E L E C T R O N I C
S E I S M O L O G I S T

Monitoring the Performance of a Geophysical Data Processing Pipeline Using Financial Analysis Inspired Tools

by William N. Junek, Charles A. Houchin, Joseph A. Wehlen III, Alan Poffenberger, and Anne S. Henson

ABSTRACT

Routine monitoring of the United States National Data Center (US NDC) geophysical data processing system's performance is critical for identifying potential or ongoing problems that may otherwise go unnoticed. Daily reviews of the US NDC system's automated network and station processing results are conducted by a human analyst using an interactive web application constructed from a collection of open-source software. A key feature of this tool is its use of time-series visualization techniques typically employed by the financial industry to analyze stock market trends. These techniques allow the user to quickly visualize the relationship between multiple time-series datasets (e.g., automatic detections, background noise, and analyst-reviewed detections). This enables the identification of a variety of problems such as the influence of seasonal noise variations on automatic signal detectors or the effect software changes have on system performance over time. The tool empowers the user to make data-driven decisions when selecting the appropriate corrective action for solving a particular problem.

INTRODUCTION

The United States National Data Center (US NDC) geophysical data processing system monitors international compliance with nuclear test-ban treaties through the real-time acquisition, processing, and analysis of seismic, hydroacoustic, and infrasonic data acquired from the United States Atomic Energy Detection System and International Data Center networks. This system produces an automatic seismic catalog using a collection of signal detection, association, and location processes that are spread across a customized hardware and software infrastructure. The automatic catalog is reviewed on a 24 hours per day, 7 days per week basis by a group of seismic analysts who rely heavily on the performance of the automatic data processing pipeline.

Routine performance monitoring of the US NDC processing pipeline is critical for identifying potential or ongoing problems with the system or network that would otherwise go unnoticed. Problems with either the data or with data processing can hinder the production of the automatic seismic catalog and subsequent interactive review. Successful performance monitoring, however, cannot be achieved through the use of automated software applications alone. This task requires manual review of processing results on a daily basis to identify problems. Areas typically reviewed include, but are not limited to:

- Station processing: Comparison of automatic and analyst-reviewed signal detection counts per station.
- Network processing: Comparison of automatic and analyst-reviewed event catalogs.
- Station quality: Review of probabilistic noise power spectral density plots for problematic stations (McNamara and Boaz, 2006).
- Magnitude residuals: Difference between network averaged and station-specific event-magnitude estimates.

For example, undocumented changes in a sensor's instrument response could produce anomalous magnitude estimates that would appear as an outlier when compared to the network averaged magnitude. Once a problem is identified, an appropriate solution or corrective action can be developed to correct or mitigate the problem.

To facilitate this work, we developed an interactive web application that displays network and station processing results on a daily basis. A key feature of this tool is its use of time-series visualization techniques typically employed by the financial industry to analyze stock and commodity market trends (e.g., by Google Finance and Yahoo Finance). The station summary display presents time-series data to the user as a set of interactive plots that allow the user to see the entire dataset at a glance. Simultaneously, the user can review a selected subset of the data in detail. Each chart includes multiple time-series plots that show station-specific information regarding the performance of automatic detectors, analyst-reviewed detections, and background noise statistics. In addition, because software maintenance actions sometimes have unintended consequences, the

 doi: 10.1785/0220150273

charts indicate maintenance events on the timeline as markers that highlight relevant software modifications, station processing parameter updates, and instrument calibration adjustments. The interactive display provides simple controls that allow the user to adjust the time frame under review. The display also provides options that allow the user to invoke an assortment of technical metrics, such as a 15-day moving average.

All station summary metrics are precomputed using a collection of processes that are continuously running in the background, and the results are stored in an Oracle relational database. As a result, the user can quickly review station summary statistics for a variety of time periods over the station's operational lifespan.

