# PROCESSING AND FILING DATA – Office procedure example

## DATA REGISTERS AND FILING

This section sets out the procedures to be followed to provide an adequate, documented record of each step in processing the hydrological data, from the time of its measurement in the field to its final archiving in the Hydrometric Database.

### General

Hydrological data are valuable in that they are relatively expensive to collect, are irreplaceable and have the potential to have very high value following certain events. To realise and maintain their value, there must be a means of verifying the accuracies and giving assurance that errors are largely absent. Thus the traceability of the data and of the methods used to collect and process them must be available in a readily followed form. Many of the provisions of this manual are aimed at achieving this traceability, in conjunction with the efficient processing of the data while preserving and verifying its integrity.

The data processing system includes provisions to

* register the data after collection to confirm its existence and track its processing
* keep backups of the data in its original form
* positively identify the individual batches of data at the various stages of processing
* identify the status of data as to its origin and whether it has been verified as fit for use
* present and store evidence of any modifications to the data
* file all field observations which verify the data
* control the amount and type of editing which can be performed and the authorisation to do this.
* present the data in a number of ways for checking and auditing by trained persons who are independent of the process.

**Registers**

As soon as batches of data reach the field office they are entered in one of a set of registers, organised as to station and data type, and in chronological order.

These registers are on hard copy in a folder in the office or spreadsheet based (Excel) document on a server, and are updated as batches of data arrive. Initially this involves noting the start and finish times of the data batch, and progresses with confirmation of editing, checking and updating to the archive. The requirement for these steps to be signed and dated is deliberate and mandatory, in order that staff take responsibility for and “ownership” of their work and its progress.

The registers thus contain a verified chronological record of data processing activities in the field office. There are separate register forms for:

     (a)  Water-level data;

     (b)  automatic raingauge data;

     (c)  flow gaugings;

     (d)   original record and back-up data;

     (e)   Updates to the database including checks made at this stage and stored in a folder or directory of all correspondence and forms.

These registers are to be updated as data progresses through the system. All columns within the registers must be filled in with the required information including the initials of the processor where required.

**Identification of Original Records**

All data are to be permanently identified with the site numbers and the dates of observation. These can be written directly on chart records and gauging cards, and on original plots of telemetered data (in the event that the plotting process does not do this adequately). Older charts require a stamp or label to prompt for the dates and times, staff gauge readings, etc. for both chart on and chart off.

Electronic data files, whether the primary file for updating or a backup file, should have their site identified in the header line(s) and be filed in an appropriate directory. However, if these are binary files, they can often only be identifiable by the file name, e.g. StationNumber\_03.csv, but the register will associate these files to a date and time.

**Plots of Original and Edited data**

At initial processing on local PC, the first main step is to produce plots of the data. These “original plots” are both evidence of the original state of the batch of data and a backup for all subsequent data processing. They are kept as archived documents in site folders. The date of processing and the initials of the processor shall be hand-written on the plots, along with explanatory notes on any real or apparent anomalies and on any editing done. Because a plot has the potential to illustrate many of the various types of errors possible, and because they are a convenient place to write notes, these original plots together with subsequent ones following editing provide the main evidence of traceable data integrity.

Any backup data should also be processed and included as evidence.

Plots should show evidence of the previous batch to verify the continuity of data, e.g., an over-plot of several days previously, taken from the archive.

This evidence can also be filed electronically in a Word document or pdf and kept on the local server under “Site files”.

**Telemetry data logging**

Daily data retrieval schedules are essential to rapidly identify any stations that have either developed problems with recording or have lost communication links. They are necessary to rectify faults immediately and reduce the likelihood of missing record. The telemetry log is also useful in determining when telemetry problems were encountered. Alarms can be set up in Flosys to provide text or email warnings of gaps, steps, low voltage, etc.

