UCRL-JC-128101
PREPRINT

# LLNL Middle East and North Africa Research Database

S. D. Ruppert
T. F. Hauk
R. Leach

This paper was prepared for submittal to the
19th Seismic Research Symposium on Monitoring a
Comprehensive Test Ban Treaty
Orlando, FL
September 23-25, 1997

July 15, 1997

Lawrence Livermore National Laboratory

# LLNL Middle East and North Africa Research Database

Stanley D. Ruppert, Teresa F. Hauk, Richard Leach

*Geophysics and Global Security, Lawrence Livermore National Laboratory*
*University of California*

Sponsored by DOE CTBT R&D Program[1]

## Abstract
The Lawrence Livermore National Laboratory (LLNL) CTBT R&D program has made significant progress assembling a comprehensive seismic database (DB) for events and derived parameters in the Middle East and North Africa (ME/NA). The LLNL research DB provides not only a coherent framework in which store and organize large volumes of collected seismic waveforms and associated event parameter information but also provides an efficient data processing/research environment. The DB is designed to be flexible and extensible in order to accommodate the large volumes of data in diverse formats from many sources in addition to maintaining detailed quality control and metadata. Researchers can make use of the relational nature of the DB and interactive analysis tools to quickly and efficiently process large volumes of data. Seismic waveforms have been systematically collected from a wide range of local and regional networks using numerous earthquake bulletins and converted a common format based on CSS3.0 while undergoing quality control and corrections of errors. By combining traveltime observations, event characterization studies, and regional wave-propagation studies of the LLNL CTBT team, we are assembling a library of ground truth information and event location correction surfaces required to support the ME/NA regionalization program. Corrections and parameters distilled from the LLNL research DB will provide needed contributions to the DOE knowledge base for the ME/NA region and enable the USNDC and IDC to effectively verify CTBT compliance.

**Key Words:** seismic, waveform, database, metadata

## Objectives
The primary objective of the LLNL research DB program is to efficiently provide large volumes of quality controlled seismic data and interactive analysis tools with connectivity to the DB in support of the LLNL CTBT Middle East and North Africa regionalization program. In the Middle East and North Africa, there is a hierarchy of ground truth that can be obtained. This hierarchy ranges from explosions with exact locations to carefully studied aftershock sequences, events located by dense local networks, and teleseismically constrained events such as CMT and EDR bulletin events. Our goal is the collection, quality checking, and conversion of tens of thousands of seismic waveforms from many different seismic networks such as IRIS, GEOSCOPE, USGS, USNDC, PIDC, MEDNET, CDSN) to a standard format provide the necessary framework to be able to relocate and characterize each event. Location studies require synthesis of traveltime and azimuth correction surfaces, phase onset measurements and identification and depth

[1] Research performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

determination. All of the waveforms compiled and integrated will by re-analyzed and quality checked by trained analysts. Event Characterization requires analysis of physical basis including earth structure derived from body waves, surface waves and regional geology, and the effects of phase blockage. The research DB will provide ready access and organization to the thousands of events and associated waveforms and also provide the framework for storage and dissemination of results. Sufficient metadata (including measurement procedures, codes, comments and measurement errors) are stored at each step in the analysis process to allow recreation or verification of results at any stage in the processing flow.

In detail, the LLNL research DB must facilitate scientific analysis during all phases of the CTBT regionalization program. From exploratory data interpretation, production data analysis and distillation of standardized measurements applied to all waveforms, and Quality Assessment (QA) and delivery of research products to the DOE knowledgebase. Out of the research DB will be compiled the reference event libraries and ground truth datasets showing space and time clustering of natural earthquake and mine events, phase blockage maps, event characterization parameters necessary for the monitoring CTBT compliance in the ME/NA region.

