UM | Centre for Earth Observation Science

CEOS Data Management - Seabird TSG Data Processing Cookbook

CENTRE FOR EARTH OBSERVATION SCIENCE

Document Control

0.1 Version History

Version	Author(s)	Type	Date Modified	Comments
1.0	Hunt, J., Campbell, Y.	Working Copy		

0.2 Document Location

A hard copy of the document can be found in the Lab 489 document cupboard.

A digital copy of this document can be found here:

0.3 License

Hunt, J., Campbell, Y.(2022). Cookbook, Version 1.0. Centre for Earth Observation Science. University of Manitoba.

Contents

1 Introduction

1.1 Description

This document outlines the steps for data conversion of Seabird Thermosalinograph data using SBE Data Processing software, and subsequent curation to transform them into standardized CSV files. The initial data processing steps are created specifically for processing data from the SBE 21 instrument collecting flow-through data aboard the RV William Kennedy while it is underway. This is supplementary to the SBE Data Processing manual.

You will need to download SBE Data Processing software to begin processing.

2 Initial Data

2.1 .hex files

Files containing raw data of each cast. Files names should include cruise ID (WKxx_dd-mm-yy) and the download date of the leg (WKxx_dd-mm-yy). For example, WK01_01-08-22 indicates the first leg of cruise #1 was completed on August 1, 2022.

2.2 .xmlcon files

.xmlcon are configuration files needed to convert to the .hex data to a usable format (.cnv).

The .xmlcon file is created by Seabird when the instruments are sent for calibration. It stores the calibration coefficients of the sensors. You do not need to edit it unless there is a change to an auxiliary sensor. Keep it with the raw data.

2.3 creating .cnv files

Open SBE Data processing. Select "Run", "1. Data Conversion", "File setup". Select appropriate instrument configuration file (.xmlcon). Select the input directory where your raw .hex and .xmlcon files are stored. Select all casts as the input files. Select your desired output directory... (see "suggested directory setup?" make a table?–>file source and location...)

Open the "Data Setup" tab. Leave the default settings except for "source for start time in output .cnv header". For this select "System UTC" (Figure 1).

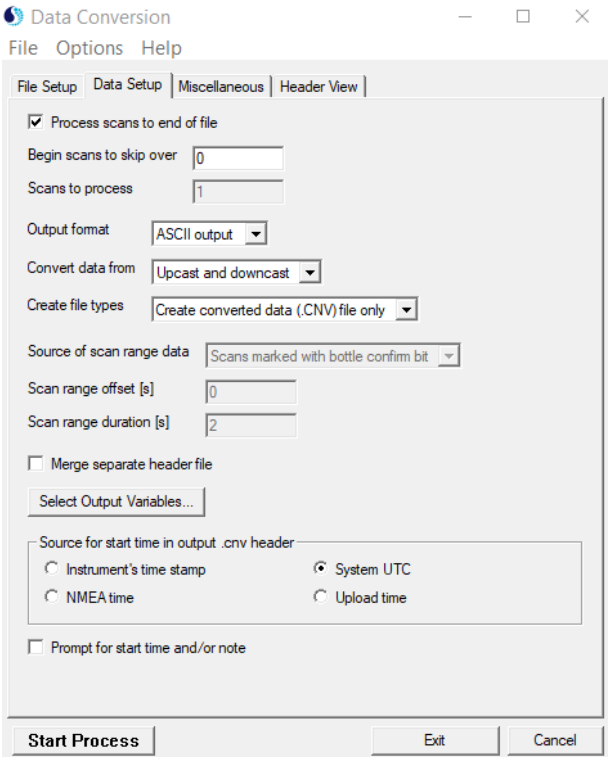

Figure 1. Thermosalinograph data conversion settings

Select output variables (table 1.1).

Table 2.1: Variables to select when creating .cnv files from TSG .hex files

Variable
Scan count
Latitude
Longitude
Temperature (ITS-90, deg C)
Salinity (psu)
Conductivity (mS/cm)

Then click the "Start Process" button. The .cnv files will be placed into the output file location.

3 Intermediate Data

The intermediate process involves converting the CNV files produced by the SBE Data processing to CSV files. This was done by using a python script.

For more details on the script, see the TSG codebook on the Datahub here, and the scripts can be found here.

3.1 Overview of script

1. Create a loop to read all the cnv files. The following steps are performed for each file.
2. Begin reading the CNV file line by line
3. Extract the metadata into a dictionary. Metadata are preceded by * in the CNV file
4. Write the metadata with XML form into an XML file. XML metadata are preceded by # in the CNV file
5. Check for the start date/time tag: *start_time* and extract the date
6. Check for the time interval (tag containing the word *interval*)
7. Extract the variable names
8. Get the actual data
9. Create a data frame from the variable names (column headers) and the data
10. Calculate a range of dates starting with the start date and interval taken from the cnv file
11. Add as a column in the data frame
12. Save as a csv file

3.2 File names and variables

Intermediate File Name(s)

- leg1.csv
- leg2.csv
- leg2.1.csv

- leg3.csv
- leg4.csv
- leg5.csv

Intermediate Dataset Variables

- scan
- latitude
- longitude
- t090C
- sal00
- date_time

4 Final Data

4.1 Vocabulary standardization

Column headers are standardized in accordance with the NERC Vocabulary, and units added after an underscore(_). See the converted headers under Dataset Variables below. Once the variables were standardized, a template was created for the final output files, and a python script was used to convert all the intermediate files to the format of this template.

