Aircraft Noise Model Validation Study

HMMH Report No. 295860.29

January 2003

Prepared for:

National Park Service
Denver Service Center

Aircraft Noise Model Validation Study

HMMH Report No. 295860.29

January 2003

Prepared for:

National Park Service
Denver Service Center
12795 W. Alameda Parkway
Denver, Colorado

Prepared by:

Nicholas P. Miller
Grant S. Anderson
Richard D. Horonjeff
Christopher W. Menge
Jason C. Ross
Marc Newmark

Harris Miller Miller & Hanson Inc.
15 New England Executive Park
Burlington, MA 01803

ACKNOWLEDGEMENTS

It is difficult to acknowledge all the individuals who contributed to the development, implementation and documentation of this study. Staff from the National Park Service, the Federal Aviation Administration, Volpe National Transportation Systems Center, Wyle Laboratories, Senzig Engineering, Sanchez Industrial Design, Harris Miller Miller & Hanson Inc. and the Technical Review Committee all contributed in significant ways. We note, however, that though these individuals provided valuable assistance with study design, data collection, modeling and analysis, and many of them provided detailed comments at various stages of the report writing, the final interpretations and words are those of the authors. The following list attempts to acknowledge the primary participants.

National Park Service
Mike Ebersole
Rick Ernenwein
Tracey Felger
Tom Hale
Wes Henry
Nick Herring
Marv Jensen
Bill Schmidt
Dan Spotsky
Howie Thompson
Ken Weber

Federal Aviation Administration
Barry Brayer
Tom Connor
Howard Nesbitt
Jon Pietrak

Volpe
Gregg Fleming
Cynthia Lee
Amanda Rapoza
Dave Read
Chris Roof
Judy Rochat
Paul Valihura

HMMH Team
Grant Anderson
Kyle Donnelly
Dick Horonjeff
Chris Menge
Nick Miller
Marc Newmark
Jason Ross

Wyle Laboratories
Kevin Bradley
Micah Downing
Chris Hobbs
Ken Plotkin
Eric Stusnick

Sanchez Industrial Design
Gonzalo Sanchez

Senzig Engineering
David Senzig

Bedford Associates
Jack Giurleo

Out of the Box Productions
Trevor May

Terrapin Acoustical Services
Ken Polcak

Technical Review Committee
Jim Barnes
Dave Keast
Bob Lee
Kåre Liasjø
Allan Piersol
Andy Powell
Lou Sutherland
Sheila Widnall

Page Intentionally Blank

EXECUTIVE SUMMARY

The National Parks Overflights Act of 1987[1] tasked the National Park Service (NPS), and the Federal Aviation Administration (FAA), with developing a plan for tour aircraft use of Grand Canyon airspace that will succeed in "substantially restoring the natural quiet in the park." NPS defined substantial restoration of natural quiet as occurring when "50% or more of the park achieve[s] 'natural quiet' (i.e., no aircraft audible) for 75-100 percent of the day." Hence, a method was required to determine when substantial restoration of natural quiet is achieved. Because only through computer modeling is it practical to assess whether or not natural quiet has been substantially restored, this report presents the methods and results of a study that examines which of four computer models best calculates tour aircraft audibility in the Grand Canyon.

The assessment method was developed by consultant firms in coordination with NPS and FAA staff, and with comments, advice and review by a team of recognized experts in acoustics, statistics and scientific methods, called here the Technical Review Committee or TRC. The computer models were assessed by collecting appropriate acoustic, tour flight operations, and meteorological data in the Canyon, using these data in each of the four computer programs to predict air tour audibility, and then comparing the computed results with audibility data collected in the Canyon by trained observers. An additional acoustic metric, the "hourly equivalent sound level" was also computed and compared with measured values.

All data were collected simultaneously by ten teams over three days at the Grand Canyon, with 301 hours of acoustic data collected at 39 sites, operations data collected at a site directly under the tour flight corridor, and meteorological data collected at five temporary and two permanent sites in the Canyon.

The four models examined were two versions of the FAA's Integrated Noise Model, INM; a model developed for the NPS, the National Park Service Overflight Decision Support System or NODSS; and a model that is a derivative of the U.S. Air Force program NOISEMAP, called the NOISEMAP Simulation Model or NMSIM.

Measured hourly tour aircraft audibilities (and equivalent sound levels) were compared with computed audibilities (and equivalent sound levels) in three primary ways. Statistical comparisons, with confidence ranges, were calculated for overall model error, bias, and scatter. Numerical values of these measures were developed, as were associated 95% confidence ranges where appropriate.

Overall, NMSIM proved to be the best model for computing aircraft audibility, because it is shown to have the most consistent combination of low error, low bias and low scatter for virtually all comparisons. (See Section 1.9.1 or Section 8.) The authors recommend that it be used for future modeling of tour aircraft audibility in the Canyon since its computed results best match the measured results. The INM versions generally have higher error and scatter than NMSIM, but tend to also show low bias in computing audibility. NODSS tends also to have higher error and higher bias than NMSIM, but with scatter comparable to or slightly lower than NMSIM.

The results also suggest that, if used with realistic values for ambient noise levels, both NMSIM and the tested INM versions do not show significant bias in computing audibility in the Canyon on average. Thus, though INM scatter is relatively greater when computing audibility levels at any

[1] Public Law 100-91, August 18, 1987.

specific location, if a parkwide average is computed for the Canyon, both NMSIM and INM results are likely to produce relatively small errors.

Both INM versions and NMSIM were roughly equivalent in computation of equivalent sound levels, with INM having slightly less bias, and both may be used for computation of these sound levels.

The report contains recommendations about how the models may be used for modeling tour aircraft overflights of National Parks, detailed analyses of the possible sources of error in the models, and suggestions for making improvements to the models. If any of the models are changed in a way that might improve or alter the predictions, the altered models should be tested with the data and techniques used in this study to identify the effects of the changes.

Section 1 of this report gives a detailed summary of the study, results, conclusions and recommendations, and may provide sufficient information for most readers. The remainder of the report, with appendices, provides a detailed step-by-step description of all methods, data and results.

1. STUDY SUMMARY

This section summarizes the study and its results. The results, though summarized, are provided in considerable detail so that many readers may obtain sufficient information from just this section. The remaining sections and appendices provide detailed information on all phases of the study.

1.1 Study Goal

This report presents the methods and results of a study that examines the overall error, the accuracy and the precision of four computer models used to calculate tour aircraft audibility in the Grand Canyon.[2] In the National Parks Overflights Act of 1987, Congress tasked the National Park Service (NPS) and the Federal Aviation Administration (FAA) with developing a plan for tour aircraft use of Grand Canyon airspace that will succeed in "substantially restoring the natural quiet in the park."[3] NPS defined substantial restoration of natural quiet as occurring when "50% or more of the park achieve[s] 'natural quiet' (i.e., no aircraft audible) for 75 – 100 percent of the day."[4] Computer modeling is the only practical means for assessing whether or not natural quiet has been substantially restored in accordance with this definition.

Models that compute when aircraft are audible over large land areas have not been widely used, and none has been tested through comparison with measured values of audibility. Consequently, NPS and FAA elected to conduct this validation study based on the concept that model validity be determined through detailed comparison of computer results with measurements made on-site in the Grand Canyon.[5] Hence, this study was designed and conducted with the primary goal to:

Determine the degrees of accuracy and precision that existing computer models provide, in comparison with field measurements, in the calculation of the percent of time tour aircraft are audible in the Canyon, and calibrate one or more of these models to provide a tool for computation of air tour audibility in the Canyon.[6]

This goal was achieved by having trained listeners keep detailed logs of when tour aircraft could be heard (were audible) at different locations in the Canyon, and simultaneously logging all tour operations that could have been heard. The computer models were then run using the logged operations to compute how much of the time the tours could have been heard. The computed results were then compared with the audibility logs, to determine how well the computed results agreed with the audibility results of the listeners.

Because achieving this goal reveals how well each of the models performs quantitatively, ranking of the models' performance is inevitable. This report clearly identifies which model best matches the

[2] Descriptions of overall error, accuracy and precision appear in Section 1.9.1.2 and those that follow it.
[3] Public Law 100-91, August 18, 1987, § 3. (b) (3) (A).
[4] U.S. DOI, National Park Service, "Report on Effects of Aircraft Overflights on the National Park System," Report to Congress, July 1995, Section 9.2.1, p. 182.
[5] Throughout this report, both audibility data and sound level data that have been collected in the canyon are called "measured" data, whether the data are measured with instruments (as are sound levels) or observed by trained staff (as are audible durations of tour aircraft).
[6] In addition to examining the "percent of time audible", the tour aircraft "hourly equivalent sound level" L_{eq} was also examined. This equivalent sound level is a measure of the total sound energy produced by tour aircraft during an hour. It is similar to the metrics generally used in Environmental Assessments, Environmental Impact Statements and other common types of environmental analyses that address noise effects on residential and commercial land uses.

measured levels according to the goal, and is, in the version tested, most appropriate for modeling audibility of tour operations over the Canyon.

1.2 Tour Aircraft Audibility

"Audibility" as used in this study begins at the instant that an attentive human listener can detect the presence of the sound produced by a tour aircraft and lasts as long as the listener continues to hear the aircraft. Though the audibility of a source can vary from listener to listener, on average, humans without significant hearing loss are able to identify the presence of a source in a given background sound environment at similar sound levels. Thus, whether or not a tour aircraft is audible is determined by both the sound level of the tour aircraft and by the sound level of the ambient or non-tour aircraft sound levels. These concepts, the mathematical form used to compute audibility, and the measured performance of the field staff that collected the audibility data are presented in detail in APPENDIX C, page 167. Very few computer models have been designed to compute audibility of sources of sound over long distances,[7] and none have undergone the rigorous testing performed in this study. The special nature of restoration of natural quiet called for by Congress, and the NPS implementation of that mandate, necessitated this unusual and complex examination of audibility.

1.3 Study Design and Review Process

The study was designed through a cooperative process involving the NPS, the FAA, the Volpe National Transportation Systems Center (Volpe), Wyle Laboratories (Wyle), and Harris Miller Miller & Hanson Inc. (HMMH). After a draft approach had been developed, a Technical Review Committee (TRC) consisting of internationally recognized experts (see Appendix A.1, page 157) reviewed and commented on the plan. Suggestions made by TRC members were incorporated into the study design. As results were produced, the full team, including TRC members, was involved in review and comment. The full team has reviewed and commented on drafts of this study report. Their comments were incorporated extensively.

1.4 Study Method

The study method involved four basic steps:

1. Acquisition in the Grand Canyon of tour aircraft audibility data, sound level data, and the associated aircraft and ambient noise modeling input data.
2. Reduction of the collected data to forms suitable for modeling and for analysis.
3. Modeling of the scenarios that were measured in order to compute values for comparison with the measured audibilities.
4. Analysis of the reduced and modeled data to: 1) compare computed and measured values; 2) assess calibration methods; 3) provide information useful for future efforts at diagnosing discrepancies between computed and measured values.

1.5 Data Acquisition

Data were acquired in the Grand Canyon over a four-day period in September 1999. Data collected included primarily the audibility logs created at some 39 sites by eight four-person teams. These logs identified the times of onset and offset of tour aircraft audibility as determined by trained

[7] One exception is described in Horonjeff, R., Fidell, S., "A Computer Program for Predicting Audibility of Noise Sources," Air Force Wright Aeronautical Laboratories, AFWAL-TR-83-3115, December 1983.

observers. Measurements generally took place between 8:00am and 5:00pm. Digital tape recordings of all sounds were also made simultaneously at 19 of these sites. One additional four-person team, located on the rim under the air tour corridor (the Zuni Point corridor), kept a log giving the time and type of each tour aircraft, recorded each aircraft's sound level, and video recorded its location. Finally, one team supervised the collection of meteorological ("met") data at five temporary sites; meteorological data were also acquired from two permanent Canyon sites. The result was 301 site-hours of audibility and modeling data, of which 192 hours had associated sound level data.

1.6 Data Reduction

The collected data were reduced to provide hourly information for modeling and analysis. The reduced data included for each hour measured:

1. Numbers, types and speeds of tour aircraft operations.
2. Source sound levels of tour aircraft, both A-weighted and by frequency (1/3 octave band).
3. Ambient sound level, both A-weighted and by frequency, by site.
4. Percent of time air tours audible by site.
5. Air tour hourly equivalent sound level, Leq, by site.
6. Wind speed and direction at the seven "met" stations.
7. Temperature, relative humidity and barometric pressure at the met stations.
8. Various site specific parameters such as distance from air tour corridor, angle of corridor visible, latitude and longitude, elevation, etc.

These data provided the information used for modeling tour aircraft audibility and sound levels for each hour of operations, and for then analyzing the results.

1.7 Modeling

Each of the four models tested were exercised with the same set of input data. The models used were:

1. The Integrated Noise Model (INM), version 5.1, which does its computations using only A-weighted levels;
2. The INM in its Research Version, which includes one-third octave band (1/3 octaves) spectral information. Both INM models, which are energy based, account for differences in site elevation, but not for shielding due to terrain.
3. NOISEMAP Simulation Model (NMSIM), which uses spectral information, accounts for park terrain, computes tour aircraft audibility, flies aircraft in the time sequence in which they occurred, and includes the directivity of each aircraft type.
4. The National Park Service Overflight Decision Support System (NODSS), which uses spectral information and was designed to account for park terrain features, and to compute tour aircraft audibility.