## Research Accomplished

### Data Collection

Data collection for the LLNL ME/NA research database began in 1996 and continues today. Initially, large events were selected from the Harvard CMT catalog and the NEIC PDE catalog. As the database fills, we are moving to smaller magnitude events and extending our coverage of the entire ME/NA area. Detecting and characterizing small magnitude events that can be associated with mining events is a research priority. In support of this, we also have collected continuous data from several stations. Analysis of the continuous data using waveform correlations, spatial and temporal clustering by LLNL CTBT researchers to detect smaller magnitude events in currently underway. The ME/NA area from which events are chosen has grown as we've developed our database and run into research issues that require data from outside the original arbitrary bounds (e.g., ray-path coverage, structure-related phase blockage or attenuation, etc.). Source-receiver distances range from local to teleseismic. Figure 1 outlines the base data acquisition flow from data request, quality checking and integration of the waveform into the DB using a common format, and access of the waveform from the database by LLNL CTBT researchers and analysis programs.

Our current area of interest (Figure 2) includes all seismic events located within and recorded by stations in the box bounded roughly by corners 10S,40W and 60N,90E. Figure 2 shows the IMS primary, auxillary, and gamma stations datacollection is focused on superimposed on NEIC seismicity between 1985 and 1995. The time period for which the most high-quality, easily accessible digital data were available from the stations chosen is from 1990 to the present. Certain selected events or datasets at stations operating prior to 1990 have been requested. ILPA (Iranian Long Period Array) data, for example, are used as a surrogate for the proposed IMS primary array, THR. ILPA was operational from 1968-69. Magnitude thresholds are determined by the catalog used to initiate the requests. We have certain regions for which all events down to mb4.0 in the PDE catalog are included in the current database. Our goal is to get all data for all stations and events at least down to mb4.0 or perhaps smaller. The compilation of local and regional seismic network bulletins is continuing as we move to smaller magnitude thresholds. We are also collecting large continuous data segments from some stations for specific research (e.g. clustering analysis, detection and identification of mining explosions in certain areas, etc.). Local bulletins sometimes include identified mining explosions.

As of July 1997, the number of waveforms in the DB exceeded 50,000 representing over 2200 events. As summarized in Figure 3, All of the waveforms currently in the database recorded by a particular station (star) and event (green circle) pair are shown as either red raypath where data has been loaded or as yellow raypath denoting where travel-time observations have been made by a LLNL or contractor trained analyst. Currently about 650 events out of the 2200 have been repicked yield over 6000 travel-time observations made available to the LLNL CTBT research teams for location and discrimination projects. Augmenting the LLNL picks, we've added all ~5 million NEIC EDR phase pick observations into the DB to be used a starting point for travel-time correction studies and correction surface generation. Statistical studies are underway comparing the variance and accuracy of the LLNL analyst picks to those from the NEIC EDR bulletins to see where using the EDR picks is most appropriate. Bulletin information many local and regional catalogs (NEIC EDR, CMT, ISC, JSOP, Greece, Turkey, Isreal, AFTAC (unclassified), Saudia Arabia ) have been incorporated into the DB to be used in forming event requests. We have also derived an initial teleseismically derived ground-truth event bulletin (Schultz et. al., 1997) to be used in event requests. The work of Sweeney (1996) shows that teleseismic events can give event locations with a known uncertainty. As part of a LLNL ME/NA location study, we derived a teleseismically constrained ground truth datasets using NEIC EDR phase pick information and AK135 velocity model. We relocated all EDR events that fit the 50-90 criterion using the this velocity model. Relocation was done using EvLoc which operates on the LLNL Oracle database and all of the AK135 relocations shown in Figure 5 have been incorporated into the LLNL hierarchical ground truth database with a 20 km uncertainty assigned to the location. Comparisons and reconciliation of events common between bulletins and those relocated using travel-time corrections for the ME/NA region is also underway as part of the location research efforts.