SOFTWARE ARCHITECTURE

Interactive visualizations are generated using the Highcharts and Highstock charting libraries developed by Highsoft AS (see Data and Resources). These open-source libraries are pure JavaScript and do not require client-side Flash or Java plugins to operate. However, either jQuery, MooTools, or Prototype Framework is needed to run the Highsoft software libraries. Here, jQuery is used extensively by the Highsoft libraries and for a variety of other purposes within the performance monitoring application. Highcharts is a general purpose charting library, which generates a number of standard charts, such as scatter plots, bar charts, and pie charts. Highstock is designed for creating interactive timeline charts like those used to analyze stock market data. The Highstock library provides a number of data navigation options such as a navigator time series, preset date ranges, datepicker, scrolling, and panning (Highsoft AS; see Data and Resources). In addition, the Highstock library provides the ability to display an event marker flag for highlighting the dates of key events that may affect the data the user is reviewing.

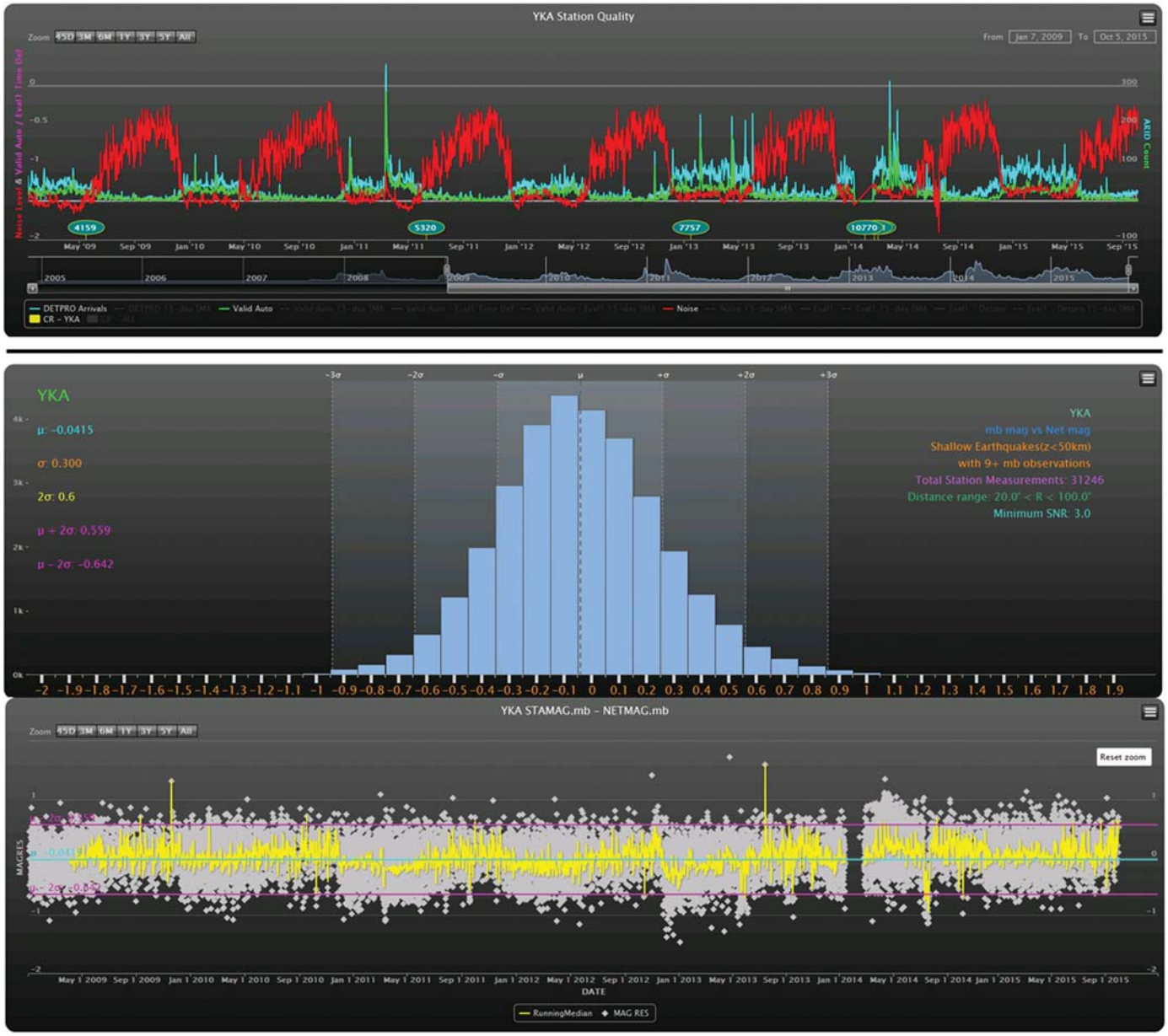
Interactive station summary charts are created in response to the client's request. Here, the web client represents the user's web browser. When the user selects a station from the performance monitoring home page, the web browser sends the request to the HTTP Apache web server. The server finds the requested file and sends it back to the client. This station-specific page is generated with JavaScript and receives supporting data by requesting files that are common gateway interface (CGI) scripts. Therefore, the server will execute the CGI process when called and send the results back to the client's web browser, where they are subsequently passed to the Highcharts or Highstock functions. In our case, all CGI scripts are written in Python, in which database transactions are carried out via the CX_Oracle Python module and time-series data analysis, and mathematical operations (e.g., moving average calculations and histogram binning) are carried out using the Pandas, Numpy, and Scipy modules. The interactive display is refreshed each time the user reloads the page, thereby maintaining the latest rendering of the data.

