

**ASSESSMENT OF DATA QUALITY OF SELECTED DATA SETS IN THE
DEPARTMENT OF ENERGY/COMPREHENSIVE NUCLEAR TEST-BAN-
TREATY KNOWLEDGE BASE**

D. N. Hagedorn, C. A. LoPresti, R. F. O'Brien, S. A. Hartley, B. G. Amidan
Pacific Northwest National Laboratory

Sponsored by U.S. Department of Energy
Office of Nonproliferation and National Security
Office of Research and Development
Contract No. DE-AC06-76RLO 1830

ABSTRACT

The U.S. Department of Energy's (DOE) Comprehensive Nuclear Test-Ban-Treaty (CTBT) Knowledge Base (KB) contains detailed regional data, from which corrections to seismic, hydroacoustic, and infrasonic signals will be generated. As the KB is populated with information and data sets detailing regional geological and geophysical structures and reference event data, questions of "How good is the information?" and "What confidence can I have in the corrections?" arise. This report documents work to-date at the Pacific Northwest National Laboratory on the development of a "toolbox" of statistically-based algorithms which may be used to assess the quality of individual data sets, and consistency across multiple data sets, both on data in the KB and prior to including new data in the KB.

Thirteen data sets (consisting of metadata, header, projection, and data files), supplied by Sandia National Laboratories, from the KB were used in this preliminary examination. The metadata files were reviewed before analysis of the data sets began. We noted that some fields were not filled in, others had very brief entries, while yet others were quite complete and informative. Comparing metadata files across data sets, we noted that the quality of the information was not consistent, there were problems with the accuracy/precision of the numerical data, processing audit trails were poor to nonexistent, and when compared to the headers in the data files, some discrepancies were noted.

Several methods were employed to evaluate the individual data sets for spurious data. We believe that because these gridded data sets in the KB are composites created from multiple sources and have been processed and smoothed, no outlier data was found. We did discover in some of the data sets that areas of constant value (algorithm default values) existed, which are not geologically reasonable. We believe that these areas were created as a result of the processing and are not valid data.

Problems with agreement between data sets were also identified. Comparing data sets was problematic because of the different grid sizes and cell locations. As an example of what can be done to evaluate agreement between data sets, we examined three Mohorovicic (Moho) Discontinuity depth maps, which contained regions in common to two or all three data sets. Depths to, and trends in, the Moho in the data sets did not agree with each other, and in some instances depths at the same location had differences of up to 20 km.

These are all serious problems, which must be rectified prior to using this KB data to generate seismic corrections. The effects of the noted discrepancies on the corrections have not yet been assessed, but we believe that corrections derived from using the different data sets would be significantly different. The cumulative effects of multiple errors could be more drastic. Certain needs for the overall KB were also identified, such as a need for well-defined criteria for accuracy of data, estimates of uncertainty to be associated with individual data, a consistent schema for gridding and combining data sets, and tolerance limits for agreement between data sets.

Treatment of uncertainty in the data and understanding the effects of that error on the correction estimates that result from using the KB data is essential. The data sets from the KB used in this study are not raw data. They have been generated from multiple sources, processed, interpolated, and smoothed, and have no estimates of error or uncertainty. Uncertainty estimates cannot be confidently derived from processed and smoothed data, as from raw data. Any estimates of error and uncertainty will need to be inferred from ground-truth events.

Key Words: Knowledge Base, Data Quality

OBJECTIVE

This work seeks to develop Matlab-based tools, in the form of an easily accessed Matlab toolbox, which can be used to assess the overall quality of the data contained in the KB and data destined for inclusion in the KB. These tools can be used both to assess the confidence in the present KB data and to make certain that future additions or “loads” (i.e., additions to or replacement of data) to the KB are both consistent with the contained data and have an understandable effect on future interpretations based on the KB data. Additionally, this 'toolbox' will facilitate the identification of areas that would benefit most from efforts to gather additional high-quality data for inclusion in the KB. This work benefits the CTBT monitoring program by building confidence in the KB, allowing identification of problem areas with poor/little data, and supporting efforts to optimize additional data collections that will enhance the KB.

RESEARCH ACCOMPLISHED

To accomplish the development of the Data Quality Assessment Toolbox (DQAT), PNNL is performing evaluations on 13 data sets currently in the KB. Data presently in the KB were obtained from many diverse sources and some data are of unknown quality. The KB data include geological and geophysical measurements and maps, seismic records, topological data, political boundaries, metadata used to qualify “hard” data, and other types of data. Multiple data sets, containing similar and overlapping data, are also a part of the KB. For this preliminary report, we examined both the individual data sets for obvious or hidden problems (inconsistencies, outlier data, mislocation errors, etc.) and the degree to which these data exhibit reasonable agreement.

Data are of two primary types: gridded data and vector data. The gridded data represent three-dimensional surfaces (i.e., contour maps), while the vector data represent two-dimensional features, such as boundary descriptions, rivers, and, in some instances, single-point features. PNNL requested representative data and SNL delivered 13 data sets: 10 gridded and 3 vector. The data sets cover the MidEast/North Africa (MENA) and/or central Asia. The data sets investigated are listed in Table 1. (Note: All references to data sets in this report are noted by use of capital letters.)