The models were run to produce for each site the hourly values of both the percent of time tour aircraft were audible and the tour aircraft hourly equivalent sound level, L_{eq}. These are the values that were compared directly with measured values, site-by-site, hour-by-hour. Of these four models, only NODSS was originally designed to compute aircraft audibility; the other three were modified to do so.

1.8 Data Analysis

Data analysis was accomplished in three parts. First, the measured and computed values were compared hour-by-hour and site-by-site and analyses done to assess the overall error, the accuracy and the precision of the models, the estimated contour error, and to assess the value of model calibration.[8] Calibration was considered as a means for using the measured data to adjust the computed results of the models so that they would better match the measurements.

Second, an analysis of the "discrepancies," i.e., differences between computed and measured values, was performed. This is a statistical analysis (multiple linear regression) that identifies associations between model discrepancies and various physical factors that were measured concurrently. The results can provide starting points for model improvements, should those be pursued.

Third, the relationship of physical factors to the measured results was analyzed. This is a multiple non-linear regression that identifies associations (or lack of association) between these physical factors and the measured results alone – independent of the computer models. This analysis also provides useful information for model diagnostics and improvements.

1.9 Results

The goal of this study is to determine the accuracy and precision of each model with respect to measured values, to investigate the utility of calibrating one or more of the models, and to provide a means for using the models. This section first presents the results of the comparison and the assessments of accuracy and precision in Section 1.9.1. Next, Section 1.9.2 discusses model calibration and why it has been rejected. Finally, Sections 1.9.3 and 1.9.4 present additional information that should be useful if further diagnostics and improvement of the models is warranted.

1.9.1 Measured versus computed Results

1.9.1.1 Overview of Comparisons

Before summarizing the results of the analyses, it is important to understand the primary comparisons of measured and computed results – the metrics compared, the data used and the quantification of these comparisons. Table 1 summarizes the comparisons made and references the primary figures in this report that make these comparisons. The following paragraphs explain the table.

All comparisons are made for two metrics: 1) the percent of an hour tour aircraft are audible, 2) the hourly equivalent sound level, L_{eq}, of tour aircraft during an hour. First, these comparisons are made using all of the individual hours of measurements collected at the 39 sites for which all the necessary data were available; that is, each hour measured at each site is used as a distinct data point. Second, the comparisons are made using site groups where the data for individual hours are averaged across sites that are located near each other and across all hours.

The comparisons using individual hours demonstrate the entire scatter of all hourly data. It should be noted, however, that this comparison is of limited interest for two reasons. First, it is rare that a

[8] Note that all analyses were accomplished using the audibilities and sound levels measured at the specific individual sites; contours were not developed for the contour error analysis. Contours are normally generated using computer-calculated data at specific sites, and hence their likely error is estimated using only the data at the specific measurement sites.

model would be used to compute results for one specific hour of operations. Second, the results of the analysis may be influenced by the number of hours measured at each specific site, which varied from site-to-site.

Table 1. Types of Comparisons Made Between Computed and Measured Results

Metrics Compared	Data Used
Individual Hours	
Percent Time Audible	192 site hours, measured ambient sound levels (Figure 34, p.85)
	301 site hours, EA ambient sound levels (Figure 2, p.9 and Figure 35, p.86)
Hourly L_{eq}	147 site hours (Figure 36, p.87)
***Site Groups* – (**for Groups, see Table 20, page 88**)**	
Percent Time Audible	12 site groups, measured ambient sound levels (Figure 37, p.90)
	13 site groups, EA ambient sound levels (Figure 11, p.16 and Figure 38, p.91)
Hourly L_{eq}	12 site averages (Figure 39, p.92)

The more useful comparisons are those done for the site groups. Noise models like the ones being tested are commonly used to compute average results for several or many hours of operations. This use is better judged by analyzing the data by site groups. Also, analysis by site group reduces the effects of having different numbers of hours measured at the different sites. The site group hourly audibility and L_{eq} are computed by first grouping the individual sites geographically, then averaging all the hourly results within each geographic group.

Finally, as shown in Table 1, two different sets of ambient sound level were used for the percent time audible comparisons. Both the sound level of the source (tour aircraft) and the sound level of the (non-tour) ambient determine when the source will be audible, and all models therefore require input values for ambient sound level. One set of ambients used was the measured ambients. These were derived from tape recordings made simultaneously at many of the locations where observers were logging the audibility times. Hence, these measured ambients are virtually the most accurate values possible for representing the ambient sound levels that occurred during the audibility logging.

The other ambients used were those derived for the Environmental Assessment of tour aircraft routes, or EA ambient.[9] These EA ambients are also based on measurements made throughout the Canyon, but at different times and locations. However, these EA ambients do provide sound levels representative of the types of vegetation and terrain conditions in which the observer logs were made. Hence, the EA ambients should be thought of as reasonable estimates of the ambient sound levels, but ones that are not directly correlated in time and by location with the audibility logs.

Note that not all comparisons use equal numbers of data points. Measured ambients could be determined for only those sites where sound levels were measured (using tape recordings). Hence

[9] "Special Flight Rules in the Vicinity of Grand Canyon National Park, Final Supplemental Environmental Assessment" Federal Aviation Administration, February 2000.

measured ambient values are available for a subset of the total site hours measured, while EA values are available for all site hours. Similarly, measured aircraft L_{eq} values are available for a subset of the total site hours.[10]

1.9.1.2 Overall Error, Accuracy and Precision

Overall error, accuracy and precision are concepts used in this study to quantify the comparisons of measured versus computed results. This section provides a brief graphical explanation of these concepts. Overall error can be separated into "accuracy" and "precision". In this analysis, accuracy is a measure of how well on average the measured and modeled results agree (also called bias error). Precision is a measure of how consistently computed results correlate with measured results (also called random error or scatter). Figure 1 demonstrates graphically the concepts of accuracy (bias error) and precision (random error or scatter). Each part of this four-part figure represents a different hypothetical relationship between measured and computed values.

As shown, accuracy and precision may independently be high or low. In general, higher precision means the data have little scatter (less random error), while higher accuracy (low bias error) means the data surround the diagonal of equality, with or without scatter. In terms of modeling noise effects, it may sometimes be preferable to have higher accuracy (little bias), whether or not there is high precision. In Figure 1, the two top panels are usually preferable to either of the two lower panels. Accurate models with scatter (low precision) may still provide reasonable estimates of audibility or sound levels, if used with care – possibly by running many cases or many alternatives that can reduce the scatter. Inaccurate models, however, will give "biased" results that can lead to incorrect decisions, by always over- or under-predicting the sound levels / audibility.

Sometimes, "calibration" can correct bias, and calibration was considered in this study (Section 1.9.2). This type of calibration simply uses the computed bias to alter the model so that the bias is removed. In Figure 1, the model would be altered so that the points are shifted to lie on the diagonal, though the scatter is not altered.

The following subsections summarize the overall error, the accuracy and precision analyses of the models and the corresponding results. For a complete discussion of these results and underlying concepts, see Section 8.

[10] The number of site hours with measured ambient and with measured aircraft L_{eq} differ because there were some locations and hours where ambient levels could be reliably derived while aircraft L_{eq} could not. The aircraft L_{eq} are more difficult to derive from measurements because, though aircraft may be audible during a given hour, their sound levels may be too close to the ambient to accurately separate and determine.

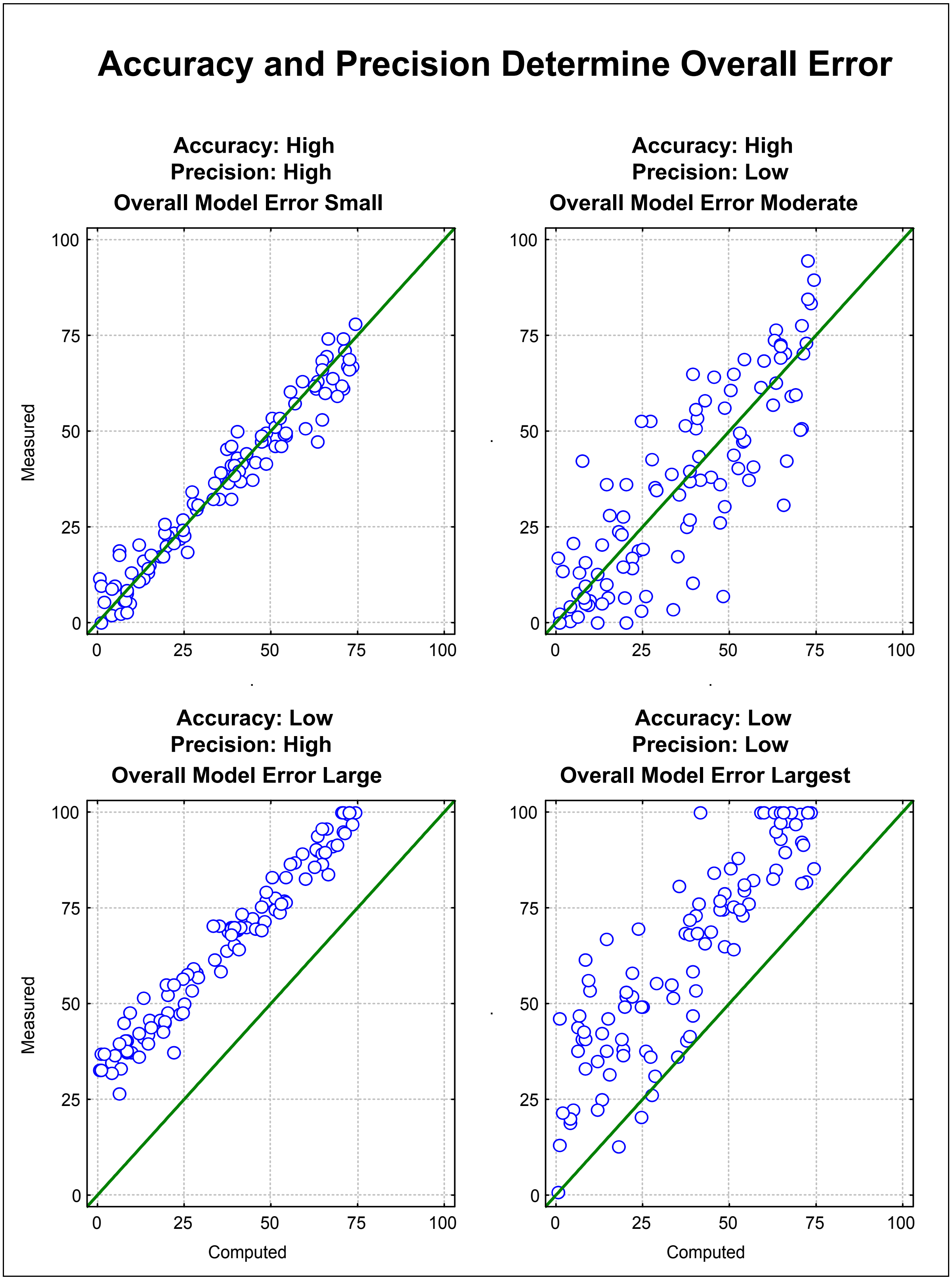

Figure 1. Illustration of Accuracy, Precision, and Overall Model Error

1.9.1.3 Overall Comparisons- Individual Hours

The simplest comparison of measured and computed values is a plot of each data point in the two dimensions of measured and computed results. (See Section 8.4.3.) Figure 2 presents this type of comparison for percent time audible, for the EA ambient for every hour measured at every site. (Figure 2 is Figure 35 from Section 8.4.3.) For these comparisons, scatter is present, and greater for some models than for others, as seen in Figure 2. The greater the scatter, the less is the precision and / or accuracy, and the greater is the error. Points in the figure are coded by location, as indicated. (Figure 22 on page 54 shows site locations, as does APPENDIX D, page 181). Also, the locations coded in these figures refer to the site groupings given in Table 20, page 88.) Figures in report Section 8.4 present the other comparisons listed in Table 1.

One way to quantify the scatter of the plotted points about the diagonal in these figures, and hence to quantify the "overall error," is to sum all the squares of the vertical distances of the points from the diagonal, divide by the number of points, and take the square root of the sum. This type of sum is the root-mean-square error and is the average vertical distance of the points from the diagonal; the squaring avoids negative values, and makes points above and below the diagonal equally important. It is a total combined measure of the accuracy (bias) plus precision (scatter), and is one way to compare model results with measured values.

Accuracy and precision are also computed, and Table 2 through Table 4 summarize all the results of this analysis of overall error, accuracy and precision. Figure 3 through Figure 10 present the same results in graphic form. Note that the figures present only the absolute value of the errors, rather than showing both the positive and negative values, since the errors are symmetrical. The resulting overall error for individual site hour data are presented in the column numbered *1* of Table 2 through Table 4. The results in Table 2 and Table 3 are given in percent of time audible. So, for example, INM (A) using the measured ambient has an overall error of 20 percent time audible as computed for the individual site hours shown in Figure 2 or Figure 35, p.86.[11] The results in Table 4 are in decibels, L_{eq}.