**Electronic data**

When downloading dataloggers, ensure that a system is used for backing up data files in the field to prevent accidental loss of original data. In many cases this will be achieved by the datalogger still retaining the data after it has been downloaded. In other cases it will be achieved by the telemetry system. However when the logger is re-programmed, either the data should be retrieved by telemetry just prior to this and the second copy made onto a laptop, or in the absence of telemetry, a second copy must be made onto a memory stick, disc, etc.

The files must not be deleted from these until the data is safely stored and backed-up on the local database and all required registers have been completed.

## INITIAL DATA PROCESSING

This section sets out the data processing steps to be applied to the different types of data.

Assembling batches of data:

Where possible plan field trips to download data monthly or after the end of the quarters (31 March, 30 June, etc.) so that data can be verified as being correct and not subject to errors from silt, fouling, slumping, mechanical issues, etc. On occasions, it is recognised that it’s not practical or economical to visit all sites this way. Examples are isolated sites such as some off-shore and alpine stations.

The obvious alternative is to update telemetry data, after checking it as far as possible from the office. When this is done, the end of the batch shall be at midnight on the last day of the quarter. This arbitrary date and time (e.g., 20110930 240000) will provide a signal that the data are non-verified.

### Adjustments to time and/or water-level

Where adjustments are to be made to either time or water-level throughout the data, the following must be answered before proceeding:

(a) Can you provide valid reasons for the adjustments?

(b) Are the adjustments consistent with the previous batch of data?

(c) Is any follow-up field work or problem-solving necessary and has this been added to the field programme task list?

Answers to (a) and (b) must be fully explained and recorded on the original plot (which will be filed in the station’s file as above). Unless the adjustments are minor, they should also have a Comment filed on the archive.

### Checking Water-level Series Data

The first check is that the start date, time and values for the batch match up (within the time increment, as applicable) with the finish values for the previous batch. This is available from the register, and quickly checked.

The initial processing should provide the maximum and minimum gauge height and their relevant times for the batch. Enter these values in the water-level register (Form WL) and check that they are reasonable. (Any unusually high or low values should be a “flag” to check that the values appear to be correct).

The next step is to plot the batch at a suitable scale and examine the hydrograph to detect such problems as:

1. silting (mainly at stilling well sites), which will round-off peaks and cause “hung-up” recessions
2. “spikes” or small numbers of values which are obviously out of context and incorrect due to an unnaturally large change in value between adjacent data points. Such occurrences can occur with errors in the digital values between the encoder to datalogger interface
3. gaps in the data
4. errors induced by the field staff, such as pumping water into a well to flush it
5. restrictions to the movement of the float/counterweight system or the encoder (perhaps caused by the cable being the incorrect length
6. vandalism
7. debris caught in the control structure, the control, or other damming or backwater condition
8. and others.

Note that such problems, if not detected on the field visit, must be investigated at the earliest opportunity. In cases where the cause has not been positively identified, full processing of the data should be delayed until it can be investigated on-site.

When the plot is complete, attach the printout containing the parameter entries and add the following by hand:

(a)  The date the data was plotted;

1. Signature of the processor;
2. Notes of all corrections to the data, plus any subsequent actions carried out that change the data in the file from that which was originally plotted (e.g. removal of spikes, inserting manual data resulting from silting, etc.).
3. Plotted with an over-plot of the previous batch from the archive to show continuity.
4. Field log sheets attached from site visits.

For stations with electronic data capture, the first plot becomes a permanent record showing the original data as collected.  If any corrections are applied, another plot is necessary to show their effects.

Both shall be filed in the station’s file as described above.

If the registers and station files are only electronic, the processing evidence will need to be stored in a Word document or pdf on the local server in a suitably named and backed-up directory.

**Backup Site Data procedure:**

1. **CSV, DAT, A01 or any raw data download files:**

Raw files for primary and backup (where it exists) data must be supplied in support of quarterly updates. Wherever practical, these should be raw logger download files (in a human readable format; i.e., not binary). Telemetry data files may be provided only where logger download files cannot reasonably be obtained.

