β_m with the Siegel (1982) method. Initially we were surprised that after a removal of an outlier the new values of α_{m+1} and β_{m+1} were not always changing significantly

AMTD
8, 4191–4218, 2015

until we realized that this is a consequence of the robustness of the Siegel method. The method is stopped when $\mathrm{rms}_m \le \mathrm{rms}_{\max}$, and the result is retained if $(n-m)/n > 1/3$. This method of finding a Langley plot we label S_{SRO}-β or S_{SRO}-α, where SRO stands for a sequential removal of outliers. Keep in mind that SRO from this section is strictly a non-parametric method unlike SRO in the Sect. 3 on LSF methods.

7 Histograms of slopes and intercepts

From values $b_{i,j}$, say in the lower triangle of the matrix $\{b_{i,j}\}$, we generate a histogram $H_\beta(b, \Delta b)$ where Δb is the width of the histogram's cells. The cell b contains points $b \le b_{i,j} \le b + \Delta b$. We find $b_{\max}$ for which $H(b, \Delta b)$ is maximum and identify the set of all x_k arguments that contributed to slopes within the cell. Points x_k are Δb-parallel. For each x_k we calculate a count c_k, which is the number of pairs the point x_k produced a slope belonging to the cell. The plot c_k vs. x_k can tell us a lot about the linearity of points from the cell. When all points are Δb-collinear then $c_k = \text{const}$ equals to $K(K-1)/2$ where K is number of points x_k. The points for which $c_k = 1$ indicate pairs that are Δb-parallel, but they are offset along the y axis from the points with large values of c_k. To draw a line through points (y_k, x_k), some points must be removed. First we remove all points with $c_k = 1$ and then calculate the median of the remaining c_k values and reject all those for which c_k is less than the median. Finally we apply the LSF to the remaining points (y_k, x_k) and calculate rms of residuals. This method defines the Langley plot with a small number of points, but is very efficient in detecting the subsets of collinear points. For this reason we did not require that the number of retained points had to be larger than $n/3$ for this method. We labeled this method H-β.

A similar process can be performed with the histogram of intercepts $H_\alpha(a, \Delta a)$. The results, however, were not as good as with the histogram of slopes. However, the histogram of intercepts can be used very effectively in filtering the set $\{x_k\}$ obtained from the histogram of slopes. From the histogram of intercepts we obtain sequences $\{x_m\}$ from its maximum cell. The cross-section $\{x_{\mathrm{km}}\} = \{x_k\} \bigcap \{x_m\}$ contains the points both

AMTD
8, 4191–4218, 2015

Δb-parallel and the ones that produce intercepts in the interval $a_{max} \le a_{i,j} \le a_{max} + \Delta a$. This filtering process was not yet implemented as an operating program, but it makes the removal of points for which $c_k = 1$ and filtering with median on c_k values unnecessary.

In Fig. 1a–d we show three cases on which the H-β method fails and one case with a large number of outliers for which H-β works correctly. The cases are shown to illustrate the usefulness of histogram analysis in prescreening cases and possibly designing a more sophisticated method that could produce, not one, but several Langley plots from one set of points P.

In Fig. 1a there are two regions: $2 \le x \le 2.12$ that contains 46 points and $3.51 \le x \le 4.16$ that contains 19 points. These two regions produced two mono-modal, narrow histograms, implying that there are many collinear points from within two regions.

In Fig. 1b there are two regions with collinear points. But each region has a different slope and it extrapolates to a different α. Both histograms are bimodal. Two distinct Langley lines could be produced in this case. Only one, if either, can be right. The question that one of them or both are anomalous Langley plots can be posed.

The Fig. 1c show a very interesting case. The histogram of slopes is mono-modal, but the histogram of intercepts is bimodal. The height of the second mode is less than half of the dominant mode. Two regions $2.25 \le x \le 3$ and $3 \le x \le 6$ have similar slopes as they produce the mono-modal histogram of slopes. But at $x = 3$ there is a step change. It is not possible that it was produced by a change in the optical depth if one excludes the change in ε responsible for the anomalous Langley. It is possible that at $x = 3$ something affected the responsivity of the instrument. In this case we will get two Langley plots that are almost parallel with different α's.

