**Chapter 6: HDSS Data Management**

**TCI Dropsonde Data Management Plan**

A large number of XDD (eXpendable Digital Dropsondes) are planned to be released during the field phase of TCI. The dropsonde data requires some quality control due to equilibrium adjustment from the aircraft, potential data dropouts from lost radio telemetry, noise in the sensors, slow response of the humidity sensor, and potential spikes or failures of the sensors. The data must also be distributed to the PIs during the field phase in usable file formats and plots, and archived for long-term access and distribution. This data management plan describes the goals for near real-time and final archival, visualization, and distribution of these data. A summary schematic of the processing chain is provided in Figure 1.



*Figure 1. Summary schematic of data processing and management for TCI.*

***File formats and processing***

The High Definition Sounding System (HDSS) initially records the raw, binary logs from the XDD radio transmissions that include telemetry and atmospheric data. There is one file per receiver and are in a proprietary binary format. These files are very large since they contain all of the transmitted information. The PIs do not need access to these files, and will only be used by Yankee. These files are hereafter referred to as “radio logs”.

A launch “origin” file is also recorded that contains data from the dispenser that initialized and launched the sondes. It is only one line of text for each sonde that provides the details of the initialization including the channel, time, latitude, and longitude. The in-flight sounding data are then differenced from these values. In the event that the origin file is lost it can be reconstructed from the radio log and decoded sounding data, and aircraft flight-track data, but this reconstruction will result in absolute time uncertainties of a few seconds, and absolute position uncertainties which correspond given the aircraft flight speed.

The on-board computers of the HDSS process the data as received (data in the radio logs). The processing involves "merging" the data from the multiple receivers to arrive at a best estimate of the transmitted bits as they have been received, then decoding the RLL-Viterbi Forward Error-correcting Code (FEC) to arrive either at transmission-error-corrected data or a rejected packet (all data at that time-step will  be NAN). Additional details on the on-board processing are provided in the XDD section below.

The first level of processing from the HDSS is a decoded, packed, binary file for each individual sounding. These files are in a “sqD” file format which can be read by software provided by Lee Harrison called “SQZ” and converted to an ascii format. The SQZ software is open source and available to scientific users. An auxiliary file is required for formatting and naming consistency, as well as the control-set data and parameters for the Viterbi FEC.

The resulting binary files are compact enough (~25 kB each) to be transmitted from the airplane in real-time over the satellite downlink. These files are hereafter referred to as “sqD files”. If the sqD files are converted to an ascii format that has no additional processing applied, then these are designated as “Level 0” files. The Level-0 output is a literal unpacking of the packed data the sonde sent, as controlled by the sqz script, such that the sqD and Level 0 files are identical other than the storage file format. Level-0 product is the calibrated output in engineering units of every sensed measurement from the sonde. Users should note that the calibrations are stored within the sonde, the data are sent calibrated by the sonde.

The process of unpacking the binary files also optionally can generate a QC report that yields important information, and also prepare quick-look graphs via GNUPLOT, including Skew-T/log-p (hereafter “skew-t”) plots.

The second level of processing involves modifications to the data to reconstruct the position data and to remove noise and gross errors from the measurements. This step is intended to produce a high quality dataset as close to the original data as possible. This processing step, hereafter referred to as “Level 1”, involves both the SQZ program and the Atmospheric Sounding Processing Environment (ASPEN).

In the first step, SQZ reconstructs a latitude and longitude and altitude for every time-step from the high data rate U, V, W-data and the lower rate latitude and longitude differences (from an initial pre-launch position which is recorded and altitude.  This is done using a G-H (aka Alpha-Beta) complementary filter (simplified relative of a Kalman Filter).  This is just integration, but it does reduce the impact of single-sample variances in the latitude/longitude/altitude data. The apparent low noise in the result is not so much a property of this filtering but is rather primarily the consequence of a newer GPS unit and better antenna/amplifier/filter for it. The filter coefficients are recorded with the data to ensure reproducibility. SQZ can also produce some derived quantities such as potential temperature, equivalent potential temperature, wind speed, and direction.

The second step in the Level 1 processing is quality control by the ASPEN program. The intermediate Level 1 ascii file produced by SQZ will be output in a comma-separated value (CSV) format that can be read by ASPEN. The format conversion and readability by ASPEN has been tested and is working properly. A minimal quality control on the pressure, temperature, relative humidity (PTH) and wind data is performed through a combination of buddy (neighbor) checks, physical limit checks, outlier checks, and low-pass filter checks. The dynamic correction for temperature and winds typically applied to RD-94 dropsondes is turned off, as is the final low-pass filter that smooths the data in the vertical.

The ASPEN program has several output formats, but here we choose the “CLASS” format which is a standard, commonly used ascii format. NetCDF and WMO Tempdrop output is also possible upon request. ASPEN also generates a skew-t plot as part of the processing. The resulting Level 1 output and skew-t plots will be archived during the field phase.