Surrogate stations were chosen for IMS primary and auxiliary network arrays and stations that have not yet been installed or for that little data is currently available. Supplemental data from broadband stations located in the vicinity of proposed IMS sites and from other stations in our general area of interest were also requested. The current list of stations for which there are waveforms in the database totals 296 (includes individual array elements). Figure 2 shows the stations we currently have acquired waveforms from in the ME/NA region. Data windows to be requested were calculated for each station so that the criteria for all segments would be consistent (begin at origin time of event and end at arrival time of seismic waves traveling at 1 km/sec). Data was requested from five major data centers: AFTAC, IRIS, GEOSCOPE, GEOFON, and the IDC. IRIS requests were submitted using the breq_fast batch method for a single station for a year-long period to keep the seed volumes small enough to transfer electronically using FTP (File Transfer Protocol). The xretrieve method was used to request continuous data segments that were provided on Exabyte tape. GEOSCOPE and GEOFON requests were similar. AFTAC requests were made using a modified autoDRM request format. Multimax, Inc., is contracted to do a portion of the data collection for the LLNL database. They obtained bulletin and waveform data from the IDC using their Req-Data request routine. Multimax is responsible for collecting recent and continually occurring earthquake data while we concentrate on the historic waveforms available.

Data arrive from the various data centers in, primarily, 2 different formats, CSS3.0 and SEED. The waveform data is processed and written out in CSS formats. The default origin definition (orid) for each waveform is the NEIC EDR monthly location solution unless a better location is available. The station location information has been found to have discrepancies between different sources. In such cases, we've gone directly to the network operators and have attempted to obtain the best information possible. There are also discrepancies in what channels are being recorded at a given station and the dates those channels are operational. Again, we have worked hard to obtain the most accurate information from the most direct sources in these instances to keep the database as accurate as possible. All available pertinent information about the event (PDE location parameters) and the station (available network station inventories, FDSN station book, IDC country book, site and sitechan tables) is examined and the verified information is added to file headers (SAC) or flat files (CSS). The processed waveform data is written to a staging disk into separate event directories prior to loading into the DB. Waveform and header data examined to verify that the resultant CSS format files are complete. Checks are made to prevent overwriting data files in the existing database. Overlap of id numbers, such as wfids, is checked to ensure uniqueness in the database. Processing logs are reviewed to ensure that all available data was processed and that there were no blatant errors. The waveform data is then moved to the disk storage system as an official part of the LLNL ME/NA research database and made available to researchers.

**Database structure**

The basic design goals of the LLNL research DB are support the representation of diverse data types, handle large quantities of data and yet be flexible, extensible and maintainable. Data stored in the DB, along with the associated metadata and quality control information provides the raw material and is of central importance to most of the LLNL CTBT regionalization programs including event location and event characterization. Because the subsets of data must be provided in a form easily accessible to many diverse research programs having different user requirements of data and metadata, the DB access tools have been designed to utilized the power of the relational database to facilitate efficient queries and data retrieval and to hide as much of the low level database calls from the end user as possible. Hence the LLNL research DB fills the role of both data archive for the assembled seismic datasets and ground truth libraries forming the deliverables to the DOE knowledgebase, but also as a analysis tool itself enabling data analysis to be done thoroughly on over 50,000 waveforms while providing data organization and visualization results.

The core of the LLNL DB is comprised of ORACLE relational database (RDB) software running on a SUN Ultra 1/170 workstation. Integral to the ORACLE server is WWW server which provides researchers, within the LLNL intranet, interactive (GUI) access to the DB contents in addition to DB access via software tools such as SAC, MATLAB, and custom data browsers based on HTML. Many seismic analysis codes such as EvLOC can connect to the DB directly from the users workstation. Storage space for the waveforms consists of two SUN disk storage arrays holding over 250 GB of seismic waveforms, DB files, and associated metadata information. The DB is connected via 100Mb/s high speed ethernet to the researchers workstations. The waveforms themselves and final deliverable parameter products are provided in CSS3.0 format utilizing AFTAC and NDC extensions where needed. Several LLNL custom tables are used to store intermediate research parameters, measurements, and metadata. The CSS3.0 standard with extensions was adopted for the purposes of compatibility.

