AIAA 2002-0842

Statistical Analysis of CFD Solutions from the Drag Prediction Workshop

M. Hemsch
NASA Langley Research Center
Hampton, VA

40th AIAA Aerospace Sciences Meeting & Exhibit
14-17 January 2002 / Reno, NV

For permission to copy or republish, contact the American Institute of Aeronautics and Astronautics
1801 Alexander Bell Drive, Suite 500, Reston, VA 20191

Statistical Analysis of CFD Solutions from the Drag Prediction Workshop

Michael J. Hemsch *
NASA Langley Research Center
Hampton, VA

Abstract

A simple, graphical framework is presented for robust statistical evaluation of results obtained from N-Version testing of a series of RANS CFD codes. The solutions were obtained by a variety of code developers and users for the June 2001 Drag Prediction Workshop sponsored by the AIAA Applied Aerodynamics Technical Committee. The aerodynamic configuration used for the computational tests is the DLR-F4 wing-body combination previously tested in several European wind tunnels and for which a previous N-Version test had been conducted. The statistical framework is used to evaluate code results for (1) a single cruise design point, (2) drag polars and (3) drag rise. The paper concludes with a discussion of the meaning of the results, especially with respect to predictability, Validation, and reporting of solutions.

Nomenclature

ANOM	Analysis of Means
AOA	angle of attack
C_D	drag coefficient
CD_PR	pressure drag coefficient
CD_SF	skin-friction drag coefficient
CD_TOT	total drag coefficient
CL	centerline
C_L	lift coefficient
C_m	pitching-moment coefficient
DPW	Drag Prediction Workshop
H	ANOM coverage factor
k	number of data groups
K	coverage factor for individual values
M_∞	Mach number
n	number of observations in a sample
pdf	population density function
RANS	Reynolds-averaged Navier-Stokes
Re	Reynolds number based on the mean aerodynamic chord
SSD	sample standard deviation
TM	turbulence model
x	value of an observation
$\overline{x}$	sample standard deviation
$\tilde{x}$	sample median
$\overline{\overline{x}}$	sample grand average
μ	population mean
$\hat{\mu}$	estimate of the population mean
σ	population standard deviation
$\hat{\sigma}$	estimate of the population standard deviation

Introduction

In June 2001, the AIAA Applied Aerodynamics Technical Committee (APATC) conducted a Drag Prediction Workshop (DPW) to determine the state of the art in the use of computational fluid dynamics (CFD), in an industrial setting, for transonic cruise drag predictions of subsonic transports. The workshop challenge was to compute the lift, drag and pitching moment for the DLR-F4 wing-body configuration[1, 2] for three sets of conditions (all at Re=3.0 x 10^6):

1. Cruise at $M_\infty = 0.75, C_L = 0.5$ (required)
2. Drag polar at $M_\infty = 0.75$ (required)
3. Drag rise at $C_L = 0.4, 0.5, 0.6$ (optional)

The DLR-F4 wing-body was chosen because of the availability of experimental data[1, 2] and because it had been used for a previous challenge.[3]

The focus of the workshop was drag prediction accuracy. For the cruise point, it was required that each submitter use one of the required grids generated prior to the workshop announcement. Submitters could use their own grids for additional

* Aerospace Engineer, Quality Assurance, Associate Fellow AIAA.

Copyright by the American Institute of Aeronautics and Astronautics, Inc. No copyright is asserted in the United States under Title 17, U.S. Code. The U.S. Government has a royalty-free license to exercise all rights under the copyright claimed herein for Government Purposes. All other rights are reserved by the copyright owner.

solutions at the cruise point and for the other sets of conditions. Additional details about the workshop may be found in References 4-9.† The DLR-F4 wing-body configuration is shown in Figure 1.

The paper is divided into five main sections. The first section presents typical customer requirements for experimental and computational simulations of performance and describes the available experimental results for the cruise point.[1, 2] In the second section, the purpose of the statistical analysis and N-Version testing is discussed and several statistical methods are presented for comparing the solutions from the various codes and users. The second section also discusses the robustness of the methods in the face of results that differ significantly from the majority. Results of the application of the statistical framework to the DPW N-Version tests are given in the third, fourth and fifth sections. The paper concludes with some final remarks.

Customer Requirements and Experimental Results

Customer requirements for wind tunnel reproducibility can vary significantly depending upon the intended use of the data, but there seems to be industry-wide consensus on the needs for performance testing of subsonic civil transports.[10, 11] Typical data quality goals are given in Table 1 for $\pm 2\sigma$ coverage.

For purposes of comparison with the above data quality goals and with the results of the DPW, typical corrected results from wind tunnel testing[1, 2] of the DLR-F4 wing-body configuration are presented in Table 2 for the primary DPW challenge condition of cruise at $C_L = 0.5, M_\infty = 0.75, \mathrm{Re} = 3.0x10^6$. The drag coefficient values in the data tables on the computer disks that are associated with Ref. 1 are rounded to 10 counts. For the purposes of the DPW, the round-off error was reduced considerably by use of an enhancement technique developed by Vassberg.[12] The results in Table 2 were obtained by linear interpolation in the lift and pitching moment curves and in the drag polars plotted as functions of the lift squared. Note that the scatter values of the wind tunnel drag and pitching moment results are considerably larger than the requirements listed in Table 1.

Multiplying the scatter in the angle of attack, for which $C_L = 0.5$, by the lift-curve slope (roughly 0.12 deg.$^{-1}$) gives the likely scatter of C_L if the angle of attack were held constant. That scatter value is ± 0.005 for $\pm 2\sigma$ coverage and is typically acceptable based on the requirements given in Table 1.

Statistical Approach for N-Version Testing

Background

For this paper, the point of view is taken that the scatter (dispersion) of the results computed by the participants for any given output coefficient at a given set of conditions represents the *reproducibility* of the computational process just as if the individual computed realizations were obtained from a replicated measurement process. In measurement, such a process is called N-th Order replication.[13] In computation, a similar process is called N-Version testing.[14-16] Reproducibility is defined[17] for measurement as *closeness of the agreement between the results of measurements of the same measurand carried out under changed conditions of measurement*. The changed conditions of "measurement" for the DPW are, of course, the different codes, solution methods, turbulence models, grids, computing platforms, observers (people who carried out the computational process) and so on.

For this type of analysis, no individual outcome (computational realization) is considered the "right" answer or "best" result. To be specific, the *collective* computational process is considered to consist of all of the individual processes used and the dispersion of the results to be noise in that collective computational process.** This viewpoint has been suggested by Youden[18, 19] of the former National Bureau of Standards for precision measurements of physical constants at different laboratories. It is also consistent with the frequentist[20] interpretation of probability and with the new international standard for reporting measurement uncertainty.[17] *For reporting purposes*, the new standard[17] rejects the categorization of errors into random and systematic in favor of two other classes: Those which are evaluated using conventional statistical methods (Type A) and those

† One of the references accessible at the website of Ref. 4 is the original talk on which this paper is based. Some of the results in that talk are slightly different in this paper because different estimators were used. The differences are not statistically significant.

** It should be noted that, for each turbulence model, a distinct true value would be expected in the limit.

which are not (Type B). With the above definition of a collective computational process, a Type A analysis is used to evaluate what would normally be thought of as systematic differences.

The above approach has several benefits:

1. Credible quantitative estimates can be made of the scatter and the mean of the virtual population of possible computational outcomes, including determining if some of the codes are not performing as well as the majority.
2. Quantitative predictions of the performance of each code can be made by using all of them together as a collective.
3. It can be determined if the scatter is small enough to be useful to designers.
4. Otherwise hidden effects and opportunities for process improvement can be revealed by "seeing into" the scatter.
5. Uncertainty measures for comparison of the *collective* CFD results to experiment can be obtained.

By doing this, together with continual improvement of the codes and processes, the possibility is created, across the design community, of making CFD predictions with credible, enduring, statements of reproducibility.

In this paper, three types of statistical graphical methods are used for analysis of the DPW N-Version testing results:

1. Running records of individual computational outcomes, together with centerlines and scatter bands.
2. Histograms of individual outcomes.
3. Analysis of Means (ANOM).

The first two methods are valuable because they display all of the data *in the presence of the scatter*.[21] One of the biggest concerns in statistical analysis is possibly drawing incorrect conclusions from aggregated data.[21-26] The advantage of running records and histograms is that they display *all* of the data and they do it in quite different ways, thus allowing the analyst to check conclusions drawn from one or the other display. In addition, several questions are asked of the first two methods:

1. Which solutions constitute a reasonable "core" set and what does that core set look like?
2. Which solutions lie outside that core?
3. What is the central location of the core solutions?
4. What is the dispersion of the core solutions?

The third method is useful because it allows discernment of possibly significant results *despite the presence of scatter*.[22]

In the rest of this section, these three graphical analysis methods are reviewed, prior to analyzing the results of the DPW challenge in the following three main sections.

Method 1 - Running Records of Individual Outcomes

In experimental measurement of repeatability and reproducibility, it is useful to display the data values on the vertical axis and the time of the data point acquisition on the horizontal axis.[24] However, since time is irrelevant in the DPW challenge, the individual solution values are plotted herein as a function of an integer index. The order of the plotting of the solution values along the abscissa was randomly assigned.

To enhance the possibility of discerning significant results in the running record, an estimate of the population mean ($\hat{\mu}$) of the plotted data points is made and shown on the graph as a *centerline*. In addition, *scatter limits* are placed about the centerline as follows:

$$\begin{aligned} \text{Lower Limit} &= \hat{\mu} - K\hat{\sigma} \\ \text{Upper Limit} &= \hat{\mu} + K\hat{\sigma} \end{aligned} \qquad (1)$$

where $\hat{\sigma}$ is an estimate of the population standard deviation and K is an appropriate coverage factor.[21-25] The area between the scatter limits is considered to be a "noise zone" within which it is not possible to discern significant results for the individual values - *at least not on the basis of statistics.*

To summarize, there are six steps for the preparation of individual value running records for graphical statistical analysis:

(1) Graphically display the outcomes, for supposedly identical conditions, in random order in a pseudo-time series.
(2) Estimate the population mean and standard deviation of the outcomes using a robust method.

(3) Display the estimate of the population mean as the *centerline* on the graph.
(4) Establish limits for the scatter of the individual outcomes based on the estimated population standard deviation and a reasonable coverage factor.
(5) Display the limits obtained in step 4 as lower and upper *bounds* on the graph.
(6) Look for significant results, i.e. outcomes that are outside the noise zone delineated by the scatter limits. Such outcomes could represent poor solutions relative to the others, significant differences in solution approaches, or opportunities for process improvement.

Robust estimation of the population parameters. The only two population parameters usually of interest in the estimation of uncertainty are the mean and the standard deviation.[24] Conventionally, these parameters are estimated using the sample average

$$\hat{\mu} = \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2)$$

and the sample standard deviation (SSD)

$$\hat{\sigma} = SSD \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (3)$$

Although, these estimators are the most efficient, they can give significantly offset estimates of the population mean and grossly inflated estimates of the population standard deviation if the sample is contaminated with values which do not belong to the population of interest.[21, 23, 27] For the purposes herein, the simplest way to achieve robustness is to use the sample median to estimate the population mean[23, 27]

$$\begin{aligned} \hat{\mu} &= \tilde{x} \\ \tilde{x} &\equiv x_{(n+1)/2} \quad (n\ odd) \\ &\equiv 0.5(x_{n/2} + x_{(n/2)+1}) \quad (n\ even) \end{aligned} \qquad (4)$$

and the sample median absolute deviation (MAD)[27,28] to estimate the population standard deviation

$$\hat{\sigma} = 1.483\sqrt{\frac{n}{n-1}}\,\text{median}\left(|x_i - \tilde{x}|\right) \qquad (5)$$

Establishing the limits. For this paper, limits for judging the significance of *individual* outcomes will be established at 100:1 odds, i.e. observations would not be found outside the limits more than one time in 100 *in the long run* by chance (99% coverage). Since there will never be enough observations for this type of application to reasonably determine the population probability density function (*pdf*), a distribution will have to be selected for use as a guide in determining the coverage factor and the Normal distribution is a convenient choice.

With the Normal distribution as a guide and with 99% coverage (100:1 odds) for infinite degrees of freedom, $K = 2.58$ for insertion into the scatter limits given by Eqs. 1.

Method 2 - Histograms of Individual Values

This statistical graphics method has been used to great effect for roughly 200 years.[26] The histogram presents a very different pattern of the data values to the eye and, hence, may reveal significant results that would otherwise be obscured in tables or a running record.[21, 26] Detailed discussions on the construction and use of histograms are given in References 23 and 25. The following descriptive comments are taken directly from Reference 23:

The most common form of the histogram is obtained by splitting the range of the data into equal-sized bins (called classes). Then, for each bin, the number of points from the data set that fall into each bin are counted. That is

Vertical axis: Frequency (i.e., counts for each bin)
Horizontal axis: Response variable

The histogram can be used to answer the following questions [qualitatively]:

1. *What kind of population distribution do the data come from?*
2. *Where are the data located?*
3. *How spread out are the data?*
4. *Are the data symmetric or skewed?*
5. *Are there outliers in the data?*

For the analysis of computational (or experimental) outcomes, There are rarely enough observations to determine skewness or the form of the population distribution itself. Of course, an attempt must be made to determine possible outliers and multiple modes. For single-event data sets such as the DPW results, outliers suggest mistakes or solution methods significantly different from the core group while multiple modes may suggest significant effects buried in the scatter.[‡] The methods of the next two subsections are designed to address the latter.

[‡] For data obtained for an ongoing process, it would be more appropriate to interpret "outliers", i.e. points outside the limits, as indications of changes in the process.

Method 3 - Analysis of Means (ANOM)

The Analysis of Means (ANOM)[22, 29] is a graphical statistical technique used for comparison of the averages of sets of supposedly similar results in the presence of noise. For comparisons of just two averages, it is identical, in numerical result, to a hypothesis test on means.[30] The ANOM method is more general, however, since it can be used for comparison of more than two averages and, of course, it is graphical.

The ANOM decision limits are given by Wheeler[22, 29] as

$$\begin{aligned} \text{Lower Limit: } & \bar{\bar{x}} - H\hat{\sigma}/\sqrt{n} \\ \text{Upper Limit: } & \bar{\bar{x}} + H\hat{\sigma}/\sqrt{n} \end{aligned} \qquad (6)$$

where $\bar{\bar{x}}$ is the grand average of the sample sets and H is a function of the coverage desired, the degrees of freedom for the estimate of the population standard deviation, and the number of sets of averages being compared, k. Note that, if the number of observations in the sets of sample averages are different, the limits will be different because of the $\sqrt{n}$ in the denominator of Eqs. 6. Note also that, once again, it is necessary to choose a *pdf* for a guide and again the Normal distribution is selected.