Discussion of Metadata

The metadata were examined for internal inconsistencies and for inconsistencies with the headers in the ASCII data, as well as for erroneous entries in the text. The metadata are a very important component of the KB, containing information about where the data came from, the processing that has occurred, the reference list, the geographic boundary data, the accuracy information, and other information.

Examination of the metadata files and cross-checking that information with the information contained in the headers of the associated data files resulted in the following problems being revealed and many questions being raised:

- differences in precision - In the boundary data, some are integers, some have eight decimal places, some three, and some one. Consistency within a single data set is often not present. Is precision to eight decimal places to be believed (i.e., to the nearest millimeter)?
- differences in overall quality - Original data were of varying quality and were often digitized from maps (already processed and smoothed), combined with data from various other sources in poorly-documented ways, and smoothed again. The resulting grid was then resampled to create the data sets received by PNNL.
- metadata not consistently filled in - Some descriptions are very complete, others are less so, and some are totally blank.
- interpolation of gridded data - All gridded data are interpolated (according to the associated metadata files); some are smoothed to a greater degree than others.
- data discrepancies with headers - One gridded set (Bouguer gravity for MENA) uses a different geodetic datum. This is contradicted by information in the header of the data set. Also, a number of columns of data in Bouguer gravity for the MENA dataset (BOUGUER) disagree with information in the header of the data file (398 vs. 460).
- incorrect interval date – The beginning interval date is wrong in the GVOLCANOES data set.

- missing data-acquisition dates - Dates referencing when original data were acquired are usually missing. For some data, this is quite important, since measurement techniques have vastly improved over the years.
- missing references - References are missing for two Cornell data sets (CORN_BASM, and MOHO).

Some of the fields lacking from the metadata files are, for our work, critical. These include locations of the raw data, estimates of the uncertainty of the raw data, contribution to the uncertainty by the processing stream, descriptive statements about how the data was processed (the steps and algorithms used), and the models used to create the final data values. Some of the fields contained in the metadata are essentially useless to the end-user from a data quality perspective.

Investigation of Individual Data Sets

Since most of the provided data sets were developed by combining and smoothing other, possibly derived, sets of data, extensive analysis of each set of data yields little information. Lack of access to the original raw data prevents this evaluation from tracing anomalies back to their sources, and thus can not be attributed to the original source or to processing, modeling, or other causes. The data sets used in this study are derived from other data sets, which may be composed of multiple sources and are processed and smoothed. Furthermore, the criteria for transforming and/or deriving each raw data set, as well as combining data sets, are unknown; therefore, the accuracy, representiveness, uncertainty, and quality of each data set are unknown.

Several types of anomalous data detection operators were applied and very little outlier data was found. This is attributed to the fact that these data have been multiply processed and smoothed, thus eliminating spurious values. One interesting observation did emerge from examination of the Bouguer gravity map: areas with constant values were discovered. We have concluded that these areas are not geologically-feasible, and are artifacts of processing and should be removed from the data sets.

Comparison of Data Sets

One problem encountered in analyzing the gridded data sets for consistency appeared when using Matlab to overlay the data with coastlines and political boundaries. The two Matlab vector data sets are not identical. While this is not an apparent problem with the gridded data sets from the KB currently under evaluation, the situation is exactly analogous to problems contained in the KB, according to reports from SNL. PNNL plans to address the KB problem in future work.

From our initial overview examination we noted that of the 13 data sets, three contained depth-to-Moho data, and there was a single area in common to all three (MOHO, COL_MOHO, and CORN_MOHO). This was an ideal area to evaluate the 'fit', or agreement, of the data in the KB. Figure 1 illustrates the geographic extent and overlap of the three Moho depth data sets.

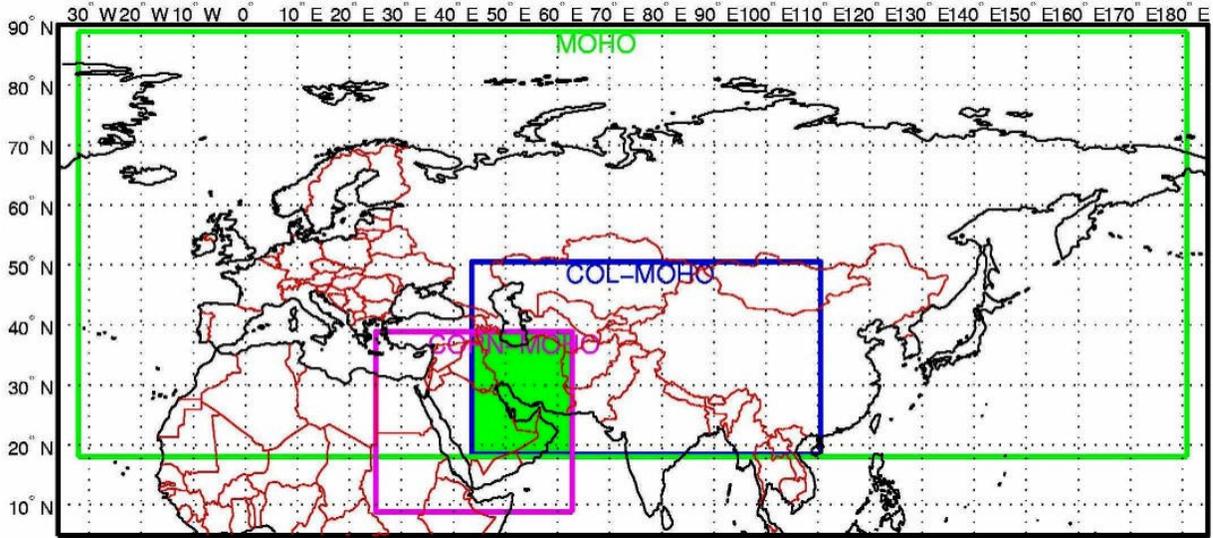


Figure 1. Geographic areas covered by data in MOHO, COL_MOHO, and CORN_MOHO depth-to-Moho data sets.