In general, NMSIM has the lowest overall error in percent time audible, whether using measured or EA ambient. Also, except for NODSS, all models have lower error when using the measured ambient. For computations of L_{eq}, both INM versions have lower overall error than either NMSIM or NODSS. Note that NODSS appears to contain some fundamental error in computation of equivalent sound levels.

[11] Note that the errors shown in the tables and figures are not entirely attributable to errors in the models. Measurements also have inherent error, and their effects are estimated in Section 8.4.5.

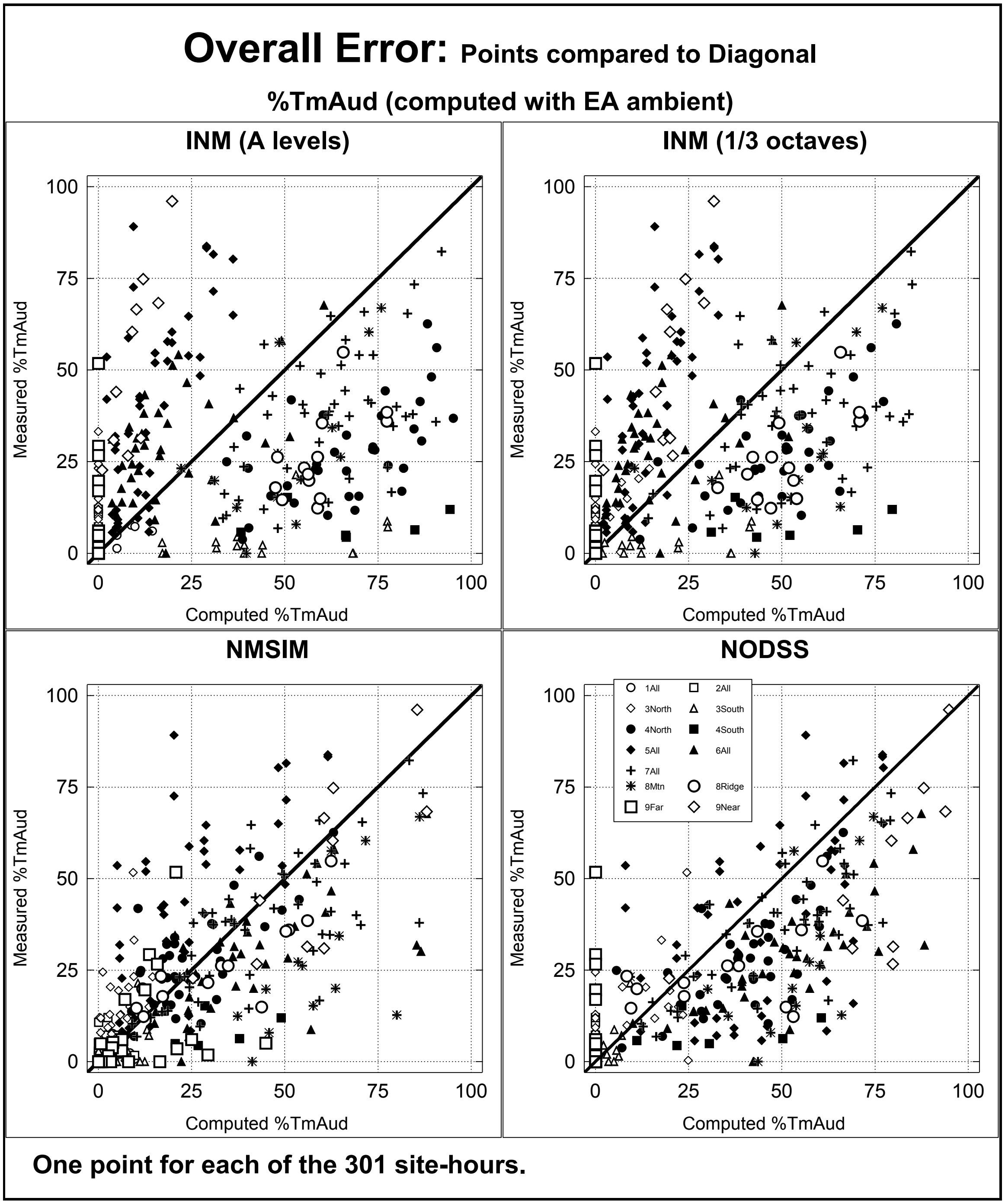

Figure 2. Individual Hours - Measured v. Computed Percent Time Audible, EA Ambient

Table 2. Summary of Error, Accuracy and Precision Results, Percent Time Audible, Measured Ambient

Model	Measured Ambient (All units except correlation coefficient are % time audible)							
	Individual Hours				Site Groups			
	1 Overall Error	*2* Accuracy	*3* Precision		*4* Overall Error	*5* Accuracy	*6* Precision	
		Bias w/ 95% Confidence	Random Error	Correl. Coeff.		Bias w/ 95% Confidence	Random Error	Correl. Coeff.
INM (A)	20	3 ±10	12	0.7	16	1 ±12	12	0.6
INM (⅓ Octave)	19	1 ±8	13	0.6	14	-2 ±10	11	0.6
NMSIM	14	1 ±4	9	0.8	7	-1 ±4	6	0.9
NODSS	22	10 ±6	10	0.7	11	6 ±5	3	0.94

Table 3. Summary of Error, Accuracy and Precision Results, Percent Time Audible, EA Ambient

Model	EA Ambient (All units except correlation coefficient are % time audible)							
	Individual Hours				Site Groups			
	1 Overall Error	*2* Accuracy	*3* Precision		*4* Overall Error	*5* Accuracy	*6* Precision	
		Bias w/ 95% Confidence	Random Error	Correl. Coeff.		Bias w/ 95% Confidence	Random Error	Correl. Coeff.
INM (A)	30	1 ± 17	17	0.3	30	5 ± 17	15	0.2
INM (⅓ Octave)	24	-2 ± 13	16	0.4	22	1 ± 13	14	0.4
NMSIM	17	-1 ± 7	12	0.7	12	2 ± 6	8	0.8
NODSS	20	10 ± 5	9	0.8	15	8 ± 6	5	0.92

Table 4. Summary of Error, Accuracy and Precision Results, Hourly Equivalent Levels

Model	(All units except correlation coefficient are decibels)							
	Individual Hours				Site Groups			
	1 Overall Error	*2* Accuracy	*3* Precision		*4* Overall Error	*5* Accuracy	*6* Precision	
		Bias w/ 95% Confidence	Random Error	Correl. Coeff.		Bias w/ 95% Confidence	Random Error	Correl. Coeff.
INM (A)	7	-2 ± 2	6	0.7	5	-1 ± 3	4	0.9
INM (⅓ Octave)	7	-2 ± 3	6	0.7	5	-1 ± 3	4	0.9
NMSIM	8	-4 ± 2	6	0.7	6	-3 ± 2	3	0.92
NODSS	18	-18 ± 3	4	0.7	19	-26 ± 8	5	0.8

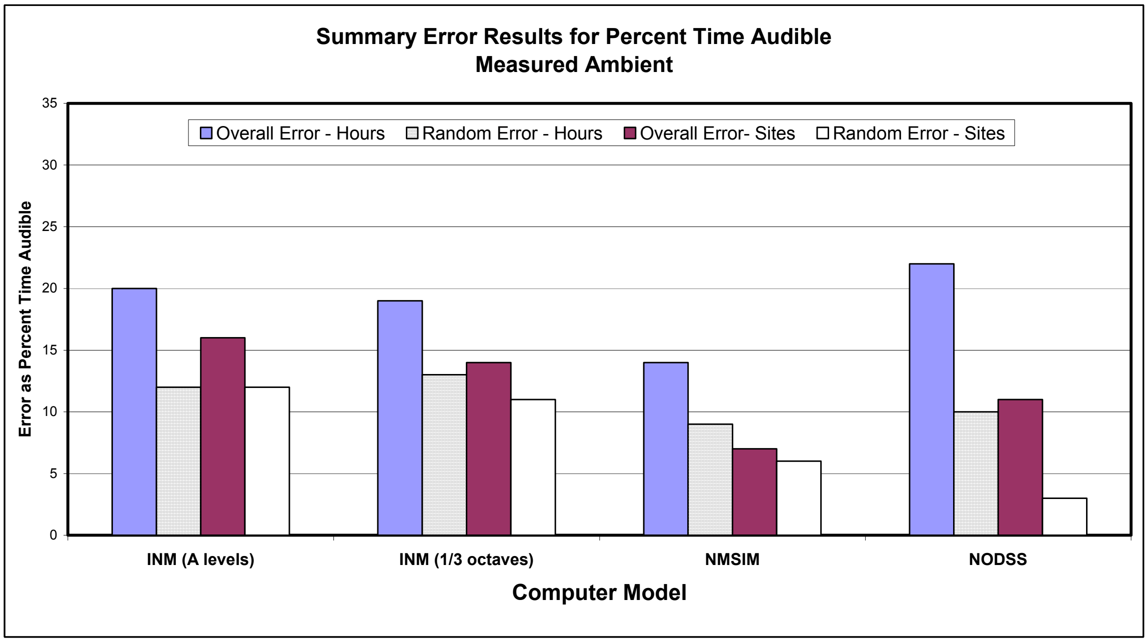

Figure 3. Summary of Error Results for Percent of Time Audible, Measured Ambient

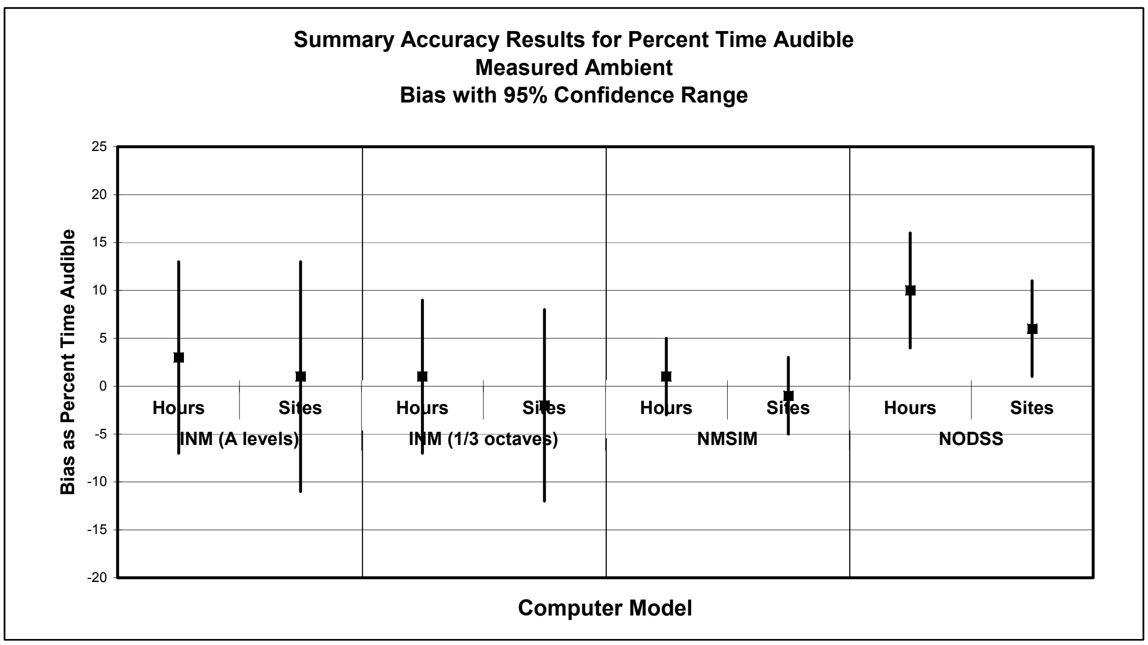

Figure 4. Summary of Accuracy Results for Percent of Time Audible, Measured Ambient