As with the interpretation for a running record of individual values, the distance between the ANOM decision limits represents the noise of the process within which the averages cannot be distinguished - *at least not statistically*. For the ANOM analyses in this paper and for the analysis of sample standard deviations described in the next subsection, an exploratory level of coverage - 90% - will be used because the intention is to look for possible opportunities for improvement. Replicating the process will determine if the effects suggested by the statistical techniques are real.[29]

Cruise Point at $C_L = 0.5, M_\infty = 0.75$

For the cruise point condition, the solution statistics are as follows[§]:

- 14 codes were used.
  - 7 structured-grid
  - 6 unstructured-grid
  - 1 Cartesian-grid (Euler + integral boundary layer)
- 35 solutions were computed
  - 24 structured
  - 10 unstructured
  - 1 Cartesian
- 3 types of turbulence models were used
  - 17 Spalart-Allmaras
    - 8 structured-grid
    - 9 unstructured-grid
  - 17 two-equation
    - 16 structured-grid
    - 1 unstructured-grid
  - 1 integral boundary layer
- 21 solutions used the provided grids
- 14 used other grids

This breakdown will be used to attempt to determine possible sources of variation in the solutions.

Analysis of all drag results

Total drag. The cruise point solutions for the total drag are shown in the running record and histogram of Figure 2 plotted at random, together with the centerline and limits obtained as described above (median and MAD) and with the experimental results from Table 2. The number shown centered just below each bin is the *upper* value for that bin interval.

The solutions whose values are very low or very high relative to the core group are identified in the histogram by their running record solution index for easy comparison. For both the running record and the histogram, it seems clear that there is a central core with a wide spread on the order of 50-80 counts, together with five solutions that are definitely outside that core. The core population parameters as estimated by the median and MAD for all of the solutions are given in Table 3.

Pressure and skin-friction drag. The cruise point solutions for the pressure and skin-friction drag are shown similarly in Figures 3 and 4 plotted with the same index as Figure 2. Note that the histograms of Figures 3 and 4 show that it is not as easy to decide which drag component solutions are in the core group or outside it when they lie close to the limits. The core population parameters as estimated by the median and MAD for all of the solutions are given in Table 3.

The breakdown on the solutions/codes that had the total drag or one of its components outside the limits is as follows:

- 7 out of 35 solutions (20%)
- 6 out of 14 codes (43%)

[§] For this paper, only the solutions provided to the DPW committee before the workshop were used.

- 4 out of the 21 solutions on the provided grids (19%)
- 3 out of the 14 solutions on other grids (21%)

Overall, there appears to be no significant difference in either location or scale between the drag solutions carried out on the provided grids and solutions carried out on grids developed or improved by the participants. However, discussion during the workshop revealed that the provided grids were not well-suited to some of the codes.

Analysis of all AOA results

The cruise point solutions for the AOA values are shown in Figure 5. The breakdown of solutions/codes that had values outside the limits is as follows:

- 7 out of 35 solutions (20%)
- 7 out of 14 codes (50%)
- 4 out of the 21 solutions on the provided grids (19%)
- 3 out of the 14 solutions on other grids (21%)

A summary of the location and scale values for the cruise point AOA solutions is given in Table 4. Again, there does not seem to be any significant effect of provided versus participant-generated grids on the medians of the two types of solutions.

Note that the difference between the estimated computed AOA location and the experimental location is 0.44 degrees which corresponds to an over-prediction of the lift coefficient of about 10%. The estimated standard deviation of 0.13 degrees corresponds to an estimated standard deviation for the lift coefficient of roughly 0.02.

Analysis of all pitching-moment results

The cruise point solutions for the pitching-moment values are shown in Figure 6. The breakdown of solutions/codes that had pitching-moment values outside the limits is as follows:

- 8 out of 35 solutions (23%)
- 7 out of 14 codes (50%)
- 4 out of the 21 solutions on the provided grids (19%)
- 4 out of the 14 solutions on other grids (29%)

A summary of the location and scale values for the cruise point pitching-moment solutions is given in Table 4. As with the drag and AOA results, there does not seem to be any significant effect of provided versus participant-generated grids on the medians of the two types of solutions.

Note that the difference between the estimated computed location and the experimental location is 0.03 which corresponds to an over-prediction of the static margin of roughly 6% of the mean aerodynamic chord.

Effect of turbulence model and grid type

If the five total-drag outlier solutions, the Cartesian integral-boundary-layer solution and the two-equation unstructured grid solution are deleted, 28 cruise-point solutions are left that break down as follows:

1. 7 solutions using the Spalart-Allmaras turbulence model on unstructured grids (3 codes with 4 observers)
2. 7 solutions using the Spalart-Allmaras model on structured grids (2 codes with 4 observers)
3. 14 solutions using a two-equation model on structured grids (6 models and 5 codes with 7 observers)

To compare these solutions and attempt to determine if there are significant effects due to turbulence model and grid type, the ANOM method described above will be used. First, the sample averages (Eq. (2)) and standard deviations (Eq. (3)) for each of the three distinct sets of values are calculated.[¶] Second, those values are pooled to obtain $\bar{\bar{x}}$ and $\hat{\sigma}$. The results are given in Table 5 for the total drag, the two drag components, the angle of attack at $C_L = 0.5$, and the pitching moment.

To complete the ANOM statistical graph, it is noted that there are three sets of averages to compare for each type of outcome and 25 degrees of freedom (28 observations divided into 3 sets). Furthermore, it is desired to compare the averages with a coverage of 90%. This gives $H = 1.76$.[22, 29] The ANOM results are shown in Figures 7-11, together with the corresponding running records (at 99% coverage). The randomly assigned index for the running records is different from the index previously used in Figures 2-6.

[¶] The robust, but less efficient, estimators of the previous analyses are not needed because the major outliers have been deleted.

Effect of turbulence model on drag. Judging from both the running record and ANOM charts of Figure 7, there does not seem to be a significant effect of the two types of turbulence model (groups 2 and 3) on the total drag. However, the pressure and skin-friction component comparisons, Figures 8 and 9, show a different story. The results from the two model types seem to be (barely) significantly offset by about 10-12 drag counts but with opposite sign. This result suggests that it would be worthwhile to investigate further.

Effect of grid type on drag. Judging from both the running record and ANOM charts of Figure 7, there does seem to be a significant effect on the total drag of the two grid types (groups 1 and 2), using the Spalart-Allmaras turbulence model. Comparing the drag component results in Figures 8 and 9, it is seen that the effect is due to a significant difference in the skin friction drag of 15 counts.

Effect of turbulence model and grid type on AOA and C_m. Considering both the running records and the ANOM charts of Figures 10 and 11, there does seem to be a significant effect of turbulence model (groups 2 and 3), but not grid-type (groups 1 and 2), on the angle of attack and pitching moment, at least for the grids used in the present study.

Drag Polar for $M_\infty = 0.75$

The required drag polars were fit in the linear range ($C_L = 0.15 - 0.40$) as follows:

$$\begin{aligned} C_L &= C_{L_0} + C_{L_\alpha}\alpha \\ C_D &= C_D@C_{L_0} + kC_L^2 \end{aligned} \qquad (7)$$

The available experimental results[1, 2] were fit as well. The running records and histograms of the fit parameters, C_{L_0}, C_{L_α}, $C_D@C_{L_0}$, k, for all of the available solutions are given in Figures 12-15. The estimated values of the population means and standard deviations are given in Table 6, together with the estimates from the three wind tunnel tests. The running records and histograms appear to be quite similar to those displayed earlier for the cruise point condition. (Note that the solution index is different from the two previously used.)

The estimated standard deviations for the computed (and experimental) fit intercepts are not surprising given the results discussed above for the cruise point condition. However, the estimates of the computational and experimental population standard deviations for the fit slopes, C_{L_α}, k, seem to be considerable larger than would be expected from conventional wisdom.

The breakdown on the fit parameters that are outside the limits given by the median and MAD for the 27 polars is as follows:

- For C_{L_0}, four (15%)
- For C_{L_α}, two (7%)
- For $C_D@C_{L_0}$, four (15%)
- For k, two (7%)

Drag Rise at $C_L = 0.4, 0.5, 0.6$

Eight drag rise solutions for all three lift-coefficient values were provided to the workshop committee: seven observers using five codes, two turbulence models (Spalart-Allmaras and Wilcox $k-\varpi$) and both grid types. The solutions, together with experimental data inferred from the drag polar fits, are shown in Figure 16. Note that one of the solution drag-rise curves is considerably outside the scatter of the other solutions. Also note that the total drag appears to be under-predicted for the highest M_∞, C_L combination (0.8, 0.6), but otherwise seems to scatter reasonably about the experimental data.

For the type of flows encountered in this challenge, it seems reasonable to ask if there is any effect of the Mach number and lift coefficient on the scatter. To that end, the SSD's of the solutions at each M_∞, C_L combination (not including the "structured Wilcox 2" solution) are shown in Figure 17.

There are several points worth discussing:

1. The computed SSD's lie within a band of 3 counts except for the two highest Mach numbers for $C_L = 0.6$.
2. The estimated population standard deviation for the drag rise points is about half that of the estimates obtained for the cruise point and drag polar analyses.
3. The estimated computational and experimental population standard deviations differ by roughly a factor of two.

The first observation suggests the possibility that some additional source of variation manifests itself

for the highest M_∞, C_L combinations. Perhaps the shock has become strong enough to create significant flow separation. This possibility could be checked by comparing the solutions for the lower and higher SSD's. The second observation is probably the result of fewer opportunities for variation to be expressed. Seven drag rise solutions were used for this analysis versus 35 and 27 for the cruise point and drag polar analyses. The third observation suggests that more work is needed to reduce the computational scatter before validation is possible to the accuracy of the experimental data.

Final Remarks

Statistical method. One aim of the analysis in this paper has been to determine if the replicated outcomes appear to have been drawn at random from a single (virtual) population. If so, it is meaningful to talk about the parameters of the population of the collective, including the mean and the standard deviation. It is also then meaningful to talk about quantifiable predictability. But, it should be remembered that the analysis herein was for a single data set. Hence, any conclusions must be considered tentative, i.e. exploratory. Confirmatory results can only be obtained by determining that the CFD results for the population parameters are *predictable*, i.e. stable, and that requires established processes and repeating the processes over time.

One can check on the stability of any process, including CFD, by making the running record of individual values into a true time series of repeated outcomes and tracking it as a process behavior chart.[21] With these caveats, it does seem that using the median to estimate the location and MAD to estimate the scale did allow discernment of outlier solutions in the running records and histograms without losing the meaning of the "core" solutions. And there do seem to be credible values of the CFD population parameters of interest for the outcomes studied. Again, whether these values are durable can only be seen by repeating this exercise.

The analysis also suggests that some set of best practices and quantitative sanity checks is needed to avoid outliers, especially those that might show up when the luxury of multiple solutions from diverse codes and models - or even grid convergence - is not available. The continued existence of such outliers would force acceptance of much bigger numbers for the scatter. For example, for $\hat{\sigma}$ for the cruise point condition, the changes would be

- CD_TOT: 51 counts versus 21 counts
- AOA: 0.37 deg versus 0.13
- CM: 0.040 versus 0.008

Unfortunately, even the scatter of the core solutions is unacceptable compared to the oft-stated requirements of the user community[10, 11] and for reasonable validation as well.

Validation. Although this paper is devoted to estimating the quality of the DPW CFD solutions as a collective, the readers may be tempted to note that the agreement of the estimated computational and experimental population means of the drag coefficient (0.0293 vs. 0.0286) is not too excessive (only 2% difference). However, consider that comparison in the light of the scatter levels. Following the suggestion of Coleman and Stern[31], define

$$\Delta \equiv \hat{\mu}_{CFD} - \hat{\mu}_{EXP} \tag{8}$$

An error propagation analysis[31, 32] for Eq. (8) gives

$$\hat{\sigma}_\Delta = \sqrt{\hat{\sigma}^2_{\hat{\mu}_{CFD}} + \hat{\sigma}^2_{\hat{\mu}_{EXP}}} \tag{9}$$

From Table 3 and excluding the outliers, the following values for the terms under the radical in Eq. (9) are obtained:

$$\begin{aligned} \hat{\sigma}_{\hat{\mu}_{CFD}} &= 0.0021/\sqrt{n_{CFD}} = 0.0021/\sqrt{30} \\ &= 0.00038 \\ \hat{\sigma}_{\hat{\mu}_{EXP}} &= 0.0004/\sqrt{n_{EXP}} = 0.0004/\sqrt{3} \\ &= 0.00023 \end{aligned} \tag{10}$$

Substituting the above results into Eq. (9) gives

$$\hat{\sigma}_\Delta = 0.00044 \tag{11}$$

Hence, the $\pm 2\sigma$ confidence interval for the difference in the population means for the cruise point condition can be written as

$$\begin{aligned} CI &= 0.0293 - 0.0286 \pm 0.00088 \\ &= 0.0007 \pm 0.0009 \end{aligned} \tag{12}$$

Thus, while it is true that the confidence interval, Eq. (12), does include zero, it must be stated that the validation is only at a level of ± 9 counts.[31]

Reporting of CFD results. Given the results of the DPW and the analysis in this paper, the recommendations in Ref. 25 for the presentation of quality control data seem to the author to apply to CFD outcomes as well:

- *Present as a minimum, the average, the standard deviation, and the number of observations. Always state the number of observations.*

- *Present as much evidence as possible that the data were obtained under controlled conditions.* [See Refs. 21-25.]
- *Present relevant information on precisely (a) the field within which the measurements are believed valid*[25] *and (b) the conditions under which they were made.*

Presenting CFD results in such a way, of course, requires obtaining results, in a collective sense, for diverse codes, grids, turbulence models and observers. It is also very likely the best way to determine the best practices needed to reduce the present levels of scatter to acceptable levels.

Acknowledgements

The author is grateful to Dr. Richard A. Wahls and the other members of the Drag Prediction Workshop Committee for suggesting this analysis and to Drs. Ollie J. Rose and James M. Luckring and Mr. Eric L. Walker for many helpful suggestions.