The first problem in trying to compare the different data sets was that each data set had different cell sizes and cell locations in the map system. Two of the data sets, MOHO and CORN_MOHO, had 0.0920762274019710° and 0.09° square cells, respectively. The third data set, COL_MOHO, was sampled on a 1° square cell size. This means that there are roughly 121 cells from each MOHO and CORN_MOHO grid that refer to the same area as one cell in the COL_MOHO data set. Furthermore, none of the three data sets had locations (cell centers) in common. To compare the data sets, it was first necessary to create a common cell grid system. The COL_MOHO data set was used as the baseline. This eliminated problems associated with creating a finely spaced grid from a coarse one and was sufficient for this overview of the data. In the areas that overlapped the three data sets, the MOHO and CORN_MOHO data sets were sub-sampled to match the COL_MOHO map coordinates as closely as possible. This was done by a “nearest-neighbor” approach: simply locating the nearest cell location in both the MOHO and CORN_MOHO data sets, coordinating with a cell location in the COL_MOHO data set, and assigning the value in that “neighboring” cell to the new grid. In Figures 2a and 2b, observe the original CORN_MOHO data set and the same data set resampled to represent 1° square cells. It can easily be seen that while some of the smaller features are absent, as expected, the sub-sampling scheme represents the original data fairly well.

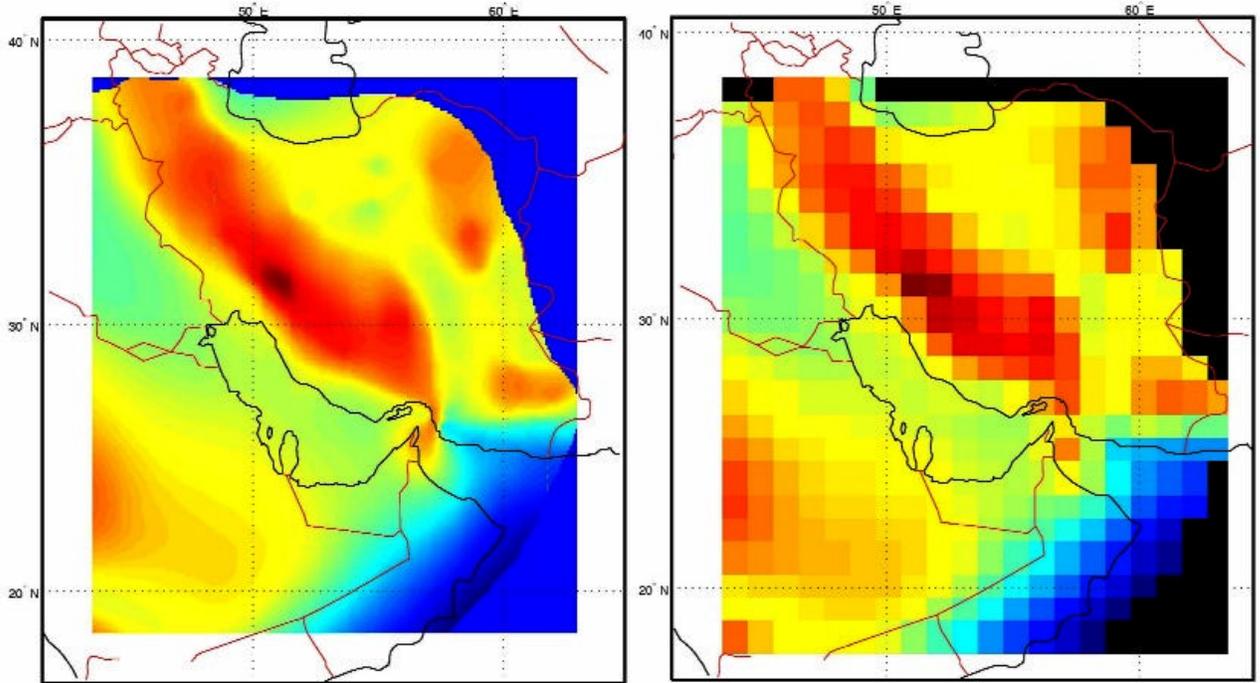


Figure 2. The CORN_MOHO data set at its original resolution (a) and the same data set resampled to represent 1° square cells (b).

Perspective maps of the three data sets that correspond to the overlapping green area in Figure 1 are presented in Figure 3 (parts a, b, and c). Note that the perspective maps for the CORN_MOHO and MOHO data sets represent the sub-sampled points from those data sets that are closest to the cell centers of the COL_MOHO data set. It is apparent from these perspective maps that the data sets differ greatly: both the MOHO and CORN_MOHO data sets have a deep trench in the Moho trending in the northwest-southeast direction, whereas the COL_MOHO data set has a bowl-like depression in the center of the map. This difference may be real (derived from the raw data) or may be an artifact reflecting how the data sets were processed and smoothed. We cannot determine the reason unless raw data are available. The degree of difference that is acceptable is not presently known and must be defined before the data are used.