Figure 1d depicts a case without major outliers. The points can be fairly well approximated with the 3rd degree polynomial (a thin line is depicted) which means that τ is the 2nd degree polynomial of air mass. Both histograms are bimodal and rather broad. One may pose a question whether histograms could be used to detect nonlin-

AMTD
8, 4191–4218, 2015

earity. The method proposed by Kuester et al. (2003) has the potential for detecting the nonlinearities, however, it seems that the authors did not explore this possibility.

8 The data set

The comparison and analysis of Langley algorithms was performed on a data set produced by the Rotating Shadowband Spectroradiometer (RSS) (Harrison et al., 1999) deployed at the Department of Energy Southern Great Plains (SGP) site near Billings, Oklahoma, USA (36.6044° N, −97.4853° W) between May 2003 and December 2008. The data set analyzed covers the period from 5 October 2003 to 30 March 2006, i.e., 1055 days. We removed many overcast days and several corrupted files that, together with the instrument down times, reduced the data set to 1023 morning or afternoon sets $P = \{(y_i, x_i) : 2 \le x_i \le 6\}$. Data from one pixel, out of 1024, at approximately $\lambda = 500$ nm are used in this analysis.

The RSS was lamp calibrated every two to three weeks with calibrators that had lamp calibrations traceable to a NIST standard. The values of V were normalized by the responsivity obtained from each lamp calibration and interpolated between the calibration days. This reduced trends in V_0 due to the instrumental instability caused by optical elements aging, diffuser degradation, and CCD response changes. Nevertheless, the RSS displayed quasi-periodic instabilities due to, what we later discovered, was an outgassing problem that led to a deposition of a thin film on the cooled windowless CCD. This resulted in a wavelength- and time-dependent etalon effect that affected the CCD's response. Lamp calibrations mitigated the effect, but did not remove it from the data completely.

9 Comparison of methods

In the previous sections we described 11 methods to identify a Langley plot. In this section we compare them at four different values of $\mathrm{rms}_{\max} = 0.010$, 0.008, 0.006, and

AMTD
8, 4191–4218, 2015

0.004 using one data set. First we look at some statistical parameters concerning $\Delta\alpha$ between each two methods, and then we look at calibration constant time series derived from each method.

In Table 1 we collected information on the number of Langley plots for each method (in the diagonal of the table) and the number of common Langley plots between the two methods (above the diagonal). Then we included some statistical parameters on $\Delta\alpha$ between each of the two methods; there are 55 combinations. Above the diagonal is the SD of $\Delta\alpha$, and below the diagonal mean and median of $\Delta\alpha$. The order of subtraction in $\Delta\alpha$ is as follows: α for the method from the row minus α for the method from the column.

The LSF methods produce the largest number of Langley plots (601 and 598) while the OA the lowest (284). The OA's α's are larger than any other method by 0.0026–0.0065 which translates to 0.26–0.65 % in V_0. For most cases medians of $\Delta\alpha$ are 5 to 10 times smaller than means of $\Delta\alpha$. This is because the main contribution to differences $\Delta\alpha$ among methods comes from the tails of $\Delta\alpha$ distributions. In other words, outliers are responsible for the main differences among the methods, however, some biases exist among them. A large median $\Delta\alpha$ indicates that the bias between the methods given by a mean is real and does not apply only to the outliers. In Fig. 2 we looked at the average of all SDs as we started eliminating Langley plots obtained after removing a large numbers of outliers. If no more than 10 % were outliers, the average of the SD is 10 times smaller than if up to 77 % of the data from which Langley plots were obtained were outliers. At the same time the average of absolute values of means of $\Delta\alpha$ drops only by factor of two. In fact, it stabilizes when no more than 25 % of the data points were outliers. This indicates the real bias between the methods. The median remains virtually unchanged. This just means that when the Langley plot consists of a smaller number of points the outcome is more method dependent.

Also, we plotted the number of Langley plots vs. the number of points remaining in the Langley plot for the LSF_{SRO}-x method.

AMTD
8, 4191–4218, 2015

The Theil and Siegel methods (T_{OSM}-β, T_{OSM}-α, S_{OSM}-β, S_{OSM}-α) produce very similar results among each other with some of the lowest means and medians of $\Delta\alpha$. In some cases medians are zero. We could not discern a difference between the "intercept first methods" (T_{OSM}-α, S_{OSM}-α) and the "slope first methods" (T_{OSM}-β, S_{OSM}-β).