References

1. Redeker, G., "DLR-F4 Wing Body Configuration", in Chapter B of *A Selection of Experimental Test Cases for the Validation of CFD Codes*, AGARD-AR-303 Vol. II, August 1994.
2. Redeker, G., Muller, R., Ashill, P. R., Elsenaar, A., and Schmitt, V., "Experiments on the DLR-F4 Wing Body Configuration in Several European Wind Tunnels", in Chapter 2 of *Aerodynamic Data Accuracy and Quality: Requirements and Capabilities in Wind Tunnel Testing*, AGARD-CP-429, July 1988.
3. Elsholz, E., "The DLR-F4 Wing/Body Configuration", in *ECARP - European Computational Aerodynamics Research Project: Validation of turbulence Models*, Notes on Numerical Fluid Mechanics, Vol. 58, 1997, p.429-450.
4. Anon., Presentations made at the June 2001 AIAA Drag Prediction Workshop, http://www.aiaa.org/tc/apa/dragpredworkshop/Final_Schedule_and_Results.html, Accessed November 27, 2001.
5. Rakowitz, M., Schwamborn, D., Sutcliffe, M., Eisfeld, B., Bleeke, H., and Fassbender, J., "Structured and Unstructured Computations on the DLR-F4 Wing-Body Configuration", AIAA-2002-0837, January 2002.
6. Mavriplis, Dimitri, and Levy, David W., "Transonic Drag Prediction Using an Unstructured Multigrid Solver", AIAA-2002-0838, January 2002.
7. Pirzadeh, Shahyar Z., "Assessment of Unstructured-Grid Software TetrUSS for Drag Prediction on DLR-F4 Configuration", AIAA-2002-0839, January 2002.
8. Vassberg, John, Buning, Pieter G., and Rumsey, Christopher L., "Drag Prediction for the DLR-F4 Wing/Body Using OVERFLOW and CFL-3D on an Over-set Mesh", AIAA-2002-0840, January 2002.
9. Levy, David W., "Summary of Data from the AIAA CFD Drag Prediction Workshop", AIAA-2002-0841, January 2002.
10. Steinle, F., and Stanewsky, E., *Wind Tunnel Flow Quality and Data Accuracy Requirements*, AGARD-AR-184, November 1982.
11. Carter, E. C., and Pallister, K. C., "Development of Testing Techniques in a Large Transonic Wind Tunnel to Achieve a Required Drag Accuracy and Flow Standards for Modern Civil Transports", Chapter 11 in *Aerodynamic Data Accuracy and Quality: Requirements and Capabilities in Wind Tunnel Testing*, AGARD-CP-429, July 1988.
12. Vassberg, John, "Enhancement of AGARD-AR-303 Data", http://www.aiaa.org/tc/apa/dragpredworkshop/Final_Schedule_and_Results.html, Accessed November 27, 2001. Presented at June 2001 AIAA Drag Prediction Workshop, Anaheim CA.
13. Moffat, R. J., "Contributions to the Theory of Single-Sample Uncertainty Analysis", Journal of Fluids Engineering, Vol. 104, June 1982, p. 250-260.
14. Hatton, Les, "The T Experiments: Errors in Scientific Software", IEEE Computational Science and Engineering, Vol. 4, No. 2, April-June 1997, p. 27-38.
15. Hatton, Les, and Roberts, Andy, "How Accurate is Scientific Software?", IEEE Transactions on Software Engineering, Vol. 20, No. 10, October 1994, p. 785-797.
16. Knight, John C., and Leveson, Nancy G., "An Experimental Evaluation of the Assumption of Independence in Multiversion Programming", IEEE Transactions on Software Engineering, Vol. 12, No. 1, January 1986, p. 96-109.
17. Anon., *U.S. Guide to the Expression of Uncertainty in Measurement*, ANSI/NCSL Z540-2-1997, October 1997.

18. Youden, W. J., "Enduring Values", Technometrics, Vol. 14, No.1, February 1972, p. 1-11.
19. Youden, W. J., "Systematic Errors in Physical Constants", Physics Today, September 1961, p. 32-43.
20. von Mises, Richard, *Probability, Statistics and Truth*, Dover 1981.
21. Wheeler, Donald J., and Chambers, David S., *Understanding Statistical Process Control*, 2nd Ed., SPC Press, 1992.
22. Wheeler, Donald J., *Advanced Topics in Statistical Process Control: The Power of Shewhart's Charts*, SPC Press, 1995.
23. Croarkin, Carroll, and Tobias, Paul (Eds.), *NIST/Sematech Engineering Statistics Handbook.* (online) http://www.itl.nist.gov/div898/handbook/. Accessed September 28, 2001.
24. Mandel, John, *The Statistical Analysis of Experimental Data*, Dover 1984.
25. Anon., *Manual on Presentation of Data and Control Chart Analysis: 6th Edition*, ASTM Manual Series: MNL 7, 1995.
26. Tufte, Edward R., *The Visual Display of Quantitative Information*, 2nd Ed., Graphics Press, 2001.
27. Mosteller, Frederick, and Tukey, John W., *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, 1977.
28. Muller, Jorg W., "Possible Advantages of a Robust Evaluation of Comparisons", J. Res. National Institute of Standards and Technology, Vol. 105, No. 4, July-August 2000, p. 551-555.
29. Wheeler, Donald J., *Understanding Industrial Experimentation*, 2nd Ed., SPC Press, 1990.
30. Montgomery, Douglas C., *Introduction to Statistical Quality Control*, 3rd Ed., Wiley, 1996.
31. Coleman, H. W., and Stern, F., "Uncertainties and CFD Code Validation", J. Fluids Engineering, Vol. 119, December 1997, p. 795-803.
32. Coleman, H. W., and Steele, W. G., *Experimentation and Uncertainty Analysis for Engineers*, Wiley 1989.

Coefficient	Increments	Absolute
C_L	0.005	0.01
C_D	0.00005 (1/2 count)	0.0001 (1 count)
C_m	0.0005	0.001

Table 1. Typical (design) customer uncertainty goals ($\pm 2\sigma$) for performance simulations.[10, 11]

Source	Variable	NLR	ONERA	DRA	Sample Mean	Sample Median	Sample Range	Use
Original data	Alpha, deg	0.153	0.192	0.181	0.175	0.181	0.039	0.18±0.04
Enhanced	C_D, counts	288	289	281	286	288	8	286±8
Original data	C_m	-.130	-.126	-.137	-.131	-.130	.011	-0.13±0.01

Table 2. Experimental results from three tunnels for $C_L = 0.5$, $M_\infty = 0.75$.[1, 2]
(The uncertainty is estimated at a coverage of $\pm 2\sigma$.)

	CD_PR	CD_SF	CD_TOT	Experiment
Estimate of the population mean	0.0166	0.0134	0.0293	0.0286
Estimate of the population standard deviation	0.0014	0.0015	0.0021	0.0004

Table 3. Breakdown of location and scale for the cruise point drag solutions. The median and MAD are used to estimate the population parameters for the computations.

	AOA, deg (CFD)	AOA, deg (Exp.)	C_m (CFD)	C_m (Exp.)
Estimate of the population mean	-0.26	0.18	-0.160	-0.13
Estimate of the population standard deviation	0.13	0.02	0.0084	0.005

Table 4. Breakdown of location and scale for the cruise point AOA and pitching-moment solutions. The median and MAD are used to estimate the population parameters for the computations.

Set	TM	Grid	CD_TOT		CD_PR		CD_SF		AOA, deg		Cm	
			$\bar{x}$	S	$\bar{x}$	S	$\bar{x}$	S	$\bar{x}$	S	$\bar{x}$	S
1	SA	Unstr.	0.0284	0.00061	0.0162	0.00053	0.0123	0.00048	-0.33	0.072	-0.161	0.0058
2	SA	Str.	0.0299	0.00108	0.0162	0.00090	0.0138	0.00024	-0.28	0.051	-0.163	0.0020
3	2E	Str.	0.0298	0.00201	0.0172	0.00141	0.0126	0.00144	-0.14	0.196	-0.153	0.0111
Pooled value			0.0295	0.00157	0.0167	0.00114	0.0128	0.00107	-0.22	0.148	-0.158	0.0086

Table 5. Values used for ANOM analysis of cruise point data.

Parameter	CFD	EXP	CFD	EXP
	$\hat{\mu}$		$\hat{\sigma}$	
C_{L_0}	0.531	0.473	0.021	0.003
C_{L_α}	0.120	0.114	0.0026	0.0019
$C_D @ C_L = 0$	0.0198	0.0188	0.0016	0.0005
k	0.0365	0.0374	0.0018	0.0040

Table 6. Results from fits to computational and experimental drag polars. (The experimental values were obtained using the sample averages and standard deviations.)

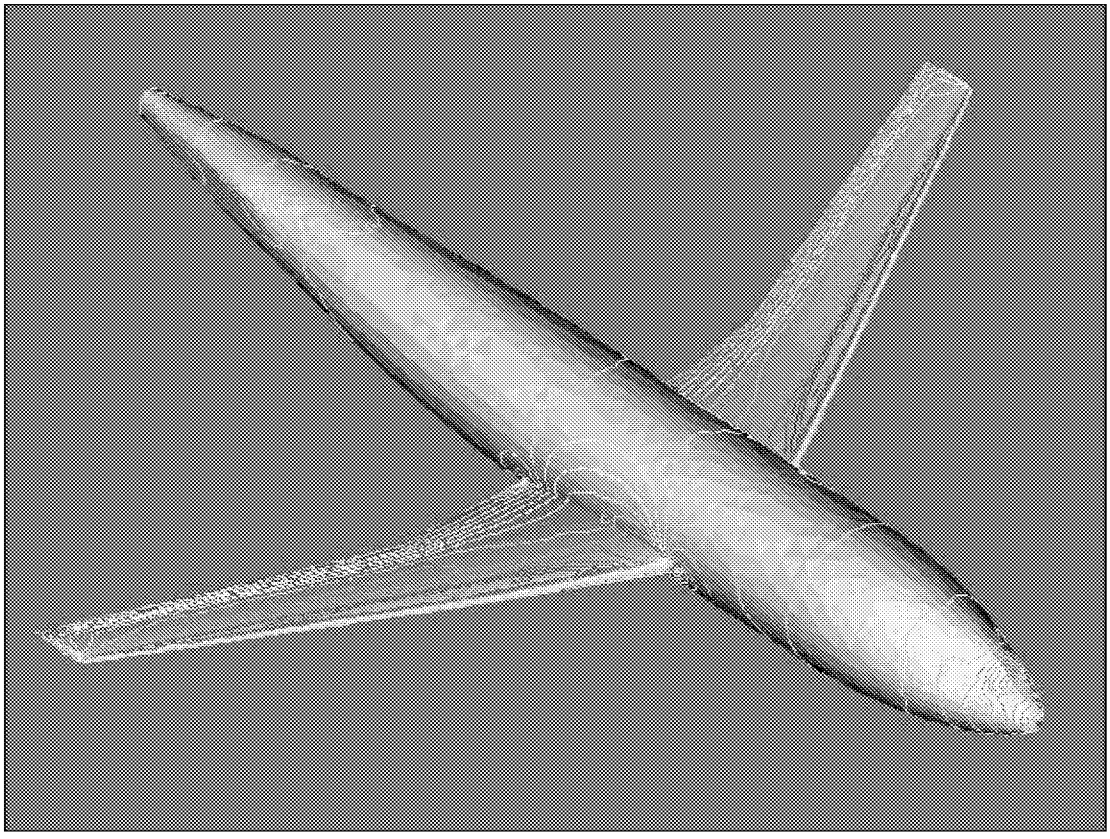
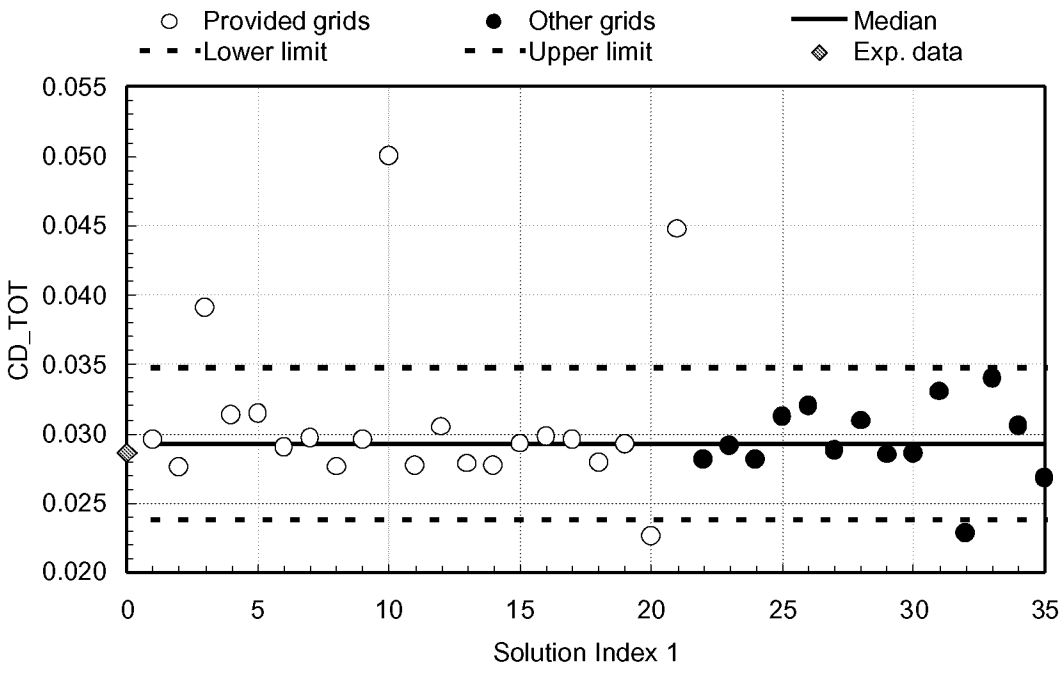
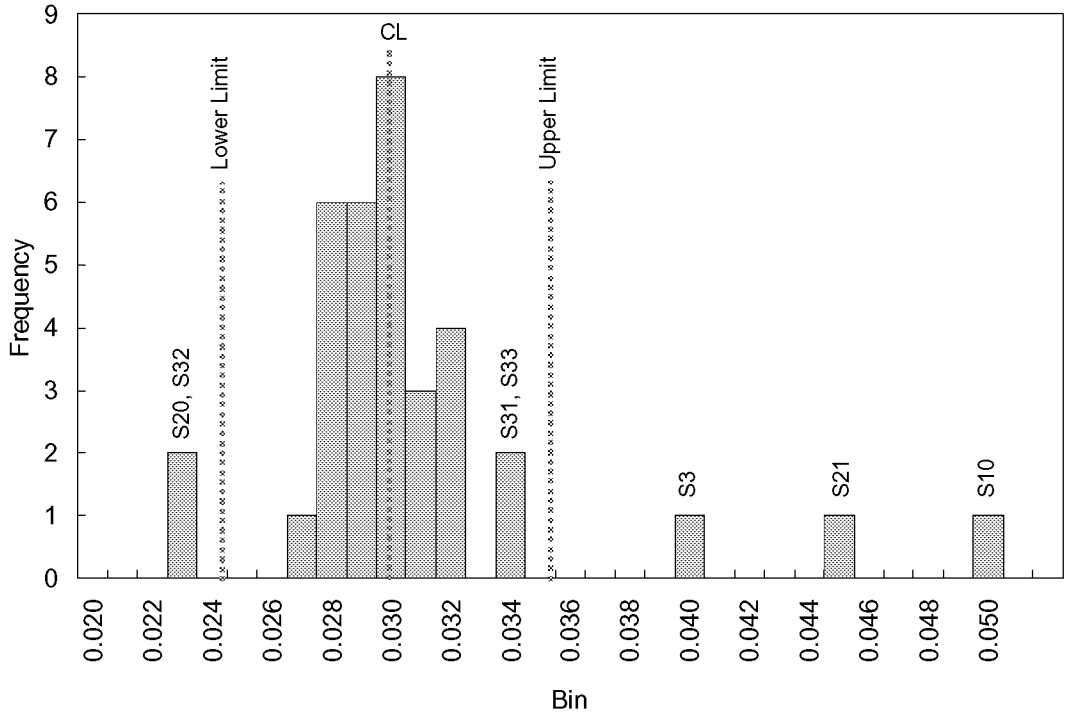


Figure 1. - Geometry of DLR-F4 model used for the N-Version testing.