The next level of processing will only be undertaken after the field phase. “Level 2” output will be the final data format suitable for publication quality results. This level of processing will include acceleration of the humidity sensor to account for sensor lag. The new thermistors on XDDs are so fast there is little reason to accelerate the temperature measurements. The RH sensor is slow however, and there are confounding interferences from sonde fall-speed mode(s) if it is not in the fast mode. An accelerated RH should be applied for final scientific use of the XDD data. Level 2 processing may also involve some additional filtering of the PTH data, but this will be assessed during the field phase.

The final Level 2 data will be targeted for release within 2-3 months after each flight. The final details of the Level 2 processing will be decided by the PIs after scientific inspection of the Level 1 data during the field phase in consultation with Yankee staff and Lee Harrison.

***Archival and distribution***

The HDSS support team will be responsible for getting the data off the aircraft in a timely manner, either in real-time via the satellite downlink or manually at the conclusion of a flight. While the PIs would like to receive the data in real-time, the requirement is for data access 2-3 hours after the end of a flight. Limited staff resources should be directed towards meeting this requirement over the lower priority real-time satellite access. Once the data is off the aircraft, the binary SQD Level 0 files will be made available on a secure Yankee FTP site. Initial SQZ processing will be performed independently by the SUNY-ASRC group and the University of Hawaii group to produce Level 0 and 1 files. It is requested that Yankee archive the radio logs in addition to the Level 0 files in the event that re-processing of the telemetry data is required.

At this point, the data will take two parallel paths. Some limited redundancy between the first and second data pathways will assure that products are available in one or more locations during and after the flights. Yankee systems and Lee Harrison will conduct the first automated pathway. Plots in the linear pressure or height space will be generated in real-time as the sonde falls for quality assessment, and will be viewable from a remote desktop application. These plots will also be distributed informally amongst interested PIs, and possibly added to the EOL field catalog if desired. Skew-t’s will be produced by SQZ and sent to the NASA MTS system shortly after splash-down of the sonde. These skew-t’s will be available to assist in situational awareness and mission planning. This data pathway has been tested during recent DC-8 test flights and appears to be working adequately.

The University of Hawaii will conduct a second automated data pathway. A script will obtain the Level 0 SQD files from the Yankee FTP site, and process through SQZ and ASPEN in an automated batch mode. The ASPEN processing differs from the normal interactive mode used when the data is destined for GTS distribution. The batch mode processing will produce the final Level 1 CLASS format and a skew-t from the QCed PTH data. If data is available via satellite link then products will be made from the real-time output, but the primary pathway will be for sondes processed at the end of each flight after the Ethernet download. If there are significant differences between the real-time and post-flight skew-t’s then multiple versions will be uploaded to the catalog with appropriate version control. If the differences are minor then the post-flight skew-t’s will just replace the real-time versions on the catalog. The UH data pathway will be tested during the June and July test flights

The Level 1 data file will then be transmitted automatically via FTP upload to the EOL data archive. These files should be clearly marked as preliminary files to ensure they are distinct from the final Level 2 files available after the project. Further distribution of the Level 1 data files to TCI PIs will be through a web interface. The skew-t plots will be transmitted automatically via FTP to the EOL field catalog and SUNY Albany. A dedicated naming convention for the files will ensure that the plots are available on the field catalog under the “Research products” section and listed by time. The catalog plots will not be geo-referenced, but will be available through a web interface for long-term access with the other model, satellite, and forecast products used during the field phase.

At SUNY Albany, the skew-t plots will be available through a web interface in a geo-referenced map overlaid on satellite imagery. This website will allow the visualization of the spatial sampling from each flight as well as the vertical profiles of PTH and wind data. The satellite overlay will enable PIs to determine where the sounding was in relation to the cloud features.

Upon conclusion of the field phase, the Level 2 data products will be put in the EOL data archive for web access and long-term archival. A documentation file will accompany the dataset and each data request that should contain the following metadata: 1) file format descriptor; 2) sensor information such as accuracy, resolution, and time constant; and 3) details on the specific steps used to process the data. These files and the accompanying documentation will represent the final XDD dataset, and should receive a Digital Object Identifier (DOI) to enable reference in scientific publications. The long-term archival and distribution will be handled by EOL as part of their custodial duties. The SUNY Albany website will be maintained for as long as possible given resource constraints.

***On-board processing of XDD data***

It is beyond the scope of this document to fully describe the issues of telemetry recovery and the properties of the FEC and decode, but some basic details are provided here for interested PIs.