All files are to be stored on local servers under primary and backup sequential numbers:

e.g., C:\Rawdata\74338 Sutton Stm at SH87\74338\_157.csv

At each quarterly update, the files downloaded for that quarter are to be sent to the NHD (prompt on Kathy’s form). Note that the period covered by the download may not perfectly match the quarter; all new data not previously supplied should be provided.

It is reiterated that these are raw records, and no translation or modifications have taken place.

To be clear, where telemetry files are used as the raw files for the quarterly update, they must be supplied with the quarterly update regardless of whether a version already exists on the RDA. The reason for this is to use an entirely consistent approach for identifying the raw data on which quarterly updates are based.

1. **Primary and Backup numbering:**

The backup numbering convention is as follows:

If the site has 4, 5 or 6 digits then the backup site is the original digits with an **001 suffix** .

If the site currently has 7 digits then the backup site is the original digits with an **01 suffix**.

eg. **74338** for the primary, **74338001** for the backup.

1. **Telemetered Sites – primary and backup data:**

Flosys retrieved data are currently stored on office servers or the Neon web server and are passed to the Real-time Data Archive (RDA) hourly. This happens automatically and is generally in a raw data form.

Telemetered backup data are handled the same way but need to have the appropriate backup numbers for the sites (see above). The telemetry data on the RDA is never to be considered the raw data supplied to the archive.

1. Non-telemetered or missing telemetered primary or backup data should be processed into Tideda and added to the Flosys data file - **C:\XXX\Tideda.mtd**

This can then be used for **over-plots with primary record for update purposes** and verification.

1. **Registers for the above:**

A separate page/sheet is needed to record backup data along with primary (separate numbers). This only needs to show the basic download to server traceability, as it isn’t being archived. Any backup record used in place of the primary will be entered fully in the primary page as before, referenced there and commented.

**Checking Rainfall Data**

Again, the first check is that the start date, time and values for the batch match with the finish values for the previous batch. This is available from the register, and quickly checked.

The initial processing should provide the recorded total rainfall for the period. Enter this in the rainfall register (Form RF) and check that it matches with the total from any check gauge.

The next step is to plot the batch at a suitable scale, preferably over-plotted with a close flow and or rainfall station, and examine it to detect such problems as

1. unusually high intensities in isolation, suggesting interference by outsiders adding water
2. nil or low rainfall compared with other stations, suggesting a malfunction in the gauge or electrical connections
3. the rainfall pattern not following the runoff pattern for the catchment or adjacent catchments
4. gaps in the data
5. errors induced by the field staff, such as extra tips when testing during a visit
6. slow response and small peaks, suggesting debris is caught in the gauge blocking the funnel, (or hail, snow, etc.)
7. and others

Note that such problems, if not detected on the field visit, must be investigated at the earliest opportunity. In cases where the cause has not been positively identified, full processing of the data should be delayed until it can be investigated on-site.

Again, the first plot becomes a permanent record showing the original data as collected.  If any corrections are applied, ***add a second plot to show the final data***.

When the plots have been fully checked, attach the printout containing the parameter entries and add the following by hand:

1. The date the data was plotted;
2. Signature of the processor;
3. Notes of all corrections to the data, plus any subsequent actions carried out which change the data in the file from that which is plotted (e.g. removal of tips)
4. The log pages attached from site visits.

Both plots shall be filed in the station’s file or electronically as described above

Finally, make sure all relevant parts of the raingauge register have been completed.

**Note:** The purpose of an automatic raingauge is to measure rainfall intensity.  Where a considerable discrepancy occurs between the recorder and any totals from a check gauge, file only those data from the recorder that are reliable.

**Editing of Water-level and Rainfall Series Data**

Tideda and other software packages provide many ways of editing or modifying data. It is often impossible to know absolutely the source and effect of any data errors after the event and, although people may have theories, these may not be the only explanation. Therefore it is essential that any modification of data is done conservatively, within strict guidelines, by or under the guidance of experienced personnel and with full Comments being filed on the archive.