LSF_{SRO}-x yields larger α than LSF_{SRO}-$1/x$ (by 0.0055). For rms_{max} = 0.010, 0.008 and 0.004, it is 0.0083, 0.0065 and 0.0039, respectively. The SD of 0.0198 is not exceptionally large or small in comparison with other methods.

The Siegel methods with sequential removal of outliers (S_{SRO}-β, S_{SRO}-α) yield significantly different numbers of Langley plots (364 vs. 475). But on the common set the mean and median of $\Delta\alpha$ are very small.

The results for the histogram method H-β do not indicate anything extraordinary. Its results are most similar to results produced by Siegel methods and LSF_{SRO}-x.

The outlier sorting method when applied to OA increases the number of Langley plots by 75 %. On the common set of data the OA_{OSM} produces smaller α's (by 0.0028). This is the opposite effect of OSM compared to NPF methods (see Sect. 5). The extra Langleys produced by the OA_{OSM} method do not necessarily indicate an improvement. Many of them are large outliers in the time series. We conclude that the OA, if it errs, it errs on being conservative, i.e., it has a fairly large missed detection rate (rejecting data sets with viable Langley plots), and at the same time the ones that are detected sometimes could be improved by a removal of few extra outliers.

When evaluating individual plots, and we looked at almost all 11×1023 of them, we found for each method cases when it went astray. There were cases of missed detection and false detection when judged by eye. However, we cannot quantify which of the algorithms has the most favorable missed and false detection rates. Out of all algorithms used, only OA deals explicitly with curvature. This perhaps might be a chief reason why it produces significantly fewer Langleys, which leads to a smaller number of large outliers in the time series.

The comparison of calibration constants that can be derived from α's obtained by each method gives us additional insight about each method as well as a strategy one

AMTD
8, 4191–4218, 2015

should use when generating the calibration constants. We compare the behavior of time series of derived calibration constants $\alpha_{cc}(d)$, where d indicates each day from 5 October 2003 to 30 March 2006. The $\alpha_{cc}(d)$ are obtained from the time series of $\alpha(d_j)$ independently for each Langley method. The $\alpha_{cc}(d)$ might be considered "the best" estimate of the calibration constant for a given day, "the best" in a sense of the method that we use to remove outliers, interpolate and smooth the series $\alpha(d_j)$.

The method consists of a moving median window of width d_{med} days that removes outliers and interpolates, which is followed by a moving a boxcar filter. For each day d the median is calculated from d_{med} number of $\alpha(d_j)$ values: $d_{med}/2$ values for $d_j \leq d$ and $d_{med}/2$ values for $d < d_j$. Then the 1055 long series $\alpha_{cc}(d)$ is smoothed with a boxcar filter of d_{smth} days. By trial and error we decided on $d_{med} = 30$ days and $d_{smth} = 25$ days.

Prior to applying the method described above, the values of α are corrected for Earth–Sun distance a by a substitution $\alpha \leftarrow \alpha + 2\ln(a)$, where a is in astronomical units. The Earth–Sun distance is calculated with the ephemeris program published by Michalsky (1988).

One of the methods of time series smoothing of V_0's to obtain the absolute calibration constants of a sun radiometer (MFRSR) was validated against calibrations at Mauna Loa by Michalsky and LeBaron (2013). The discrepancy between the time series smoothing and Mauna Loa calibration constants in terms of V_0's were always smaller than 0.6 %. This at air mass $m = 2$ translated to aerosol optical depth uncertainty of less than 0.003 and at $m = 6$ to less than 0.001. This MFRSR was located in Salt Lake City, Utah, which, in terms of number of sunny/clear sky days is somewhat superior to the SGP site. The OA method was used to identify the Langley plots and obtain individual V_0's.

In Fig. 3 we show all 11 $\alpha_{cc}(d)$ curves and individual intercepts α's from OA and H-β methods for rms ≤ 0.006. In the course of 1055 days the RSS's calibration constants vary within ±3.5 % (in terms of V_0) band. All methods follow these changes however there are differences among them. Statistically, the differences are ±1.4 % for 95 %

AMTD
8, 4191–4218, 2015

of days from the calibration constant curve that is the average of all 11 curves. For other values of rms_{max}: 0.010, 0.008 and 0.004 the differences are ±1.6, ±1.45 and ±1.4 %, respectively. The effect of the maximum rms on the differences between the $\alpha_{cc}(d)$ curves is not very dramatic. This is because the parameter d_{med} = 30 days of the median filter is relatively large.

