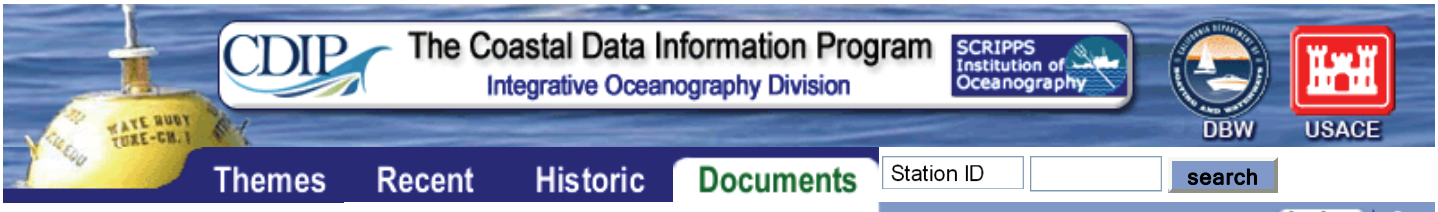


LATEST NEWS: Data updates will be affected by a planned power outage Wednesday 20 March 7am - 10am

Data Processing

Since its inception, CDIP has been committed to using a fully-automated, realtime processing system for all of the data the program collects. This emphasis allows us to provide the most up-to-date and relevant coastal data to a wide range of users, from surfers and boaters to the National Weather Service. But the data are not only timely; the processing system includes a range of rigorous quality control tests, ensuring that CDIP data are highly accurate and reliable as well.

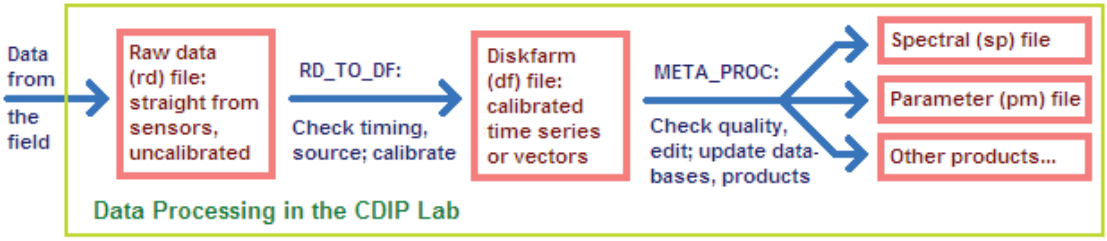
System Organization

Every hour of every day, the computers in the Lab, CDIP's central computing facility at SIO, contact all the active shore stations to collect their latest data. When these data arrive back in the Lab, they are sequentially passed, file by file, through CDIP's automated processing and distribution system. This system performs a wide range of analyses and data transformations, producing everything from error reports and diagnostic e-mail to condensed paramters and web tables.

When the data first arrive in the Lab, they are in the form of rd - raw data - files. This is data directly as read from the sensors in the field, without any significant modification or editing. The data in rd files have not been decoded or calibrated; they are effectively a byte-by-byte record of a sensor's output.

Turning this raw data into to all of the valuable products and information that are found on the CDIP website is basically a two-step process. First, after verifying the source and timing of the rd file, it is calibrated and used to produce a df - diskfarm - file. The df files constitute CDIP's core data set. An accurate record of the readings made by each sensor at each point in time - this is the essential foundation for all of the information that the program provides.

While the df files are accurate records of a sensor's readings, another major question remains: how well is the sensor measuring what it's supposed to? The second step in data processing is to address this very question. A range of quality control checks are performed on the data, to check if they are suitable for further processing. If so, a variety of calculations and transformations are performed on the data, and finally the results are distributed to all the appropriate products. (For a more detailed version of the image below, please see the **processing flowchart**.)



Software

At the heart of CDIP's processing system lie two FORTRAN programs, *rd_to_df* and *meta_proc*. These programs process hundreds of data files a day, supplying thousands of users with the information they need.

RD_TO_DF: Raw data and the diskfarm

As noted above, *rd_to_df* performs several important operations on the sensors' raw data. First, it checks that the source of the data are correctly identified. In the header of each rd file, a number

of basic sensor charactersitics are noted: the sensor type, its location, calibration factors, etc. All of this information is checked against CDIP sensor archive once *rd_to_df* begins to process the file.

