Tropospheric Airborne Measurement Evaluation Panel (TAbMEP) Introduction to Analysis and ICARTT Field Study

1. Background

The primary objectives of TAbMEP are to provide an unbiased assessment of the measurement uncertainties and measurement consistency for historical airborne observations, and to establish systematic approaches for combining airborne data sets from multiple instruments/techniques and aircraft platforms. The TAbMEP assessment is part of larger effort sponsored by NASA MEaSUREs program to make the airborne databases suitable for the assessment of global and region models. In the case of ICARTT, four different aircraft conducted extensive intercomparisons during the summer 2004 field campaign [*Fehsenfeld et al.*, 2006, *Singh et al.*, 2006]. This report is to recommend methods to combine these ICARTT data for any analysis based upon data collected on different aircraft, especially the analysis involving the comparisons and contrasts of the ICARTT data. Measurement biases between platforms can potentially confound such comparisons and contrasts. The current TAbMEP work is designed to put limits on the magnitude of possible biases and to provide objective uncertainty limits for the data collected from the instruments on all of the intercompared aircraft. The present analysis is limited to a few selected species: O_3, H_2O, CO, NO, NO_2, PAN, HNO_3, SO_2, few VOCs including CH_2O, temperature, wind, j(NO_2), j(O^1D), particle number density, volume density, and sulfate. The following chapters give the assessments of the measurements of each of these species.

2. Intercomparison Flights

This assessment is based primarily upon five intercomparison flights conducted during the field campaign. Each of the intercomparison flights involved two aircraft flying wingtip-to-wingtip at two or three different altitudes for 40 minutes to more than an hour. The flights were planned to encounter a range of conditions. Table 1 summarizes the five flights, and Figures 1(a)-1(e) illustrate the flight tracks.

Table 1. Summary of intercomparison flights

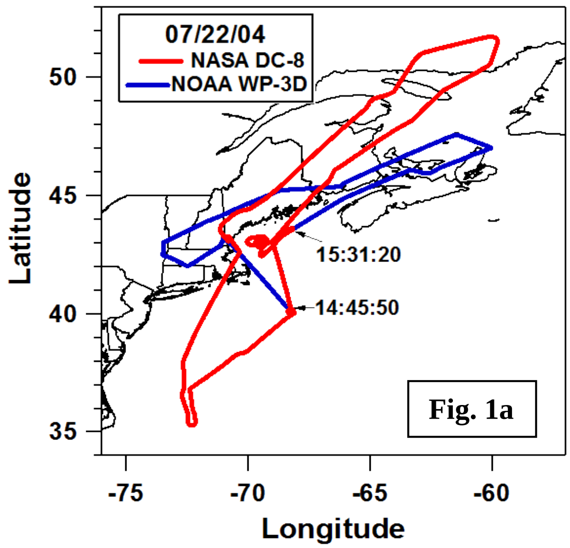
Date – Time (UTC)	Aircraft	Location
7/22/2004 - 14:45:50-15:32:14	DC-8/WP-3D	W. Atlantic – E. of Massachusetts, US
7/31/2004 - 22:52:50-23:32:10	DC-8/WP-3D	Eastern Maine, US
8/07/2004 - 21:35:10-22:19:10	DC-8/WP-3D	Bay of Fundy
7/28/2004 - 15:49:30-17:01:20	DC-8/BAe-146	Central N. Atlantic – W. of Azores
8/03/2004 - 15:08:45-16:31:00	BAe-146/DLR Falcon	French Atlantic coast