The control data will be universal to all sondes in the TCI experiment and stored on the sonde flash memory. Yankee wishes to keep the control information and algorithm for the FEC private.  A side effect of this nonlinear RLL-Viterbi is that it is effectively a permutation-substitution cipher.  Security was not a goal of this FEC, it arises naturally out of the math which creates an optimal RLL-Viterbi. Without the control information for the Viterbi the raw data in the radio log are effectively enciphered with a P-S cipher with an equivalent key of approximately 1600 bits.  At present only Lee Harrison, Todd Allen and Eric Griffen (the latter two at Yankee) hold this effective "key."  (This key is not unique, the statement of the equivalent key size in binary bits is an approximate estimate of the number of equally-good equivalent control sets).

The basic performance of the FEC for recovering data with random "flipped bits" is shown in the Figure 2. (Please note that one flipped bit = 2 randomly corrupted or lost bits; this code-rate 3/8 RLL Viterbi can correct packets with as much as roughly 1/5 of the bits randomly corrupted).

The real-time processing done on the HDSS in flight is computationally limited to guarantee that the real-time system does not overload and back-up. Todd Allen will provide benchmarks about the limits he will set with the current hardware, but they will not be high enough in real time to yield data under conditions on the right side of this graph.  
  
One of the advantages of a long-convolution Viterbi vs. a block-code based on the Galois algebra is that the block codes guarantee a correct solution up to a specified number of error bits, and unfortunately also guarantee a false solution (with no diagnosis) beyond that. (Reed-Solomon and "Turbo" codes are the best-known block codes).  
  
Viterbis have the property that the work is an exponential of the number of errors; the probability of a correct solution is very high up to a number of erring bits, where it starts to fall rapidly. This can be seen in the graph.  A work limit set below the error avalanche discriminates against false solutions, independent of the number of errors (this discrimination is not perfect, but strong.)  
  
Viterbis are trailing convolutions, the data protection is strong for early data in the packet and weakest at the end.   The organization of data in the telemetered packet (which can be seen in the attached .sqR file) puts meteorological data first, "housekeeping" data last.  In practice some decode errors in house-keeping data are tolerated, and seen, in processed data.  
  
At present there are a set of unknown nor not-fully-optimized issues with regard to the telemetry processing hardware and algorithms.  It is beyond the scope of this note to detail these, however a major and necessary reason to obtain the radio-logs is to facilitate such work.  
  
In principle, and also in practice, "missing" records (censored by the work limit of the in flight processing) may be additionally recovered by post-processing.   However in this regard, we do not have optimal reprocessing algorithms in hand at present, and there are limitations as signal-to-noise degrade that effectively produce an "error wall" beyond which no recovery can be expected.



*Figure 2. Performance of the Forward Error-correcting Code with random flipped bits.*

**HIRAD Data Policy**

If Bandwidth allows, HIRAD will likely be able to generate very rough quicklooks of uncalibrated (or poorly calibrated) brightness temperatures. After downloading the full data (post-flight), they will apply quality control and produce initial calibrated brightness temperatures (and possibly wind speed retrievals). This will probably take hours to days. The final product (fully calibrated brightness temperatures, and corresponding wind speed retrievals) should be ready a few months after flights and will be in the netcdf file format on the EOL Archive site.

**TCI On-line Field Catalog**

EOL will customize, deploy, and maintain a web-based TCI on-line Field Catalog that will be operational during the TCI field phase to support the field operational planning, product display, and documentation (e.g. facility status, daily operations summaries, weather forecasts, and mission reports) as well as provide a project summary and “browse” tool for use by researchers in the post-field analysis phase. Data collection products (both operational and research) will be ingested into the catalog in near real time beginning in July 2015. The Field Catalog will permit data entry (data collection details, field summary notes, certain operational data etc.), data browsing (listings, plots) and limited catalog information distribution. Catalog Maps is a GIS tool that is integrated into the Field Catalog and provides real-time situational awareness of project operations. EOL will provide Internet Relay Chat (IRC) chat services as the primary communications tool between the various ground-based and airborne facilities. Public access to the on-line Field Catalog will be located at: http://catalog.eol.ucar.edu/tci/

EOL Field Catalog staff will coordinate with NASA staff operating the Mission Tools Suite (MTS) system used in SHOUT to ensure both systems contain relevant products from each campaign to assist in the overall collaboration and coordination efforts. EOL will monitor and maintain the field catalog through the duration of the field deployment and also provide in-field support and training to TCI project participants. Preliminary data will not be retained as part of the Field Catalog, but will be password protected and available through the **TCI final archive**.

**Data Archive**

It is important to TCI researchers and the scientific community to maintain a final long-term archive that is easy to access. TCI will work closely with both SHOUT and IFEX to collect multi-disciplinary high-resolution observational datasets during the field campaign. Data will be collected by various airborne and ground-based facilities, and will be stored in various data formats, locations and time scales. EOL will ensure that arrangements for all required datasets will be made and develop and implement a series of Data Management web pages to provide “one-stop” access to all distributed TCI data sets, documentation, on-line field catalog products, collaborating project data archives, and other relevant data links.