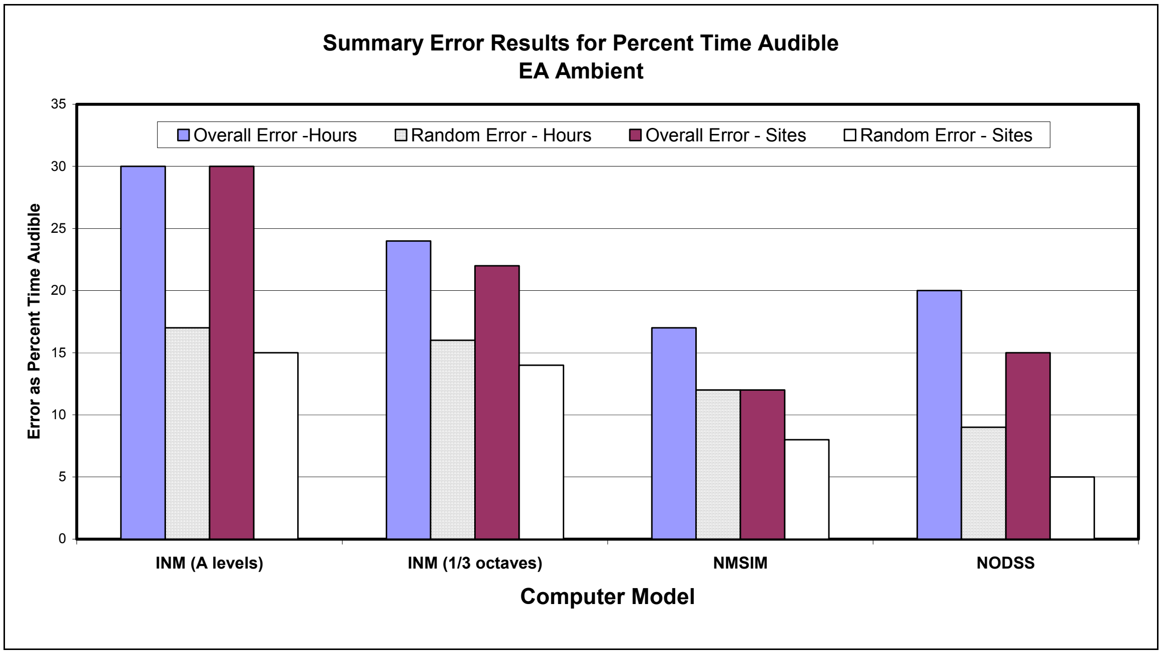

Figure 5. Summary Error Results for Percent of Time Audible, EA Ambient

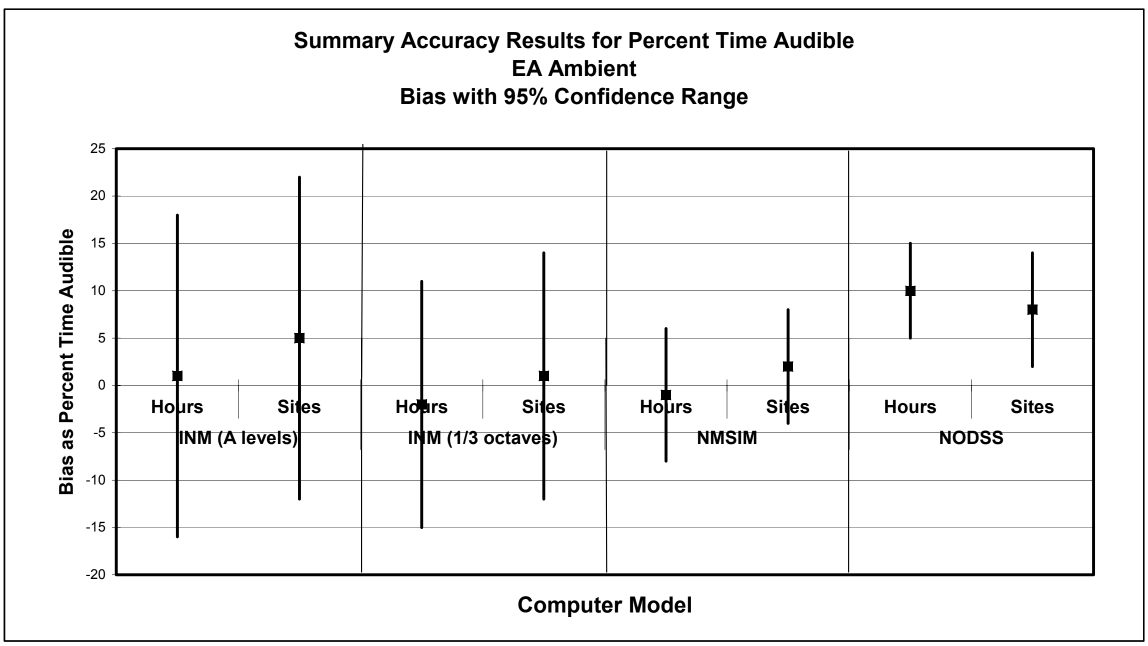

Figure 6. Summary Accuracy Results for Percent of Time Audible, EA Ambient

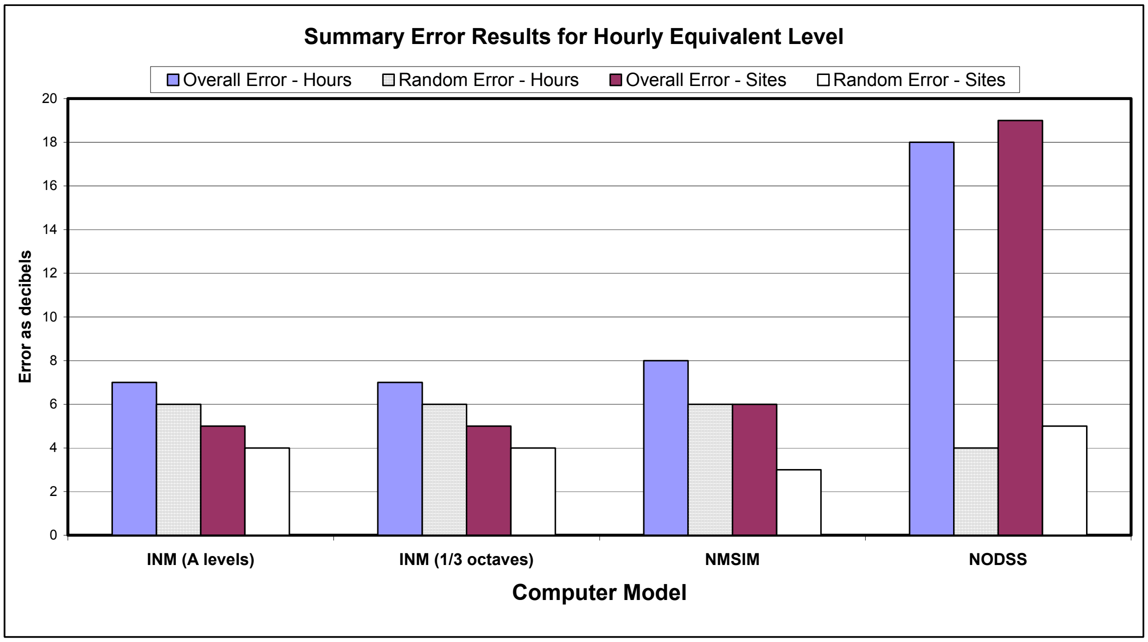

Figure 7. Summary Error Results for Hourly Equivalent Level

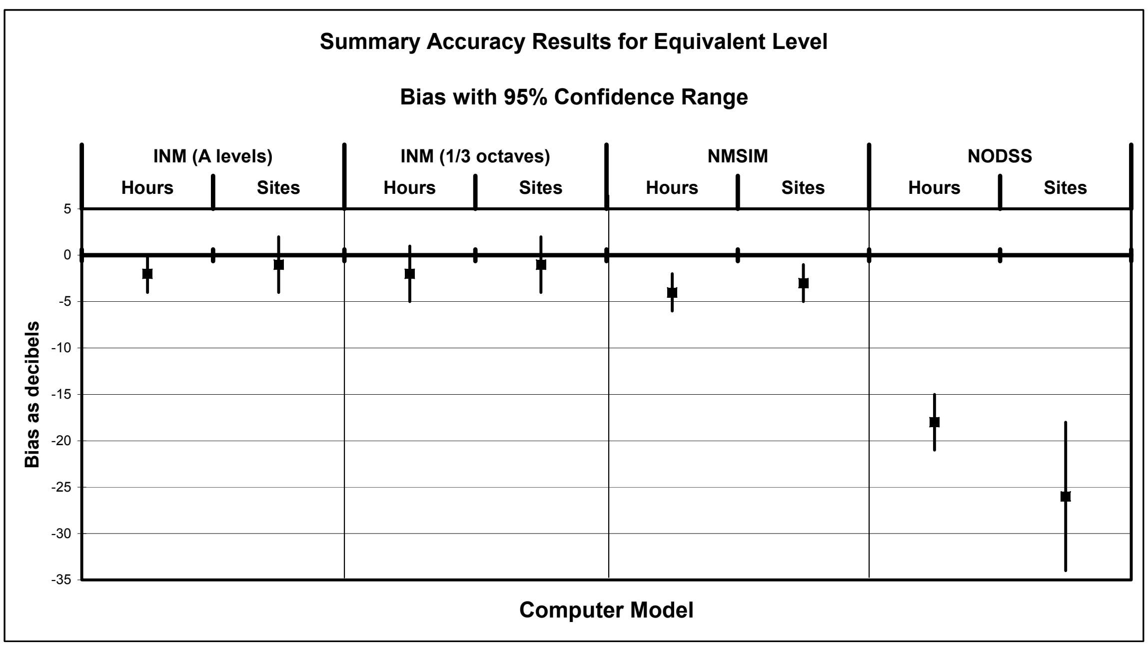

Figure 8. Summary Accuracy Results for Hourly Equivalent Level

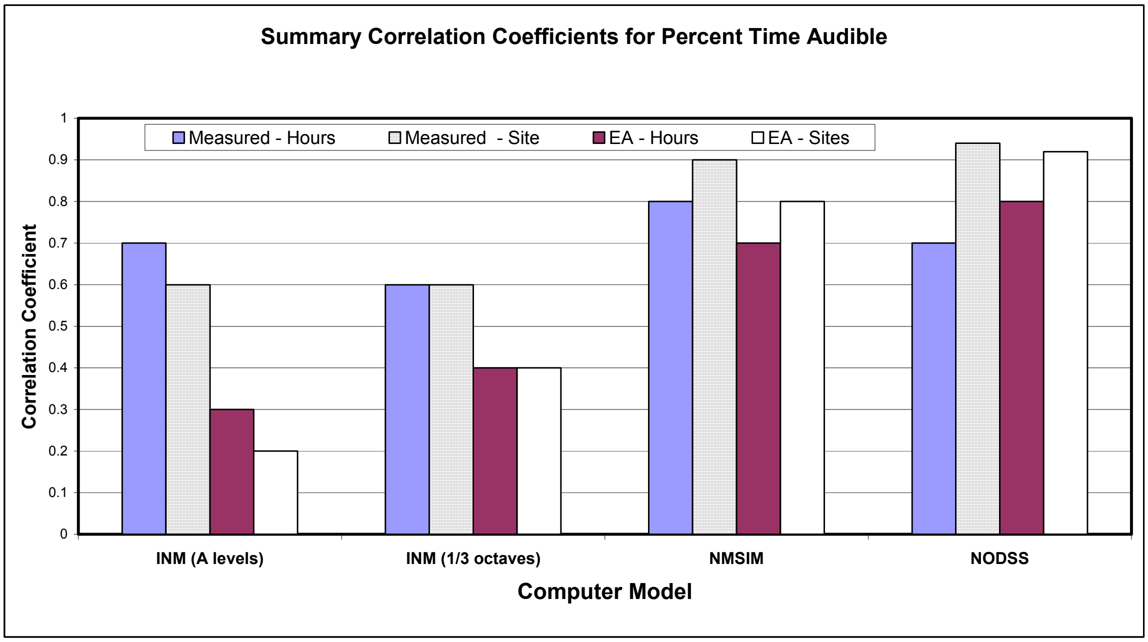

Figure 9. Summary Correlation Coefficients, Percent Time Audible

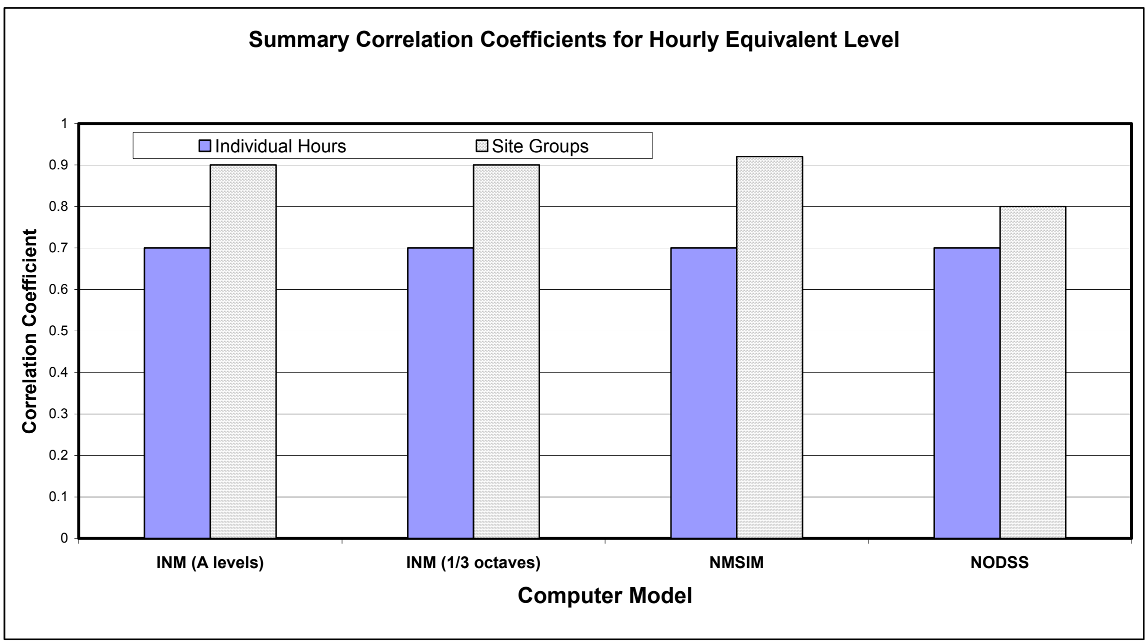

Figure 10. Summary Correlation Coefficients, hourly Equivalent Level

1.9.1.4 Overall Comparisons- Site Groups

Figure 11 plots measured and computed percent time audible for the site groups. (See Section 8.4.4.) By averaging the hourly data by site group to yield single numbers, hour-to-hour variability is eliminated, and to the extent that differences between measured and computed values are a result of this hourly variability, the overall error should be reduced - the plotted points should be closer to the diagonal. Examination of Figure 11 and column *4* of Table 2 through Table 4 shows this reduction does generally occur. For audibility using measured ambient, averaging hours by site group reduces the overall error for all models, as is the case with hourly equivalent levels, except for NODSS. For audibility computed with the EA ambient, averaging hours by site group decreases the overall error for all models except for INM (A) where it is unchanged. As with the individual site hours, these site groups have less error when using the measured ambient than when computed with the EA ambient.