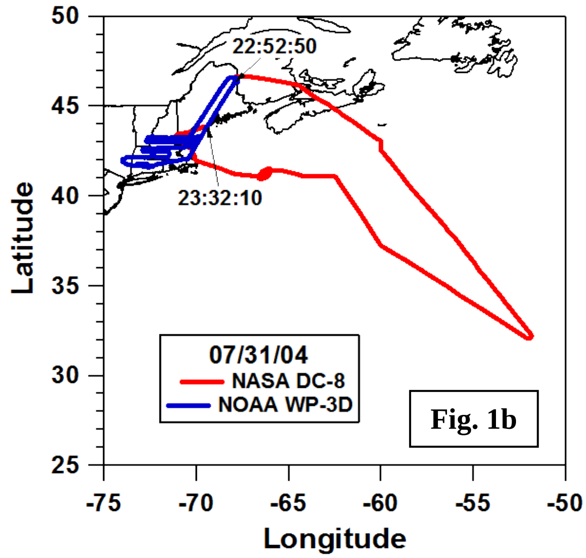
3. Analysis Techniques

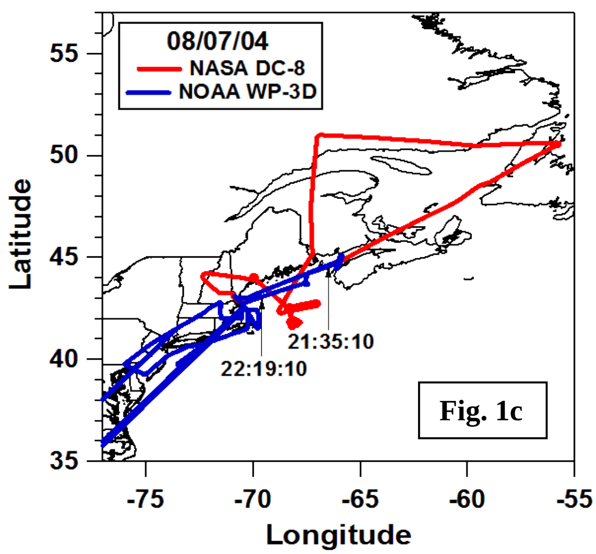
Several different analysis techniques have been utilized in this assessment of instrument precision and bias. Summaries of these techniques are given in the following sections.

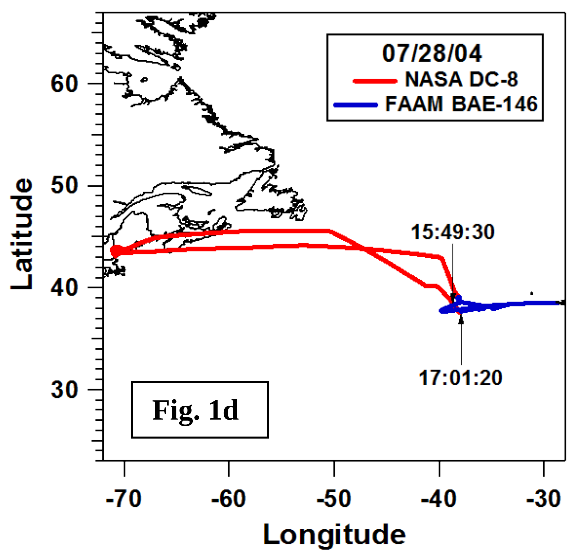
3.1. Precision Analysis

Internal Estimate of Instrument Precision (IEIP) is an objective and data-driven approach to assess absolute and/or relative instrument precisions. IEIP directly estimates, under a few assumptions, the instrument precision through the variance over a small time interval, Δt. For species x, the total variance can be expressed as:









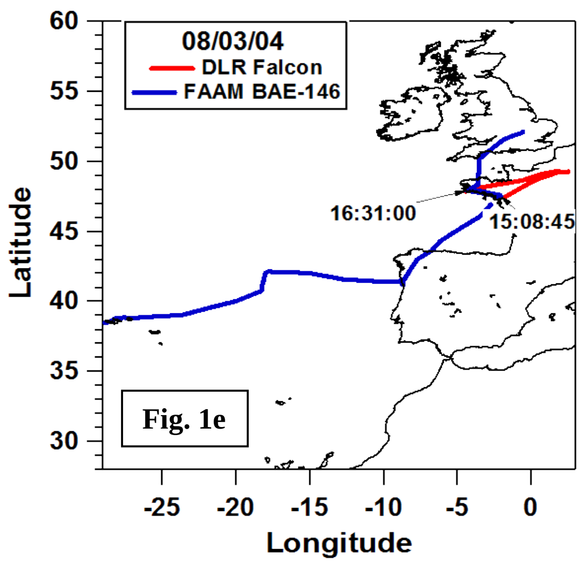


Figure 1(a), (b), (c). Flight tracks for the NASA DC-8 (red) and NOAA WP-3D (blue) intercomparison flights on (a) July 22, 2004, (b) July 31, 2004, (c) August 7, 2004.