Station summary data are stored in an Oracle relational database and retrieved as needed by the CGI scripts. Database tables are populated daily by Oracle scheduler jobs that execute a collection of queries that retrieve station summary data from the US NDC operational database. These queries precompute a majority of the station statistics, such as station magnitude residuals (i.e., residual equals station magnitude minus network magnitude), which are stored in the MAGRES table. Each query executes a merge operation that simultaneously retrieves station summary data from the US NDC operational database and inserts it into the performance monitoring database. Precomputing statistics in this manner shifts the computational load to the server-side hardware, which reduces the response time of the tool on the client side.

STATION GALLERY DISPLAY

Figure 1 shows the interactive charts as they appear in the performance monitoring tool station gallery display. The top panel shows station-specific time-series data, which includes the number of automatic detections per day (Detpro), analyst-reviewed detections per day (Valid Auto), and background noise measurements (Noise) in cyan, green, and red, respectively. Each time series can display a 15-day moving average to smooth short-term fluctuations in the time-series data, which allows the user to see underlying trends in the dataset. Maintenance events and configuration changes specific to the selected station are represented by the event marker flags (e.g., CR-YKA) along the x axis, which are shown in cyan. The numbers inside the markers represent software change request numbers that are tracked using the Rational Team Concert software application. A brief summary of the flagged event is provided to the user simply by placing the mouse cursor over an event flag. There are several additional metrics shown along the bottom of the plot that are not activated. The user can easily toggle different traces on and off to change the information shown on the chart by simply selecting the desired feature from the legend. The ability to visually identify and correlate the effects a software change has on the system's performance is a key feature of the tool. This event notation process is similar to the one used by Google Finance, which allows the user to see how a press release influences the value of a publicly traded stock.

The x-axis date range can be set several ways. In the upper left corner of the plot, a series of preset data range buttons allow the user to set the time range to 45 days, 3 months, 6 months, 1 year, 3 years, 5 years, or "All," which will display the station's complete history. In the upper right corner, two boxes allow the user to manually enter the display start and end times. Along the bottom of the chart, a blue envelope provides the user with a simple snapshot of the station's history. Under the envelope, an adjustable slider can be expanded or contracted and moved over the date range of interest. Once the user releases the slider, the station summary display will automatically update. The user can also set the x-axis range by graphically drawing a window over the desired data range by placing the mouse at the start time, clicking the left mouse button, and releasing the left mouse button at the desired end time. These data visualization and selection features allow the user to see the entire data holdings for a station, quickly set arbitrary or targeted time frames of interest, and enables the rapid investigation of anomalous behavior at any point in the data archive.