**Conditions for Editing Data**

Editing or altering data files carries with it the responsibilities of ensuring traceability and data integrity. The following guidelines shall be observed in all cases:

(a) No alterations shall be made unless the justification for the assumptions made is scientifically defensible, and recorded, as below.

(b) Such alterations must have the explanation recorded either on, or attached to, the plot of the original data in the site file or as a Comment on the database.

(c) **As a general guideline, gaps due to missing record will not be filled with synthetic data or interpolated.**  Any approximate data available shall be made available to users by inclusion or reference to it in the Comment for that gap. Exceptions to the non-use of synthetic data and interpolation are in (d) to (e) below.

(d) A gap in a water-level or index-velocity record may be filled with a straight line or curve as applicable, if all of the following conditions are fulfilled:

1. The river is in a natural recession with the water-level lower (or the same) at the end of the period, and
2. It has been ascertained that no significant rain fell in the catchment during the time of concentration which would relate to the gap period, and
3. The catchment is known to be free of abstractions and discharges which modify the natural flow regime (e.g., wastewater, power station , irrigation), and
4. The resulting plot of the data shows consistency with the data on either side
5. In some situations, (e.g. power stations) an adjacent station may measure the same data or almost the same data. In the former case, the record can be filled in as if it were a backup recorder. In the latter, the data may be filled in if the uncertainty is less than that in the standard or if the correlation between stations for that parameter and that range can be shown to be 0.99 or greater. A comment containing the details of the relationship must be filed. There are Tideda PSIM programs to help with this.
6. The station is not a lake which normally has seiche or wind tilt, or a sea-level site (These are often studied, and synthetic record will not be able to re-create the phenomena.)
7. A Comment is filed (in standard format) that explains the reason for the missing record and comprehensively covers and justifies the above points.

(e) Where the conditions do not meet these criteria, but trained personnel were on-site for the whole period (e.g. flushing) and recorded manual observations, the gap may be filled with these values and interpolated accordingly.

(f) Filling a gap in the original data record with synthetic data derived by correlation is ***not permissible***. It can be used, however, for supplying data requests where the user is informed of the uncertainty involved.  Such data must be carefully controlled to ensure they are not erroneously filed on either the local or central archives.

(g) A gap in a rainfall record may be “interpolated” only if it can be established that no rain fell during the period, by means of correlation with other gauges within or without the catchment for which there is an established correlation and with a correlation coefficient of 0.99 or higher.

In all such cases, the Comment that is filed must comprehensively justify and explain the interpolation so that users may have confidence in these data.

**Telemetry Data**

As telemetry data arrive on the computer system and are available for use immediately without checking (albeit as “unverified” data), they are subject to the same potential errors as other data. Users who are provided with data at this stage need to be made aware of its status.

Despite the speedy arrival, the data’s verification is still subject to later on-site checking of the sensor’s operation by hydrological staff. Therefore telemetry data are to be checked, processed and verified from site visits in the same manner as other data retrieved via the regular (e.g., quarterly) site visit regime.

For stations with telemetry, a regular (hourly or daily) retrieval and checking schedule is required for all stations, with a minimum daily back-up procedure to the RDA to eliminate loss of recorded data.

The following steps are carried out to achieve this:

1. Initially, set up the hourly or daily data schedule to collect data into a “telemetry” directory
2. Daily, check that all stations have replied and all data have been received. The plots should be checked for large gaps or irregular steps. A record should be kept of all stations not responding or showing problems (e.g. a log book or plots with notes attached).
3. Resolve all problems by checking through the telemetry log file for likely problems, then trying to retrieve the data manually and, finally, scheduling an urgent trip to the affected station.
4. An automatic back-up procedure shall be in place to ensure against the loss of telemetry data.
5. The back-up shall be done at least daily and is now done via Flosys and TDClient to servers in Christchurch and Wellington. When the field visits have been made, those data can be retrieved from the telemetry file to undergo the normal data processing regime of plotting, editing if required, transfer, etc.