Figure 1(d). Flight tracks for the NASA DC-8 (red) and FAAM BAE-146 (blue) intercomparison flights on July 28, 2004.

Figure 1(e). Flight tracks for the DLR Falcon (red) and FAAM BAE-146 (blue) intercomparison flights on August 3, 2004.

$$V(x) = \sigma_x^2 + \sigma_{\varepsilon-x}^2 \quad (1)$$

where $\sigma_{\varepsilon-x}^2$ is the random instrument variability (instrument precision) for measurement of species x and σ_x^2 represents the natural ambient variability.

$V(x)$ can be a reasonable estimate for $\sigma_{\varepsilon-x}^2$ if Δt is small enough such that σ_x^2 is negligible compared to $\sigma_{\varepsilon-x}^2$. At the same time, Δt must be large enough to minimize the effect of autocorrelation. $\sigma_{\varepsilon-x}^2$ can be assessed by following procedures listed below:

1. Compute standard deviation over Δt and generate frequency distribution or histograms.
2. Vary Δt and repeat the previous step, then look for the values of the modes, which are relatively constant over a limited range of Δt values.
3. How long should Δt be? In principle, it should be long enough to overcome any significant autocorrelation impact and short enough such that σ_x^2 is negligible.
   - Δt depends on temporal and spatial variability of the species or parameter of interest.
   - Δt depends on instrument sampling rate.
   - Δt determination requires expert judgment.

IEIP analysis is typically applied over an entire flight and/or a large segment of data with fairly constant values. Points below the limit of detection (LOD) may introduce bias in IEIP analysis. However, in the situation where the instrument LOD points are not flagged in the data, IEIP absolute precision may be used to estimate the limit of detection if a significant amount of data is in the low range near the LOD. It should be noted that this approach may or may not be feasible for measurements with long integration times and/or significant gaps between the data points. IEIP analysis may also be problematic when measurement precision is strongly dependent on the ambient values.

IEIP Example: O_3 instrument precision assessment

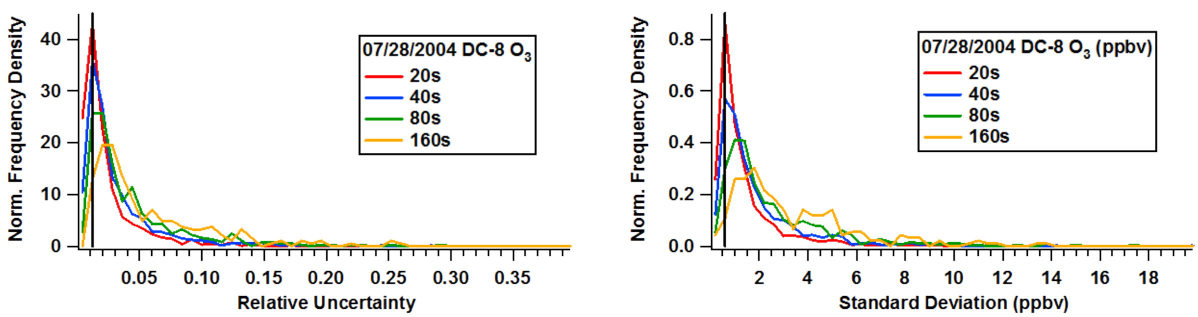


Figure 2. Example of IEIP analysis of NASA DC-8 O_3 observations during INTEX-A/ICARTT.