These site group results mean that for contour modeling, the overall error can be substantially reduced by averaging hours together. Contours can be computed by running the model for many different hours, then averaging the results and from these averaged results, deriving contours of equal exposure. This type of averaging reduces or eliminates measured versus computed differences to the extent the differences result from hour-to-hour differences. Large site-to-site variability, however, cannot be corrected in the contouring process.

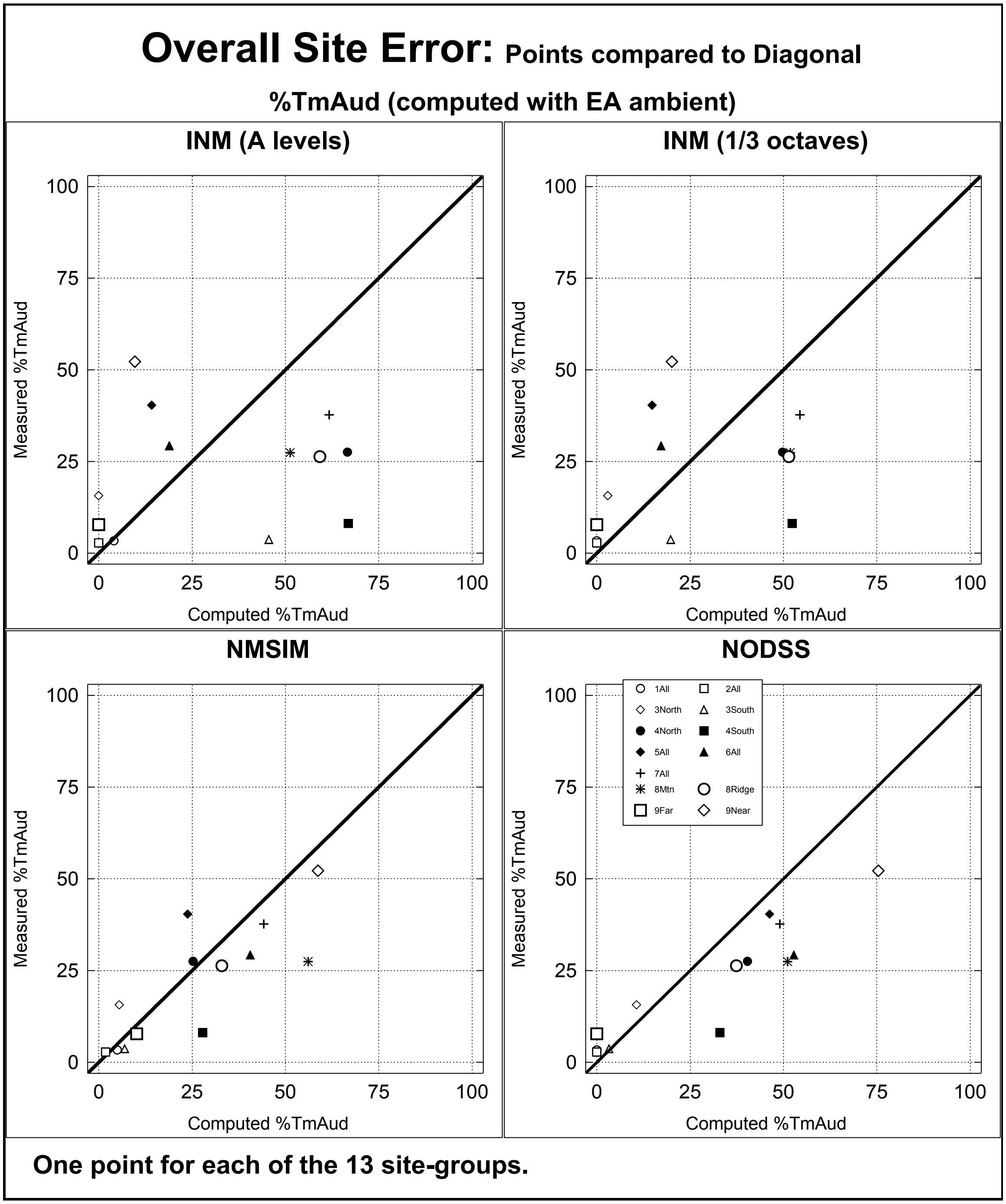

Figure 11. Site Groups – Measured v. Computed Percent Time Audible, EA Ambient

1.9.1.5 Accuracy

Accuracy here is determined in two ways. First, by computing a single number bias and the associated 95% confidence range for that bias. Second, by computing the best fit regression line for the measured versus modeled data, including the regression's 95% confidence regions, and comparing that line and limits to the diagonal line of equality. (See Section 8.5.)

Columns *2* and *5* of Table 2 through Table 4, and Figure 4, Figure 6, and Figure 8 present the results of the first type of accuracy analysis. These tables and figures give the bias value and associated 95% confidence interval. For these comparisons, the closer the bias is to zero, and the smaller the 95% confidence range, the more reliable the model is in computing results that match the measurements. Thus, judging any model's accuracy depends upon two aspects of the results: (1) Does the 95-percent confidence range include zero bias? (2) How wide is the 95-percent confidence range? For percent of time audible, all models except NODSS can compute unbiased results, though NMSIM is more likely to do so than either of the INM versions (NMSIM has a smaller 95% confidence interval). For equivalent levels, both INM versions are likely to compute unbiased results, and NMSIM is likely to compute biased results, but only slightly so. Again, NODSS clearly contains some fundamental error in its computation of equivalent levels.

For the second method, the closer the best-fit line is to the diagonal, the greater the model's accuracy (Section 8.5 presents a full discussion of this analysis and results.) How well the best-fit line matches the diagonal can also be judged by computing the confidence region around the best-fit line. In this analysis, the 95% confidence region is computed. If the collection of measured values were repeated over and over again and each collected set compared with corresponding modeled results, the regression line would lie in this confidence region for 95% of the comparisons.

Figure 12 through Figure 14 present the results of this type of analysis. In these figures the narrow line is the regression line, while the heavy curved lines show the 95% confidence regions. The diagonal heavy line is the line where computed equals measured. Note that the audibility regressions and confidence regions are curved in these figures. This curvature results from the type of regression analysis used, which was chosen due to the nature of the audibility metric. This type of regression analysis guarantees that neither the regression line nor its 95% confidence region are ever less than 0% or greater than 100%. This type of analysis also recognizes that at the limits of this region, all models should be very accurate. That is, for high enough numbers of tour aircraft and/or close enough to the corridor, all models should compute 100% of the time audible; for zero traffic or at very large distances from the corridor, all models should compute 0% of the time audible.[12]

Also note that confidence regions encompassing the diagonal do not necessarily mean the model is accurate. If the confidence regions are wide, and enclose much of the diagonal, the implication is that the model may be unbiased, but there is low confidence that this is so. Conversely, if the confidence regions are narrow and do not enclose the diagonal, the conclusion is that the model is biased with a high degree of certainty.

For audibility, all models have less bias, are more accurate, when the measured ambient levels are used for computations. NMSIM for each case lies closest to the diagonal over the greatest range of values.

[12] Note that because of site locations, neither measurements nor computations produced results of 100% time audible. Hence, there is insufficient data at this high level of time audible to result in a regression line or confidence regions that collapse around 100%.