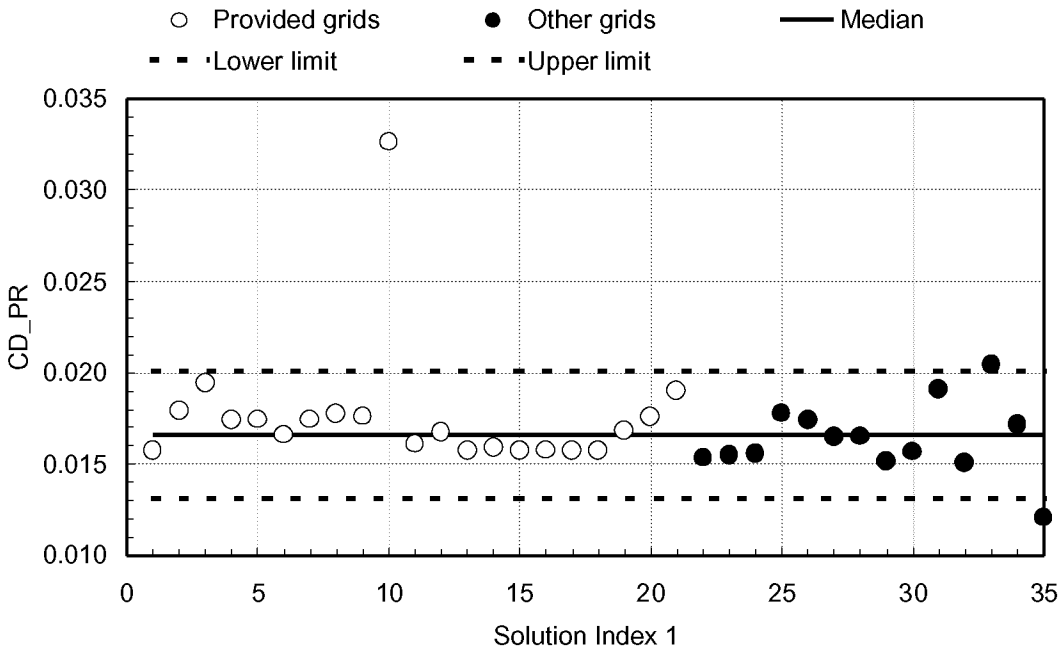


(a) Running record.

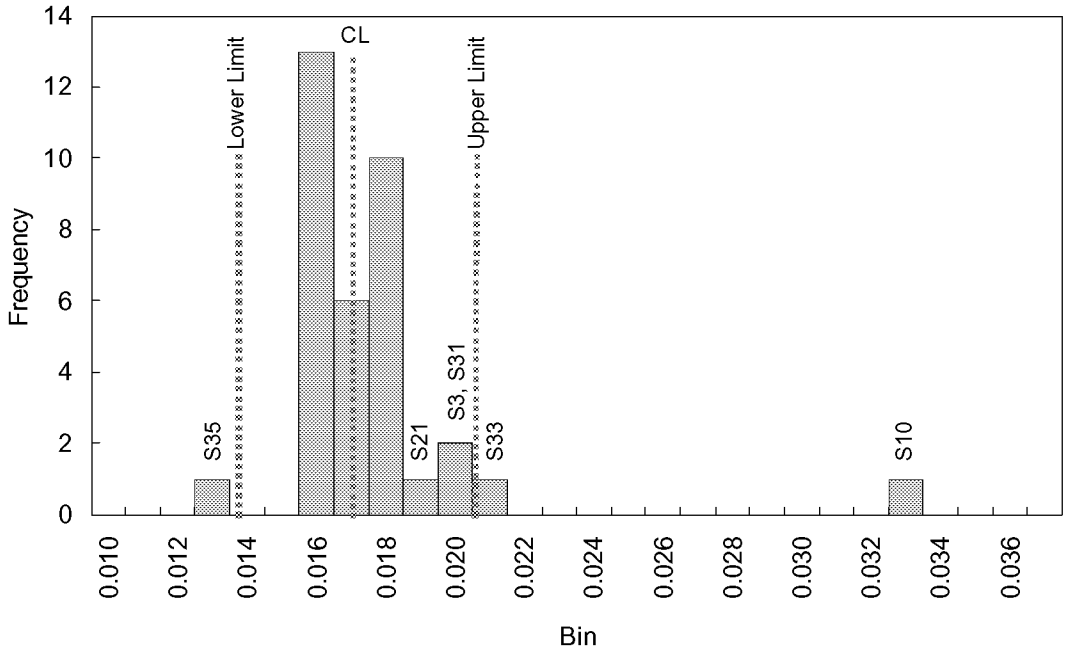


(b) Histogram (Each bin covers 10 drag counts.)

Figure 2. All total drag solutions at $C_L = 0.5$, $M_\infty = 0.75$. Median + MAD + 100:1 limits.

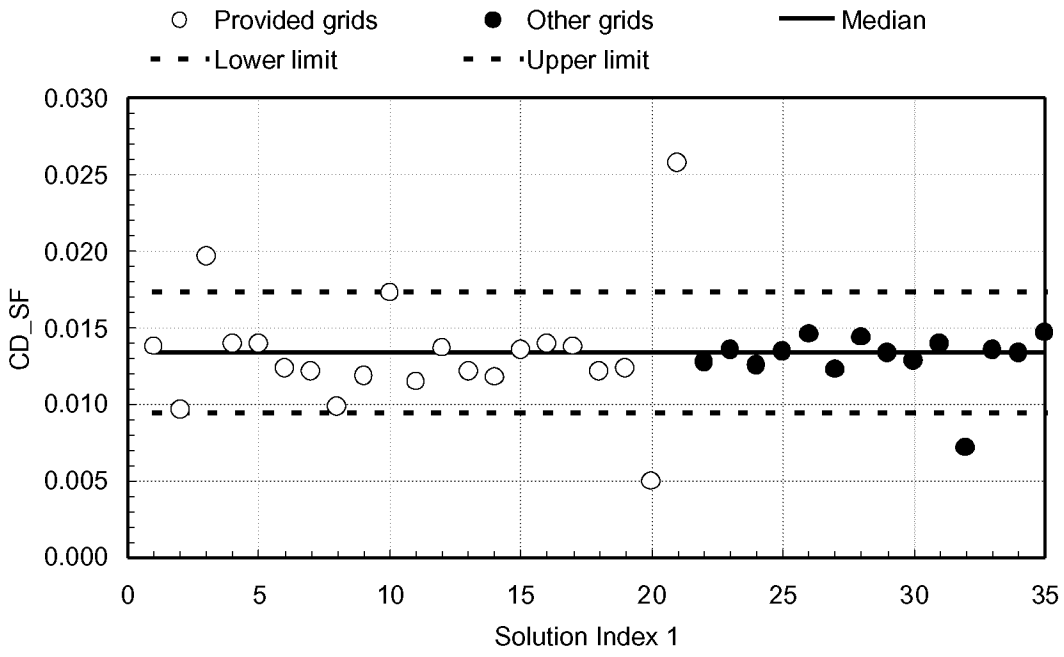


(a) Running record.

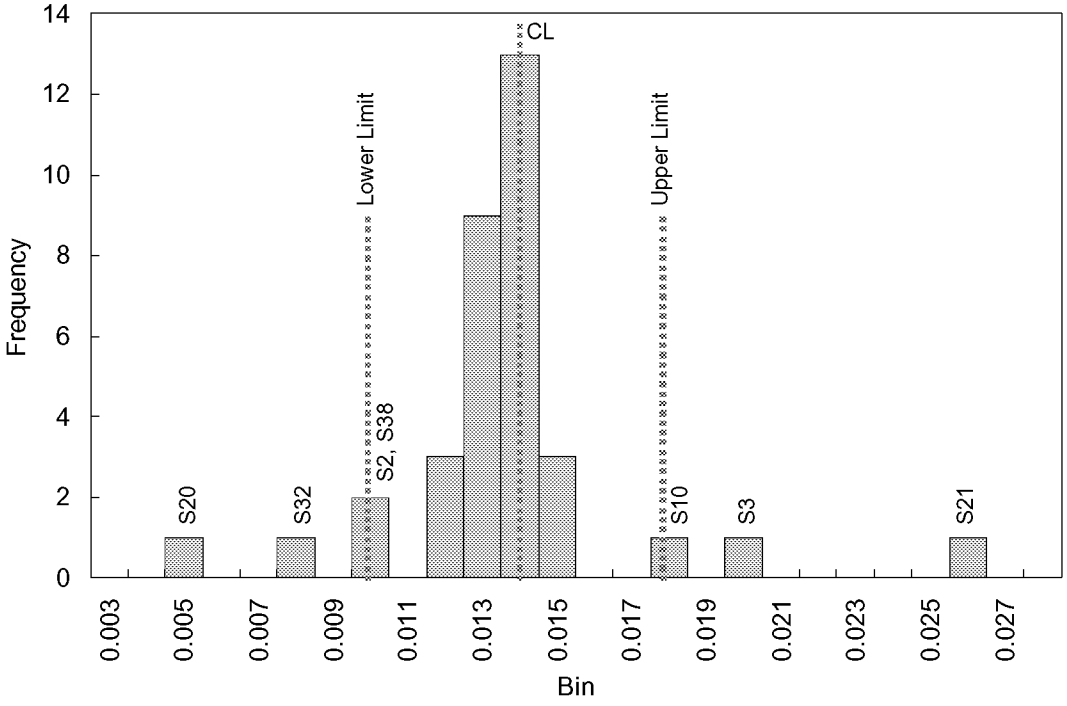


(b) Histogram (Each bin covers 10 drag counts.)

Figure 3. All pressure drag solutions at $C_L = 0.5$, $M_\infty = 0.75$. Median + MAD + 100:1 limits.

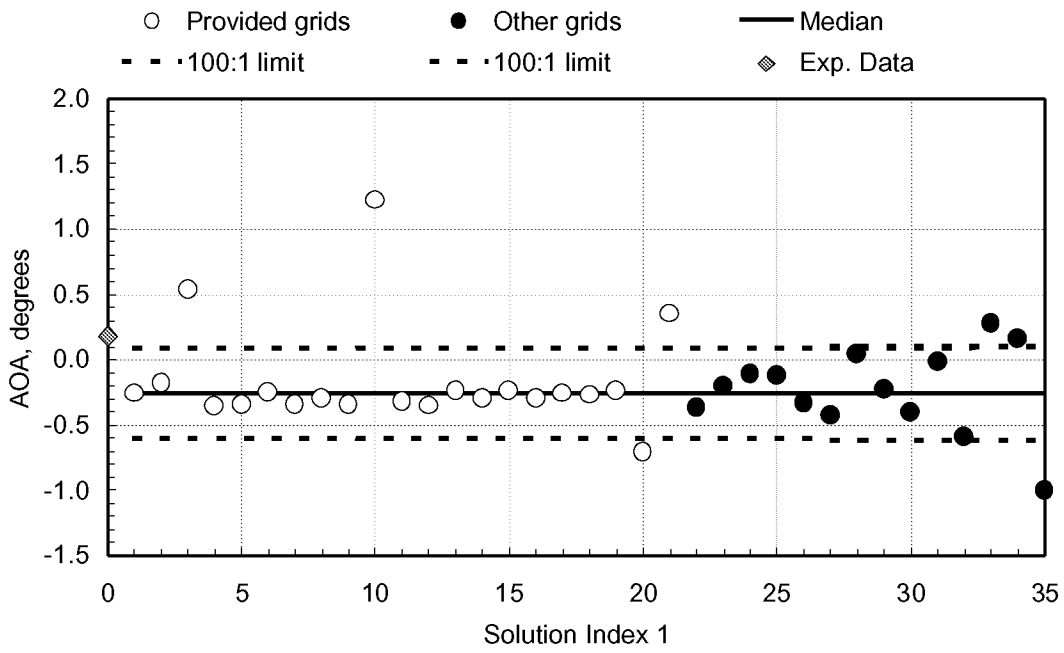


(a) Running record.

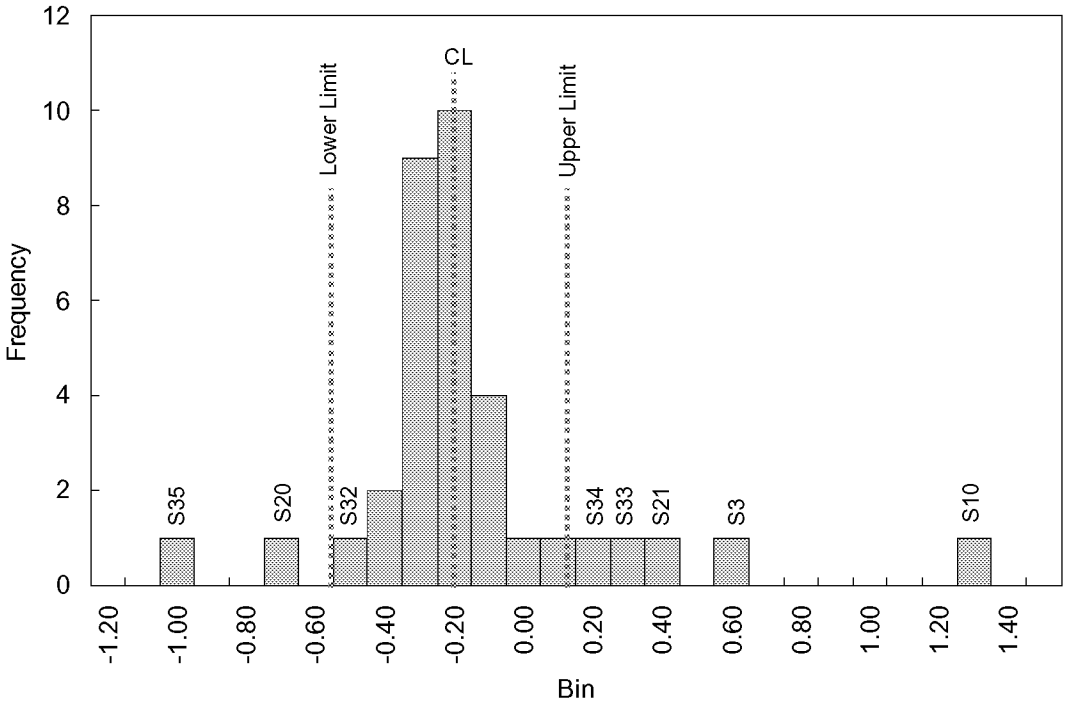


(b) Histogram (Each bin covers 10 drag counts.)

Figure 4. All skin-friction drag solutions at $C_L = 0.5$, $M_\infty = 0.75$. Median + MAD + 100:1 limits.