### Manually entered series data

Manual input of series data can be used to:

* Make corrections to existing data;
* Input data for which there is no alternative method (e.g., manual staff-gauge readings, storage raingauge data, flow gaugings, temperature and ater quality values);
* Input data read manually from chart recorders, (very rarely these days);
* Input synthetic data to cover gaps in records. Note that there are strict conditions for filing synthetic data; see above.

### Comment data

Associated with the Tideda database is the Comment database. This provides the facility for text comments to be filed for a station against a date/time stamp. Comments are an essential part of any database. They are sometimes referred to as “meta-data”. They provide information about aspects of the data that cannot be easily quantified, such what conditions are like at the measurement site.

Tideda provides several ways to build a database of comments. Comments can be entered into an unstructured file of text. However, if the comments are entered into an ODBC (Open DataBase Connectivity) database they can be accessed by other special database programs such as Microsoft Access. Therefore this method has been followed.

For instructions on adding comments to the Comment database, see Appendix.

The comments filed can be read directly from the database, but Tideda also has the ability to display relevant comments and indicate, for example on plots, that comments are filed at specific date and times.

Comments are very useful to explain unusual features or events in the record that users of the data should be aware of. In addition, routine comments are required for key information. These include:

* All recording stations require an initial comment detailing station information and parameters measured.
* “Recorder comments” are required for both water-level and rainfall stations, giving instrument and sensor types and their resolutions and accuracy
* Flow stations require a “rating comment”, to provide information on the ranges covered by flow gaugings.
* Comments relating to gaps, missing record, synthetic record, or any specific time-related event, are also required in a standard format. They should be filed one time interval (e.g. 15 minutes) into the missing (gap) or synthetic record period or at the time the specific event takes place.

Comments covering the accuracy of data and gaps in records should be informative, coherent and identify the period(s) for which the data are suspect or missing. Examples of these types of comment formats are given in Appendix A.

## Data archiving

Up to this point, all the new data have been imported, checked, edited if necessary and stored in working directories or in a update file on the local server (e.g. update.mtd). Once the checking and verifying is complete, the files are passed on to or picked up by the database manager for updating to the Hydrometric Database. The steps in this process are:

* The data are assembled in a designated transfer directory on the server or shared directory designated for this purpose. No other batches shall be contained in that directory.
* Register pages are filled in for all the data batches with all columns filled in and signed off.
* The data supervisor checks the register pages and batch listing for completeness, and that the filed end of batch data (time and stage) follows on to new updated batch. Data are listed to check this; then one plots on-screen the data for these and the previous months to check for continuity of time series (absence of gaps), continuity of values and apparent interpolations between the batches.
* All batches without apparent errors are copied to the dispatch file (containing the update date) for updating to the Hydrometric Database, and any that have apparent errors or anomalies will be returned to the data processor.
* The Database Manager receives the files, updates them, and returns the form and a copy of the updated archive to the data supervisor.
* The data supervisor files the returned batch listing, remedies any anomalies and archives the new replacement archive file to the local archive, then updates the registers to show that this has been done

### Archives

All the data collected are transferred to network drives as soon as possible, as these drives are routinely backed up according to a regime maintained by the local IT support, usually nightly. Staff using data should take an active interest in this backup operation to ensure it is done adequately.

Protection needs to be put on all local archives so they are read-only and cannot have data copied to them by mistake. The data supervisor is in charge of this.

### Transfer of data batches

As detailed above, it is necessary to keep a record, in chronological order, of all data movements from the field office to the Hydrological Database. A paper or electronic listing of transfer directory contents must accompany all transfers, and a duplicate copy of the files is held in the office and archived when transfer is completed.