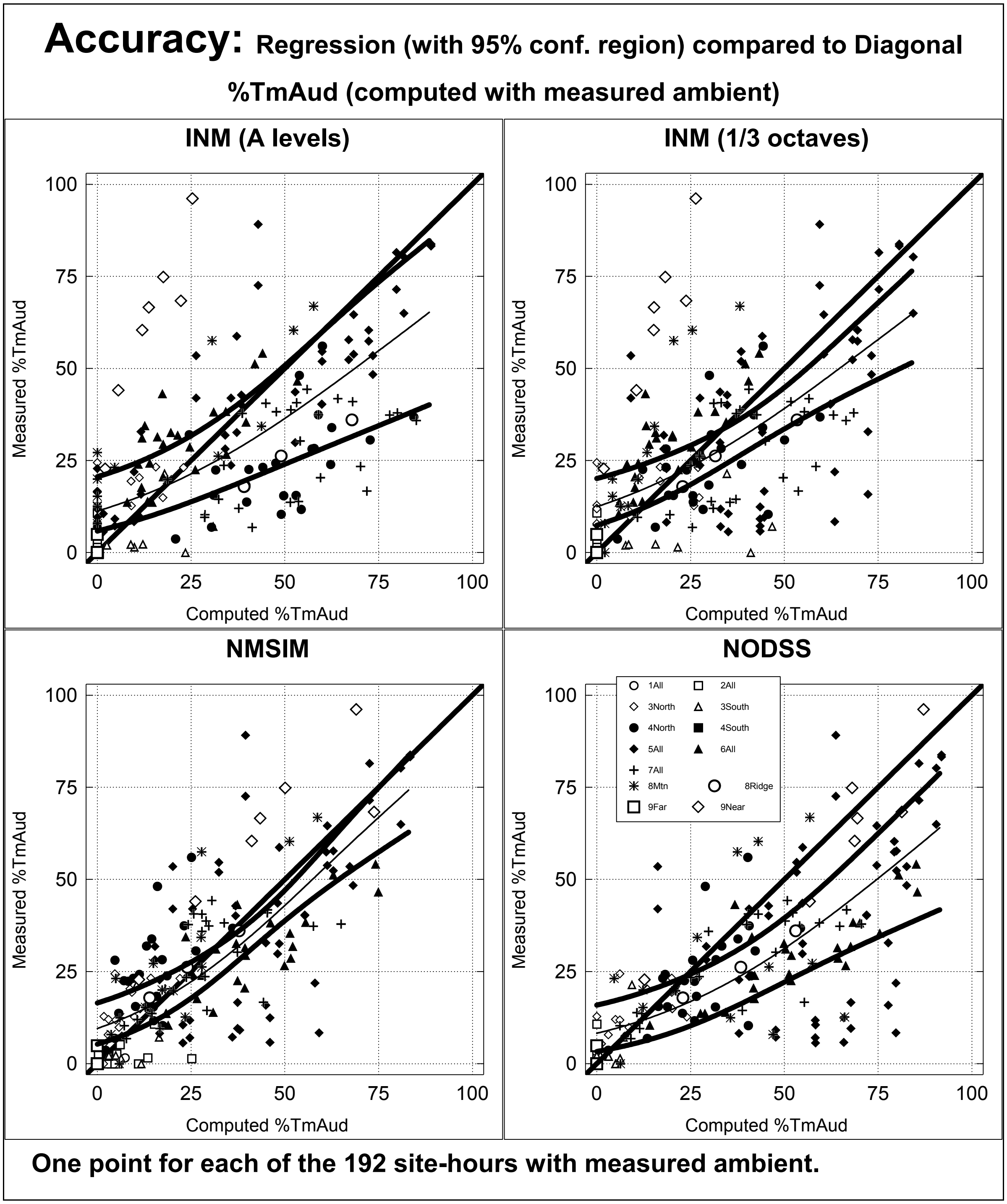

Figure 12. Accuracy – Percent Time Audible, Computed with Measured Ambient

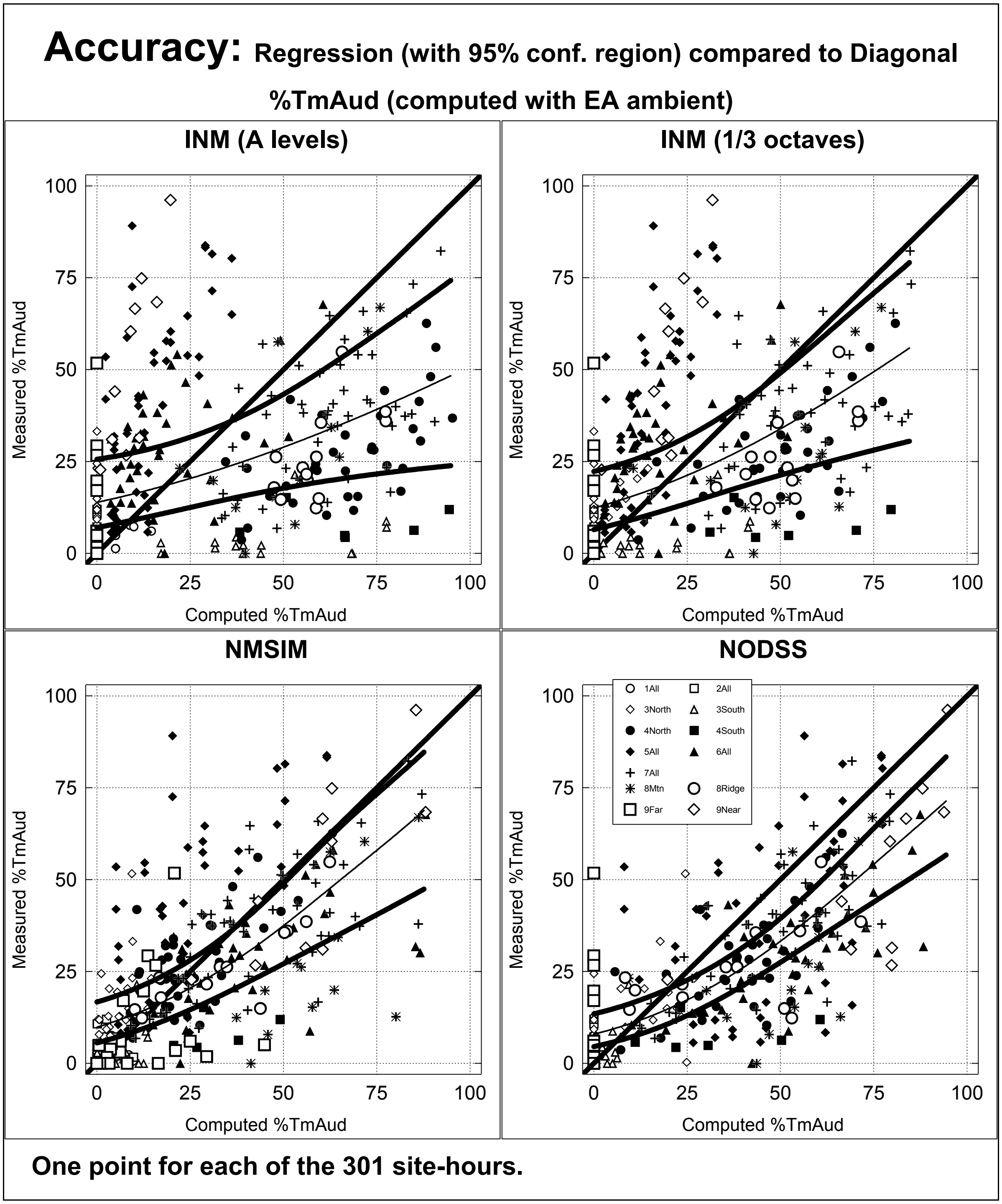

Figure 13. Accuracy: %TmAud, Computed with EA ambient

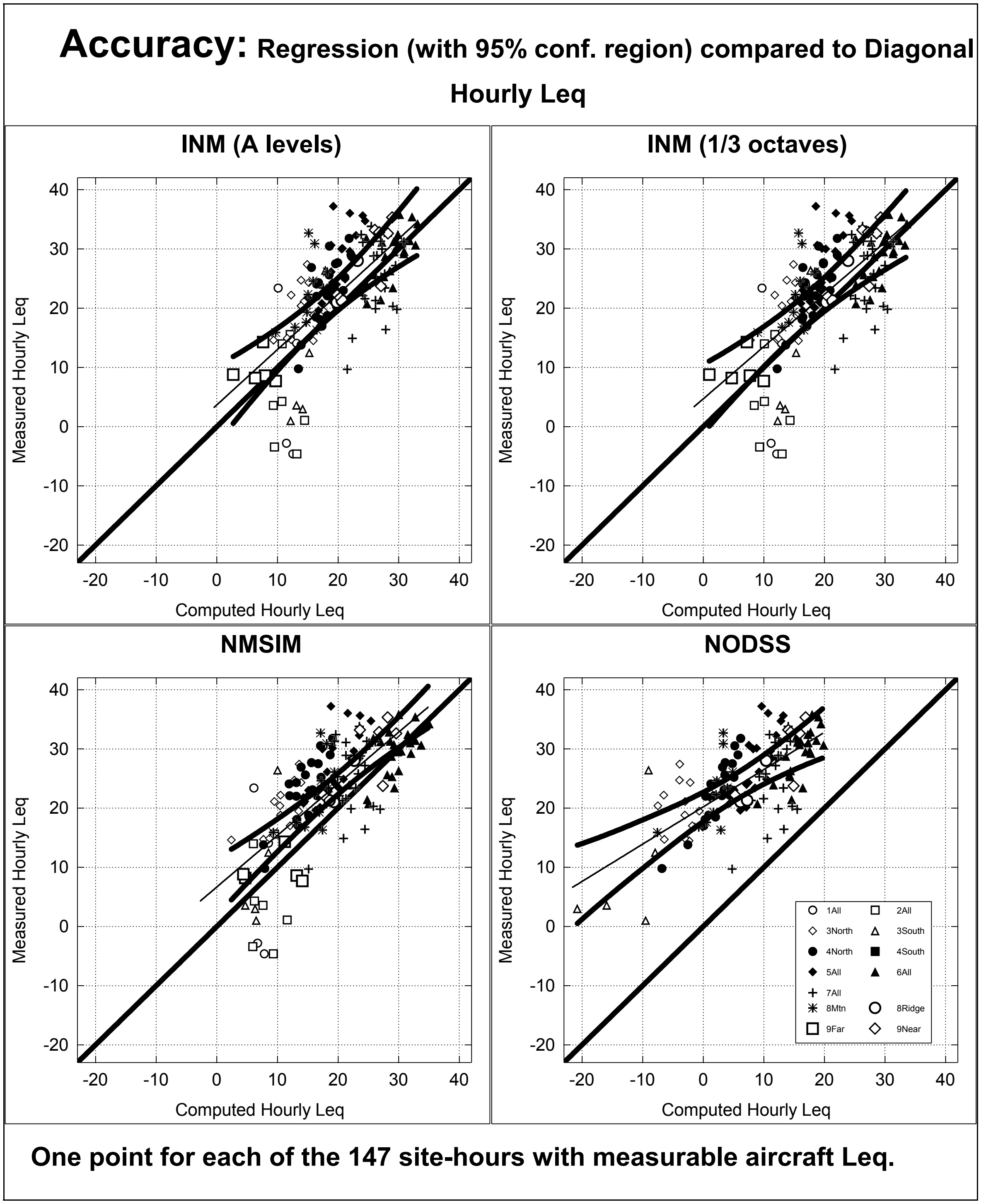

Figure 14. Accuracy – Hourly Equivalent Level

For hourly equivalent levels, both INM versions and NMSIM show similar accuracy though NMSIM shows slightly more bias, and lie closer to the diagonal than does NODSS.

Note that this analysis of model accuracy depends upon both the hour-to-hour correlation and the site-to-site variability in the data. Both of these aspects contribute mathematically to the size of the 95% confidence limits shown in these three figures. Thus, these figures show the complete statistical representation of accuracy. Had hour-to-hour correlation been excluded from the analysis (and each hour treated as completely, but incorrectly independent of all other hours) the 95% confidence limits would have been very much narrower due to the effective increase in number of independent data points.

1.9.1.6 Precision

The discussion of Figure 1 described precision as being a measure of the scatter or random error of the data about the regression line. Two quantities, the root mean square random error about the regression line (computed like the overall error, but as distances from the regression line, rather than distances from the diagonal) and the correlation coefficient quantify this scatter. The larger the random error, the greater is the scatter. The correlation coefficient varies between zero and unity. Correlations close to zero indicate the data are widely scattered about the regression line, and that the regression line cannot represent them very well. A value close to unity occurs when the data very closely approximate the regression line, and that line provides a good generalization of the data. Columns *3* and *6* in Table 2 through Table 4 present these random errors and the correlation coefficients. Figure 9 and Figure 10 also graph the correlation coefficients. (See also Section 8.6, page 117.)

Site groups in all cases, except for NODDS equivalent levels, have random error equal to or less than the random error for the individual site hours. For percent time audible, using measured ambient levels generally reduces this error compared with use of EA ambients. For audibility, NMSIM and NODSS generally have higher correlation coefficients than those for either INM version. For Leq, correlation coefficients are approximately equal across all models.

1.9.1.7 Overall Comparisons - Contours

Because the models will be used to develop a Canyon-wide (or parkwide) depiction of tour aircraft sound in the form of contours – lines of equal percent time audible or of equal equivalent level – it is useful to estimate the error likely to be associated with such contours. (See Section 8.7.) Models determine contours by first computing values at many points, then interpolating the contour locations from these points. In this analysis, in a similar manner, sites were grouped by distance from the corridor, and differences between measured and computed values for each grouping determined. These differences are representative of magnitudes of the differences that would result between contours computed at these distances, and actual measured values. In this analysis, contour error is quantified as the 95% confidence interval on specific contours values.

For each model, for audibility and equivalent sound level, 95% confidence intervals are determined as a function of both distance from the flight corridor, and contour value. Figure 15 through Figure 17 present contour confidence intervals at different site distances from the aircraft track and for different computed values. The contour values and distances are chosen to be representative of audibility percents or equivalent levels that might occur at the identified distances. These figures present examples of what the confidence intervals would be for the selected audibilities, hourly equivalent levels and distances. For audibility, NMSIM has the lowest error, followed by NODSS, and the INM versions having the highest. For hourly L_{eq}, the INM versions and NMSIM have comparable errors to about 7 miles from the corridor, beyond which the INM versions have slightly increasing errors. NODSS has the highest errors for hourly equivalent levels. The values plotted are derived from Figure 56, through Figure 59.