Cell-by-cell pair-wise comparisons of overlapping areas of CORN_MOHO (Cornell University), MOHO (Cornell University), and COL_MOHO (University of Colorado) reveal cell value differences ranging from approximately -17,000 to +17,000 meters. While these differences appear large, the mean and median cell differences were negligible. Using exploratory analysis techniques, it appears that while both mean and median differences (biases) between Moho depths (over large areas: 16° of longitude and 19° of latitude) between maps may be small, there are smaller areas of the maps where values differ significantly and may be of concern.

Differences in the maps may be due to different methods used to derive the map data (original data), possible errors or inaccuracies of measurements, different smoothing algorithms used for each map, and/or the effect of combining different data sets into each of the individual data sets. Determination of the origin of the errors was beyond the scope of this study, as the raw data are not available.

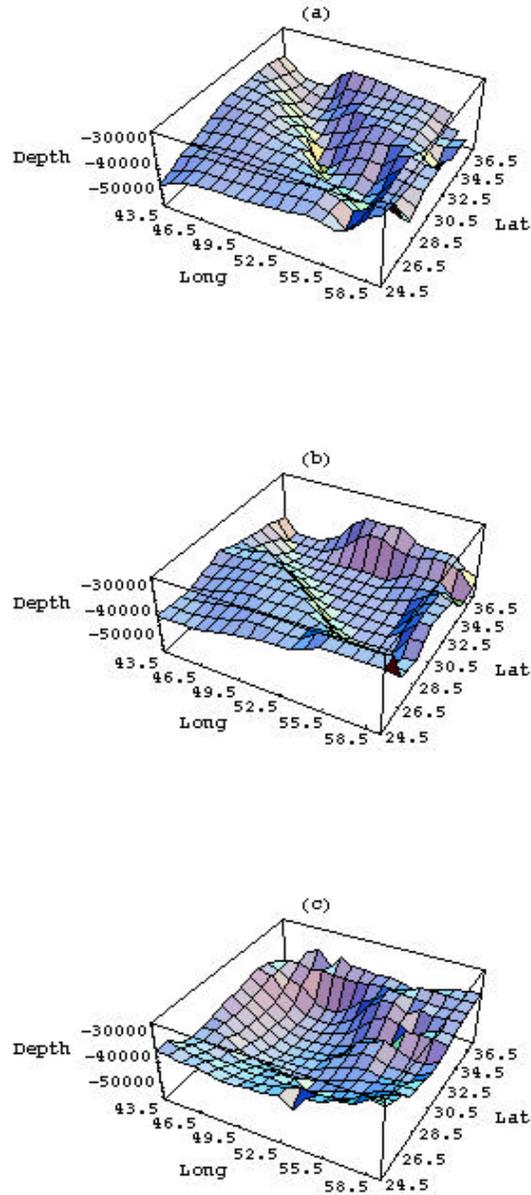


Figure 3. Perspective plots of the depth-to-Moho values in: (a) CORN_MOHO, (b) MOHO, and (c) COL_MOHO.

Histograms of the depth-to-Moho values, made for each of these data sets, are shown in Figure 4. The most striking feature of these histograms is seen in Figure 4 (b), representing the MOHO data set. The figure shows several abnormally large bars at 30,000, 35,000, 40,000, 45,000, and 50,000 meters. It is postulated that these values are the result of digitizing a contour map from the Institute of Physics of the Earth, Moscow, Russia. Such anomalies appearing at 5000-meter intervals are likely due to data inside artificial contour boundaries that are local minima or maxima areas of Moho depth. That is, after smoothing the data sets, local minima or maxima within the contour remain constant. Simply knowing that these regions exist can help explain some of the anomalies that were observed. It can also be noted that the spread of the histograms differs, mainly due to the truncation of the MOHO data at 25,000 and 45,000 meters.

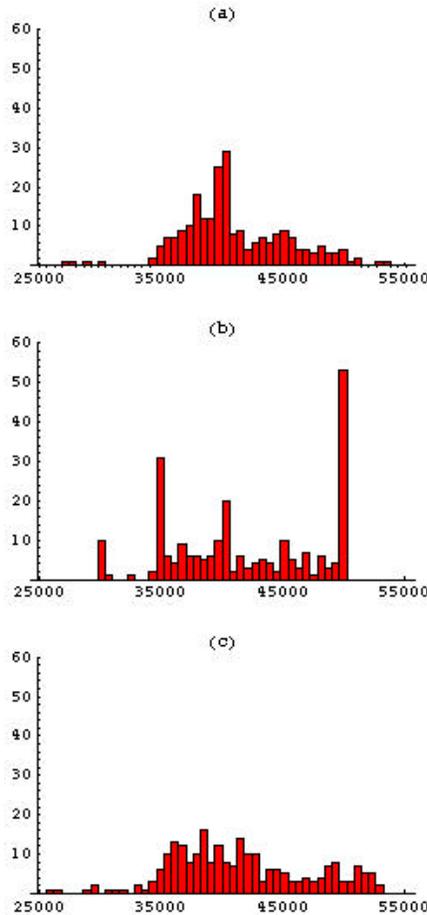


Figure 4. Histograms of the depth-to-Moho values in: (a) CORN_MOHO, (b) MOHO, and (c) COL_MOHO.