Figure 2 shows an example of IEIP assessment for O_3 for both relative and absolute uncertainties. Note that the modes of the distributions (i.e., the location of the peaks) are relatively constant over the range of Δt from 20-40 seconds. The standard deviation increases with longer Δt times, which is likely due to the O_3 natural variability. The resulting relative uncertainty for this DC-8 flight is about 1.2% and absolute uncertainty is 0.6 ppbv.
This procedure is an effective method to estimate so called "short-term" precision, which accounts for signal variation during a short period of assumed constant measurements. Because this assumption is not always valid, the IEIP estimate tends to provide an upper limit of the

instrument short-term precision. Over longer time scales, however, some instruments are subject to lower precision (i.e. larger variability), which includes variability that arises from uncorrected changes in the zero level or sensitivity of the instrument. These additional contributions to the variability are not likely reflected in the IEIP derived precision, but the intercomparison flights do provide a reasonable check on their influence. This effect was examined through the comparisons of the "expected variability" and "observed variability" in the individual species assessments. “Expected variability” is defined as the quadrature sum of the individual IEIP precisions for the paired instruments. “Observed variability” is derived from relative residual plots, an example of which is shown in Figure 3.

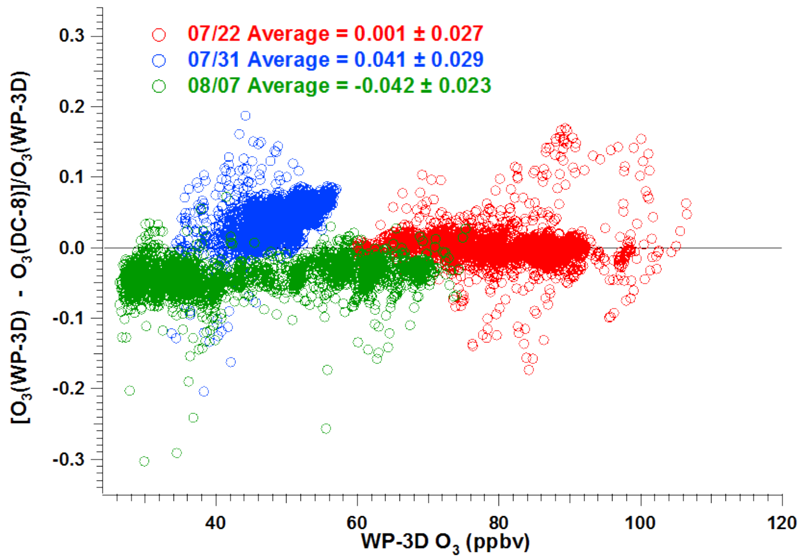


Figure 3. Relative residual for NOAA WP-3D and NASA DC-8 O_3 during the ICARTT campaign.

Table 2. Example of IEIP analysis results

Flight	Platform	IEIP Precision	Expected Variability	Observed Variability	Adjusted Precision
07/22	O_3 DC-8	1.2%	1.8%	2.7%	1.8%
	O_3WP-3D	1.4%			2.1%
07/31	CO DC-8 DACOM	0.8%	1.7%	1.6%	0.8%
	CO WP-3D VUVF	1.5%			1.5%

The “observed variability” (in percentage) equals 100 $\times$ standard deviation (or 2.7% on 7/22 in this example). Each standard deviation (or “observed variability” value) should equal the “expected variability”. When the “observed variability” is larger than the “expected variability”, the IEIP derived (short-term) precision needs to be adjusted to reflect the longer term fluctuations (see 07/22 O_3 example in Table 2). This “adjusted precision” is obtained by proportionally scaling the IEIP estimates so that the “expected variability” value equals the “observed variability.” When the "observed variability" is smaller than the “expected variability”, the "adjusted precision" is set equal to the "IEIP precision" (see 07/31 CO example in Table 2). Ambient variability should not pose a problem, since it should have been sampled by both instruments during the intercomparison period. A key assumption made here for the precision adjustment is that the long-term precision is proportionally scaled with the short-term precision. This assumption may not reflect the actual instrument performance and may lead to either an underestimate or an overestimate of the long-term precision for each of the given instruments. The final adjusted precision estimates are required to be reviewed by TAbMEP measurement experts.