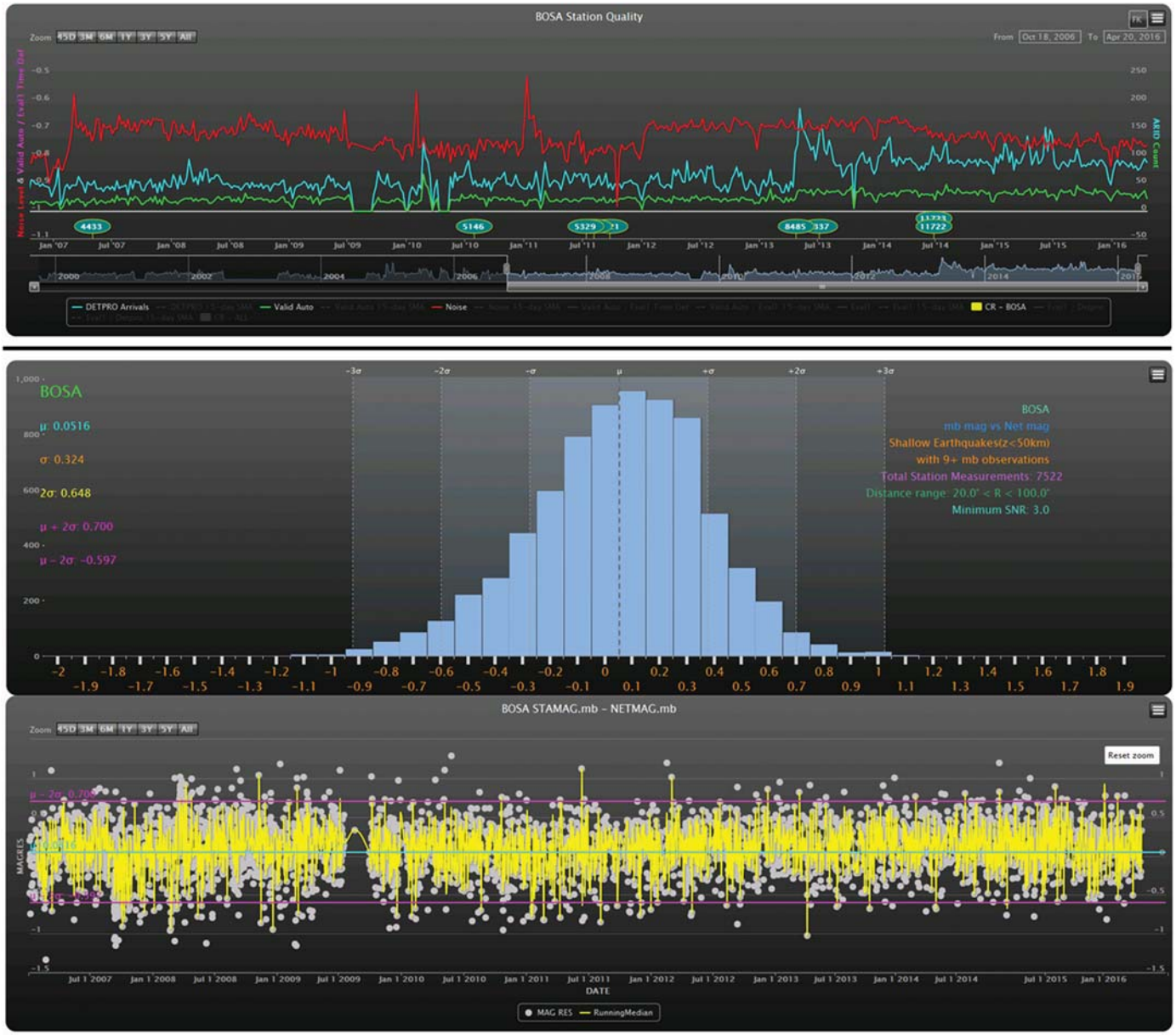
▲ **Figure 1.** Station gallery display showing performance monitoring metrics for the Yellowknife Seismic Array (YKA), located in Yellowknife, Canada. Top panel displays station summary statistics. The middle panel shows a histogram of magnitude residuals and other pertinent statistics, such as the mean and standard deviation, as text along the side. The bottom panel shows a time series of magnitude residual measurements as gray dots, and the running median is represented by the yellow line.