Quality and metadata information is carefully maintained for the collected and corrected waveforms, on all event relocations, and on all event characterization and identification measurements. Metadata allows the processing history and quality to be assessed at any step during the processing flow (Figure 1) and will be an integral part of the derived parameters and corrections delivered to the DOE knowledgebase. Design of the metadata tables was undertaken in collaboration with Sandia National Laboratory and Los Alamos National Laboratory CTBT programs to ensure compatibility and completeness of the final deliverable products. LLNL metadata tables for internal use provide frequent cross-checks of the data collected to data and allow for efficient processing of 1000's of waveforms arriving monthly from disparate global and regional networks as well as from contractors.

An example of two types of metadata gathered automatically is as follows. First, any defined process can generate metadata across all the signals in the database. For example, signal-to-noise (SNR) is an important parameter in many research studies. We have generated pre and post-P wave rms signal power measurements for all the signals in the database. The instantaneous normalized power *p(t)* associated with a time-varying signal *x(t)* is:

$$p(t) = x^2(t) \tag{1.1}$$

The average normalized power P is then obtained by a time average of p(t) over a period T.

$$P = \frac{1}{T}\int_0^T x^2(t)dt \tag{1.2}$$

The root-mean-square (rms) value $X_{rms}$ is defined as the constant value that produces the same average power as the given time-varying signal. which leads to:

$$X_{rms} = \sqrt{P} = \sqrt{\frac{1}{T}\int_0^T x^2(t)dt} \tag{1.3}$$

Using 1.3, we have generated signal and noise rms power measurements for all the signals in the database. The results along with the metadata (algorithms used and waveforms measures) were placed into the database to allows other researchers to use these results to readily obtain SNR values for any signals of interest. This was recently done on over 50,000 signals presently in the database.

Secondly we have developed an automated algorithm which can proved tailored meta information on defined subset of signals in the database. This is extremely important in areas with sporadic or poorly recorded data or in aseismic areas. For example, if the focus of a study is the P and S arrival, parameters can be set to focus on a desired time including and in the vicinity of those arrivals. A specific station or set of stations can also be chosen. The algorithm perms 18 separate error detections on the signals requested only during the time of interest. These checks include timing errors such as a negative time axis and non-existing data. They also check for zero slope areas of the data and discontinuities. Finally, in the case of multi-channel data, median filters are used to determine if the signal power, noise power, or SNR is reasonable relative to the other channels. This is useful for array data as well as three-component signals. The metadata gathered is stored in database tables for review. By applying this algorithm only during times of interest, many more events can be identified which might have been rejected using a meta filter across the entire signal. For example, in Figure 6, 92 ABKT events are shown (darkened ) which were used in a recent study. The light events are events recorded in the database, but not used because they were identified with the automated detector as having one or more of the 18 error checks. This process can save the enormous amount of time which is often spent carefully reviewing events by hand.

## References


Grant, L., J. Coyne, F. Ryall, 1993, "Ground-Truth Database: Version 1 Handbook", Technical Report C93-05, Science Applications International Corporation, Center for Seismic Studies.

McCornack, M. 1996, "AFTAC Extensions Database Design Document, Version 2.0, February 1993", Sandia National Laboratory.

Sweeney, J.J., 1996, "Accuracy of teleseismic event location in the Middle East and North Africa", Lawrence Livermore National Laboratory, UCRL-ID-125868.