3.2. Linear Regression Techniques

The results of instrument intercomparisons are often reported as the linear regression of the measurements of one instrument as a function of the measurements of another, e.g. [*Hoell et al.*, 1985]. Orthogonal distance regression (ODR) is a regression technique similar to ordinary least squares (OLS) fit with the stipulation that both x and y are independent variables with errors. ODR minimizes sum of the squares of the orthogonal distances rather than the vertical distances (as in OLS). ODR is generally equivalent to

$$\min_{\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{\varepsilon}} \frac{1}{2}\sum_{i=1}^{n}\left(w_{\varepsilon_i}\varepsilon_i^2 + w_{\delta_i}\delta_i^2\right)$$

subject to $\boldsymbol{y_i} + \boldsymbol{\varepsilon_i} = \boldsymbol{f}(\boldsymbol{x_i} + \boldsymbol{\delta_i}; \boldsymbol{\beta})$ where $\boldsymbol{\varepsilon_i}$is the error in $\boldsymbol{y}$, $\boldsymbol{\delta_i}$ the error in $\boldsymbol{x}$, $\boldsymbol{w_{\varepsilon_i}}$and $\boldsymbol{w_{\delta_i}}$ weighting factors, and $\boldsymbol{\beta}$ a vector of parameters to be determined (i.e. slope and intercept in this case), [*Zwolak et al*., 2007]. Note that a weighted ODR ($\boldsymbol{w_{\varepsilon_i}}$ and $\boldsymbol{w_{\delta_i}} \neq \boldsymbol{1}$) is necessary when observations $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$ are heteroscedastic (variance changes with $\boldsymbol{i}$, or observations have point by point uncertainties), [*Boggs et al.*, 1988].

It has been shown that ODR performs at least as well and in many cases significantly better than OLS, especially when $\boldsymbol{d} = \boldsymbol{\sigma_\varepsilon}/\boldsymbol{\sigma_\delta} \leq \boldsymbol{2}$, [*Boggs et al.*, 1988]. Boggs et al. have shown that ODR results in smaller bias, variance, and mean square error (mse) than OLS, except possibly when significant outliers are present in the data, [*Boggs et al.*, 1988]. For the bias of the parameter, ($\boldsymbol{\beta}$), and function estimates, $\boldsymbol{f}(\boldsymbol{x_i}; \boldsymbol{\beta})$, OLS is statistically better only 2% of the time while ODR is significantly better 50% of the time. Results for the variance and mse of the parameter and function estimates were similar; ODR variance and mse were smaller than that from OLS about 25% of the time. OLS results were significantly better than ODR only 2% of the time, [*Boggs et al.*, 1988].

For our application, the data from the instruments being compared is merged to the same time base then plotted and fit using ODR. No LOD points should be included in regression. Points that have significantly large ambient variability need be examined to determine if they should be included in the regression analysis. Normally the observations are not heteroscedastic. In addition, an accurate estimate of measurement uncertainty is not often available on point by point basis. Therefore, in the interest of treating all the intercomparisons uniformly, we use $\boldsymbol{w_{\varepsilon_i}}$and $\boldsymbol{w_{\delta_i}} = \boldsymbol{1}$. The coefficient of determination, $\boldsymbol{R^2}$, is used to evaluate the robustness of the regression. Under some circumstances, we do not use ODR analysis. This occurs when the range of the data is small or there are very few data pairs (typically for some VOCs). The general rule to assess the range is when variability of the data set is less than 5 times the uncertainty, ODR is not used. In that case we present the data but do not subject it to additional analysis.

It is not uncommon that a few points are far apart from the bulk of the data points. In this case, a test should be performed to determine if any point exists that could be considered as an influential point. For TAbMEP analysis purposes an influential point is defined as a point that causes the slope or intercept of a regression line to statistically significantly differ with removal of the point. This check is to ensure robustness of regression, i.e., no one point has significantly large weight in the determination of the regression line. If large scatter exists it may be necessary to check the goodness of fit of the linear regression model. A F test should be considered to validate the linear hypothesis.