4.2 File names and variables

Final File Name(s)

- leg1_processed.csv
- leg2_processed.csv
- leg2.1_processed.csv
- leg3_processed.csv
- leg4_processed.csv
- leg5_processed.csv

Final File Location

Processed files can be found on CanWIN's Datahub here.

Final Dataset Variables

Table 4.1: Variables in the final data files

Header	Description	Data Type	Units
scan	Scan count	numeric	None
latitude_north	Latitude	numeric	Decimal degrees
longitude_east	Longitude	numeric	Decimal degrees
WC_temp90_degC	Temperature	numeric	Degrees Celsius
P_sal_psu	Practical Salinity	numeric	PSU
Cond_S_m	Conductivity	numeric	S/m
Date_time	Date and Time	Timestamp	UTC

A Reference Tables

A.1 Data Levels

Level 0 – Raw data: unprocessed data and data products that have not undergone quality control. Depending on the data type and data transmission system, raw data may be available within seconds or minutes after real-time. Examples include real-time precipitation, streamflow, and water quality measurements

Level 0.1 – First pass QC: A first quality control pass has been performed to remove out of range and obviously erroneous values. These values are deleted from the record. E.g: Online Environment Canada stream-flow data, laboratory data

Level 1 – Quality Controlled Data: Data that have passed quality assurance procedures such as Level 0.1 and have been further quality controlled by data provider before being submitted to CanWIN (e.g. Idronaut data with only downwelling (upwelling data removed) data included.

Level 1.5 – Advanced Quality Controlled Data: Data have undergone complete data provenance (i.e. standardized) in CanWIN. Metadata includes links to protocols and methods, sample collection details, incorporates CanWIN's or another standardized vocabulary, and has analytical units standardized. Note: Process still under development in CanWIN (as of May 13, 2020).

Level 2 – Derived Products: Derived products require scientific and technical interpretation and can include multiple data types. E.g.: watershed average stream runoff derived from stream-flow gauges using an interpolation procedure.

Level 3 – Interpreted Products: These products require researcher (PI) driven analysis and interpretation and/or model-based interpretation using other data and/or strong prior assumptions. E.g.: watershed average stream runoff and flow using streamflow gauges and radarsat imagery

Level 4 – Knowledge Products: These products require researcher (PI) driven scientific interpretation and multidisciplinary data integration and include model-based interpretation using other data and/or strong prior assumptions. E.g.: watershed average nutrient runoff concentrations derived from the combination of stream-flow gauges and nutrient values.

Content retrieved from https://lwbin.cc.umanitoba.ca on July 06, 2020.

A.2 Result Value Qualifiers

ADL	Above Detection Limit
BDL	Below Detection Limit
FD	Field Duplicate
LD	Lab Duplicate
$	Incorrect sample container
EFAI	Equipment failure, sample lost
FEF	Field equipment failed
FEQ	Field Equipment Questionable
FFB	Failed. Field blank not acceptable
FFD	Failed. Field Duplicate
FFS	Failed. Field spike not acceptable
H	Holding time exceeded
ISP	Improper sample preservation
ITNA	Incubation time not attained
ITNM	Incubation temperature not maintained
JCW	Sample container damaged, sample lost
NaN	Value is missing and reason is not known
NC	Not collected
ND	Not detected
NR	Sample taken/measured on site but information in this field not recorded
NS	Sample collected but not submitted
OC	Master Coordinate List Used
P	Analysis requested and result pending
prob_good	probably good value. Data value that is probably consistent with real phenomena but this is unconfirmed or data value forming part of a malfunction that is considered too small to affect the overall quality of the data object of which it is a part
prob_bad	probably bad value. Data value recognised as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena
Interpolated	This value has been derived by interpolation from other values in the data object
Q	Below limit of quantification (LOQ). The value was below the LOQ of the analytical method. The value in the result field is the limit of quantification (limit of detection) for the method

B Glossary of Options and Packages

B.1 R Packages

Visit https://cran.r-project.org/web/packages/available_packages_by_name.html to learn more about R packages

- **Package 1** - Description
- **Package 2** - Description

B.2 Python

B.2.1 Python Script-Specific Options

- **Option 1** - Description
- **Option 2** - Description

B.2.2 Python Packages

Visit https://docs.python.org/3/library/ to learn more about python packages

- **Package 1** - Description
- **Package 2** - Description

Example: Section 2.1 from Victory's semi-hemi codebook