Summary statistics for each of the data sets in Table 1 reveal that mean Moho depth in the three maps differs by up to 1000 meters, while the median differs by at most about 550 meters. Notice also, that while the range of the MOHO data set is smaller than the others; its standard deviation is larger due to the large number of values at the extreme truncation points of 30,000 and 45,000 meters.

Maps	CORN_MOHO	MOHO	COL_MOHO
Mean	40,927.7	41,979.0	41,435.4
Median	40,128.0	40,680.0	40,674.0
25% Quantile	38,100.0	36,590.0	37,237.0
75% Quantile	43,857.0	48,790.0	45,261.0
Standard Deviation	4,318.7	6,146.6	5,625.4
Range	26,623.0	20,150.0	27,112.0
Interquartile Range	5,757.0	12,200.0	8,024.0

Table 1. Summary Statistics of COL_MOHO, Sub-Sampled CORN_MOHO, and MOHO Data Sets

Cell-by-cell differences demonstrate the magnitude of the differences that exist between the maps. Table 2 gives brief summary statistics of the differences. The overall mean cell-by-cell differences are only slightly biased, ranging from in value -1,015 to +543 meters. However, while the mean difference in cell-by-cell differences of Moho depth between maps is small, this does not imply that the maps are similar. The

dissimilarity in the maps is suggested by the rather large variability of the differences, as can be seen by examination of the histograms and the standard deviations of differences.

Mapped Differences	Mean (km)	Median (km)	Standard Deviation
CORN_MOHO - COL_MOHO	-507.7	-197.0	5,724.6
MOHO - COL_MOHO	543.6	356.0	7,145.4
CORN_MOHO - MOHO	-1,015.3	-146.0	4,505.6

Table 2. Cell-by-Cell Relative Differences of Moho Depth Between Maps

It can be determined from the data that roughly 50% of all relative differences are off by more than 10% of the baseline value. The largest relative differences seem to occur at the lower and upper values of the baseline data sets, while the maximum differences seem to occur most frequently at roughly the 30,000- and 50,000-meter depths. Thus, bias between all maps appears to be a function of depth to the Moho.

The reasons for the differences in the data sets are certainly varied and may be due to accuracy of the actual measurements used, the smoothing techniques used, and other modeling assumptions (both documented and undocumented). The determination of which data set is most accurate at this point is not a statistical issue, but one of basic geological science.

Detailed Comparison of MOHO Versus COL_MOHO Data Sets

In this section, we examine a larger geographic area than the area in common to all three Moho depth data. We sought to discover whether the disagreements observed in the smaller region extended to larger areas. To do so, we examined the agreement between the MOHO and COL_MOHO data sets.

For reasons described earlier, a map of 1° cell size was created from the MOHO data set, with cell locations corresponding to those in the COL_MOHO data set. A nearest-neighbor approach was again used. The MOHO value located closest, in a great-circle distance, to the mapped location in the COL_MOHO data set, was taken as the value at that new mapped location. As may be observed in Figure 1, the region of overlap is quite substantial, covering about 28° of latitude and about 65° of longitude, which is nearly all of the COL_MOHO region and approximately 16% of the region covered by the MOHO data set. The regions in common that were used in this analysis are presented in Figure 5 (MOHO, binned to a 1° grid) and Figure 6 (COL_MOHO).

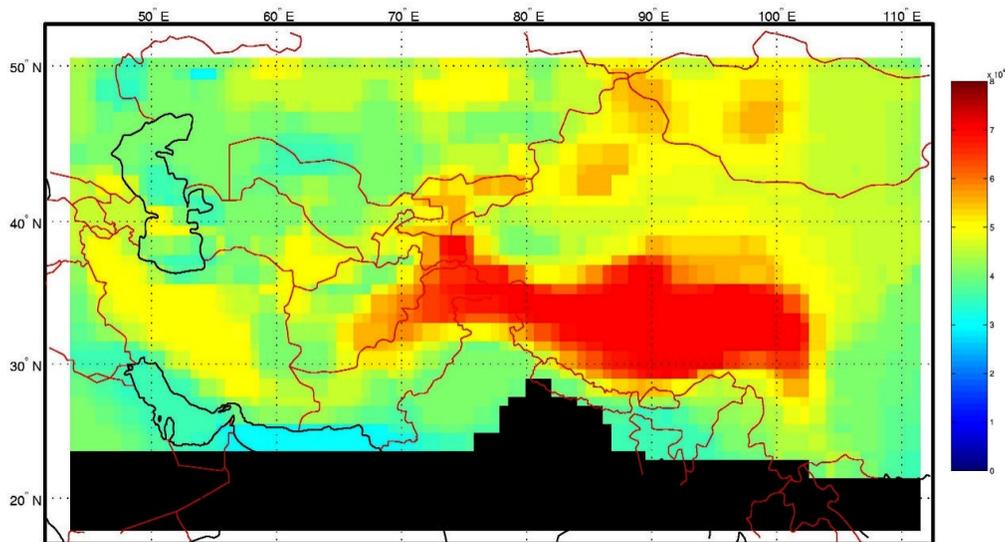


Figure 5. Depth-to-Moho values in MOHO data set, binned to a 1° grid.