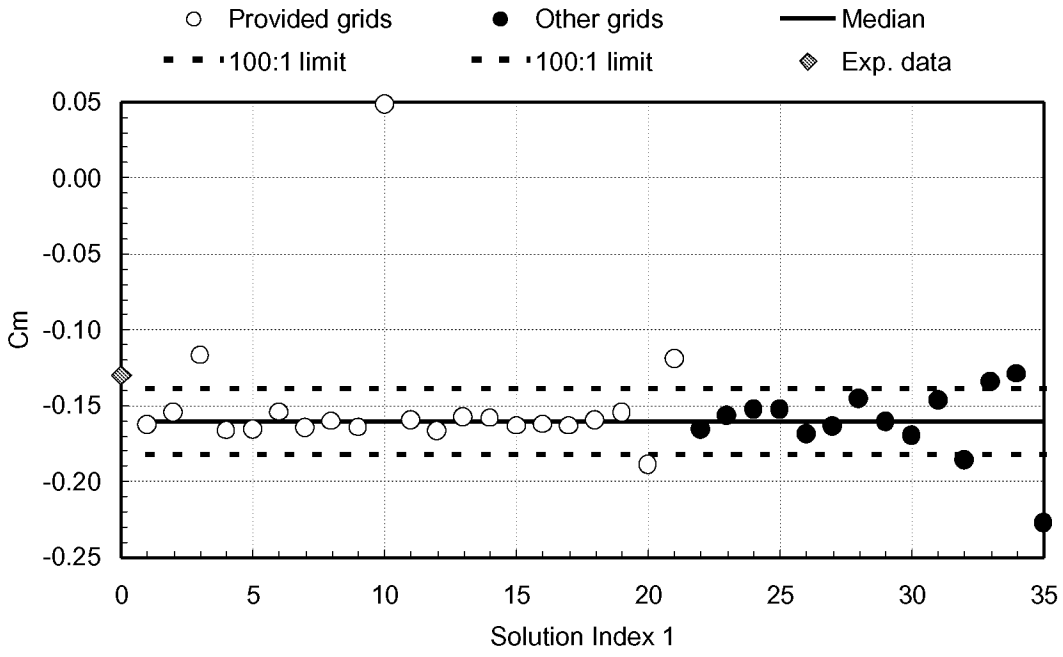


(a) Running record.

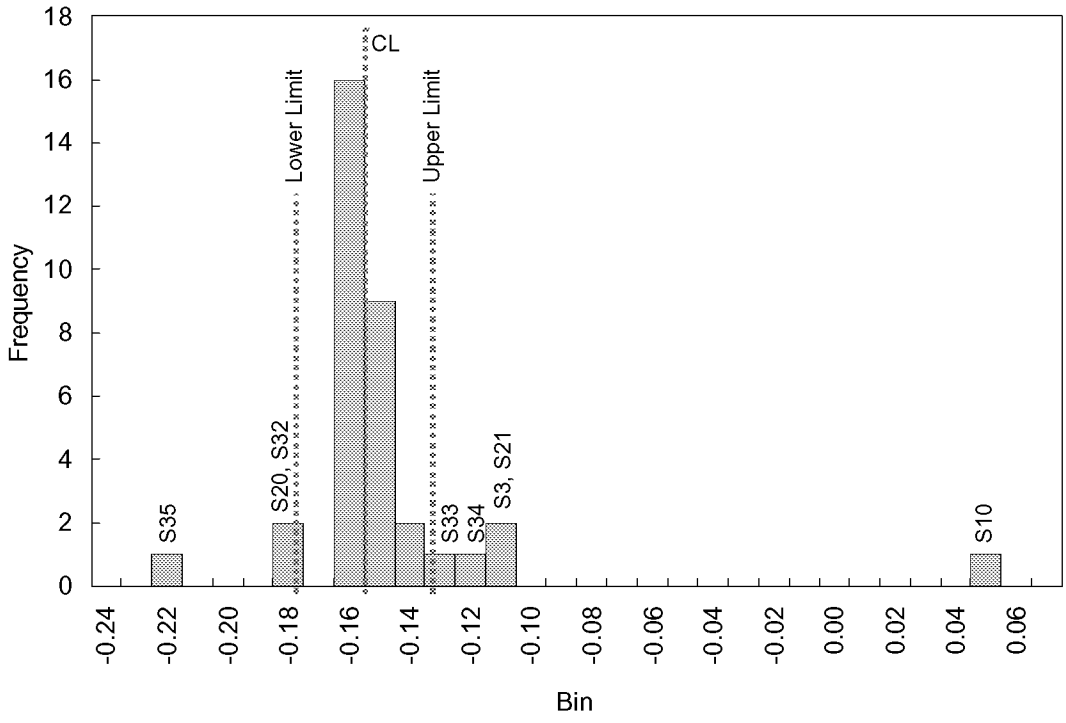


(b) Histogram. (Each bin covers 0.1 degrees.)

Figure 5. All AOA results at $C_L = 0.5$, $M_\infty = 0.75$. Median + MAD + 100:1 limits.

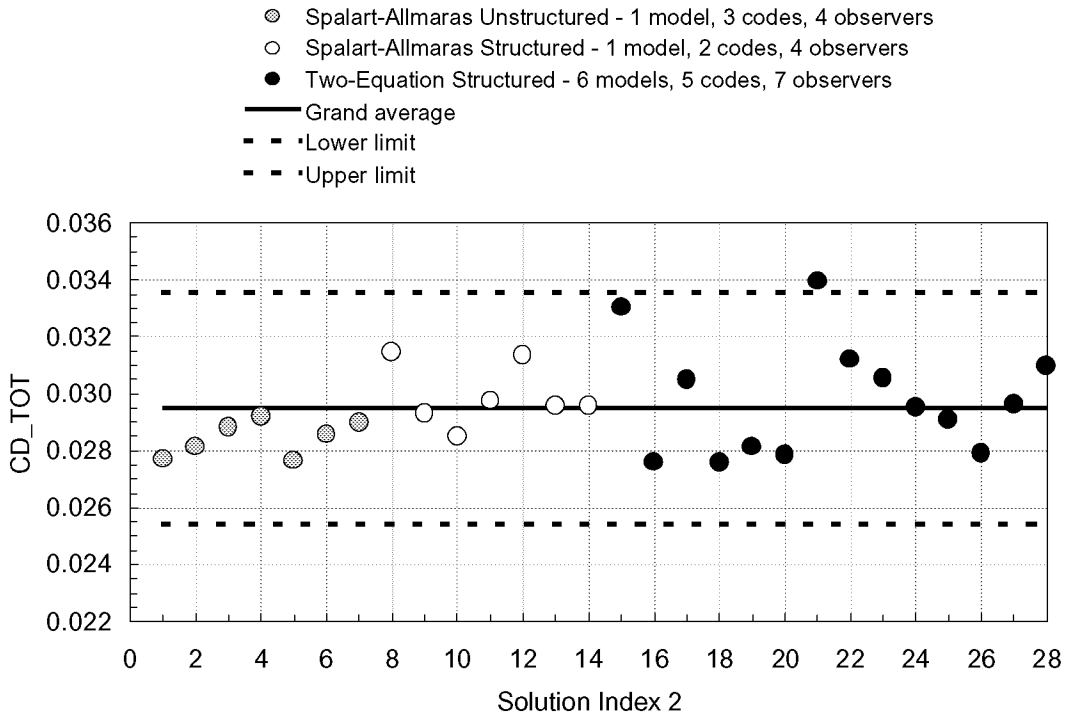


(a) Running record.

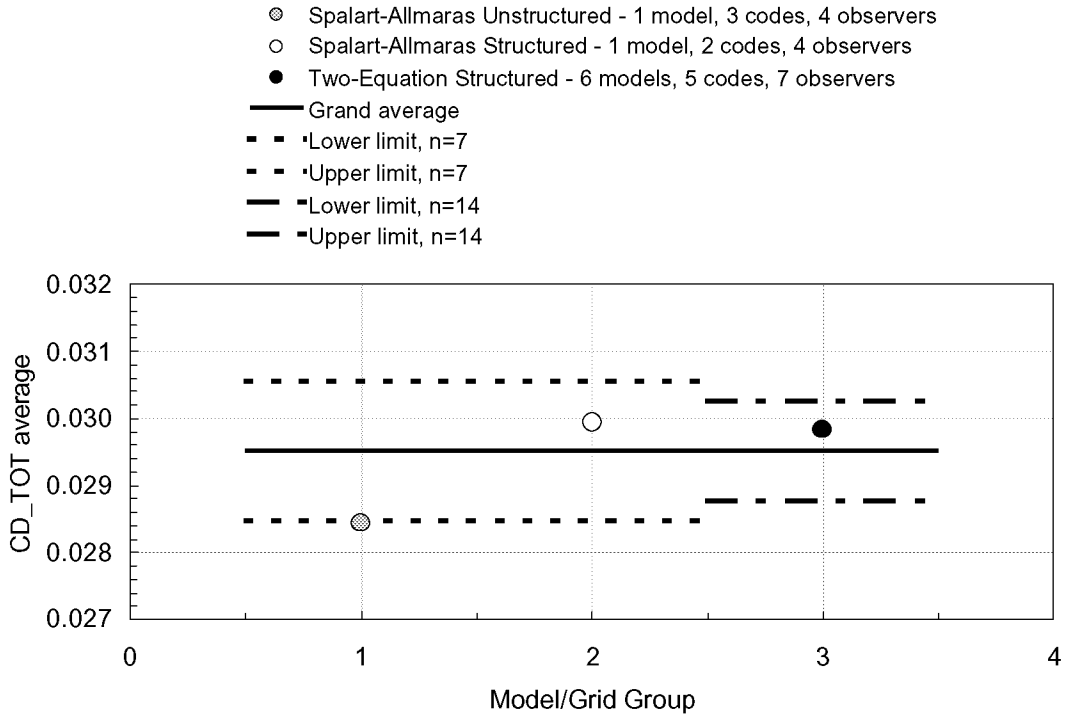


(b) Histogram. (Each bin covers a delta CM of 0.01 degrees.)

Figure 6. All pitching-moment results at $C_L = 0.5$, $M_\infty = 0.75$. Median + MAD + 100:1 limits.

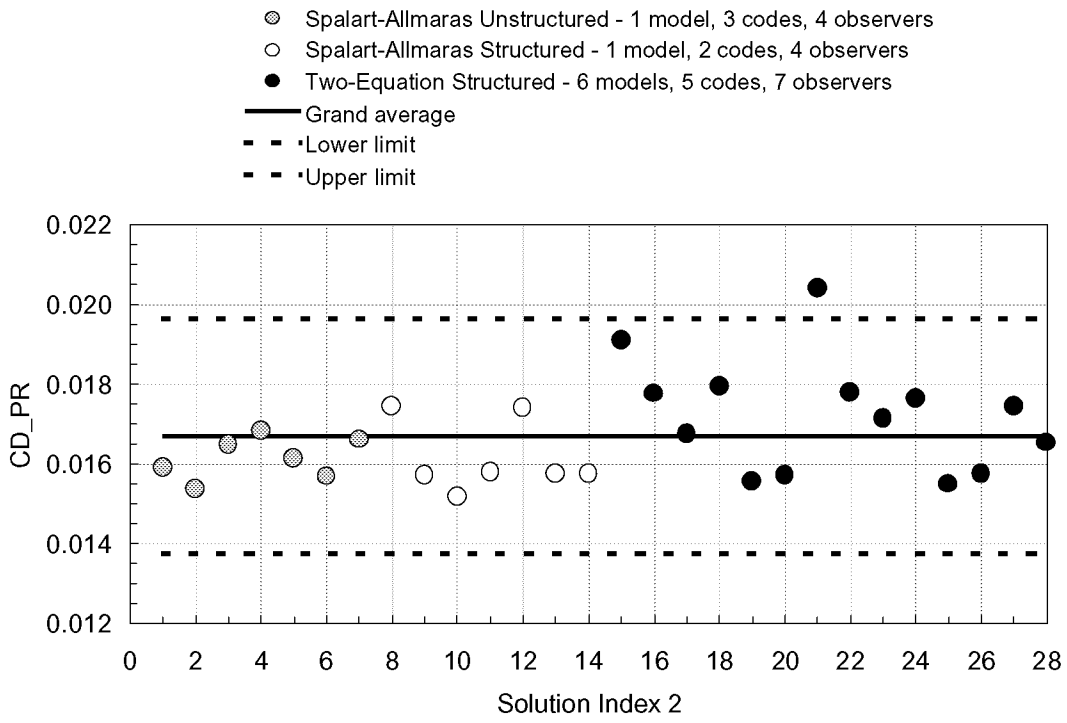


(a) Running record.

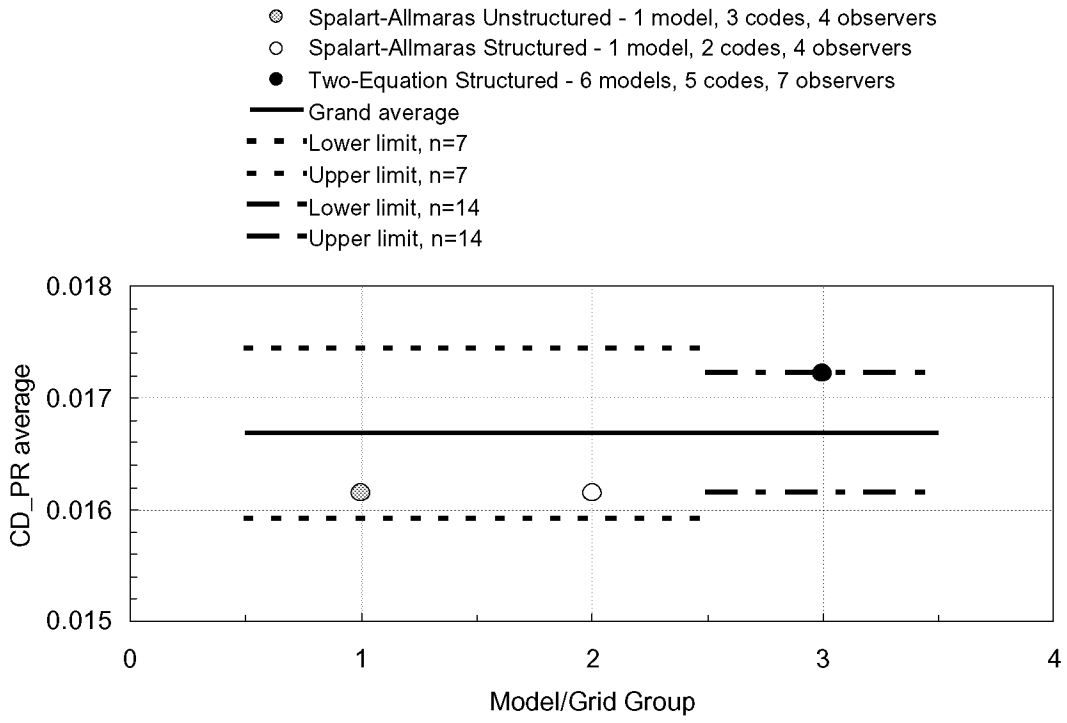


(b) Analysis of Means.

Figure 7. Total drag. Analysis of turbulence model and grid-type effects at $C_L = 0.5$, $M_\infty = 0.75$.

