**GRAPHS Lascar CO Data Validation Plan**

Validation of CO data will proceed according to information provided by the following three variables:

* Calibration Factor Confidence (high/low/none)
* Visual validity assessment (1 = looks valid; 2 = suspect; 3 = looks invalid)
* Duration (1 = hours ≥ 44; 2 = hours between 18 and 44; 3 = hours < 18). 18 hours is 75% of 24 hours.

**OVERALL VALIDITY**

The overall validity of a CO file will be coded as follows:

**1** if all of the following:

* CF confidence = high
* Visual Validity = 1
* Duration = 1

**2** if all of the following: OR if all of the following:

* CF Confidence = high • CF Confidence = high
* Visual Validity = 2 • Visual Validity = 1
* Duration = 1 or 2 • Duration = 2

**3** if all of the following:

* CF Confidence = low
* Visual validity = 1 or 2
* Duration = 1 or 2

**4** if ANY of the following:

* Visual Validity = 3
* OR Duration = 3
* OR CF Confidence = none

**CALIBRATION FACTORS AND CF CONFIDENCE**

Calibration factors are calculated using data collected during calibration sessions with the 50ppm CO tank. Calibration factor (CF) is the measured value/ expected value. Each data point will be assigned two variables associated with calibration:

1) a numeric CF;

2) a “CF confidence flag” indicating high/low/no confidence in application of the numeric CF to the data.

1. **Validity Criteria for Calibration Factors:** 
   1. We will generally consider calibration factors to be valid if 0.6 ≤ CF ≤1.2.
2. **Application of CFs to Lascar Session Data:**
   1. A CF will be generated for each month-year combination for each month that a Lascar unit was in use (e.g. there will be a CF for Lascar SN = 11115 for October 2014).
   2. Each session’s CO data is corrected using a constant CF value. This value is determined by the date of the last sample (e.g. a 72-hour session that starts on Oct. 30, 2014 and ends on Nov. 1, 2014 will be assigned the “November 2014” calibration factor for the Lascar used in that session.)
   3. NOTE: the **overall mean CF** across all units over all calibration sessions in 2014, excluding calibration sessions outside the 0.6-1.2 range, is **0.85**.
3. **Determining monthly CFs:**

Determining initial CFs:

* 1. Data collection for GRAPHS began in October 2013, yet no calibration sessions were held until February 2014. Thus we will set a “virtual” CF for October 2013 at the first measured CF value. All months between this “virtual” measurement and the first actual measurement will be assigned this constant CF.

Determining CFs for the “middle” of the data:

1. Determining monthly averages: If two CFs are measured within the same month, take the average of the two as the monthly CF value.
2. Interpolating CFs: Calculate CF by month and year according to monthly linear interpolation between measured (monthly averaged) CFs.
3. (See Special Note 1 below for dealing with data preceding “zero” CFs).

Determining CFs for the end of the data (extending beyond the last calibration session):

1. Use constant interpolation from the last measured CF through the end of the data. Color the tail end of data purple and assign “medium” validity if it has been more than 8 months since the last validation.
2. **Confidence Flags for Calibration Factors:**

Each monthly CF receives a “high/low/none” flag for our assumed confidence in applying the CF to the data. General rules for applying confidence flags are as follows:

* 1. Assume “high” confidence if the CF is in the range 0.6 ≤ CF ≤ 1.2
  2. Assume “low” confidence if the CF is > 1.2 or if 0.2 ≤ CF ≤ 0.6
  3. Assume “no” confidence if the CF is < 0.2 (this is a CF of essentially zero).
  4. **Special case:** If there are no two adjacent valid monthly averaged CFs for a particular unit, (including the first “virtual” data point as one of the CFs to use in making this determination), assign “low” CF confidence to any points that would otherwise have received “high” confidence. We have low confidence in the data generated by this Lascar unit.

SPECIAL NOTE 1: **If a zero (defined as CF < 0.2) is generated during a calibration session**, do not linearly interpolate between the previous value and the zero. Rather, for the months between the last recorded CF and the zero, apply constant interpolation from the last CF up until the month of the zero. Assign these months before the zero “low” confidence.

SPECIAL NOTE 2: **If a Lascar Unit was never calibrated**: Assign the mean CF to the entire dataset. Assign “low” CF confidence to the entire dataset.