In Fig. 4 we show 9 calibration constant curves and intercepts α's for LSF_{SRO}-x and S_{SRO}-α methods for rms ≤ 0.010. In this case we used α's from Langley plots that had no more than 10 % outliers. The OA and H-β results did not pass the filter of 10 % outliers only. The differences between the calibration constant curves are ±0.8 %. For 40, 30 and 20 % outliers, the calibration constant curves are within ±1.2, ±0.9 and ±0.87 % bands, respectively. So, the effect of number of outliers removed to obtain a Langley plot has a larger effect on the spread among the methods than the effect of rms_{max}.

We note that the OA and H-β calibration constant curves from Fig. 3 are marginally within the band defined by the curves in Fig. 4.

The majority of points in Fig. 4 are outliers and they are defined by Langley plots with 90 % or more points. By the criterion rms ≤ 0.01 the points are collinear. Nevertheless, they are off and some by more than ±5 % (in terms of V_0). In our opinion, the majority of the outliers are cases of anomalous Langley plots. The topic of anomalous Langley plots will be pursued in another paper.

10 Conclusions and summary

Eleven Langley plot methods were compared. Two of them were the least square methods and nine were non-parametric methods which included the objective algorithm (OA) method by Harrison and Michalsky (1994).

We developed two methods to terminate the non-parametric methods in order to determine the existence of the Langley plot: the outlier sorting method (OSM) that was applied to two Thiel (1950) and two Siegel (1982) methods, and a new non-parametric

AMTD
8, 4191–4218, 2015

method of sequential removal of outliers (SRO) was applied to two Siegel methods resulting in two new iterative Siegel methods.

We found that analysis of histograms of slopes and intercepts can be an excellent tool to prescreen a data set for Langley-plot viability. The histogram of slopes was used to generate Langley plots that produce lines defined by a small number of points. The histogram method offers a possibility to extract Langley plots when outliers dominate and to find all subsets of collinear points.

The OA method turned out to be robust though conservative. It identifies the lowest number of Langley plots. It produces intercepts slightly larger than all other methods.

The Siegel (1982) and Thiel (1950) methods with OSM produce very similar results. The two least square methods yield the largest number of Langley plots, with expected bias between them.

The largest differences among methods are on Langley plot cases that turn out to be outliers in terms of the calibration constant curve. Predominantly these are the cases that produce Langley plots, but with a small number of points. In cases that are close to the calibration constant curve the differences are small, but there are systematic biases.

We have no way of determining which of the methods produces results closest to the truth. In fact, the answer may depend on the data set. When the number of outliers in a Langley plot is small, all methods tend to produce similar results.

The metrics used to define the Langley plot was rms of residuals. The effect of the value of rms, whether it was 0.10 or 0.06 had no great impact on the calibration constant curves: all methods produced calibration constant curves within a band between ±1.4 to ±1.6 for 95 % of days. It is the number of outliers in the data set that has a greater impact. The calibration curves generated using a smaller number of Langley plots with each Langley defined by a larger number of points produce calibration constant curves that are less dependent on the method. For instance when Langley plots retain 80 % of the points all calibration constant curves are within ±0.9 % band for 95 % of days.

AMTD
8, 4191–4218, 2015

The outliers from the calibration constant curves are predominantly caused by anomalous Langley plots when the optical depth has a hyperbolic component as a function of air mass. This effect cannot be detected from the data set, and no Langley plot method can determine if this hyperbolic change with air mass is occurring. This effect at difficult sites like the SGP ARM site in Oklahoma sets the ultimate limit of accuracy of in situ calibrated sun radiometers.

Acknowledgements. We want to express our gratitude to Robert Evans of NOAA, Boulder, Colorado for providing the G. M. B. Dobson 1958 report; Bruce Forgan of Bureau of Meteorology, Melbourne, Australia for providing information on his approach to Langleys; Alberto Redondas of Izaña Atmospheric Research Centreto, Tenerife, Spain for providing information on Langleys in the Brewer network; Jim Schlemmer of ASRC, SUNY at Albany, NY for running the OA Langley method on RSS data.

References

Augustine, J. A., Cornwall, C. R., Hodges, G. B., Long, C. N., Medina, C. I., and DeLuisi, J. J.: An automated method of MFRSR calibration for aerosol optical depth analysis with application to an Asian dust outbreak over the United States, J. Appl. Meteorol., 42, 266–278, 2003.