The center and bottom panels of Figure 1 show magnitude residual calculations. These charts are linked to the datepicker options in the upper left corner of the bottom chart, which ensures the scatter plot and histogram displays remain in sync. The bottom plot allows the user to set the desired data range using the same methods as described above for the station summary chart. The middle chart presents the user with a histogram showing the magnitude residual distribution over the selected time frame. The histogram chart was constructed using the Highcharts bar chart display. Relevant statistics describing the distribution, listed along the left side of the chart, provide the user with a set of simple diagnostics that would quickly highlight a magnitude bias or potential calibration problem over the selected time period. Event selection constraints used for compiling magnitude residual statistics are displayed along the right side of the chart. The bottom chart shows a time-series plot of the magnitude residuals over the same time frame, in which the dots represent residuals for individual events, and the yellow line is the running median. This display places the entire or selected time history of magnitude residual measurements at the user's fingertips, which allows the user to employ time-trend analysis techniques to identify potential instrumentation problems at any point in the data archive. Lines highlighting the mean of the time series and departures from the

mean by twice the standard deviation are shown in blue and magenta. The user can choose the datasets to display by toggling on and off the legend text for the desired data type.

APPLICATIONS

The tool's ability to allow the user to easily view long-term station summary statistics as a function of background noise variations provides valuable insight into how different stations perform over time. An example of this is shown in Figure 1, in which the relationship between the strong seasonal variations in background noise and signal detection statistics at Yellowknife seismic array (YKA) is apparent. Because YKA is located in northern Canada, strong seasonal fluctuations in background noise are typically observed, due to freezing and thawing cycles in the region. During winter months, background noise levels at YKA are low, due to arctic temperatures that cause the land and water in the region to freeze. Over summer months, noise levels at YKA are high, due to thawing occurring in the area. Detector performance is inversely correlated to background noise conditions, for which Valid Auto detection rates are low in the summer and autumn when background noise levels are high and elevated during winter and spring when background noise conditions are low. A similar seasonal variation is also observed



▲ **Figure 2.** Station gallery display showing performance monitoring metrics for the three-component seismic station located in Boshof (BOSA), South Africa. Top panel displays station summary statistics. The middle panel shows a histogram of magnitude residuals and other pertinent statistics, such as the mean and standard deviation, as text along the side. The bottom panel shows a time series of magnitude residual measurements as gray dots, and the running median is represented by the yellow line.

in the magnitude residual plot shown in the bottom panel, which is the result of larger signal amplitudes produced by the elevated noise levels present during the thawing cycle. This information is extremely useful for empirically tuning station processing parameters for optimizing performance.