The sensor archive is a database that holds complete descriptions of all the sensors that have been deployed in the field to collect data for CDIP. Serial numbers, deployment and recovery dates, water depths: they are all detailed in the sensor archive. The *rd_to_df* code checks the header information and data format in the rd file against the description in the sensor archive. If any discrepancies are noted they are recorded in the processing error logs.

Sample information from the sensor archive

```
STATION    : POINT REYES, CA
NUMBER     : 029

CHAN       :  1
Frm  Start Date       End Date              Gauge Type              Depth    Serial     Latitude   Longitude
       (UTC)            (UTC)                                       (m)     Number      dec min    dec min
 1  04/12/1981 08  06/17/1982 16  Datawell non-directional buoy      73.2  WG7533-9     37 56.300  123 03.800
 2  07/03/1982 08  05/02/1983 16  Datawell non-directional buoy      73.2  WV67681      37 56.300  123 03.800
 3  12/05/1996 21  09/20/1997 12  Datawell Mark 2 directional buoy  570.6  WV30210      37 56.800  123 28.400
 4  10/13/1997 21  06/25/1998 17  Datawell Mark 2 directional buoy  543.2  WV30253      37 56.700  123 28.200
 5  06/25/1998 17  10/21/1998 03  Datawell Mark 2 directional buoy  543.2  WV30254      37 56.700  123 28.200
 6  11/04/1998 08  09/12/1999 07  Datawell Mark 2 directional buoy  543.2  WV30261      37 56.700  123 28.200
```

Next *rd_to_df* checks that the rd file's time is correct and and confirms that all the data in the file can be assigned a definite time. Some sensors' data are recorded with regularly-spaced sync words and time tags injected into the data stream; *rd_to_df* decodes the tags to ensure that the timing is accurate. Other sensors, such as Datawell buoys, include sync words and counters, which are similarly examined. Minor timing problems are noted and corrected; major problems may result in the immediate cessation of processing for that file.

Once the source and timing of the data have been definitively established, *rd_to_df* is ready to create diskfarm (df) files. For stations with more than one sensor in use, the rd file will contain data from several different sensors. For instance, an rd file from Scripps Pier (Station 073) may include wave energy data from an underwater pressure sensor, wind data from an anemometer, and air temperature readings from a temperature sensor. So the first step in creating df files is separating out the data from the different sensors. Once this is done, the values from each sensor can be calibrated. The calibration factors recorded in the sensor archive are then applied to the data as appropriate, and the resulting values are written to single-sensor df files.

There are two general formats for diskfarm files. For directional buoys, the df files are composed of Datawell 'vectors', 10-byte lines containing error, spectral, displacment, and parity information. For all other sensors, the df files are decoded time series; depending on the sensor type, these time series values may be vertical displacements, water column heights, temperatures, etc. In both of these forms, the df files constitute CDIP's core data repository.

Further Documentation:

Full rd_to_df documentation

Sensor archive description

DF file format: time series data

DF file format: Datawell vectors

META_PROC: From the diskfarm to our users

While the diskfarm files form a very detailed and comprehensive database, they are but the starting point for CDIP's analyses of wave climatology and ocean conditions. The FORTRAN program *meta_proc* takes the df files, checks the quality of their data, performs a range of complex calculations - such as spectral and directional wave analyses - and then produces a variety of output, such as spectral (sp) and parameter (pm) files.

As the program name somewhat pretentiously implies, *meta_proc*'s first task is to figure out how each df file should be processed. What quality control checks should be applied to the file? Which other files should it be grouped with? What calculations should be performed, and what products should be generated?

For *meta_proc*, the key to answering these questions is CDIP's processing archive. Like the sensor archive, the processing archive is a database containing a number of time frames for each station. These time frames give complete processing instructions for a station at any point in its history. For example, for an array of pressure sensors, the processing archive will state which of the sensors should undergo "Gauge comparisons", a rigorous set of quality control checks designed to confirm that the data from closely-associated pressure sensors are in agreement. The processing archive will also state which sensors should be used for directional processing. For a station like Harvest Platform (063), with eight pressure sensors in close proximity, the sensors used for directional processing may vary considerably over the life of the station.