Birknes, D. and Dodge, Y.: Alternative Methods of Regression, Wiley-Interscience, ISBN: 978-0-471-56881-0, 1993.

Campanelli, M., Nakajima, T., and Olivieri, B.: Determination of the solar calibration constant for a sun-sky radiometer: proposal of an in situ procedure, Appl. Optics, 43, 651–659, 2004.

Dobson, G. M. B. and Normand, C.: Determination of Used in the Calculation of the Amount of Ozone from Spectrophotometer Measurements and the Accuracy of the Results, International Ozone Commission (I.A.M.A.P.), October, 1958

Forgan, B.: Practical Sun Spectral Radiometer Calibration Methods, International Pyrheliometer Comparison, PMOD-World Radiation Center, Davos, Switzerland, 2000.

Harrison, L. and Michalsky, J.: Objective algorithms for the retrieval of optical depths from ground based measurements, Appl. Optics, 33, 5126–5132, 1994.

AMTD
8, 4191–4218, 2015

Harrison, L. Michalsky, J., and Berndt, J.: Automated multi-filter rotating shadowband radiometer: an instrument for optical depth and radiation measurements, Appl. Optics, 33, 5118–5125, 1994.

Harrison, L., Beauharnois, M., Berndt, J., Kiedron, P., Michalsky, J., and Min, Q.: The rotating shadowband spectroradiometer (RSS) at SGP, Geophys. Res. Lett., 26, 1715–1718, 1999.

Harrison, L., Kiedron, P., Berndt, J., and Schlemmer, J.: Extraterrestrial solar spectrum 360–1050 nm from Rotating Shadowband Spectroradiometer measurements at the Southern Great Plains (ARM) site, J. Geophys. Res-Atmos., 108, 4424–4432, doi:10.1029/2001JD001311, 2003.

Herman, B. M., Box, M. A., Reagan, J. A., and Evans, C. M.: Alternate approach to the analysis of solar photometer data, Appl. Optics, 20, 2925–2928, 1981.

Ito, M., Uesato, I., Noto, Y., Ijima, O., Shimidzu, S., Takita, M., Shimodaira, H., and Ishitsuka, H.: Absolute calibration for Brewer spectrophotometers and Total Ozone/UV radiation at Norikura on the Northern Japanese Alps, Journal of the Aerological Observatory, 72, 45–55, 2014.

Jaeckel, L.: Estimating the regression coefficients by minimizing the dispersion of residuals, Ann. Math. Stat., 43, 1449–1458, 1972.

Kasten, F.: A new table and approximation formula for the relative optical air mass, Arch. Meteor. Geophy. B, 14, 206–223, 1965.

Kendall, M. G.: A new measure of rank correlation, Biometrika, 30, 81–93, 1938.

Kuester, M. C., Thome, K. J., and Reagan, J. A.: Automated statistical approach to Langley evaluation for a solar radiometer, Appl. Optics, 42, 4914–4921, 2003.

Kiedron, P. W., Michalsky, J. J., Berndt, J. L., and Harrison, L. C.: Comparison of spectral irradiance standards used to calibrate shortwave radiometers and spectroradiometers, Appl. Optics, 38, 2432–2439, 1999.

Long, C. N. and Ackerman, T. P.: Identification of clear skies from broadband pyranometer measurements and calculation of downwelling shortwave cloud effects, J. Geophys. Res.-Atmos., 105, 15609–15626, 2000.

Marenco, F.: On Langley plots in the presence of a systematic diurnal aerosol cycle centered at noon: a comment on recently proposed methodologies, J. Geophys. Res.-Atmos., 112, D06205, doi:10.1029/2006JD007248, 2007.

Michalsky, J. J.: The Astronomical Almanac's algorithm for approximate solar position (1950–2050), J. Sol. Energ.-T. ASME, 40, 227–235, 1988.