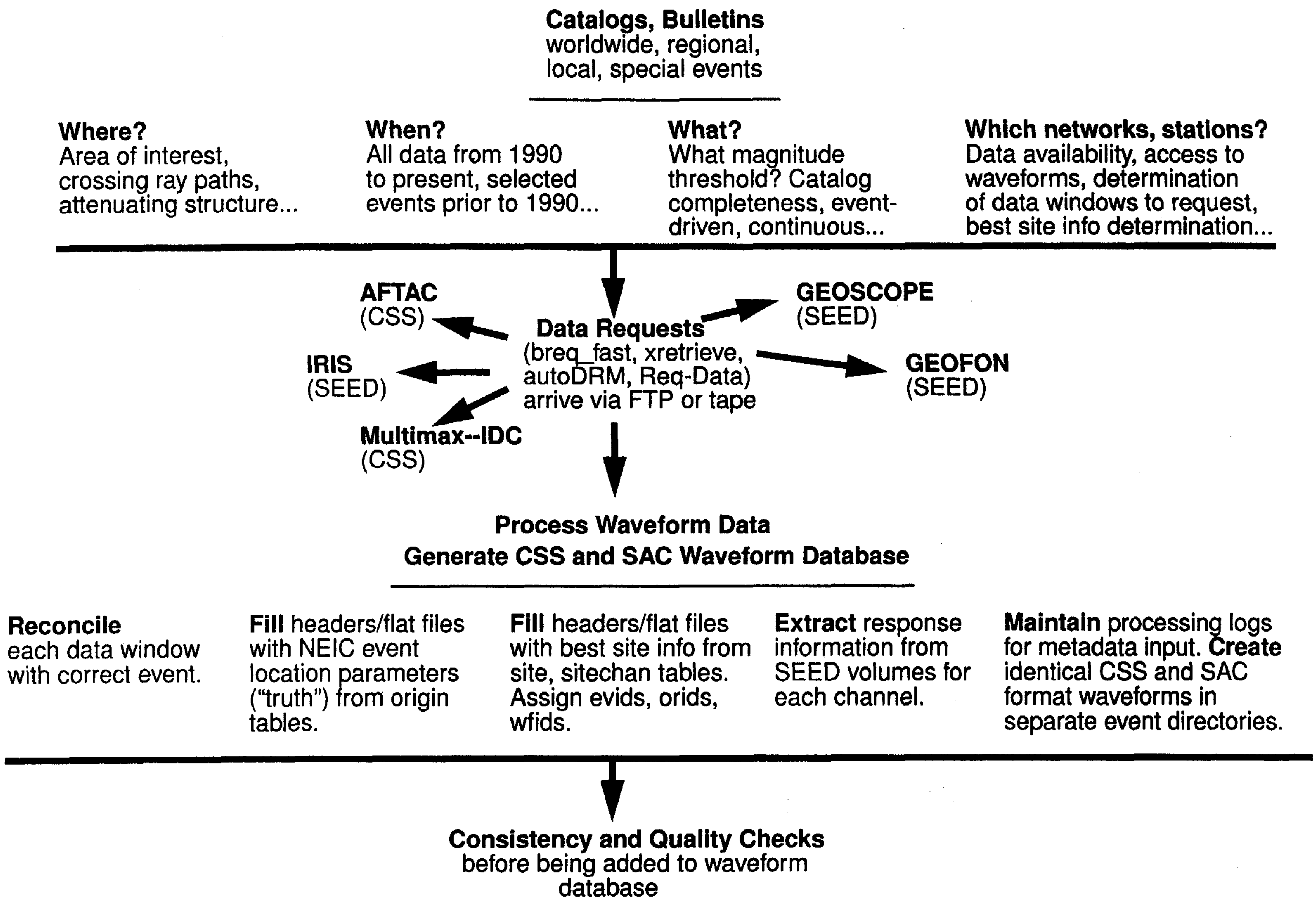


**Figure 1.** Flowchart of data collection and research data population and analysis.

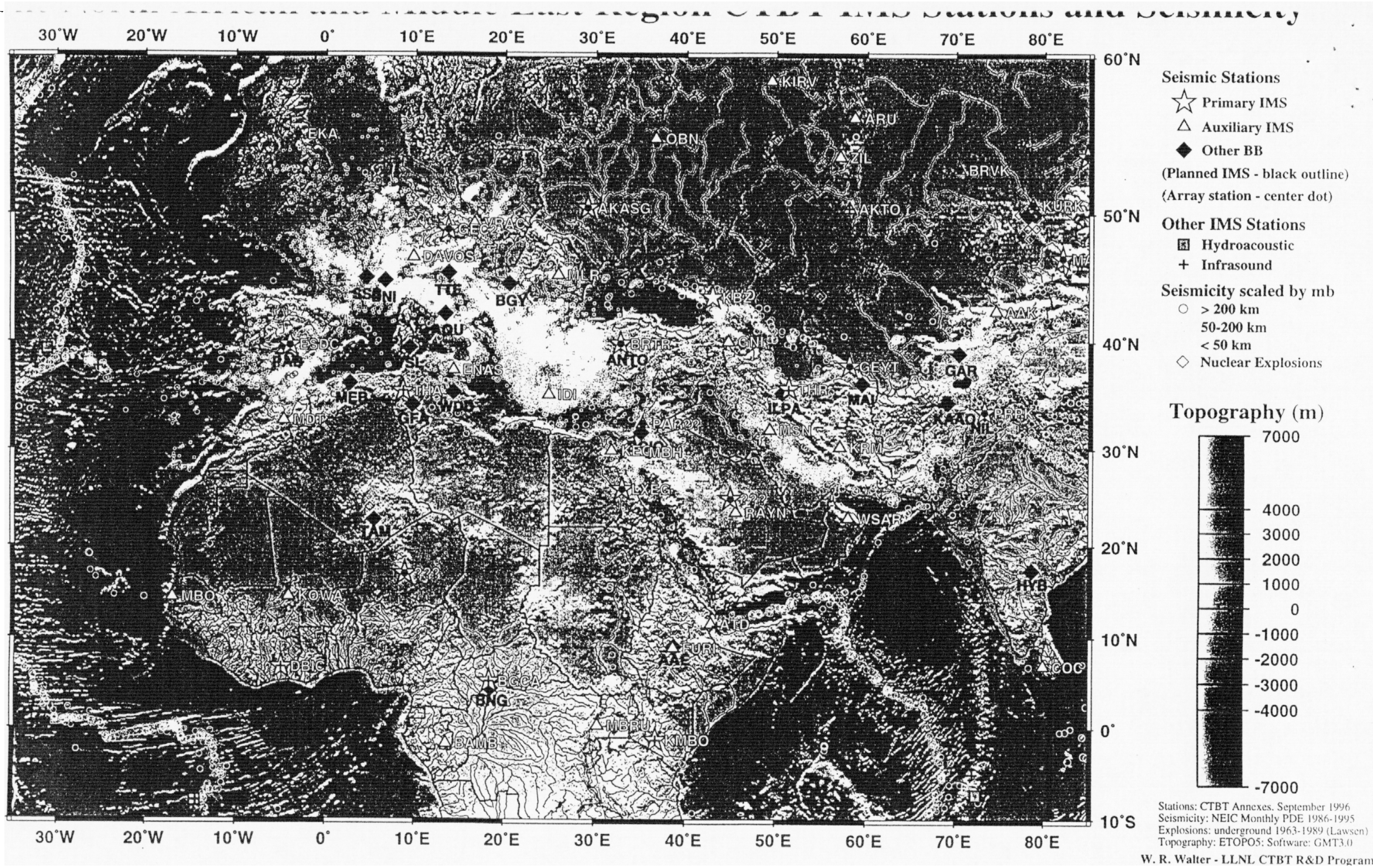


**Figure 2.** Stations and typical seismicity within the LLNL CTBT Middle East and North Africa (ME/NA) study region.

**Figure 3.** Database inventory plot as of summer 1997. All of the waveforms currently in the database recorded for all station (star) and event (green circle) pairs are shown as either red raypaths where data has been loaded or as yellow raypaths where travel-time observations have been made. As of summer 1997 over 50,000 raypaths from 2200 events and 6000 travel-time picks have been quality checked and made available to the LLNL CTBT research teams for location and discrimination projects.