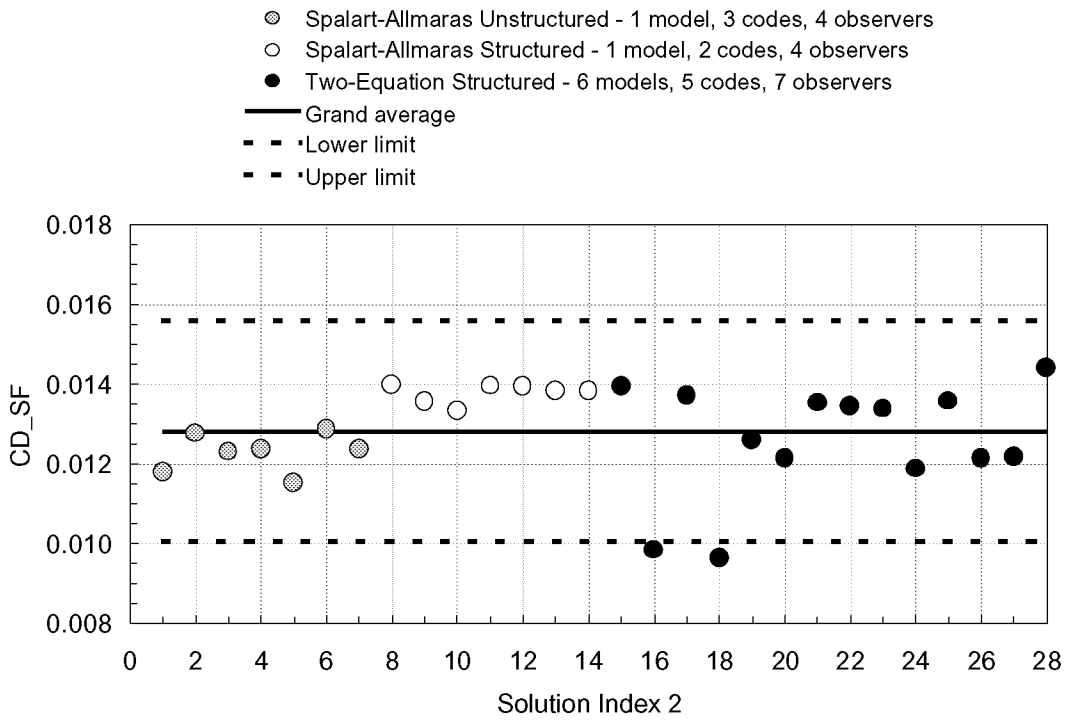


(a) Running record.

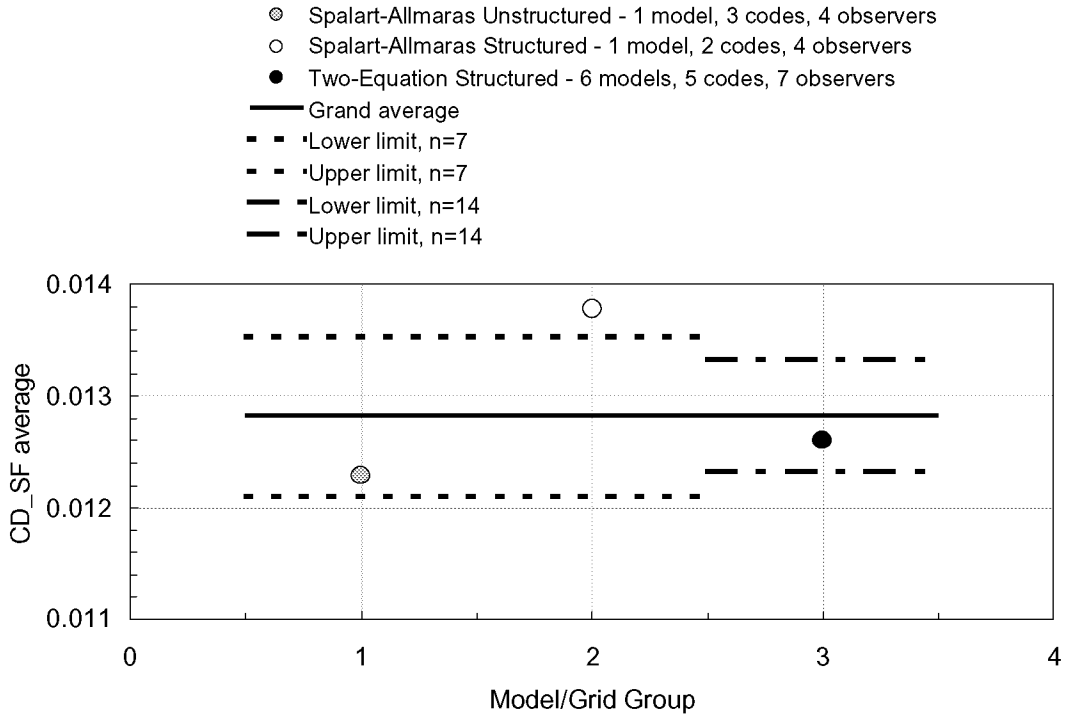


(b) Analysis of Means.

Figure 8. Pressure drag. Analysis of turbulence model and grid-type effects at $C_L = 0.5$, $M_\infty = 0.75$.

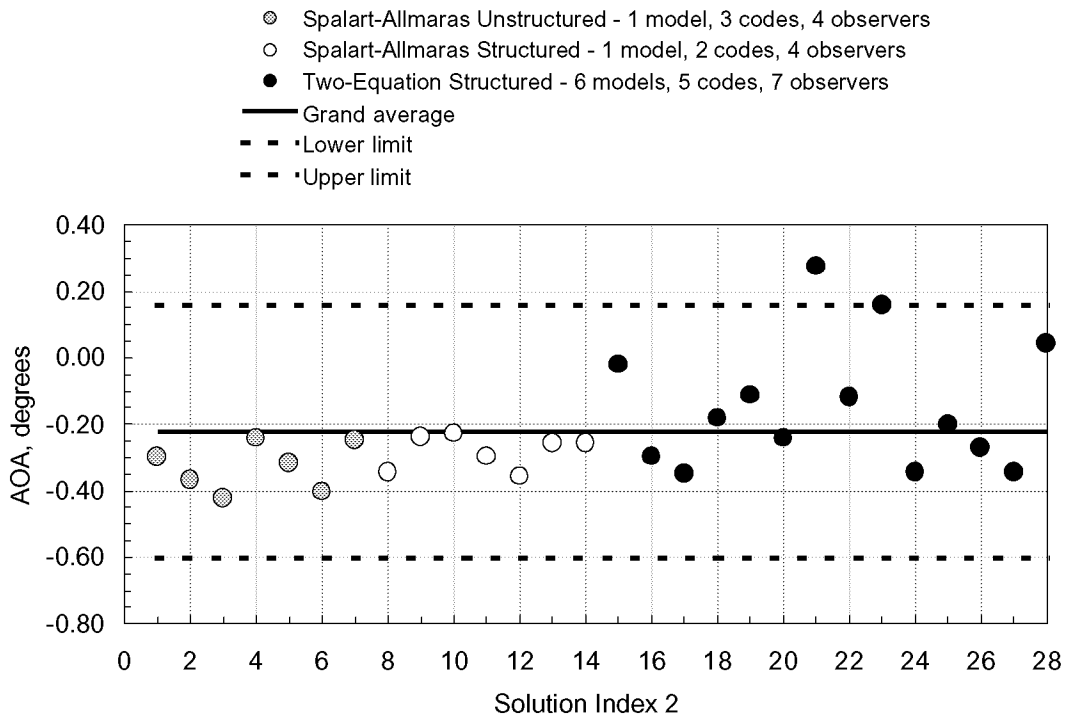


(a) Running record.

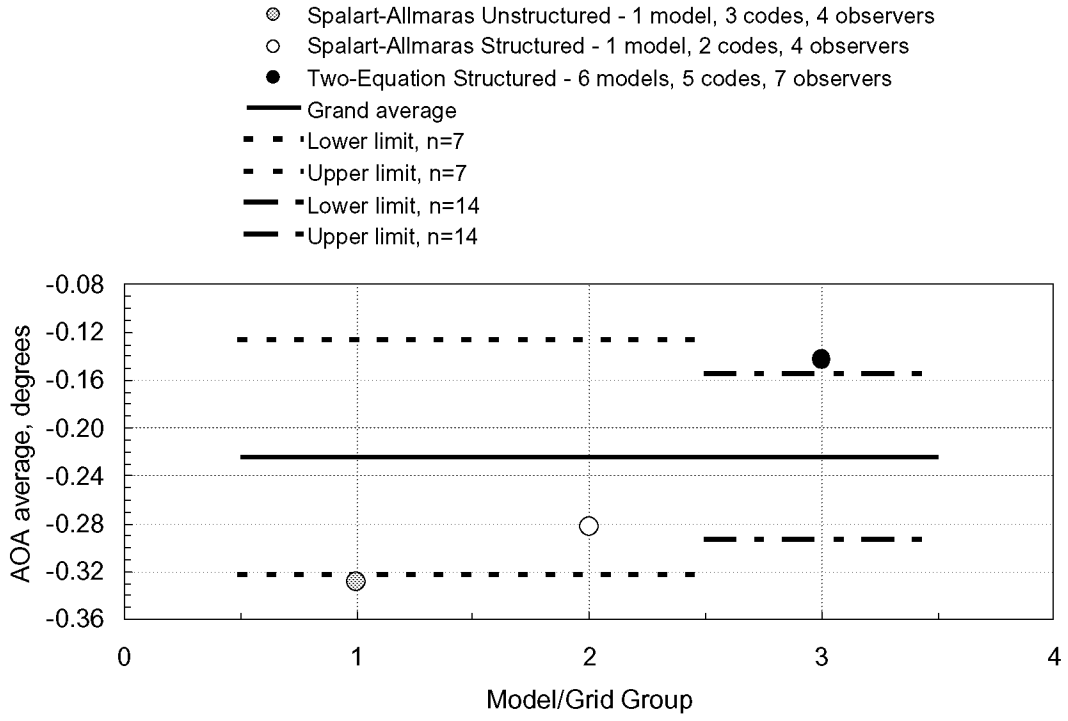


(b) Analysis of Means.

Figure 9. Skin-friction drag. Analysis of turbulence model and grid-type effects at $C_L = 0.5$, $M_\infty = 0.75$.

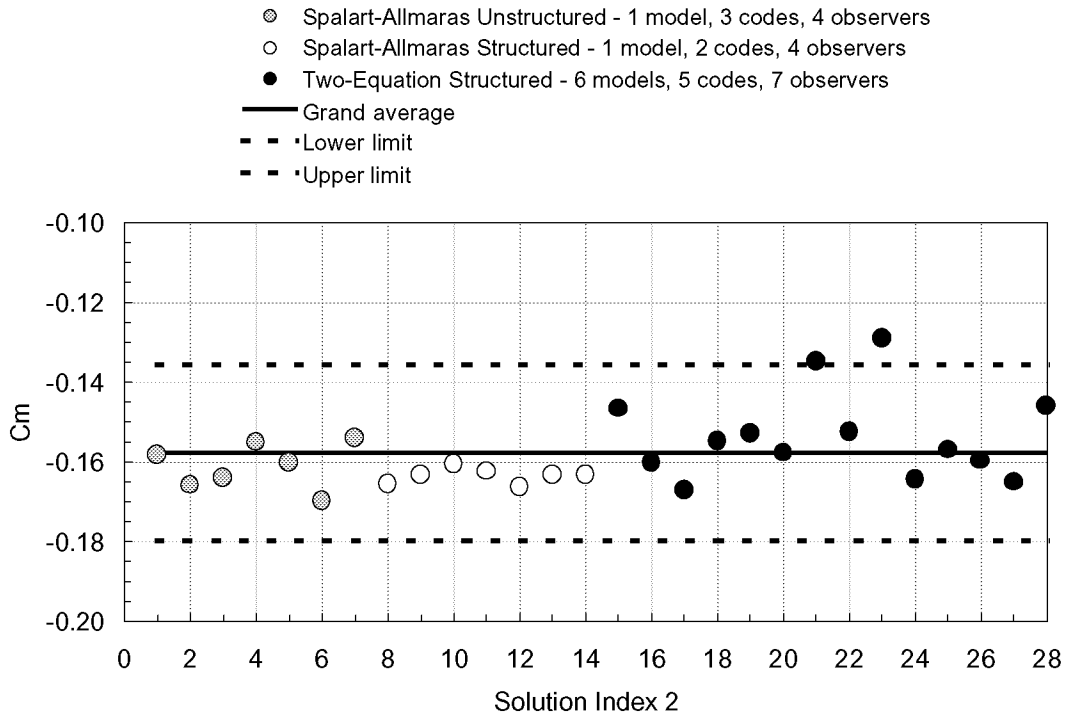


(a) Running record.

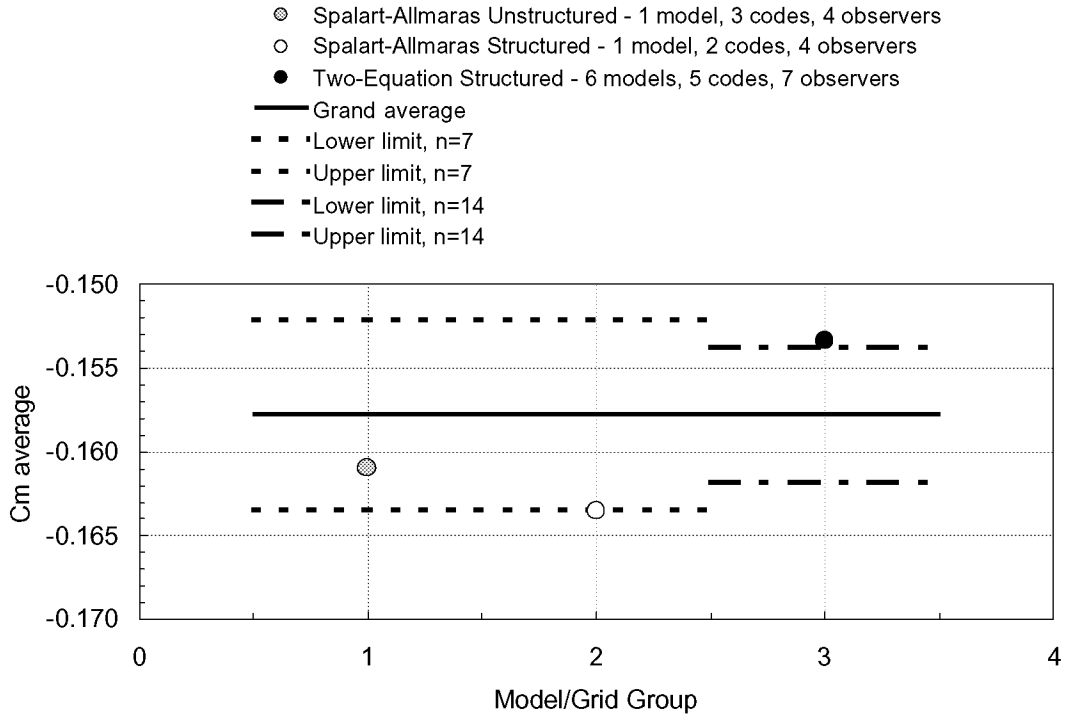


(b) Analysis of Means.

Figure 10. Angle of attack. Analysis of turbulence model and grid-type effects at $C_L = 0.5$, $M_\infty = 0.75$.