AMTD
8, 4191–4218, 2015

Michalsky, J. and LeBaron, B.: Fifteen-year aerosol optical depth climatology for Salt Lake City, J. Geophys. Res.-Atmos., 118, 3271–3277, doi:10.1002/jgrd.50329, 2013.
Mlawer, E. J., Brown, P. D., Clough, S. A., Harrison, L. C., Michalsky, J. J., Kiedron, P. W., and Shippert, T.: Comparison of spectral direct and diffuse solar irradiance measurements and calculations for cloud-free conditions, Geophys. Res. Lett., 27, 2653–2656, 2000.
Nieke, J., Pflug, B. G., and Zimmermann, G.: An aureole-corrected Langley-plot developed for the calibration of HiRES grating spectrometers, J. Atmos. Sol.-Terr. Phy., 61, 739–744, 1999.
Redondas, A.: RBCC-E ozone absolute calibration, langley regression method, in: The Ninth Biennial WMO Consultation on Brewer Ozone and UV Spectrophotometer Operation, Calibration and Data Reporting, 2005.
Schmid, B. and Wehrli, C.: Comparison of Sun photometer calibration by use of the Langley technique and the standard lamp, Appl. Optics, 34, 4500–4512, doi:10.1364/AO.34.004500, 1995.
Shaw, G. E.: Error analysis of multi-wavelength Sun photometry, Pure Appl. Geophys., 114, 1–14, 1976.
Siegel, A. F.: Robust regression using repeated medians, Biometrika, 69, 242–244, doi:10.1093/biomet/69.1.242, 1982.
Tanaka, M., Nakajima, T., and Shiobara, M.: Calibration of a sunphotometer by simultaneous measurements of direct-solar and circumsolar radiations, Appl. Optics, 25, 1170–1176, 1986.
Theil, H.: A rank-invariant method of linear and polynomial regression analysis, I, II, III, Nederl. Akad. Wetensch. Proc., 53, 386–392, 521–525, 1397–1412, 1950.

AMTD
8, 4191–4218, 2015

Table 1. The comparison among 11 Langley methods for rms ≤ 0.006. Number of successful Langley plots is on diagonal. Above the diagonal the number of common Langley plots for two methods and a SD of differences $\Delta\alpha$. Below the diagonal the mean and median of $\Delta\alpha$. The order of subtraction in $\Delta\alpha$ is as follows: α for the method from row minus α for the method from the column.

	LSF_{SRO}-x	LSF_{SRO}-$1/x$	OA	OA_{OSM}	S_{OS}-β	S_{OS}-α	T_{OSM}-β	T_{OSM}-α	S_{SRO}-β	S_{SRO}-α	H-β
LSF_{SRO}-x	**601**	588	284	485	537	539	522	525	361	450	414
		0.0198	0.0071	0.0141	0.0190	0.0188	0.0190	0.0190	0.0096	0.0135	0.0164
LSF_{SRO}-$1/x$	−0.0055	**598**	283	479	535	537	523	526	361	448	414
	−0.0002		0.0143	0.0258	0.0203	0.0228	0.0217	0.0210	0.0192	0.0171	0.0201
OA	+0.0043	+0.0065	**284**	284	281	281	278	279	264	277	277
	+0.0029	+0.0035		0.0037	0.0118	0.0108	0.0103	0.0107	0.0071	0.0075	0.0077
OA_{OSM}	+0.0051	+0.0111	−0.0028	**499**	475	477	463	466	340	411	361
	+0.0009	+0.0017	−0.0026		0.0178	0.0163	0.0142	0.0148	0.0085	0.0133	0.0113
S_{OSM}-β	+0.0029	+0.0075	−0.0044	−0.0045	**564**	561	536	538	351	436	398
	$+2\times10^{-5}$	+0.0012	−0.0031	−0.0010		0.0040	0.0103	0.0100	0.0141	0.0144	0.0373
S_{OSM}-α	+0.0038	+0.0088	−0.0041	−0.0039	+0.0008	**564**	537	539	353	437	398
	+0.0004	+0.0014	−0.0028	−0.0009	0		0.0096	0.0095	0.0132	0.0137	0.0372
T_{OSM}-β	+0.0036	+0.0085	−0.0047	−0.0042	+0.0009	+0.0001	**541**	541	348	428	390
	+0.0003	+0.0013	−0.0029	−0.0007	0	0		0.0019	0.0097	0.0110	0.0369
T_{OSM}-α	+0.0036	+0.0086	−0.0048	−0.0043	+0.0010	+0.0003	+0.0002	**544**	350	430	391
	+0.0005	+0.0015	−0.0029	−0.0005	0	0	0		0.0101	0.0110	0.0367
S_{SRO}-β	−0.0005	+0.0036	−0.0048	−0.0029	−0.0005	−0.0007	$+4\times10^{-5}$	−0.0001	**364**	352	328
	−0.0007	−0.0002	−0.0046	−0.0023	−0.0011	−0.0013	−0.0010	−0.0010		0.0043	0.0061
S_{SRO}-α	−0.0004	+0.0038	−0.0054	−0.0053	−0.0011	−0.0016	−0.0011	−0.0012	−0.0008	**475**	375
	−0.0007	+0.0006	−0.0045	−0.0031	−0.0014	−0.0015	−0.0011	−0.0011	−0.0004		0.0064
H-β	+0.0006	+0.0031	−0.0046	−0.0035	−0.0039	−0.0046	−0.0042	−0.0043	$+4\times10^{-6}$	+0.0006	**434**
	−0.0006	+0.0001	−0.0037	−0.0023	−0.0010	−0.0014	−0.0011	−0.0011	+0.0007	+0.0011	