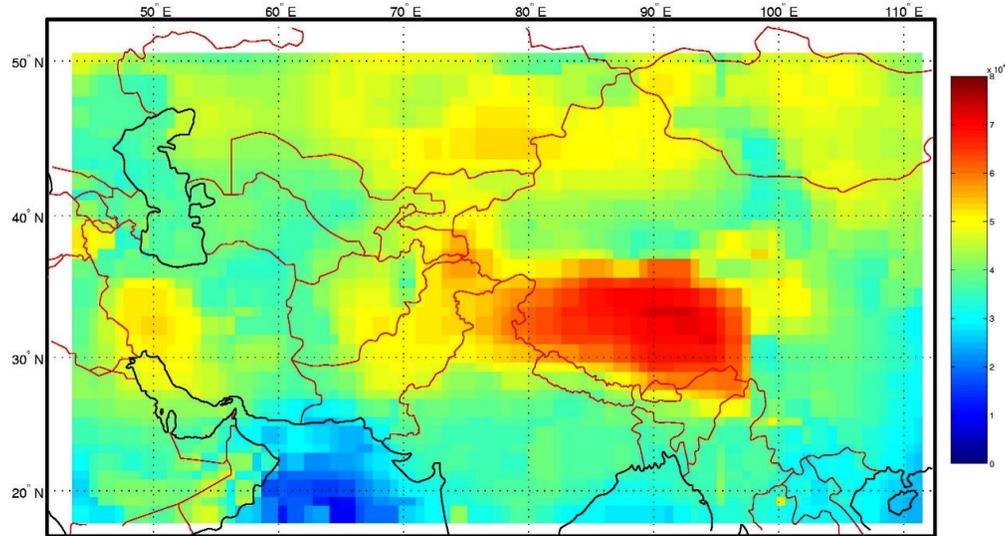


Figure 6. Depth-to-Moho values in COL_MOHO data set, at the original 1° grid.

The two areas were differenced and percentages created from those residuals. Figure 7 shows the spatial distribution of the relative percent differences. The red area, at the bottom of the map, represents missing values. Residuals range from -16 km to +35 km, and relative percent difference ranges from -30% to over +100%. Notably, different geographic regions show patterns of “underprediction” or “overprediction” relative to the reference map. For instance, over the Persian Gulf, the dark blue shows that the MOHO map claims a shallower depth-to-Moho than does the COL_MOHO map. The raw MOHO map indicates a typical depth-to-Moho in that region of 34 to 38 km, and the COL_MOHO map indicates 45 to 50 km. Figure 7 shows a few orange and reddish areas; in these areas, the MOHO map suggests Moho depths up to twice that suggested by the COL_MOHO map.

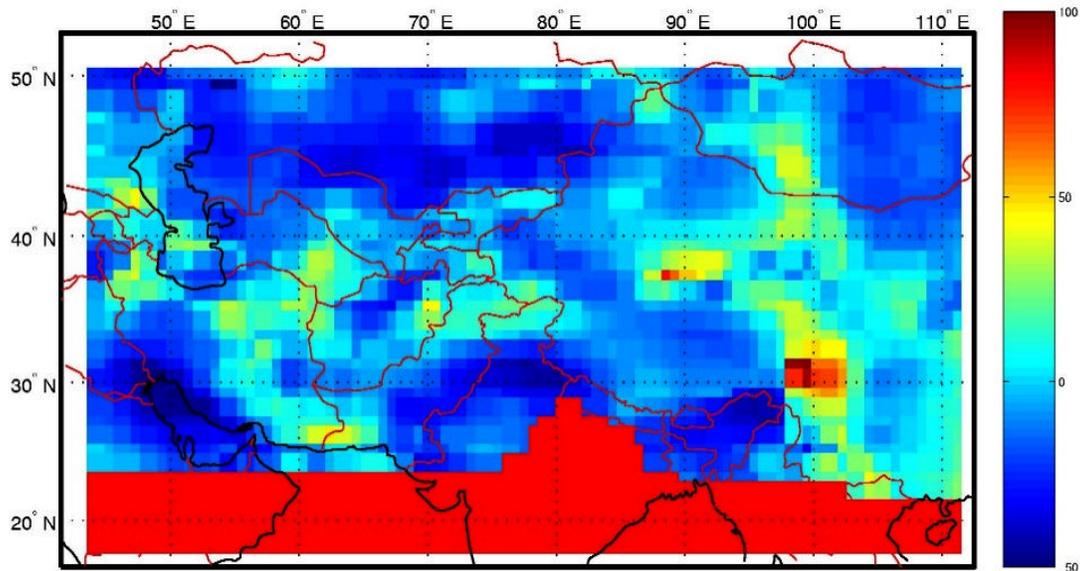


Figure 7. Spatial distribution of the relative percent differences in depth-to-Moho values resulting from MOHO - COL_MOHO.

Figure 8 presents the distribution of relative percent differences as a histogram. Although the center of the distribution is around 0% difference, a large fraction of the differences is substantially different from 0 km. Table 3 presents a statistical summary of the raw differences. Significant variations are apparent: the mean of the MOHO data is nearly 4600 km deeper than the mean of the COL_MOHO data, and the range of depths for MOHO is 8,565 to 70,140 km versus COL_MOHO depths of 29,930 to 73,697 km. While these statistics are enlightening, they are only a summary of the facts.