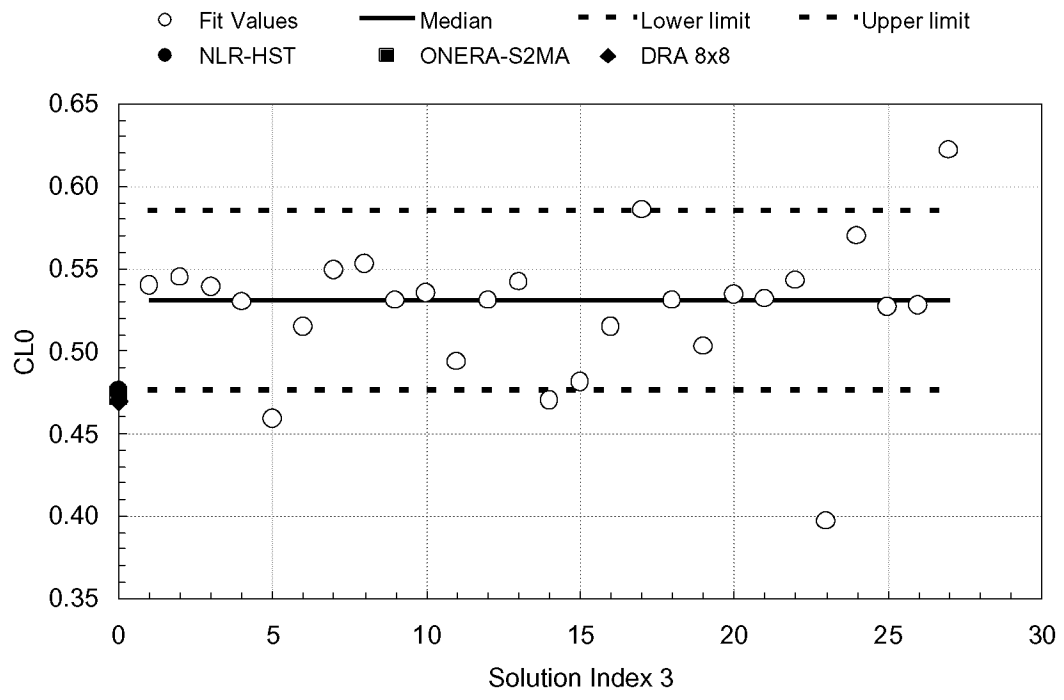


(a) Running record.

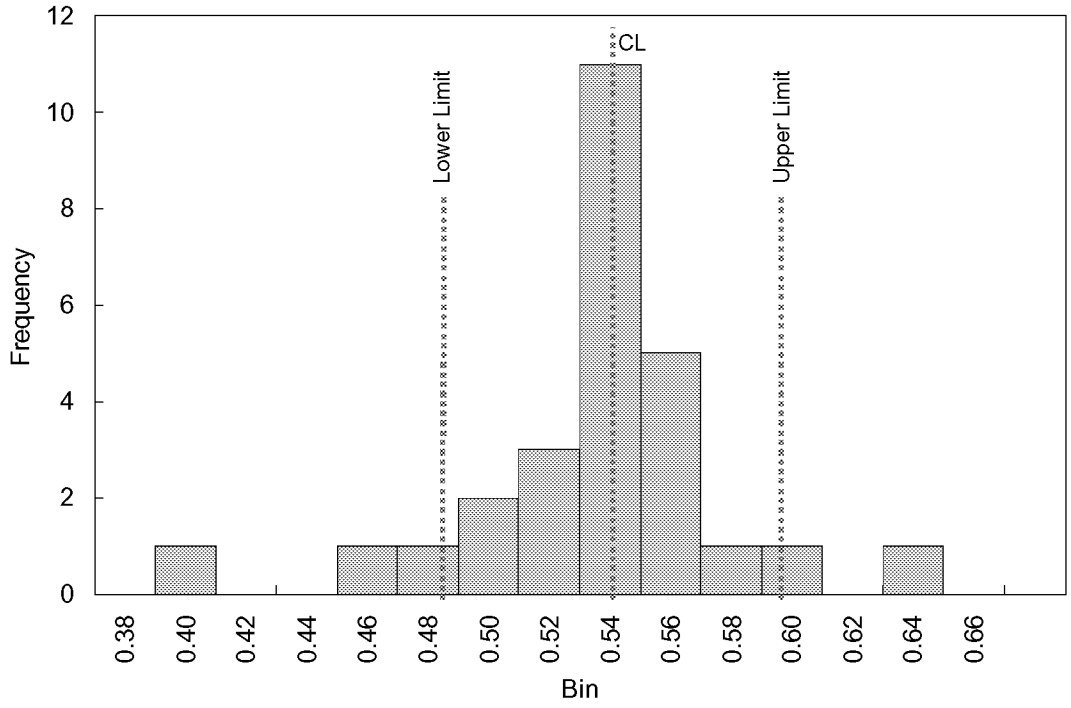


(b) Analysis of Means.

Figure 11. Pitching moment. Analysis of turbulence model and grid-type effects at $C_L = 0.5, M_\infty = 0.75$.

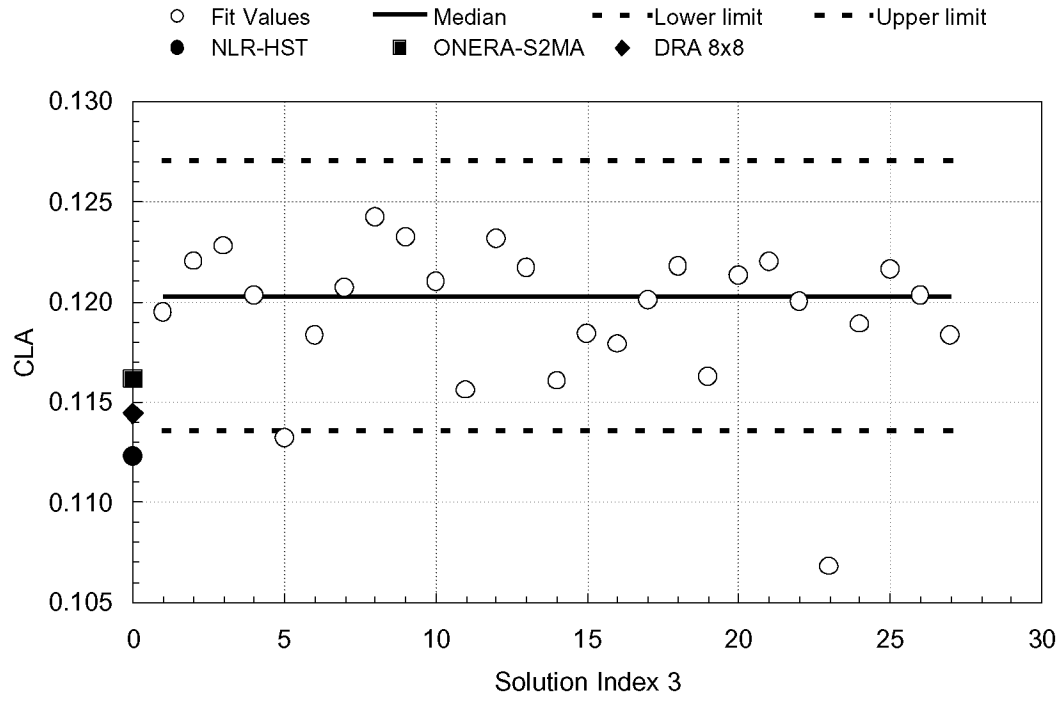


(a) Running record.

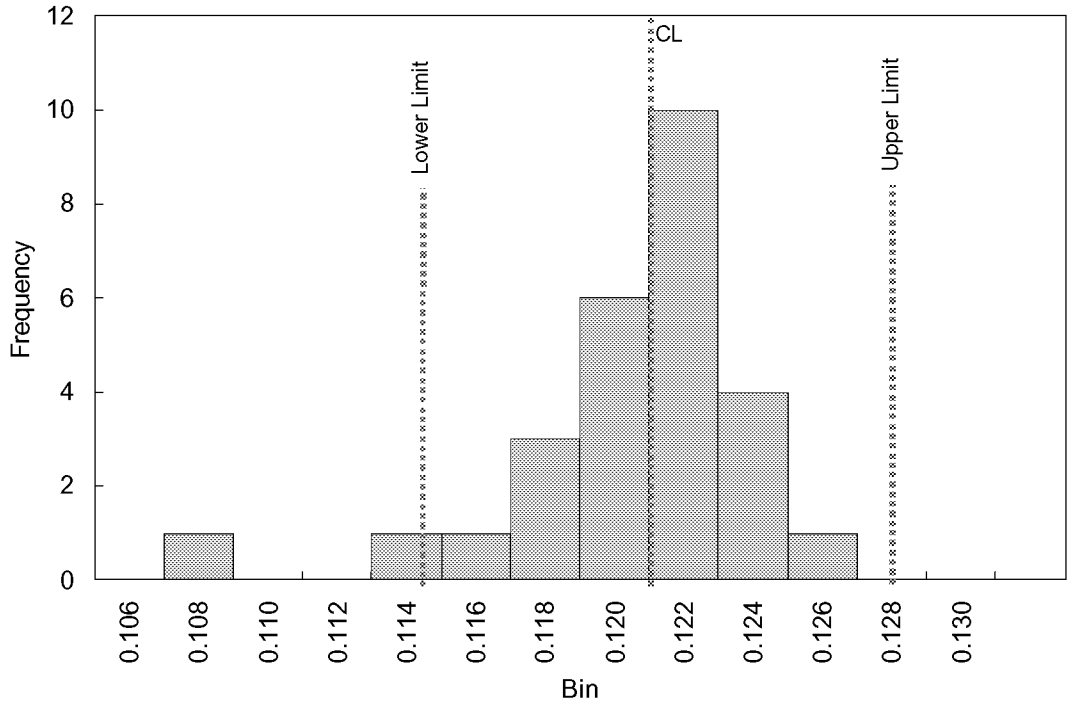


(b) Histogram. (Each bin covers a delta CL0 of 0.02.)

Figure 12. Lift-curve intercept. Drag polars at $M_\infty = 0.75$. Median + MAD + 100:1 limits.

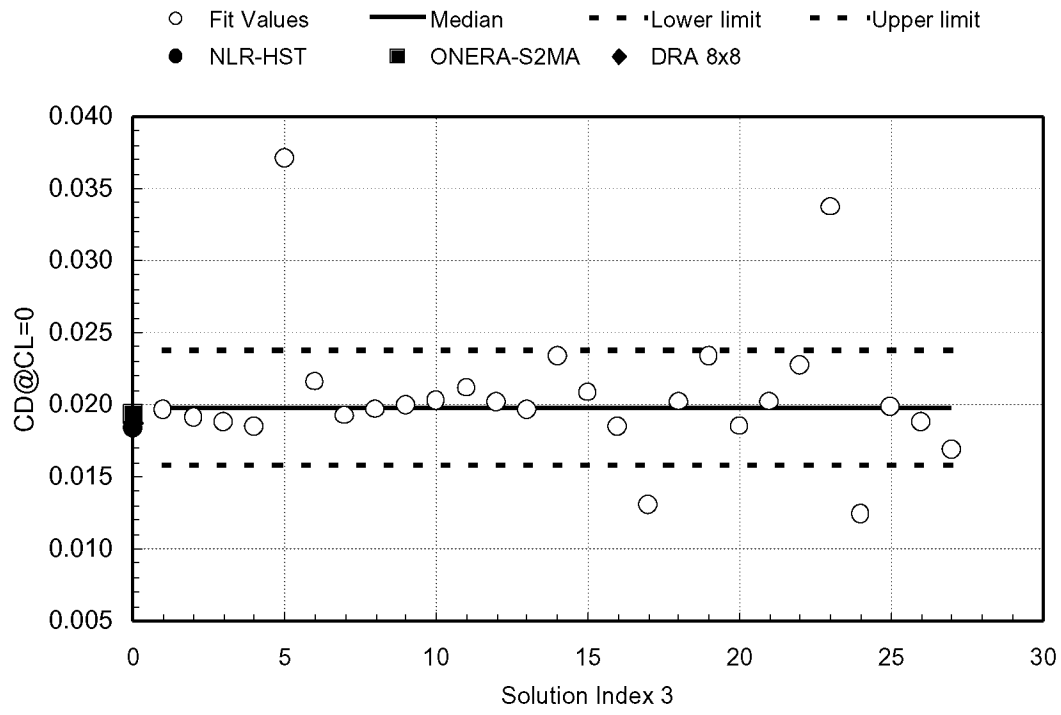


(a) Running record.

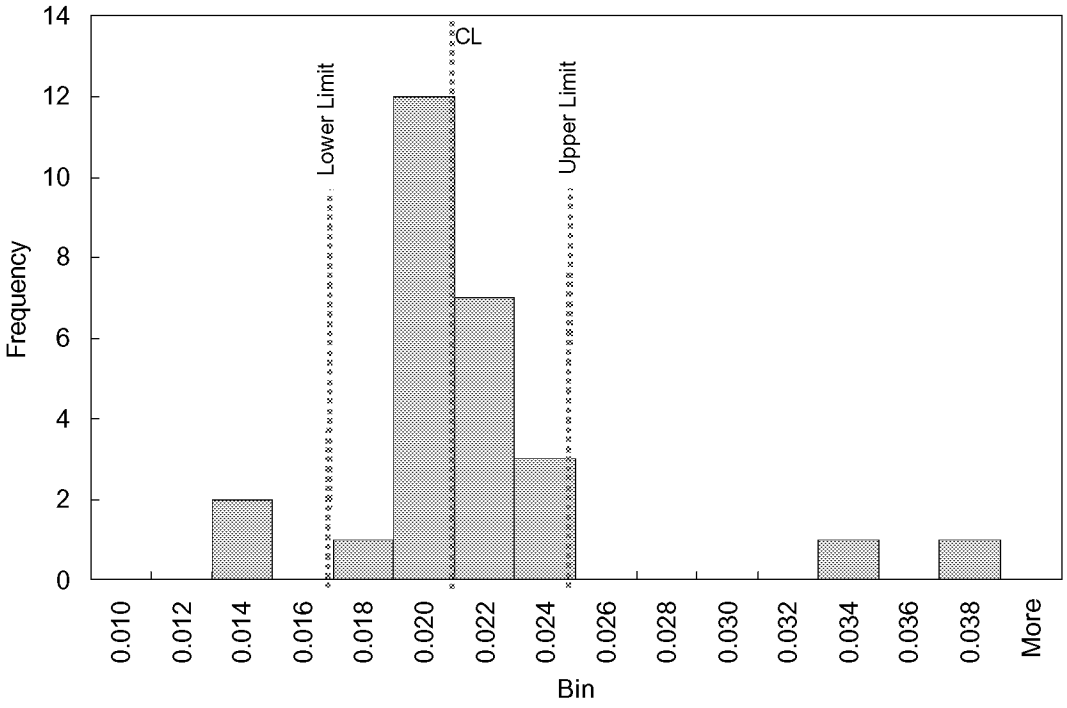


(b) Histogram. (Each bin covers a delta CLA of 0.002.)

Figure 13. Lift-curve slope. Drag polars at $M_\infty = 0.75$. Median + MAD + 100:1 limits.