AMTD
8, 4191–4218, 2015

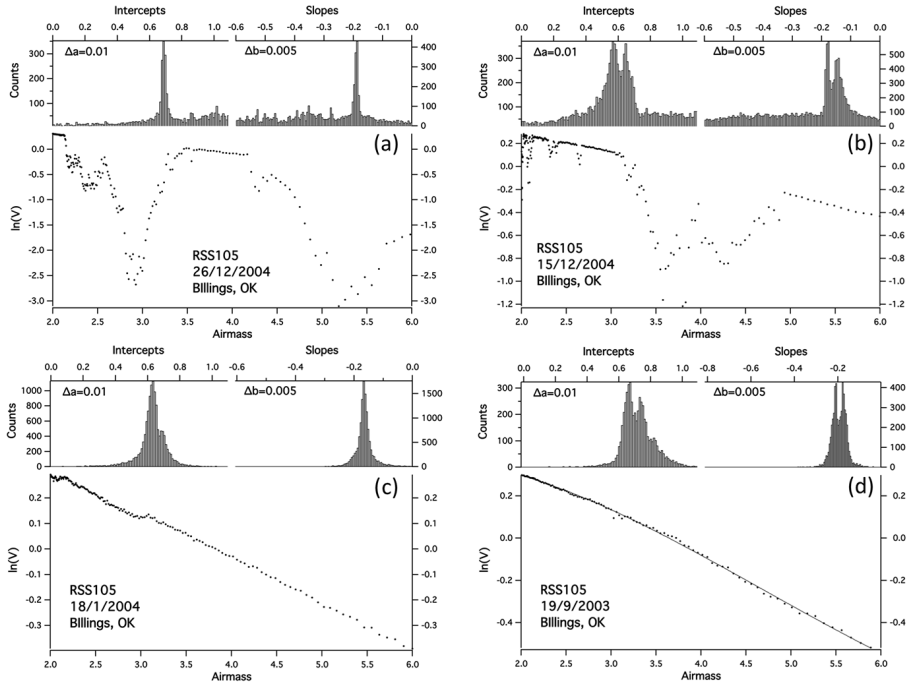

Figure 1. Four cases that illustrate how slope and intercept histograms can aid in analysis of points for the extraction of the Langley plot. In **(b)** and **(d)** both the slope and intercept histograms are bimodal. In case **(c)** only the histogram of intercepts is bimodal. Case **(a)** has many outliers, but both histograms are mono-modal indicating the existence of a single Langley plot. Case **(d)** has no large outliers, but y vs. x is nonlinear (3rd degree polynomial).