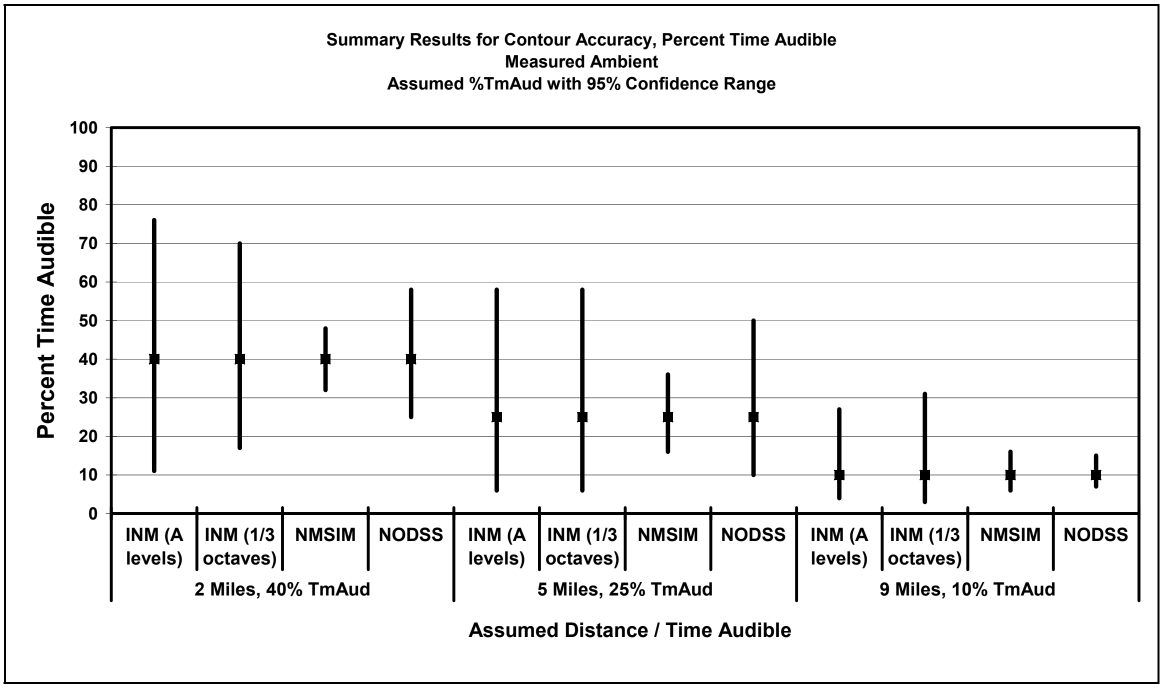

Figure 15. 95% Confidence Intervals for Time Audible Contours, Measured Ambient

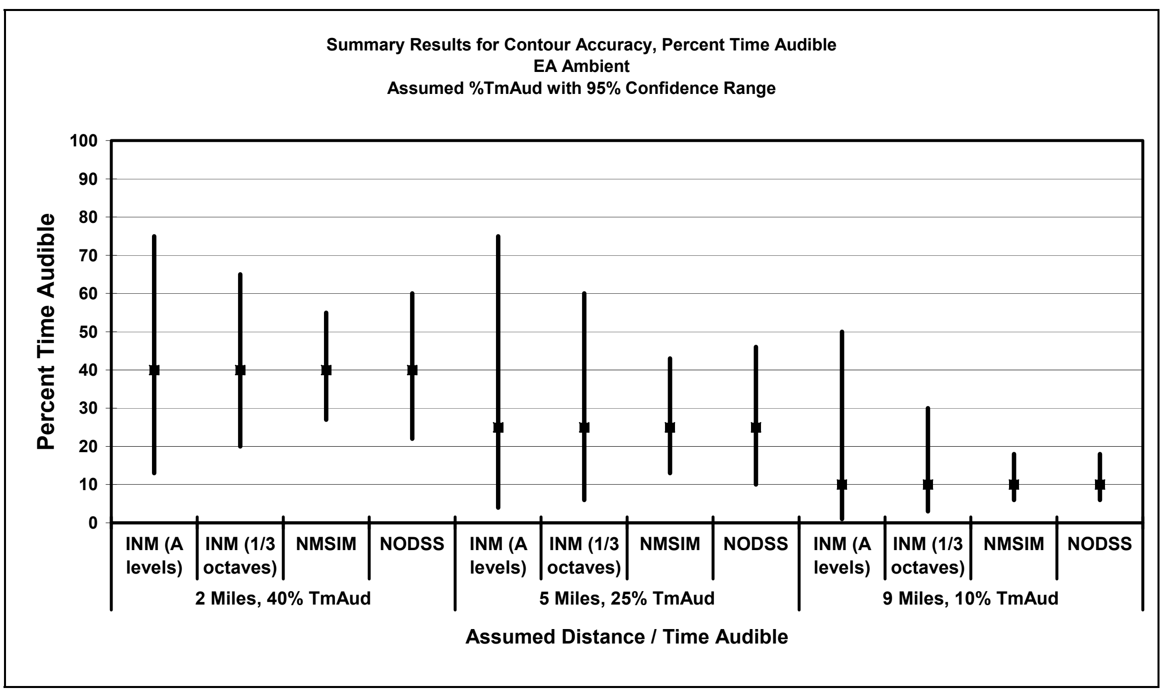

Figure 16. 95% Confidence Intervals for Time Audible Contours, EA Ambient

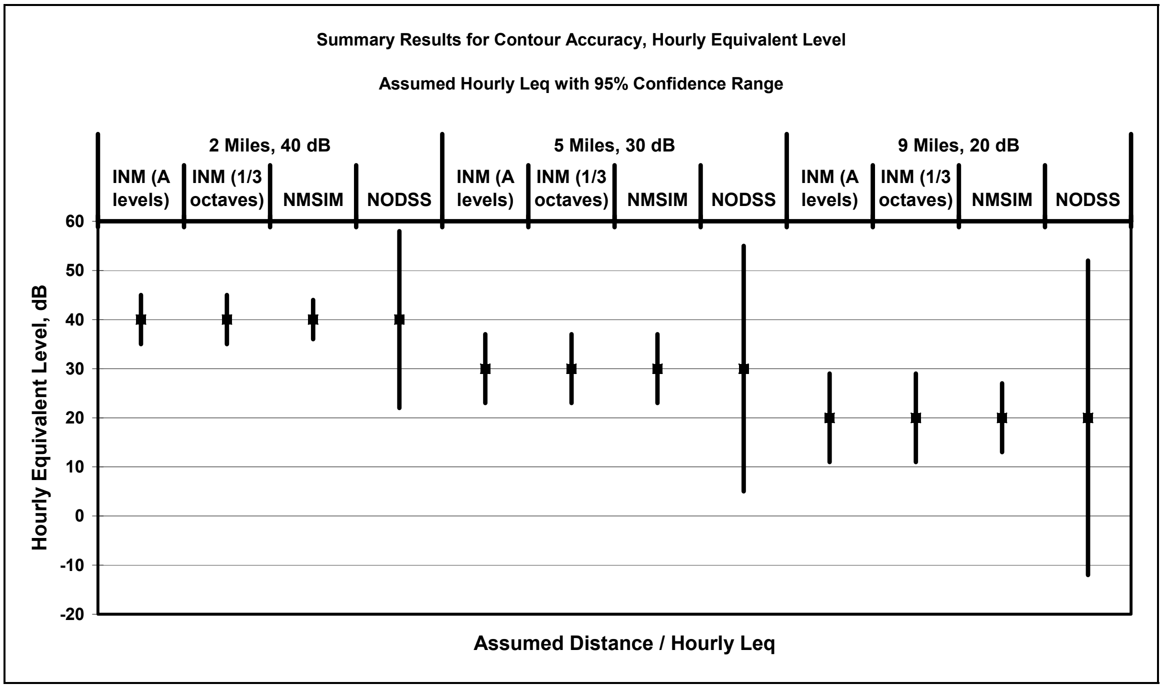

Figure 17. 95% Confidence Intervals for Hourly Equivalent Level Contours

1.9.2 Calibration of Models

Calibration was originally a part of this study's goal. Calibration, as discussed in Section 1.9.1.1 with respect to Figure 1, is the forced removal of bias in a model. However, due (1) in part to some of the models providing what is judged to be reasonable levels of accuracy and precision, but (2) due mainly to the shortcomings of resorting to this type of calibration, calibration is not recommended. This type of calibration must rely solely on the data used and on the model to be calibrated, and takes no account of possible reasons for discrepancies. Hence, a calibrated model provides little certainty that its use for different conditions or for different parks will provide realistic results.[13] It is recommended that rather than resorting to calibration, models be used as they currently are configured, or that improvements be made to the models as appropriate. (Section 1.11.2 or Section 11.2 summarizes the areas of the models suggested for examination and possible improvement.)

1.9.3 Analysis of Discrepancies

Multiple linear regression was conducted to identify which physical factors may be statistically significant (at the 90% level) in relation to the differences between computed and measured values for all models, for both percent of time audible and hourly L_{eq}. (See Section 9.2.) Some eleven

[13] Calibration is often acceptable when it is based on physical reasons. For example, the appropriate value for one of the variables in a model may be unknown, such as sound attenuation due to forests. If measurements are taken in such a way to yield a valid comparison of forest and non-forest attenuation, then the results might be used to quantify the forest attenuation and hence "calibrate" it for forests. Both the INM and NODSS as applied in this study, use a type of calibration. Neither model internally accounts for overlapping sound of closely spaced aircraft; the audibility time for each aircraft is computed independently. To account for this possible over-prediction of audibility, an empirical adjustment was applied to INM and NODSS results (see APPENDIX J page 243).

factors are significant, and the report summarizes these, quantifies the effect on each model's discrepancy, and provides some insights about these results. As discussed in the report, these results should be regarded as inexact, since the analysis forces a linear form on all the relationships. Nevertheless, the results provide useful input into model diagnostics, should model improvement be pursued.

1.9.4 Relationship of Physical Factors to Measured Results

Non-linear regression was used to determine the physical factors that affect measured tour aircraft audibility, and the magnitude of their effects. (See Section 9.3.) The results of this analysis are completely independent of the modeling and are based solely on the measured values obtained in this study. Specific significant results are:

1. The Vistaliner (a specially quieted Twin Otter / DHC6) can reduce audibility - on average, multiply percent time audible by 30% – if only quiet aircraft similar to this are used. (See Table 17, page 70 for a complete list of aircraft types measured.)
2. Terrain shielding is significant, accounting on average for 13dB reduction of sound levels across all measurements; hence, its affect on audibility can be significant.
3. Wind speed and direction can affect audibility from hour to hour, but these effects tend to average out over time.
4. Vertical temperature gradients (decreasing temperature with altitude) reduce tour audibility hour to hour, generally more in the afternoon than in the morning. This effect will *not* average out over time, because it is always a reduction of sound level and audibility.
5. Using the limited data available in this study (7 of the 39 sites), local shielding, due to a local boulder, trees or small cut, cannot be shown to be significant when compared to overall variability due to other factors. Most local shielding, in any case, will have only local effects.

The regression also reveals the effects of using less accurate input in each computer model. In general, ignoring wind speed and direction has little effect on results. On the other hand, using generalized ambients, such as the EA ambients tends to reduce the precision of a model. Also, terrain is significant, and its omission from a model is likely to produce over-prediction of audibility.

1.10 Conclusions – Preferred Models

This section presents the conclusions about the models that the authors draw from the analyses presented in this report. It discusses the preferred models for use and the reasons for our preferences.

We consider NMSIM to be the model most suited for use in computing percent of the time tour aircraft are audible. Either version of the INM is suited for computation of hourly equivalent sound levels, and NMSIM performs almost as well. The following paragraphs review the basis of these recommendations.

1.10.1 Overall Error

For the computation of audibility, NMSIM provides the lowest overall error, whether for the measured or EA ambient, or for the hourly data or the site group data. Additionally, the comparisons of these overall errors for the different ambients and data sets give results that are logical and favorable for use of NMSIM in computations.

The NMSIM overall errors for measured ambients are smaller than for the EA ambients. In these comparisons of measured and computed values, it is useful to keep in mind the differences between the measured and the EA ambients as described in Section 1.9.1.1. The measured ambients were measured at the times when, and at the sites where, the audibility logging was conducted, while the EA ambients are generalized ambients based on earlier data. Hence, use of the EA ambients in computing audibilities should give results similar to those computed using the measured ambients, but with somewhat less accuracy and precision. NMSIM demonstrates this trend.

It is unlikely in future modeling of the Canyon or of other parks that ambient levels will be as widely and thoroughly measured, as were the measured ambients of this study. The ambient levels will have to be generalized from limited measurements.[14] Thus the results using the measured ambients should reveal the "best" that the models can do, given the "best" ambients, while the results using the EA ambients provide what might be considered a more realistic application of the models. The two ambients may be considered as testing the various models' sensitivities to different assumptions about ambient levels, and in this sense can provide additional insight about model performance.

For all models except NODSS, use of measured ambients produces less scatter (less overall error) than use of the EA ambients, and the scatter is in both cases least for NMSIM, and greater for NODSS and for the INM versions.

From this perspective, for audibility, NMSIM provides what we judge to be the best-behaved transition from measured to EA ambient; the data become more scattered, for both hourly and site group data, but still reasonably surround the diagonal of equality. The scatter of the data for the other models changes appreciably from measured ambient to EA ambient, suggesting that the calculations of these other models are more dependent on the specific ambient sound levels that are used.

It is especially desirable that the site group overall error be relatively small. Sites (that is, averages over several hours) are what will generally be used in examining tour operations. First, hour-by-hour operations are unlikely to be known, and in most cases, the goal of modeling will be to examine average operations, rather than the operations of a single specific hour. Second, it is likely that modeling will be used to examine the effects of air tour sounds on specific park locations. Finally, if the model results are to be checked for reasonableness or again validated with measurements, the model with the lowest site error will require the fewest measurement sites. For audibility, NMSIM has the lowest overall site group error.

For computation of hourly Leq, both INM versions have the same and the lowest overall error. Whether for individual hours or for site groups, the INM versions have lower errors than do either NMSIM or NODSS (Table 4). The INM was originally designed primarily for computation of equivalent levels, and the results of this test tend to confirm the versatility of that design for even the complex geometries and terrain of the Canyon.

1.10.2 Accuracy

Audibility

[14] For example, to model the entire Canyon, generalization of the ambients is necessary and one method is provided in APPENDIX F, page199. It would be valuable to rerun each of the models with these generalized ambients to determine how overall error is affected. Such a run would provide a scenario more typical of an actual park application than that provided by using either the measured or EA ambients, see Section 1.11.3.1.

For the single number bias and confidence ranges (Figure 4 and Figure 6), NMSIM has the narrowest confidence ranges that always include zero (no bias), and a bias that is the same as or smaller than that of the other models (except for EA ambient, Site Groups, where its bias is 2±6 and INM (⅓OB) is 1±13). NMSIM is the model most likely to produce unbiased results. Using the best fit regression line and confidence regions (Figure 12, Figure 13), whether for measured or EA ambient, the NMSIM results agree best with measurements – its regression most closely follows the diagonal, and is closest to it, compared with the other models.

Hourly Equivalent Level

For the single number bias and confidence ranges (Figure 8) the INM versions have the smallest bias with 95% confidence ranges that also includes zero. From the regression fit, both INM versions are equally accurate, and NMSIM slightly less so (Figure 14). NODSS is clearly faulty in its calculations of equivalent levels.

1.10.3 Precision

In general, precision comparisons among models behave the same as the comparisons of overall error, discussed above in 1.10.1. NMSIM and NODSS have less random error than the INM versions for all percent time audible comparisons, and INM and NMSIM versions have similar random error for hourly equivalent level. For audibility, NMSIM and NODSS have higher correlation coefficients (meaning the model results lie closer to the regression line – have less scatter) than those of the INM versions. For Leq, the INM versions and NMSIM have similar correlation coefficients, while the NODSS coefficient is lower. The corresponding degrees of scatter may be seen in Figure 11 and in Figure 12 through Figure 14.

1.10.4 Contour error

For many future analyses, one or more of the models will be used to generate contours of equal percent time audible or of equal hourly equivalent level. This analysis estimated the error that is likely to be associated with these contours (see Figure 15 through Figure 17 and Section 8.7, page 119).

Audibility Contours

Since the distance of the contour from the corridor will vary for different corridors, it is desirable for the model's error to be relatively independent of this distance, and as low as possible. NMSIM provides the lowest contour error of the four models, and that error is relatively independent of distance from the corridor.

Hourly Equivalent Contours

Both INM versions and NMSIM compute hourly equivalent level contours with comparable errors. Beyond about 7 miles, the INM error increases to about ± 9 dB to ± 10 dB, while NMSIM error remains at about ± 7 dB, see Section 8.7.5, page126.