3.3. Bias Calculations

The Reference Standard for Comparison (RSC) is introduced to quantitatively evaluate the bias for each individual measurement. Conceptually, RSC can be derived through a weighted average of measurements from instruments on spatially and temporally co-located aircraft platforms. This can mathematically be expressed as

$$RSC(t,\overrightarrow{L}) = \sum_{i=1}^{n} m_i\,(t,\vec{L}) * \; w_i$$

where RSC is a function of a given time *t* and location $\vec{L}$, intercomparison data from the i^{th} instrument is given by m_i, there are n total number of instruments involved in a the field study, and the normalized weighting factor assigned to the i^{th} data set is w_i (i.e., $\sum_{i=1}^{n} w_i = 1$). The values of the w_i are determined by consensus of the TAbMEP measurement experts. It should be noted that the RSC is only used for measurement comparison purposes.

In practice, the RSC is difficult to find since each intercomparison was conducted between only two aircraft at a time. However, the field campaign measurement comparison strategy was designed to guarantee that each instrument could be related to any other instrument through paired intercomparisons. The approach prescribed below is an effective way to arrive at a reasonable approximation to the RSC.

Data from n instrument intercomparison flights will first be analyzed via pairwise orthogonal distance regressions (ODR) over intercomparison periods when pairs of planes were flown in wingtip-to-wingtip formations. The regressions will yield best-fit curves

$$m_i = a_{i,j} + \; b_{i,j} m_j \quad (1)$$

where m_i and m_j are data from instruments i and j, respectively. Once data from each instrument is related via best-fit line to at least one other instrument, the system of best-fit equations can be manipulated to express the data sets $\{m_i\colon\ 1 \le i \le n\}$ as linear functions of a single chosen data set m_1. The slopes and intercepts may be directly obtained from the regression or indirectly through algebraic manipulation of regression results. We note that the choice of instrument to serve as the independent variable m_1 will not affect the final RSC. Thus, m_1 should be chosen for convenience to correspond to the instrument with the highest number of direct intercomparisons against other instruments.

The RSC can easily be written as a linear function of the chosen independent variable, m_1 :

$$RSC = \; A_1 + B_1 m_1 \quad (2)$$

where $A_1 = \sum_{i=1}^{n} w_i a_{i,1}$, $B_1 = \sum_{i=1}^{n} w_i b_{i,1}$, $a_{1,1} = \; 0$, and $b_{1,1} = 1$.

Using the original set of regression equations the RSC can be expressed as a function of the data from any instrument (m_i). Thus, the Best Estimate Bias for the ith instrument can then be expressed in terms of m_i, i.e.,:

$$Best\ Estimate\ Bias_i = m_i - RSC$$

It is acknowledged that this approach provides a reasonable estimate of the average bias from the available intercomparison data, however, the accuracy of this estimate is limited, to a large extent, by the robustness of the regressions between the intercomparison data sets.

References

Boggs, P. T., et al. (1988), A computational examination of orthogonal distance regression, *J. of Econometrics*, *38*, pp 169-201.

Fehsenfeld, F. C., et al. (2006), International Consortium for Atmospheric Research on Transport and Transformation (ICARTT): North America to Europe—Overview of the 2004 summer field study, *J. Geophys. Res.*, *111*, D23S01, doi:10.1029/2006JD007829.

Hoell, J. M., et al. (1985), An intercomparison of carbon monoxide measurement techniques, *J. Geophys. Res., 90*, D7, 12881-12889.

Singh, H. B., et al. (2006), Overview of the summer 2004 Intercontinental Chemical Transport Experiment-North America (INTEX-A), *J. Geophys. Res.*, *111*, D24S01, doi:10.1029/2006JD007905.

Zwolak, J. W., et al. (2007), Algorithm 869: ODRPACK95: A weighted orthogonal distance regression code with bound constraints, *ACMTrans.Math. Softw.*, *33*, 4, Article 27 (August 2007), 12 pages. DOI = 10.1145/1268776.1268782 http://doi.acm.org/10.1145/1268776.1268782.