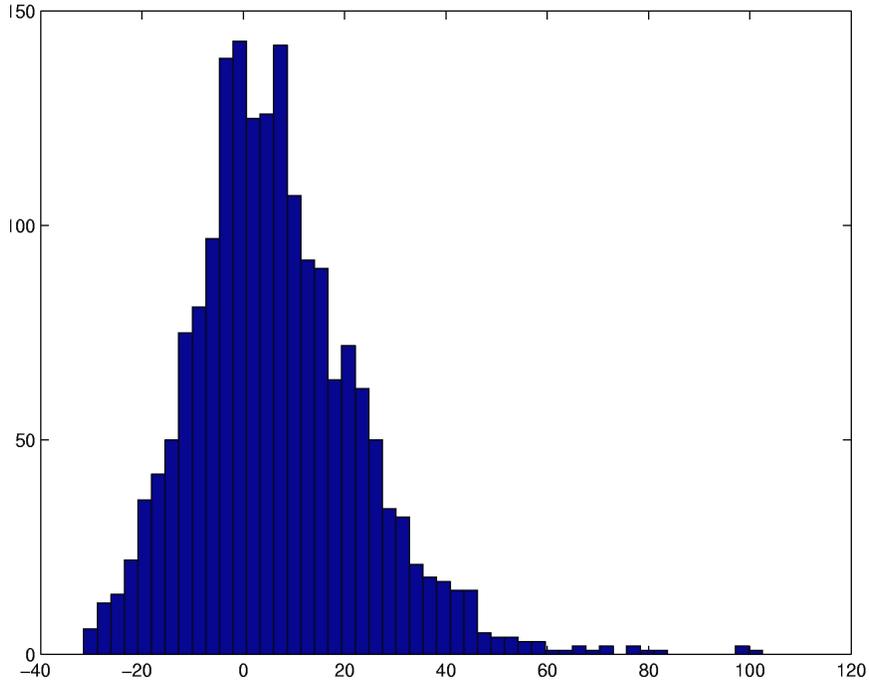


Figure 8. Histogram of Relative Percent Residuals Resulting from MOHO - COL_MOHO. Note the skewness to the distribution.

	MOHO, 1 ^o grid	COL MOHO
Mean	46,804 km	42,207 km
Median	45,000 km	41,524 km
Standard Deviation	8,998 km	9,118 km
Maximum	70,140 km	73,697 km
Minimum	8,565 km	29,930 km

Table 3. Results from statistical analysis of the two Moho depth data sets (MOHO and COL_MOHO)

Table 4 presents the percentiles of residuals and the relative percent differences between the MOHO and COL_MOHO maps. With regard to Moho depths greater than COL_MOHO depths, 10% of the data pairs have a raw difference of +10.9 km or greater, which translates to a relative percent difference of +27.5% or greater. On the other side of the distribution, with MOHO values less than COL_MOHO values, 10% of the data pairs have a raw difference of at least -6 km, which also translates to a relative percent difference of at least -12.6%. To illustrate the extent of these differences, observe that only 25% of the data pairs have a residual of less than ± 2 km (24th to 49th percentiles), and only 27% of data pairs have a relative percent difference of less than $\pm 5\%$ (23rd to 51st percentiles). This large area comparison shows significant disagreements between the MOHO and COL_MOHO maps. The amount of variation observed when comparing all three Moho depth data sets over a small region in common to all three appears to extend to the entire data set. The amount of disagreement exhibits a geographic dependence. Certain geographic

regions are systematically overpredicted while others are underpredicted. We present a short example of this type of analysis below.

Percentile	Difference (km)	Relative Difference (%)
Minimum	-16.2	-31.5
1	-13.4	-25.9
5	-8.7	-18.0
10	-6.0	-12.6
25	-1.9	-4.3
50	+2.2	4.8
75	+6.7	15.9
90	+10.9	27.5
95	+13.8	36.1
99	+22.4	55.9
Maximum	+35.4	102.5

(a) For corresponding plots, see Figure 23 for distributions of residuals and Figure 24 for relative differences between MOHO map depths versus COL_MOHO map depths.

Table 4. Percentiles of distributions of residuals and relative differences between depths of MOHO data set versus COL_MOHO data^(a). Total number of cells, 1832.

Analysis of MOHO and COL_MOHO Data Sets by Subregion

The MOHO and the COL_MOHO maps differ in important respects in various geographical subregions over the overlapping portions. Explanations for these differences no doubt lie in the different pedigrees of each map and possibly even the definitions of the physical measures used to calculate estimates of the values for depth-to-Moho.

To further understand the observed spatial inconsistencies between the MOHO and COL_MOHO depth-to-Moho data sets, the region in common to both data sets was subdivided into 32 subregions, and each of these was then statistically examined for agreement. Two of the subregions contained too few data pairs for analysis. Table 5 presents correlation coefficients for equally sized subregions of the overlap, in terms of numbers of 1° cells. The higher the value of $|r|$, the more closely the two surfaces agree with each other. Recall that correlation coefficients indicate tendency of variables to vary in step with each other, but say nothing about equality of values being compared.