1.10.5 Calibration

Calibration was considered as a possible solution for improving the accuracy of the models. However, not only do we believe that current models are sufficiently accurate for application to parks (however see Section 1.11.2 for areas of possible model improvement), but calibration depends entirely on the available data and makes questionable any wider use of the calibrated model for other park applications.

1.11 Recommendations

1.11.1 Recommended Application of Models

This section presents the authors' recommendations about how the various tested models would be used to achieve the most realistic computed values, based on the results of this study. We realize that both NPS and FAA may have their own requirements and criteria for modeling tour aircraft sounds in parks, and these recommendations are made without consideration of such requirements.

1.11.1.1 NMSIM

Of the four models, NMSIM is the most likely to compute realistic values of tour aircraft audibility in the Canyon. It can be used to model air tours throughout the entire Canyon by separately modeling twelve to twenty different hours of tour operations randomly chosen from the tour period of interest. The results of these runs should be averaged together, and then audibility contours computed from the averages. Using more than about 12 hours in this process will maximize the probability that the results are realistic, based on the contour error analysis of Section 8.7.4. That section, and Figure 56 and Figure 57 show that the narrowest confidence limits are achieved when many hours of operations are averaged.[15]

NMSIM may be applied to other parks. Though this study has used only Grand Canyon data, the important features of terrain, distance, number of operations, temperature and wind gradients have been included in the analysis and demonstrated no significant biasing of NMSIM results. However, local park ambients should be used, and some type of reasonableness tests of model results should be included for applications to other parks. Ambient levels used will depend upon judgments of what sound levels are appropriate, likely based either upon what ambient sound levels are intruded upon, or on what ambient sound levels affect air tour audibility.[16] These ambients should be adjusted to account for the effects of the human threshold of audibility (see Section 6.1.5.1, page 67). Note that use of NMSIM requires spectral data for both ambient and aircraft sound levels, including directivity information on the aircraft.

Applications to other parks should include tests for "reasonableness" if not strict validation testing. The type of validation provided in this current study is far too demanding of resources to be practical at additional parks. Rather, we propose that 1) careful measurements be made of any tour aircraft used at the park that were not measured in this study and that those measured levels be included in the modeling process; 2) that sound monitoring together with collection of observer logs be done at

[15] The data show that with increased number of hours averaged, the 95% confidence limits tend to reduce asymptotically, and above about 12 to 15 hours used for the average, these confidence limits are likely to be within a few percent of the minimum, see for example Figure 58. Naturally, the more hours averaged, the narrower the limits, though with diminishing returns. If the variability in the number of tours per hour during the period of interest is higher than encountered in this study (2 tours per hour to 14 tours per hour), it may be useful to average more hours – perhaps a percent of total hours such as 10%.
[16] This model validation analysis used for the measured ambient, the L_{50}s of periods at each site that the observers identified as natural, see Appendix C.3, page 168. It should be noted that future modeling of the entire Grand Canyon might first be preceded by running the model(s) to be used with the ambient levels derived in APPENDIX F, page 199. This run would show how well the models perform with these new generalized ambients. See also Section 1.11.3.1.

several sites exposed to tour aircraft noise, and that these measurements be compared with modeled results. Exact procedures for such measurements and comparisons need to be developed.

NMSIM may also be used to compute hourly equivalent sound levels for tour aircraft over parks, though the INM versions performed slightly better. Proper spectral data are needed for the aircraft, and reasonableness testing is recommended.

1.11.1.2 INM, either version

Either version of the INM can be used to compute realistic hourly equivalent sound levels for tour aircraft over the Canyon and for other parks. As discussed in the previous section, 1.11.1.1, several hours of operations (this study suggests more than 10 to 15 hours, see Figure 59) should be randomly selected from the tour period of interest, run in the model, then averaged and used to determine contours, if appropriate. Or equivalently, for hourly L_{eq}, air traffic can be averaged over many hours and then the model run just once. Proper tour aircraft sound level data are needed and, as with NMSIM, reasonableness testing is recommended when the INM is used for other parks.

1.11.2 Suggested Improvements of Models

Analysis of how physical factors (such as wind speed and direction, ambient levels, etc.) relate to differences between measured and modeled results, as well as analysis of how these factors relate to the measured results, helps to identify which factors may produce model error. Such factors are candidates for inclusion or for further examination in the model. The following suggestions are offered by the authors as initial areas to investigate for improvement and are based on the results of these analyses.

1.11.2.1 NMSIM

►NMSIM currently does not account for additional attenuation that may result from heavily forested areas. Further development of NMSIM should consider how this additional attenuation, could be included in the model. The analysis showed NMSIM tends to over-predict for these forested areas. This type of attenuation is likely more important for computation of percent time audible than for hourly equivalent levels.

►NMSIM shows a slight bias toward under-prediction of equivalent levels. This under-prediction does not appear to be a result of wind or temperature gradients. Examination of single event sound levels may suggest some possible causes.

►NMSIM generally under-predicts audibility for the "9Near"[17] sites. These sites are about the same distance from the corridor as the 6 and 7 sites, which are not under-predicted. Possibly, the complex flight tracks near the 9Near sites affect NMSIM computations adversely.

1.11.2.2 INM Models

►Both INM versions compute zero percent time audible for the "9Far"[18] sites when tour aircraft were audible, which suggests these models might be improved through examination of their: 1) assumptions for long-distance propagation, since both models apparently predict levels so low at

[17] 9Near sites are sites 9C and 9F, Table 20, page 88, which are about 2 miles from the corridor (see Figure 22, page 54 and Table 11, page 55).
[18] 9FAR sites are sites 9A, 9B, 9D and 9E, Table 20, which were 11 to 15 miles from the flight corridor (see Figure 22, page 54 and Table 11, page 55).

these distances that they are determined to be inaudible, 2) computation of audibility when aircraft sound levels are low, and 3) computation when only a small portion of a flight track contributes to the sound levels.[19]

►Both INM versions uniformly underestimate time audible at 9Near sites, suggesting that how these models treat curved flight tracks might be examined, since these sites are likely to receive sound from several portions of the track that curves into and out of the Little Colorado.

►The INM over-predicts audibility close to the corridor (within 0 – 6 miles) when shielding is present (visible angle is small), but under-predicts at these distances when little shielding is present (visible angle is large), see Figure 44, page 105 and Figure 45, 106. The former result is likely due to the fact that the INM does not account for the shielding effects of terrain, while the latter effect may be the result of how the model treats the various parameters associated with audibility, such as the source directivity assumptions. Hence, inclusion of terrain shielding should be considered. Also, for the INM 1/3 octave band version, the components of audibility calculations, especially source directivity should be examined.

►As with NMSIM, the INM versions do not include attenuation of tour aircraft sound levels due to expanses of forested areas. The analysis showed that the INM ⅓ octave band model tends to over-predict audibility for these areas.

1.11.2.3 NODSS

►NODSS computations of equivalent levels should be examined. All NODSS results show a clear bias toward under-prediction of hourly equivalent results. NODSS was designed to compute *total* hourly equivalent level, including the contribution of the natural ambient. Since such results would not provide an appropriate comparison with measured results, NODSS input was modified, see Section 3.4.2. This modification may have caused the significant under-prediction of computed equivalent levels, though currently, no explanation has been determined.

►NODSS also appears to over-predict audibility in the forested areas. Inclusion of adding this type of attenuation should be considered.

►NODSS computes zero audibility for the distant 9Far sites where aircraft were audible. As with the INM versions, reasons for this under-prediction should be examined.

1.11.2.4 Factor Not Recommended for Inclusion

One factor that may have some significance, vertical temperature gradient, is not recommended for inclusion in any of the models. Though absence of this factor in the models could result in some over-prediction at large distances, particularly with respect to equivalent levels, see Figure 47 through Figure 49, the complex relationships between this factor, distance and terrain shielding makes derivation of the exact importance of this factor virtually impossible with the current data. Moreover, in terms of audibility, all models tend towards slight under prediction at these larger distances, so that the net effect of temperature gradient as evidenced by the available data suggests that temperature gradient did not have a dominant effect on the measured results. Finally, acquisition of temperature gradient information for incorporation in future modeling, whether at the Canyon or other parks, is likely to be well beyond the resources available for data collection.

[19] From the location of 9Far sites, the flight corridor would subtend a relatively small angle, less than 45 degrees.

1.11.3 Suggested Possible Further Analysis

1.11.3.1 Run Models Using Generalized Ambients

New generalized ambient levels have been developed that can be used throughout the Canyon.[20] APPENDIX F, page 199, provides the derivation of these ambients. At the discretion of NPS / FAA, these values may be used first to run any of the models that will be used to compute audibility Canyon wide. Results would be compared with measured audibilities, and model performance determined. Such application will provide a realistic assessment of how well the models perform when carefully measured, but generalized ambients are used.[21] This approach recognizes that ambients similar to the "measured ambients" used in this study will rarely, if ever, be available for modeling purposes. After this run and analysis, model performance using the generalized ambients will be known.

1.11.3.2 Additional Analysis of Quiet Aircraft

One of the primary reasons for conducting the regression analysis of the measured results was to determine whether quieter aircraft, such as the Vistaliner (a specially quieted Twin Otter / DHC6 using Raisbeck designed modifications to the fuselage and quiet propellers) could have a statistically measurable effect on tour aircraft audibility (see Section 9.3). The analysis shows that the Vistaliner audibility was, on average, 30% that of other tour aircraft. (See Table 17, page 70 for a complete list of aircraft types measured.) If aircraft like the Vistaliner replaced the other aircraft measured here, they would very significantly reduce audibility of tour aircraft in the Canyon.

FAA has a congressional mandate to identify "quiet technology" aircraft that could be used as tour aircraft. Congress has designated such quiet aircraft as eligible for special consideration in use of tour routes over national parks.[22] It is possible that the resulting FAA research efforts on quiet technology aircraft could benefit from further detailed analysis of this study's data to determine whether the tour aircraft types might be rank-ordered by their relative contributions to audibility. Such rankings might be useful in FAA's efforts to define "quiet technology" as required by law.

1.11.3.3 Model Testing Procedure

The National Parks Air Tour Management Act of 2000 establishes a public process for development of Air Tour Management Plans (ATMP's). It is likely that the ATMP development process will require modeling of tour aircraft at other National Parks. For these applications, it will be useful if some basic procedures are defined for testing the reasonableness of the modeled results for the park under examination. Procedures would include methods for measurement of aircraft types not already measured for the present study, and collection of data for comparison with model results.

1.11.3.4 Computation of Parkwide Metric Error

It is likely that any single-number, parkwide impact metric will have considerably lower overall error than the values reported here for hourly or site error. A parkwide metric is an average over a large

[20] These generalized ambients are similar in concept to the EA ambients; they apply throughout the Canyon based on vegetation zone. These new generalized ambients, however, are derived from the data acquired as part of this study, unlike the EA ambients that were derived from previous measurements, see references in footnote 43 page 59.

[21] The authors judge, however, that the performance would likely be between that found in this study using the measured ambients and that found using the EA ambient.

[22] Title VIII of Public Law 106-181, National Parks Air Tour Management Act of 2000.

number of computed sites throughout a park. Hence, it can average out both the hour-to-hour error, and the site-to-site error reported here. Over-predicted sites tend to balance out under-predicted ones.

An important single-number parkwide metric in the Canyon is the computed fraction of land area where tour aircraft are audible more than 25% of the time. This study's computer programs could be used to compute this value, and this study's results then used to determine the confidence interval of this single-number metric of impact. We suggest that this computation of error be done, due to the importance of this metric in determining restoration of natural quiet in the Canyon. By applying propagation of error techniques (mathematical methods that combine the uncertainties of multiple factors into the resulting uncertainty of a single function of those factors) to the site errors for each model, it would be possible to estimate the error associated with a model-computed area exposed to tour audibility in more than 25% of the time.

1.11.3.5 Use Measured Data to Test Detection Algorithms

The measured data (which includes second-by-second 1/3 octave band levels and associated second-by-second observer logs) represents virtually the best data source possible for testing automated identification of "natural" and "aircraft" sound levels. Ultimately, most sound measurements in parks will probably need to be collected with unattended, long-term monitoring. It will be extremely advantageous if these unattended data can be reliably used to quickly determine the sound levels of the natural ambient and the number and sound levels of intrusions. The measured data collected for this study provide the means for testing and checking the reliability of various detection algorithms with respect to human determination of audibility.

1.11.3.6 Rerun NMSIM with Equally or Randomly Spaced Aircraft

For the study, NMSIM ran the aircraft flights with the actual timings that they flew. In modeling of future studies at other parks, the exact timing and spacing of tours will probably not be known. The model could be run with aircraft at equal spacings and at random spacings to determine the magnitude of the error such approximations can produce. These runs could also help determine how best to select tour aircraft spacings for modeling when the actual spacings are unknown.

1.11.3.7 Revise "Compression" Algorithm

Neither the INM versions nor NODSS account for the overlapping of aircraft audibility when aircraft fly in close succession. These models compute the audibility duration for each aircraft separately, and then add all durations together. Such an approach will over-predict total audibility when aircraft fly close enough to result in audibility of more than one aircraft at a time. A "compression" algorithm was derived empirically from previous measurements to reasonably reduce these computed audibilities, see APPENDIX J, page 243. The data from this current study could be used to develop an up-dated compression algorithm that might be applicable to more situations and more parks.

Page Intentionally Blank