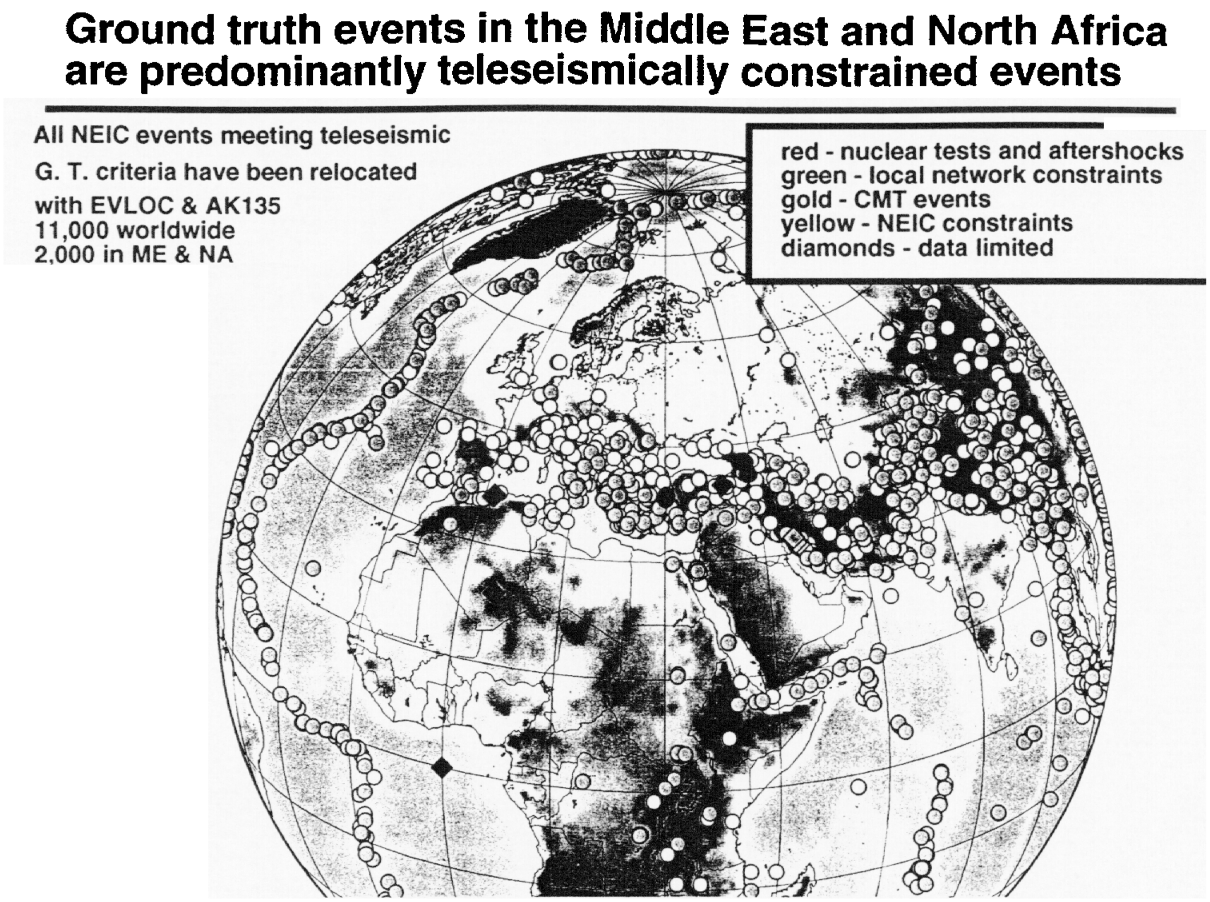


**Figure 4.** Map view of regional relocations plotted over the teleseismically constrained locations for the case that the event was within the footprint of the recording stations (AzGap < 180°). The locations are shown for the case of no station calibration and the case of an optimal static station calibration. At least three regional P recordings were required for a location and station calibrations were applied only if 10 or more independent events were available for the correction.