Sample information from the sensor archive

```
STATION   : 006 - SANTA CRUZ HARBOR, CA
p1: ARRAY
p4: SURGE

Stream  Processing        Frm    Start Date      End Date      Stream/channel list
          Type                     (UTC)          (UTC)
 p0     Default            1    08/24/1977 08  09/23/2001 00   01*02*03*04*05

 p1     Parameter          1    08/24/1977 08  09/23/2001 00   p2

 p2     Waves              1    08/24/1977 08  07/10/1979 08   01*02*03*04
 p2     Waves              2    02/06/1980 08  10/30/1982 16   01*02*03*04
 p2     Waves              3    10/30/1982 16  12/04/1983 20   01*02*03*04
 p2     Waves              4    12/10/1986 16  09/23/2001 00   01*02*03*04

 p3     Gauge comparison   1    08/24/1977 08  09/23/2001 00   01*02*03*04

 p4     Parameter          1    08/24/1977 08  09/23/2001 00   05
```

The processing archive also spells out which sensors should be used to produce all of the plots, tables, and other products that are generated from the diskfarm and made accessible on the web. By setting up a number of "parameter streams" in the processing archive, different products can be generated for a single station and presented to web users as separate data sets.

Once *meta_proc* has consulted the processing archive and assembled all the instructions for handling a station's data, it proceeds through three main stages of processing. First comes quality control and editing. The data are subjected to a range of tests - checks for extreme values, spikes, abnormal distributions, etc. (For a more detailed description of the tests used, please see the following section, on **quality control**.) When problems are found in the data, the values may be edited - perhaps removing a spike from a time series - or they may be rejected as unfit for further processing.

If the data pass the QC tests, processing continues to the next stage: decoding and transformation. This is where FORTRAN's number-crunching facilities are used to great advantage, as the bulk of the calculations occur at this point. In wave analysis, for instance, the time series from pressure sensors undergo spectral and directional analyses, and a range of algorithms transform the data into spectral coefficients, condensed parameters (Hs, Tp, etc.), and the like.

At this stage of the processing - as at most others - Datawell directional buoys are treated somewhat differently than other sensors. Datawell buoys perrform spectral and directional analyses internally, and the buoys output these spectral data along with the buoy's displacement time series. For this reason, CDIP does not actually do any number-crunching for the Datawell buoys. Instead, *meta_proc* simply decodes the spectral information produced by the buoy itself.

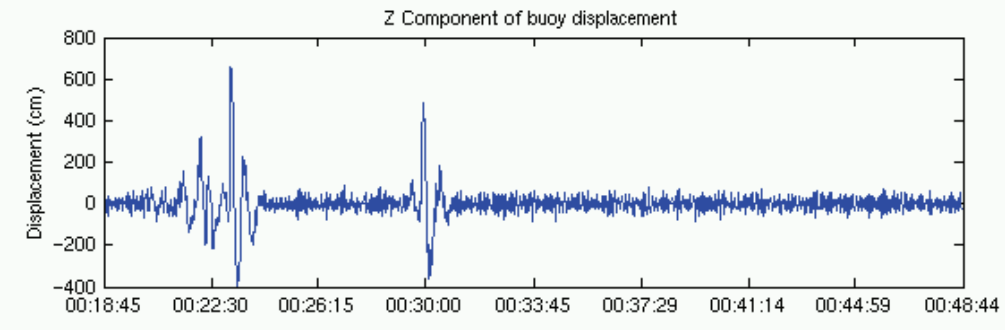
Once all of the calculations have been completed, further QC is performed: are the results reasonable? If so, there is one remaining stage to the processing: distributing the results to all of the relevant products. Since a single station's sensors may be used to generate a number of different data sets at any point in time, *meta_proc* carefully follows the processing archive's instructions to ensure that its results are added to all of the appropriate files and databases. Once this last stage is complete, all of CDIP's standard products are easily accessible to web users, the National Weather Service, and a number of research insitutions.

Further Documentation: **Full meta_proc documentation** **Processing archive description**

Quality Control

CDIP needs to provide its users with data which are not only timely, but accurate as well; this is a responsibility that is taken very seriously. Rigorous quality controls are implemented at several stages in the processing, and catch the vast majority of problematic files.