Locations of relevant files:

Datasets of Calibration Factors: ~/Dropbox/Ghana\_exposure\_data\_SHARED (1)/CO\_calibration\_files/Calibration Factors/Datasets/

* calib\_factors\_bymonth (uninterpolated monthly-averaged CFs)
* calib\_factors\_bymonth\_interp (interpolated CFs)

Plots of Calibration Factors: /Users/ashlinn/Dropbox/Ghana\_exposure\_data\_SHARED (1)/CO\_calibration\_files/Calibration Factors/Plots/

* Lascar\_Calibrations\_All (plot of all Lascars calibrated by calibration session)
* Lascar\_Calibrations\_Each (separate plots of each Lascar’s calibrations)
* Calib\_factors\_bymonth\_interp (plots of the interpolated CFs for each lascar)

**VISUAL VALIDATION OF LASCAR SESSION DATA**

Lascar session data (original and corrected according to the CFs decided above) is plotted here: ~/Dropbox/Ghana\_exposure\_data\_SHARED (1)/CO\_files\_processed/. (As of Dec. 21, 2014, the most recently processed data is in the 20Dec2014 folder within this directory).

Humans will look at all the plots and make a determination of its validity in a form. A folder of these forms is available here: ~/Dropbox/Ghana\_exposure\_data\_SHARED (1)/CO\_files\_processed/Validation Forms/

1. The CO\_VALID column of the form will be filled out according to the following codes:

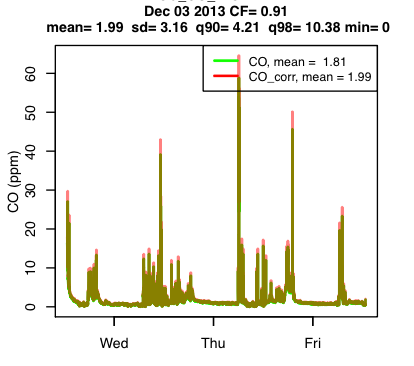
1 = visually valid

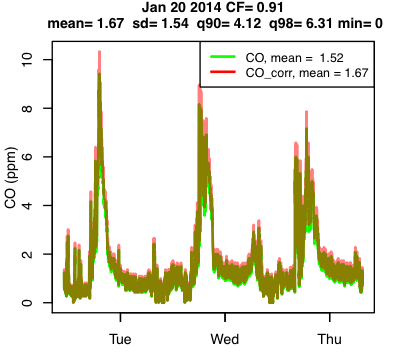
2 = visually suspect

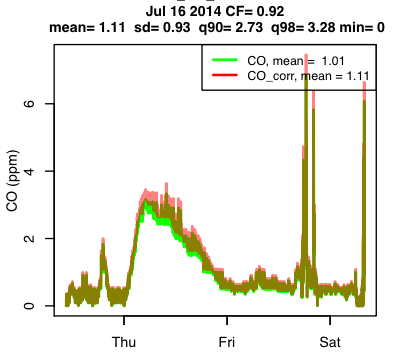
3 = visually invalid

1. The plots at the link in (3a) above have been “flagged” according to initial automated criteria as follows:
   1. with blue titles (in the plot) and coded “3 = visually invalid” (in the form, column “duration\_valid”) if the duration of sampling was < 44 hours,
   2. with red titles (in the plot) and initially coded “2 = suspect” (in the form, column “co\_valid\_init”) if any of the following criteria have been met:
      1. 98th percentile of CO = 0 (these files have unusually long stretches with no measured CO)
      2. 90th percentile of CO > 20 (this pulls files that have an odd "plateau" or "hill" in the CO tracing)
      3. minimum CO value > 10 (many files have a minimum CO value slightly above 0 , see below - but the ones above 10 usually look seriously messed up)
      4. sd of CO > 60 (this overlaps with some of the others but pulls some additional files that also have odd-looking shapes).
2. Note that the computer-generated red flags / code 2s are merely a suggestion that the data should be examined more closely by humans.
3. Criteria that will be assessed by human eye include:
   * + 1. Pattern of “spikes” looks as expected (generally a zero baseline with occasional high values: no high plateau, no “hill” of consistently increasing or decreasing CO values, no “mountain” or “ski-slope” shapes, no “stair-step” blockiness in the data) – note that these visual patterns will be influenced by the scale of the data (hills, mountains, etc may be permissible if the CO range is small).
       2. Elevated baseline – The majority of CO readings hover above 0 (above 1-2ppm, this is unexpected and probably indicates a problem. Note that this behavior may occur even if the “min” CO value is 0). Files were given a “2” for visual validity if the baseline was between 3-10 ppm and a “3” if the baseline was above 10ppm.
       3. Long periods of “flatline” at 0 may indicate problems. These will be evaluated on a case-to-case basis (e.g., periods of flatline at 0 while CO spikes still occur may not be problematic, but a sudden change from more responsive data to sudden flatline should be flagged as 2 or 3). Files demonstrating flatline for longer than 36 hours at any point during the = session were given a “2”.
       4. Individual data plots should be considered in the context of the plots that precede and follow them – if the instrument clearly fails at a certain point, closer examination of the data that precedes the failure may be necessary (and any deviations from normalcy in this data may trigger a validation flag of 2 or 3).
4. Duration: Files < 44 hours in duration are automatically flagged with a “3” in the Duration\_valid column. Visual validity will be assessed irrespective of session length; e.g. for files that are < 44 hours in duration, those whose CO data looks ok will be marked as “valid” (1) in the CO\_valid column.