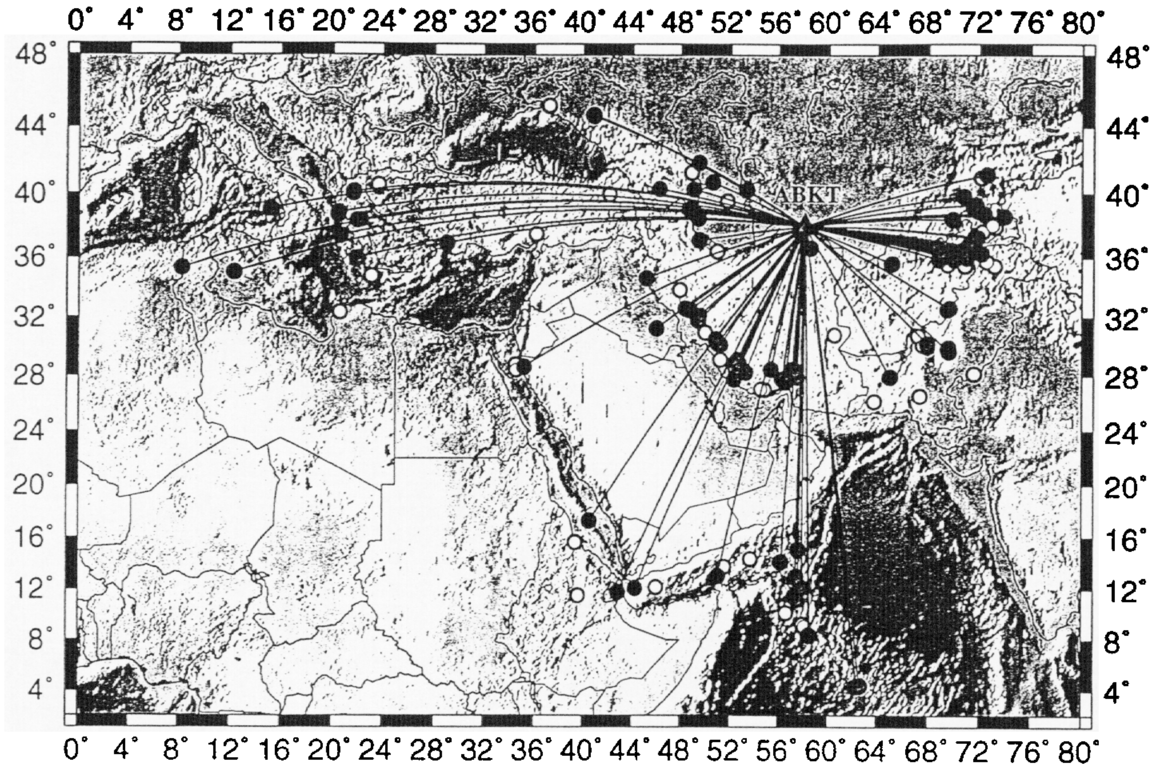


Figure 1, 92 ABKT events are shown (darkened) which were used in a recent study. The lighter events are events in the data base, but not used because they were identified with the automated detector as having one or more errors. (glitches, etc.)

**Figure 5.** Figure showing example of waveform quality control processing and the metadata collected.

*Technical Information Department* **• Lawrence Livermore National Laboratory**
**University of California • Livermore, California  94551**