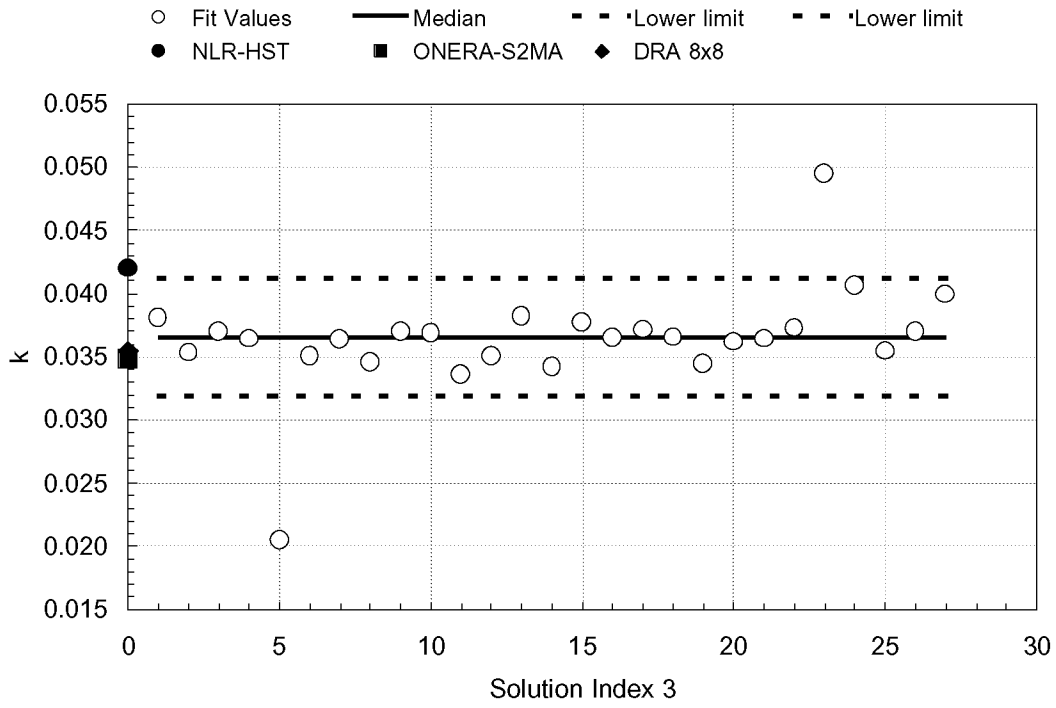


(a) Running record.

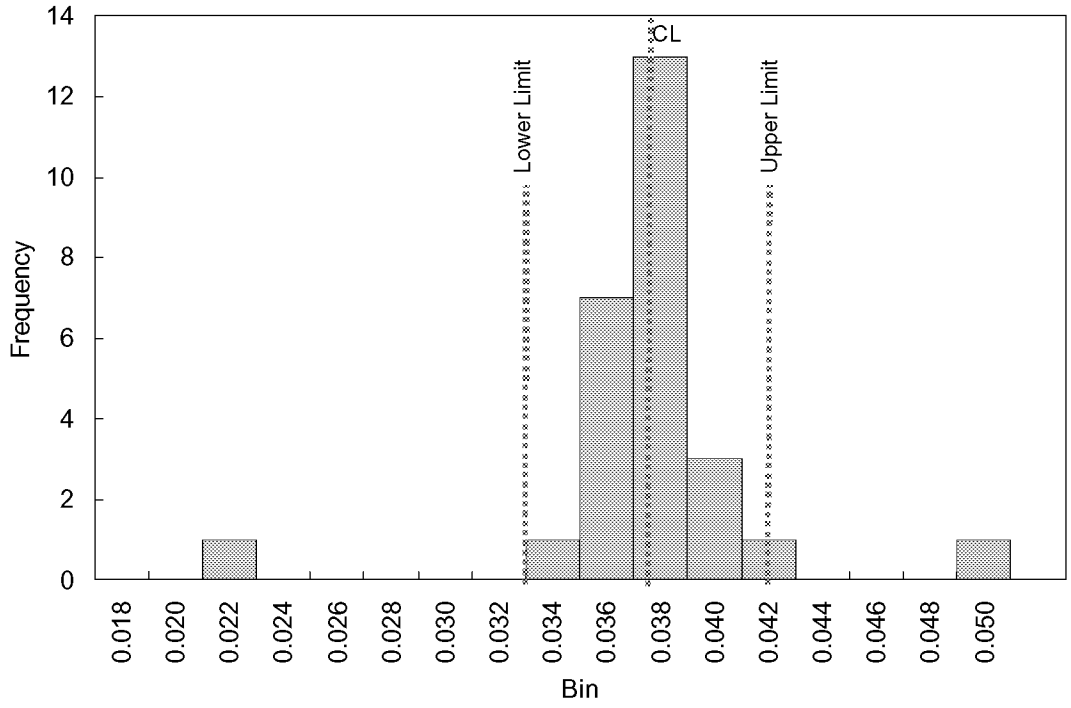


(b) Histogram. (Each bin covers a delta CD@CL=0 of 0.002.)

Figure 14. Drag-curve intercept. Drag polars at $M_\infty = 0.75$. Median + MAD + 100:1 limits.

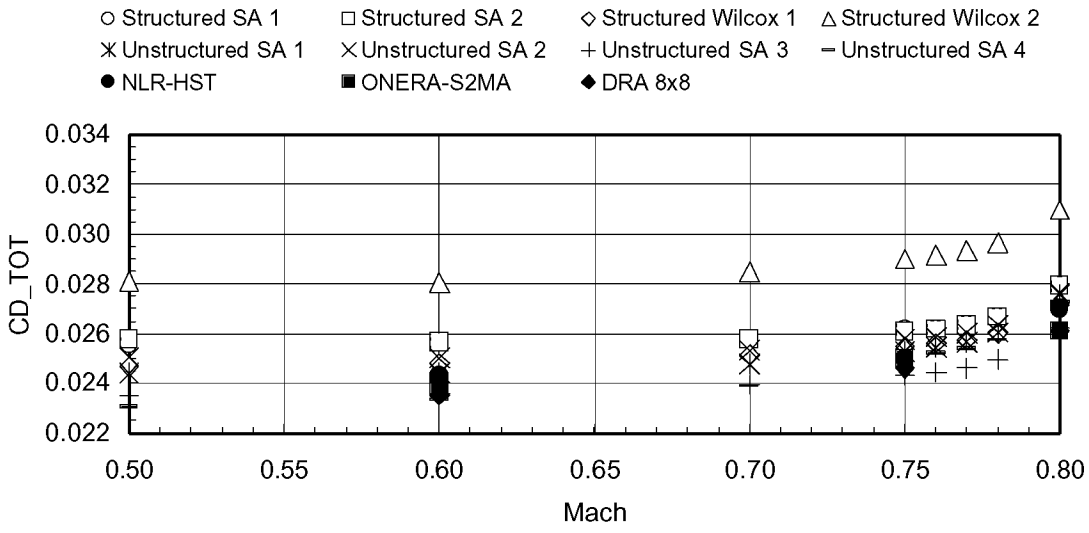


(a) Running record.

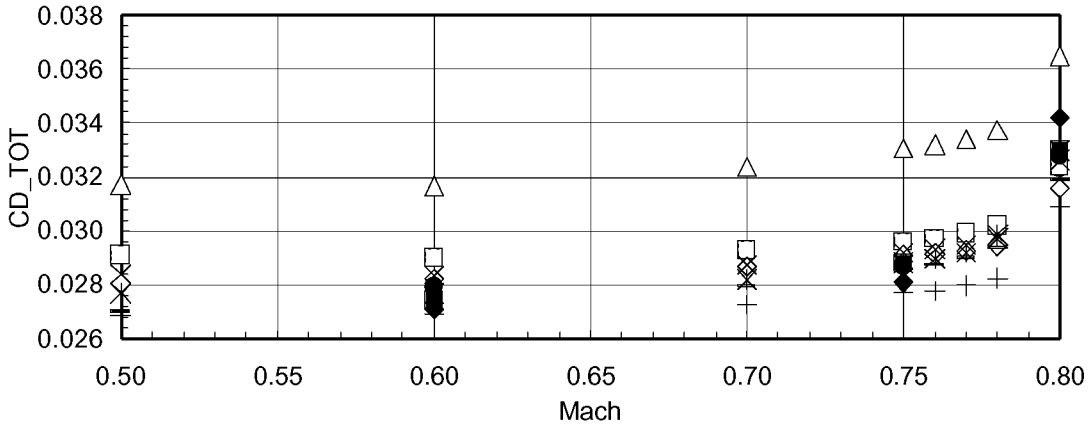


(b) Histogram. (Each bin covers a delta k of 0.002.)

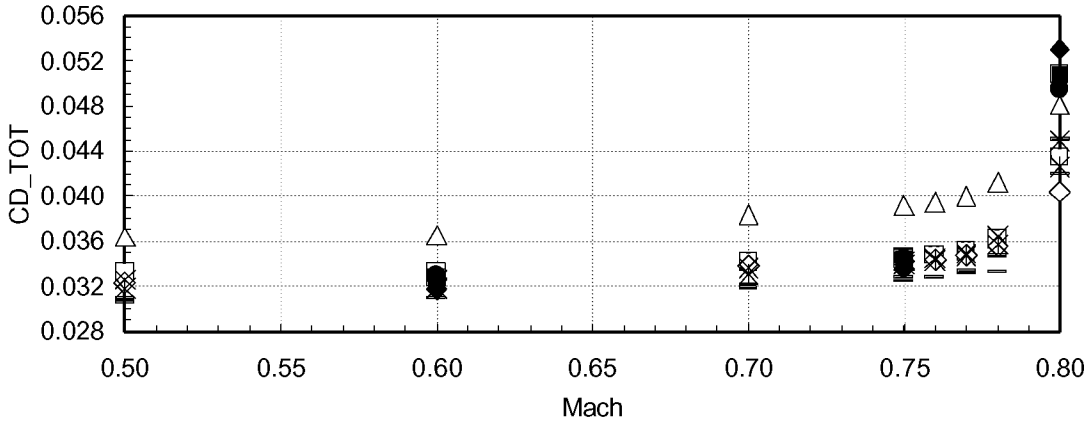
Figure 15. Drag-curve slope. Drag polars at $M_\infty = 0.75$. Median + MAD + 100:1 limits.



(a) CL = 0.4



(b) CL = 0.5



(c) CL = 0.6

Figure 16. Comparison of drag rise solutions.

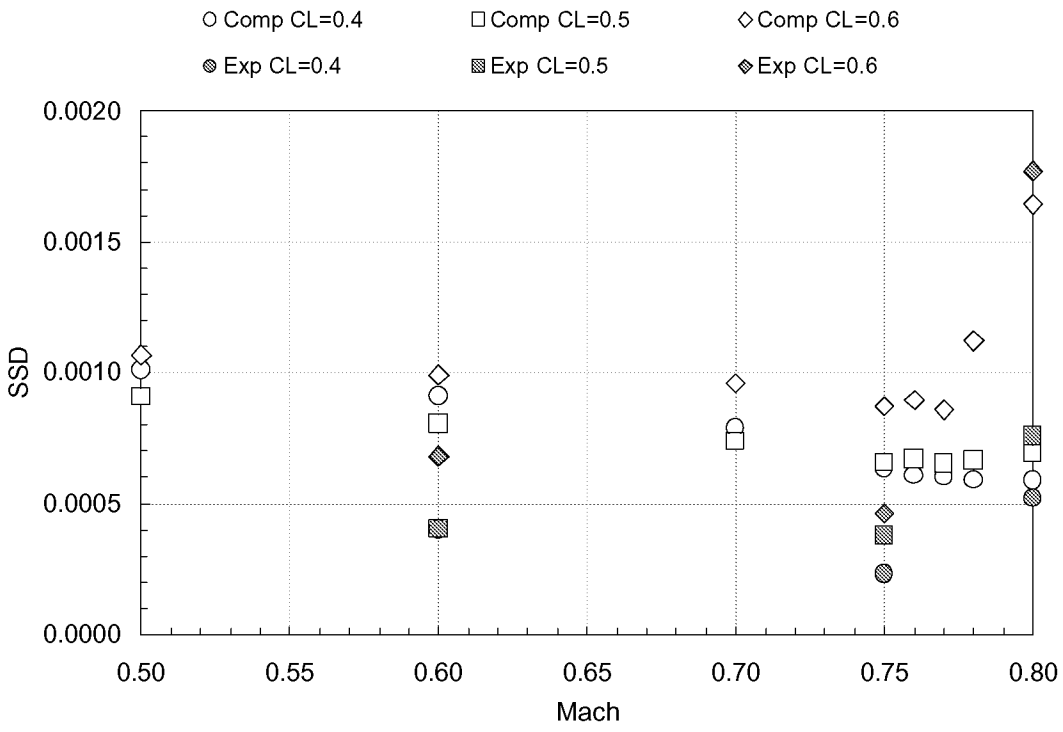


Figure 17. Running record of estimated standard deviations from drag rise studies.

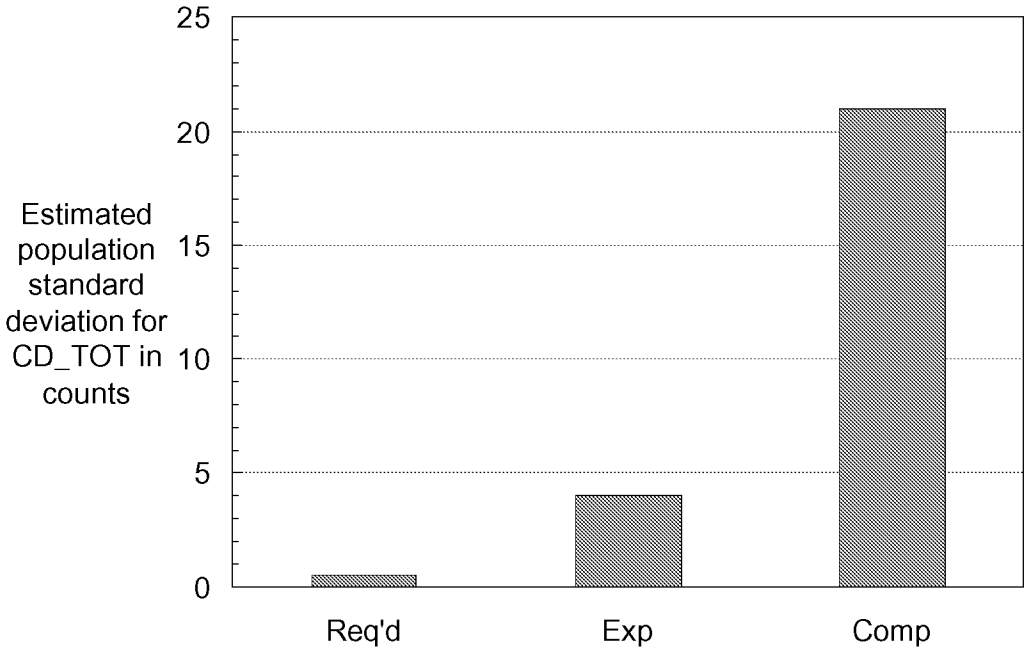


Figure 18. Comparison of required, experimental and computational estimated population standard deviations.