AMTD
8, 4191–4218, 2015

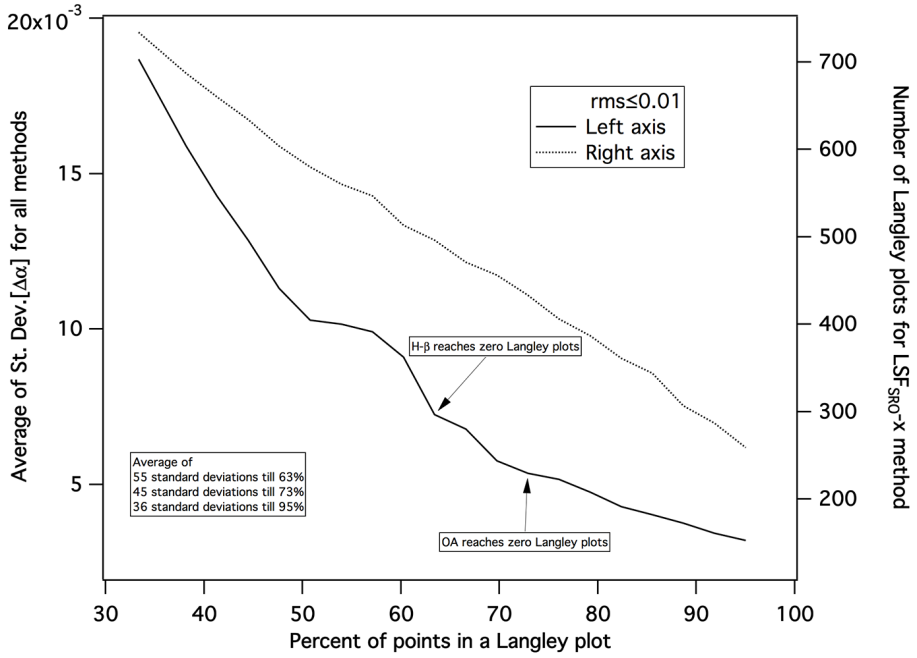

Figure 2. The average over all methods of SDs of $\Delta\alpha$ as a function of number of points in a Langley plot.

AMTD
8, 4191–4218, 2015

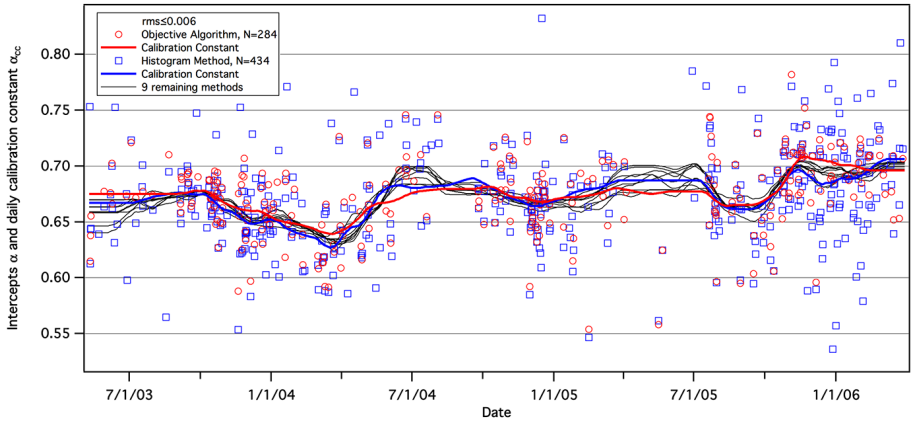

Figure 3. Calibration constant curves for all 11 methods (rms < 0.006) and individual intercepts α for OA and H-β methods.

AMTD
8, 4191–4218, 2015

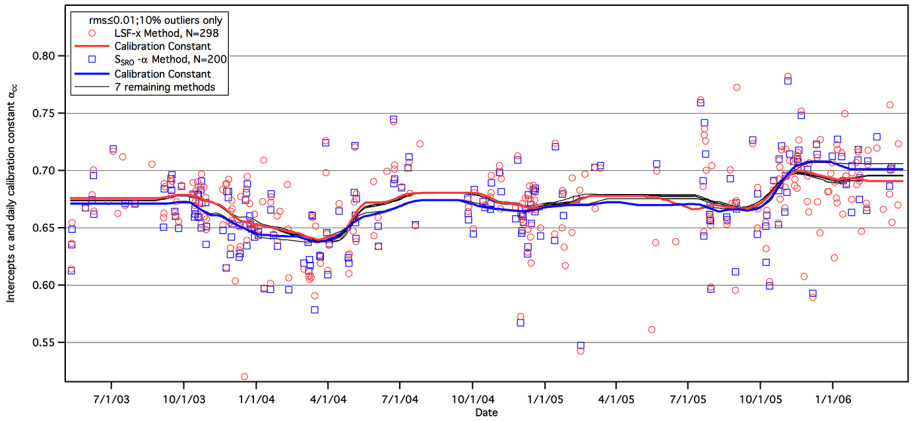

Figure 4. Calibration constants curves for 9 methods (rms < 0.01) from Langley plots with no more than 10 % outliers and individual intercepts α's for LSF_{SRO}-x and S_{SRO}-α methods. (The OA and H-β results did not pass the filter of 10 % outliers only.)

AMTD
8, 4191–4218, 2015