Field Data Management and Compensation

Dongseok Yang[1], Laura Bowling[2], Keith A. Cherkauer[†]

1) Graduate student, Dept. of Agricultural and Biological Engineering, Purdue University, West Lafayette, 47907
2) Professor, Dept. of Agronomy, Purdue University, West Lafayette, 47907
†) Professor, Dept. of Agricultural and Biological Engineering, Purdue University, West Lafayette, 47907 (Corresponding author)

Date: 2023-05-02 (Final)

Project Summary

Abstract (Synopsis)

This study aimed to build a management program from the collected field data. Management in this program is consisted with the four steps. Step 1, Read the raw data and distinguish between waterlevel and barometric data. Step 2, compensate the raw-waterlevel data, which is based on the pressure to the metric waterlevel. Step 3, conduct outlier removal and integrate all the data in one single file per each station. Step 4, conduct time series check to find empty period of the data. While implementing these processes, the program will provide multiple metric analysis and graphical analysis of the data. From Step 2, user will receive timeseries line graph of the compensated waterlevel. And from Step 3, timeseries line graph and boxplot graph will be provided for each monitoring period (one file for station of each monitoring). And simple metric analysis for cleaned waterlevel data will be also provided.

Description

What is expected from this project?

This project is in Python language and requires admin level authority for implementing the script. After implementing the script, in the script directory, multiple .png files will be saved and multiple .csv files will be updated.

Why this project is important?

The goal of this project is to make the process of compensation and data-analysis including statistics and cleaning more simply and easier compared to the past using manufacturer's software. Therefore, this project is including a few automatic data handling processes. The data being handled in this study is monitored data from the wetland which is study site. And this data is important for analyzing the water balance and its behavior within watershed.

In the previous study, the data collection and data quality checking had implemented separately using different software or method. Specifically, researchers compensated the data collected from the field using manufacturer's software and inspected to see there is outlier or not. After inspection, they put secured all the data in one file copying and pasting to the data management software such as Microsoft Excel manually. For the graphical analysis, they also used other graphical software such as Microsoft Excel or Sigmaplot. However, the program built from this study is importing all the raw-data in certain directory and import updated files for compensating. After compensating, it will automatically conduct outlier removal and integration to the one file. Also, during this process it will show or save the graphs for user.

Keywords

Atmosphere monitoring with geodetic techniques (6952)
Data assimilation, integration and fusion
Data management, preservation, rescue
Decision analysis (4324, 6309)
Wetlands (1890)

Overview

Source Data

Input source data are in the three types.

1) Water level data (Field scale)
2) Barometric data (Field scale)

For 1) Water level data (Field scale) and 2) Barometric data (Field scale), they are observed using Solinst water level logger (Junior 5) and Barometric logger recording in 15-min resolution. They are collected manually and saved to the local directory in .csv and .xle format. In the case of the .csv file, it is for the process and .xle file is for backup and for the post process of the paired program provided by Solinst.

The length of the data will depends on the period of the observation, duration between two field monitoring plan, and could be vary from few days to few weeks. The data for the processes in the project will be using only .csv files.

For this project, Pandas library based on the Python language will be used to handle data in dataframe/series form to manage data in order of timeseries. The Pandas library is powerful in handling timeseries data and concatenating the data in column or row.

Raw data (Level data)

Date	Time	ms	LEVEL	TEMPERATURE
3/31/2023	3:15:07 PM	0	10.137	8.8
3/31/2023	3:30:07 PM	0	10.128	7.8
3/31/2023	3:45:07 PM	0	10.131	7.8
3/31/2023	4:00:07 PM	0	10.131	7.8
3/31/2023	4:15:07 PM	0	10.122	7.8
3/31/2023	4:30:07 PM	0	10.119	7.7
3/31/2023	4:45:07 PM	0	10.119	7.7
3/31/2023	5:00:07 PM	0	10.116	7.7
3/31/2023	5:15:07 PM	0	10.107	7.7
3/31/2023	5:30:07 PM	0	10.101	7.7
3/31/2023	5:45:07 PM	0	10.098	7.7
3/31/2023	6:00:07 PM	0	10.098	7.7
3/31/2023	6:15:07 PM	0	10.098	7.7
3/31/2023	6:30:07 PM	0	10.098	7.7
3/31/2023	6:45:07 PM	0	10.098	7.7
3/31/2023	7:00:07 PM	0	10.092	7.7
3/31/2023	7:15:07 PM	0	10.089	7.7
3/31/2023	7:30:07 PM	0	10.086	7.7
3/31/2023	7:45:07 PM	0	10.08	7.7
3/31/2023	8:00:07 PM	0	10.074	7.7
3/31/2023	8:15:07 PM	0	10.074	7.7
3/31/2023	8:30:07 PM	0	10.068	7.7
3/31/2023	8:45:07 PM	0	10.062	7.7
3/31/2023	9:00:07 PM	0	10.083	7.7
3/31/2023	9:15:07 PM	0	10.08	7.7