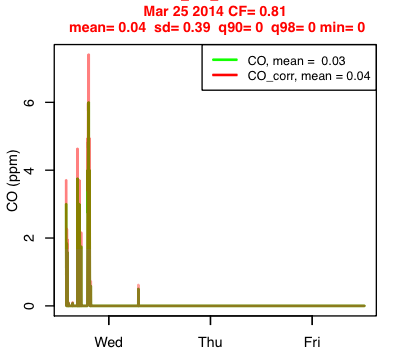
**Examples of “good” tracings (coded 1):**

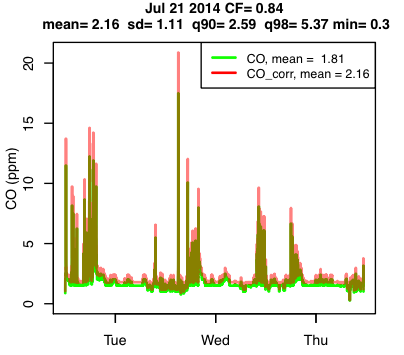


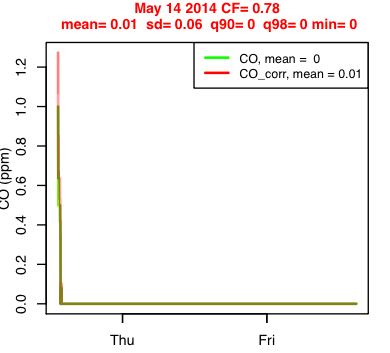
Note the small y-axis range

 Although a “hill” exists in the data, it is at a low CO level and is likely to represent real exposure.

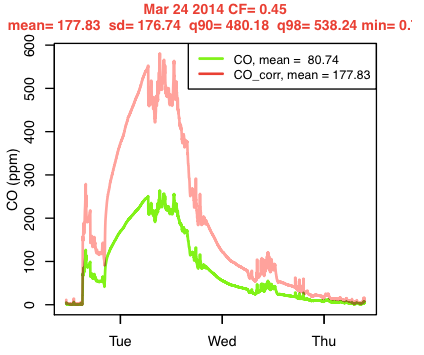
**Examples of tracings coded 2:**

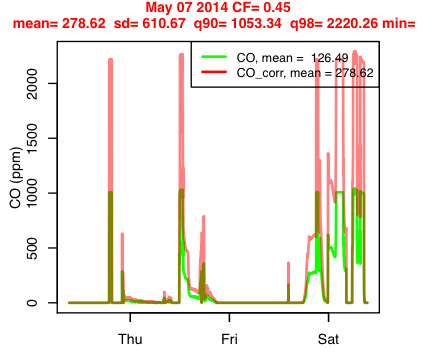
This plot flatlines at 0 after initial response. The following plots also show problematic tracings and the unit stops logging after 5 subsequent sessions.

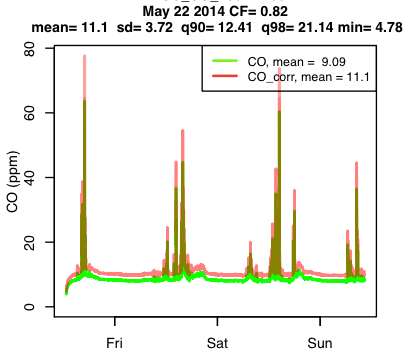
Note slightly elevated baseline.

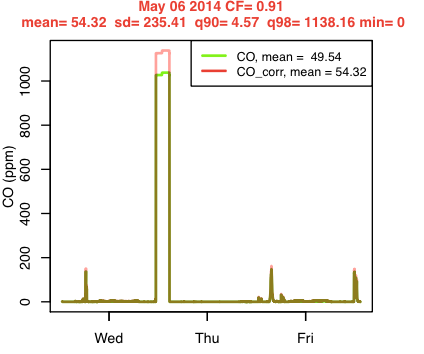
Note that virtually no CO was recorded (but the plots before and after this one were good)

**Examples of tracings coded 3:**

Note “mountain” shape in data (staying at high values of CO for hours/days). CF is also questionable.

Note repeated high plateaus. CF is also questionable.

Note significantly elevated baseline.

Note high plateau.