As described above, the first quality control checks ensure that each data file is properly attributed, with its full provenience - both time and place - accounted for. Then, as the data are processed by *meta_proc*, CDIP's full suite of QC algorithms and analyses is deployed. For time series data, a wide range of analyses are used, with different tests applied to different data types. For water column and vertical displacement time series - i.e. wave measurements - the checks include: extreme values test, spike test, mean shift test, flat episodes test, mean crossing test, equal peaks test, acceleration test, and period distribution test. Some of the tests edit the time series, cleaning up the data where possible; others simply flag it bad. Where multiple sensors are deployed in close proximity, the above tests are followed by a battery of comparison tests, to ensure that the sensors are in agreement. For a full description of the tests used and the data types to which they are applied, please refer to our **QC documentation**.

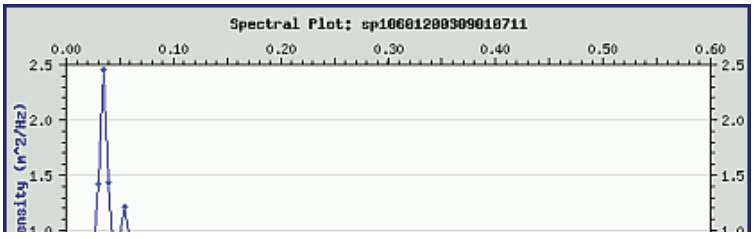


Time series from a buoy with a bad hippy (heave-pitch-yaw) sensor

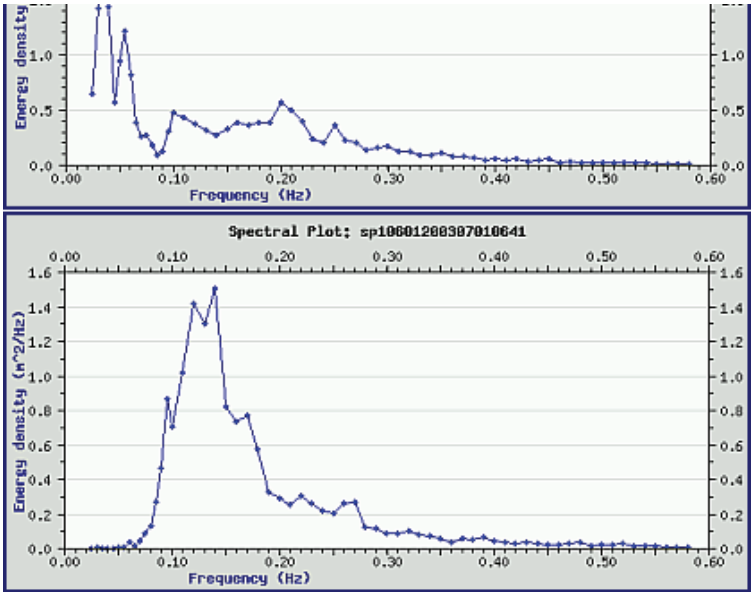
- **Sample time series from directional buoys**

Once the time series have been processed, the resulting values - condensed parameters and spectral information - also undergo QC checks. For Datawell buoys these checks are quite extensive, since the buoys perform their own time-series handling and spectral processing internally. For example, the time series above shows a problem - spikes in a large, long-period waveform - that CDIP's time series tests could easily identify. But since these data are from a Datawell buoy, the time series was processed internally, so no CDIP editing was applied.

Nonetheless, CDIP's post-processing checks correctly identify the problem with this file, since the resulting spectral distribution



and parameter values (in this case Tp) are skewed. Here the long-period spikes result in a spectral shift to lower frequencies, and in an unnaturally high Tp value (approximately 28 seconds).



Two spectral plots. The upper plot shows the shifted distribution of the time series data above; the lower plot shows a more typical spectral distribution.

In addition to the checks outlined above, there is one more full stage of automated QC applied to CDIP data. While all the previous analyses are applied to a single file's data, the final QC tests address a station's data over longer time periods, dozens of files. Once per day, all of a station's recently acquired condensed parameters are compiled and compared. Once again, a range of tests check for spikes, unusual values, and the like, notifying CDIP staff via e-mail if anything seems amiss. (Please read our **post-processing documentation** for more details.)

Only data which pass all of these QC tests are publicly released. Data which are flagged as suspect or bad at any point along the way are archived at CDIP but not publicly disseminated except when specifically requested. All of these automated measures, combined with periodic visual inspections, are very effective in preventing the distribution of erroneous data.

Further Documentation:

Automated QC **Non-automated QC** **QC summary table** **Time series editor** **Direction buoy check factors**

Back to top

CDIP's major funding contributors are the US Army Corps of Engineers and the California Department of Boating and Waterways.