The effect software modifications have on system performance is highlighted in the second example shown in Figure 2. The figure shows nine years of performance statistics for the Boshof (BOSA) three-component broadband station in South Africa. As before, automatic detection statistics are illustrated in the figure's top panel by the cyan trace. Between January 2007 and April 2013, the mean detection rate remained relatively constant at ~45 per day. On April 2013, a BOSA processing parameter change was implemented on the US NDC system that caused the mean detection rate to increase to nearly 75 per day. The detection rate increase is a direct result of implementing optimized process parameters that were derived through a rigorous detector-tuning process that is described in VanDeMark *et al.* (2013). Note the blue bubble at the bottom of the time-series plot with the number 8485. This flag denotes the reference number of the software change request that implemented the parameter update. The user can retrieve vital information regarding the software modifications made under this change request by reviewing the documentation stored in Rational Team Concert that is linked to that number. This documentation includes detailed information regarding the nature of the change, name(s) of the software engineer(s) implementing the modifications, a list of modified files, and the design of the implemented changes. Timely access to this information is vital when trying to troubleshoot a data processing problem that is adversely affecting an operational system.

CONCLUSION

The web-based performance monitoring application was promoted to US NDC operations in mid-2013. Since its deployment, the tool has been used on a daily basis to evaluate the performance of the US NDC data processing pipeline results and has led to the discovery of numerous problems that may have gone unnoticed otherwise. As performance issues are discovered, information regarding the issue is distributed to the appropriate personnel for analysis and resolution. The use of financial visualization techniques for analyzing the performance of the US NDC geophysical data processing system has been a great success. The YKA and BOSA examples demonstrate the ease at which subtle problems that are difficult to identify can be diagnosed and resolved. Our interactive web application enables users to rapidly assess the state of their system, provides a means to quickly analyze the nature of anomalous events at any point in the system's history, and provides a simple way to immediately visualize the effect a software change has on station or network processing.

Development of the US NDC performance monitoring tool is ongoing. Future work will include the incorporation of an interactive heat map that highlights the back azimuth of persistent noise sources. In addition, information regarding station data latency and database query efficiency will also be included.

DATA AND RESOURCES

Seismic data from the Yellowknife array and the Boshof (BOSA) three-component station can be obtained from the Incorporated Research Institutions for Seismology (IRIS) Data Management Center at www.iris.com (last accessed November 2015). Station processing results shown in Figure 1 were computed using the United States National Data Center (US NDC) software and are not available to the public. Details of Highcharts and Highstock charting libraries are available at http://www.highcharts.com/ (last accessed October 2015).

ACKNOWLEDGMENTS

This work would not have been possible without the successful collaboration of a multidisciplinary team of geophysicists, computer scientists, and seismic analysts. The authors would like to acknowledge the efforts of Mike Jezard (Leidos Corporation) and Randall Caldwell for their work regarding the performance monitoring database infrastructure. Their efforts have allowed the tool to quickly extract large datasets from the United States National Data Center (US NDC) operational database that has contributed to the seamless operation of the tool. In addition, the authors would like to recognize the efforts of the US NDC software test and geophysics teams: Ryan Stutzman, Fred Ward, Glen Meldrum (Leidos Corporation), Thomas VanDeMark, Brian Pope, and Gene Ichinose. The results presented in this article are solely the opinion of the authors; they do not represent the official position or policy of the United States Government.

REFERENCES

McNamara, D. E., and R. I. Boaz (2006). Seismic noise analysis system using power spectral density probability density functions: A stand-alone software package, *U.S. Geol. Surv. Open-File Rept. 2005–1438*.

VanDeMark, T. F., W. N. Junek, B. M. Pope, A. Henson, F. Ward, G. A. Ichinose, A. Poffenberger, and R. C. Kemerait (2013). Overview of detector tuning methodology at the United States National Data Center, *CTBTO Science and Technology Conference*, Vienna, Austria, 17 June 2013.

William N. Junek
Charles A. Houchin
Joseph A. Wehlen III
Alan Poffenberger
Air Force Technical Applications Center
10989 South Patrick Drive
Patrick Air Force Base, Florida 32925-3002 U.S.A.
william.junek@us.af.mil

Anne S. Henson
Leidos Corporation
Melbourne, Florida 32901 U.S.A.

Published Online 29 June 2016