Latitudes	Longitudes							
	43.5° – 50.5°	51.5° – 58.5°	59.5° – 66.5°	67.5° – 74.5°	75.5° – 82.5°	83.5° – 90.5°	91.5° – 98.5°	99.5° – 106.5°
42.5° - 49.5°	-0.14	0.11	0.37	-0.18	-0.20	0.24	-0.11	-0.06
34.5° - 41.5°	0.41	-0.16	-0.13	0.80	0.88	0.82	0.79	0.55
26.5° - 33.5°	0.47	0.02	0.72	0.67	0.82	0.96	0.46	0.60
18.5° - 25.5°	-0.80	0.74	-0.23	0.54	NaN ^(b)	NaN ^(b)	0.29	0.10

(a) Each cell (subregion) in the northernmost three rows represents 64 depth pairs. Each cell (subregion) in the bottom row represents from 16 to 29 data pairs. The highlighted cells indicate those subregions that are statistically similar (strong correlations). The significance of the correlation coefficient, r , at $p < 0.01$, is dependent on the number of data in the regression. $|r| > 0.33$ for the top three rows.

(b) Not enough data pairs for a reliable correlation coefficient.

Table 5. Product-Moment Correlation Coefficients, r , over 30 Equal-Sized Subregions of the Region of Overlap Between COL_MOHO and MOHO Data^(a)

Table 5 indicates substantial spatial variation in the tendency of the two maps to indicate similar trends in Moho depths. Although many of the subregions show strong positive correlations, indicating that the two depth surfaces more or less follow each other, at least 13 of the 30 subregions do not have statistically significant correlation coefficients. The significance of the correlation coefficient, r , is dependent on the number of data in the regression, thus the values in Table 5 should not be viewed as absolute indicators of fit. Small coefficients indicate that, the two surfaces vary independently of each other. The subregion in the extreme southwest corner is negatively correlated (-0.80) indicating opposite surface trends with respect to Moho depth. The problem with this subregion may be related to the quality of the data near the edge of the COL_MOHO data set. This region may contain sparse raw data, and thus the uncertainty may be extremely high. This is only a hypothesis and cannot be proved or disproved without associated error estimates or by examination of the raw data. In spite of this regional variation and disagreements, the overall correlation coefficient (r) for all 30 of the data pairs was very strong at +0.67 (n=1832).

Interpretation of the significance of the correlation coefficients may be guided by the following: for the northernmost subregions with 64 data pairs, a value of $|r| > 0.33$ would be considered statistically significant at the 1% level (two-sided test). The cells in the bottom row represent 16 to 29 pairs, such that $|r|$ needs to be >0.45 to 0.60 to indicate statistical significance at 1%. The critical values of r are presented as general guidelines rather than as formal statistical tests. These subregions also were not optimized to the detected features of disagreement. A more in-depth examination should reveal more detail about the nature of the disagreement, whether region-dependant, artifacts of the method for data set generation, or resulting from the types of processing applied to the data.

CONCLUSIONS AND RECOMMENDATIONS

The preliminary investigation into what “tools” would be of most use in assessing the quality and consistency of the data in the KB has resulted in more questions and new directions, which should be pursued. The initial premise that statistical tools could be used to examine individual data sets for erroneous values became less important as we began to understand the genesis of these data. Most of the gridded measurement data has been highly processed and smoothed. Thus, erroneous values have been either eliminated prior to processing or have been averaged away by the processing and smoothing. Additional data sets, based directly upon and containing raw data, that will be added in the future, will be more receptive to the use of statistical tools.

Problems were found within and among data sets, identified by several methods, and include the following:

- metadata incompleteness and disagreement with header file data,
- significant disagreements among overlapping data sets reporting the same parameter,
- a lack of any error/uncertainty estimates accompanying data sets,
- different data sets reporting the same parameter receiving vastly different processing and smoothing operators,
- missing values and constant values that complicate analysis and bias results,
- a need for a consistent schema for gridding and combining data sets and a lack of any consistent reference points,
- a need for well-defined criteria for: the accuracy of data, the degree of error that is tolerable for correction calculations, and tolerance limits for agreement among data sets.

Key future tasks will address the following questions:

- What measures of uncertainty or error estimates need to be included with the data?
- What are the issues associated with assessing data in light of cultural and geologic features (rivers, cities, roads, faults, coastlines, topography, etc.)?
- What issues are associated with sequential kriging, and multiple sampling and smoothing of the data?
- What are the problems associated with combining vector and gridded data types?
- What is the optimum method for fusing multiple maps into a single data set?
- How do we best propagate and estimate the cumulative uncertainties in the KB-based corrections?

The tools under development are being designed to enable evaluation of the degree of agreement between different data sets, sequential “loads” (updated data sets), and adjacent data sets. The tools will:

- be easy to use and provide understandable graphical/textural results,
- provide robust evaluations of the status of the KB,
- be built upon statistical tools designed to accurately describe the completeness, uncertainties, and limitations of the state of the KB and/or input data “loads”,
- help identify what data would be most beneficial to include in future data sets,
- assist in the development of data-quality criteria for the incorporation of future data, based upon the current content of the KB (i.e., the new data needs to be able to be “melded” or “fused” into the existing data), and
- provide a basis for prioritizing data improvement efforts.

REFERENCES

(none)