A transfer register shall be maintained, comprising the paper batch listing which has been sent for archiving and signed off and returned by the NHD manager. Each such transfer listing shall be given a unique incremental number. This is to ensure that there is no duplication or confusion with the data transfer process.

All relevant forms to be filed in for an update are located on the Field Teams’ intranet page.

A quarterly schedule may be adequate for all new data to be updated to the NHD. It is important to have the discipline of carrying this out at regular intervals

### Deleting data from the archives

The deletion of any data from archives will only be undertaken from written requests (appropriate form) to ensure full traceability of data movements are maintained by the NHD manager.

### Marking gaps in records

Tideda inserts gap markers to prevent interpolations being made between values where this may not be valid.

Each batch of series data provides a record from a start time to a finish time. Being a series, interpolation between successive values within a batch is assumed to be valid. However it is not always valid to interpolate between the value at the end of one batch and the value starting the next batch. Tideda safeguards this potential error by automatically entering a gap maker when a new batch is first read. This directory entry automatically prevents interpolation between its start time and any earlier data that may already be filed. Thus, a missing record between two batches is automatically marked, and interpolation between the values is prevented.

Also, when data values are deleted, a gap is automatically introduced. So when a batch of data includes values that are in error, using the Tideda process of deleting those values will introduce the required gap marker.

A gap marker that has been introduced automatically and is not required can be removed using the Tideda process Manage/Gap.

### Storage of original records

All original records, whether in written form (gauging cards, survey books, etc.), or electronically recorded (telemetry and solid-state) are to be kept in a safe environment against damage and loss (including theft) at all times.

Fire-proof storage with a minimum two-hour fire rating is required. All data not in use or not being processed are to be held here at all times. The door to these areas must be kept closed, and locked as appropriate.

Where the original data is in the form of a data file on electronic media, copies of theunaltered but identifiable files must be stored on the local server which are backed up, automatically and daily to another location (e.g., Christchurch or Wellington main servers).

# Data checking techniques:

## General

This section covers the checking of data in a longer time context than is generally done while processing and updating batches of data. While the techniques can be applied in an annual data review process (see next section) they are also applied during routine verification, and occasionally in situations where data needs re-checking before delivery to users.

Good office practice, as outlined in Section 3, will minimise the amount of additional quality control checking required and reduce any further work required.

## Flow checking

Water-level data, gaugings, ratings and their derivative, flow, require checks and comparisons in a longer context than the relatively short batches that they are collected in. The main reason for this is that rating changes are often not immediately evident or confirmed within a few weeks or months.

The following guidelines outline the minimum checks required to demonstrate the quality of the archived hydrometric data. The resulting documents and records should be kept in the site files ***indefinitely*** as a quality “audit trail” record.

### Water-level data

Water-level plots either on-screen or on paper at appropriate scales, give a “picture” of the data that can be interpreted to identify a number of potential errors:

* “Non-gapped” missing records; look for straight lines in the data that may be wrongly interpolated across a gap.
* Errors associated with inconsistent recorder/staff gauge readings.
* Steps or spikes and inconsistent recession shapes in the data.
* Differences between filed gauge heights for gaugings and filed series data. These may need further checking and correcting;
* Poorly defined water-level hydrographs created by either synthetic record or inaccurate digitising. If these have resulted from synthetic data, they should be removed and a gap inserted with the relevant comment filed;
* If synthetic data have been filed, it should be clearly identified and the method of construction and confidence limits detailed in a filed comment. Synthetic data should only be constructed for short durations and when the river is in a known recession, as defined in the previous section;
* Time corrections that have been applied are not instances where the logger has stopped recording for a brief period;
* Silting problems in wells or around transducers have been correctly identified and corrected or commented;
* Water-level data should be inspected for bed level changes indicating rating changes during periods of infrequent gaugings;
* Comments for missing records, and on data quality;
* Maximum and minimum recorded levels are within the range of the instrument;
* Data is not over-compressed. Automatic “zero” compression should only be used.