Barometric data

Date	Time	ms	LEVEL	TEMPERATURE
2/15/2023	4:34:57 PM	0	98.7934	13.61
2/15/2023	4:39:57 PM	0	98.7825	13.177
2/15/2023	4:44:57 PM	0	98.7801	12.833
2/15/2023	4:49:57 PM	0	98.7746	12.368
2/15/2023	4:54:57 PM	0	98.7747	12.026
2/15/2023	4:59:57 PM	0	98.7841	11.729
2/15/2023	5:04:57 PM	0	98.7929	11.45
2/15/2023	5:09:57 PM	0	98.8011	11.146
2/15/2023	5:14:57 PM	0	98.8093	10.838
2/15/2023	5:19:57 PM	0	98.8072	10.622
2/15/2023	5:24:57 PM	0	98.8019	10.456
2/15/2023	5:29:57 PM	0	98.8089	10.231
2/15/2023	5:34:57 PM	0	98.8122	9.936
2/15/2023	5:39:57 PM	0	98.8207	9.604
2/15/2023	5:44:57 PM	0	98.8184	9.318
2/15/2023	5:49:57 PM	0	98.8291	9.025
2/15/2023	5:54:57 PM	0	98.8327	8.73
2/15/2023	5:59:57 PM	0	98.8445	8.444
2/15/2023	6:04:57 PM	0	98.8408	8.159
2/15/2023	6:09:57 PM	0	98.8573	7.874
2/15/2023	6:14:57 PM	0	98.8617	7.598
2/15/2023	6:19:57 PM	0	98.8675	7.309
2/15/2023	6:24:57 PM	0	98.8733	7.013
2/15/2023	6:29:57 PM	0	98.8839	6.735

Figure 1. Example of the data
The date and time will be handled with Pandas datetime function in this project.

Methodology

Input source data are in the three types.

1) Water level data (Field scale)
2) Barometric data (Field scale)

The program developed from this study is consisted with two scripts for better modification and structural management. First is 'compensator_func' and this script is including all the process in form of function that can be imported from any script. And second is 'main' and this script is for implementation of the functions in 'compensator_func'. From the 'main' script, user can check or edit the process of this project more easily. By the date 2023-05-01, this project is handling relatively small amount of data but in the future, if more data and process are required for the study, all the functions will be implemented from different script for convenience.

For the field scale data, they are consisted with the date and data. Before compensation water level data is recorded in terms of the pressure. Therefore, this project designed automated compensation tool to convert pressure data to the water level using barometric data collected from the field or observation point. Additionally, this project let users to input the height of the logger that could be changed for every monitoring, but this could be set to single number unless there is a change to the sling connected to the logger.

After compensation, this project will automatically download the graph of the water level for each station. Currently the stations are ILA, ILB, CEN, OUT, F63 and these stations could be changed according to the users need.

As a last process of the project, the project will integrate the whole files into single file to obtain continuous data. For this process, 'history.log' file will be working as a list of the files considered in the integrated file. Therefore, any files in the history.log will be excluded from the integrating process.

The project is consisted with the nine parts for total and they are following below:

```
1. graphy : draw graph for the data from pandas dataframe
2. history : update history.log file for future management of the data files
3. readdata : read data file, make it as dataframe format.
4. intwrite : read all the compensated data, write it to the integrated file
5. comp : compensation function for the loggers using barometric
        from this function, the barometric data(ATM) will be added to the dataframe
        of the level logger
6. oldetector : outlier detector for various situation
7. set_read : this file will read environmental variables for the compensation
```

Graphical data analysis (visual data quality checking)

For the graphical data analysis, this project is providing outlier removal, compensation, and data integration results. For outlier removal, timeseries plot and boxplot for before and after process will be provided. And for the compensation, compensated waterlevel and barometric will be provided in the graph to compare waterlevel behavior according to the atmospheric pressure.

This will let users to assess compensation. Lastly, for the data integration, the graph of the all the time period will be provided to the user. Users can visually check the absent of the data from this process. And the program will provide empty period of the data if it is longer than 24 hours.

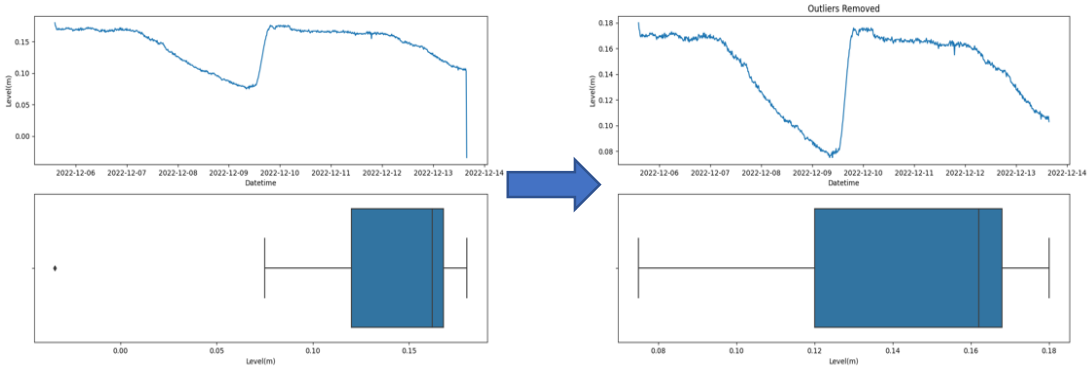


Figure 2. Graphical analysis for the outlier removal
Data below 0.00 which is outlier while logger stayed out of the water for collection got erased from outlier removal process.

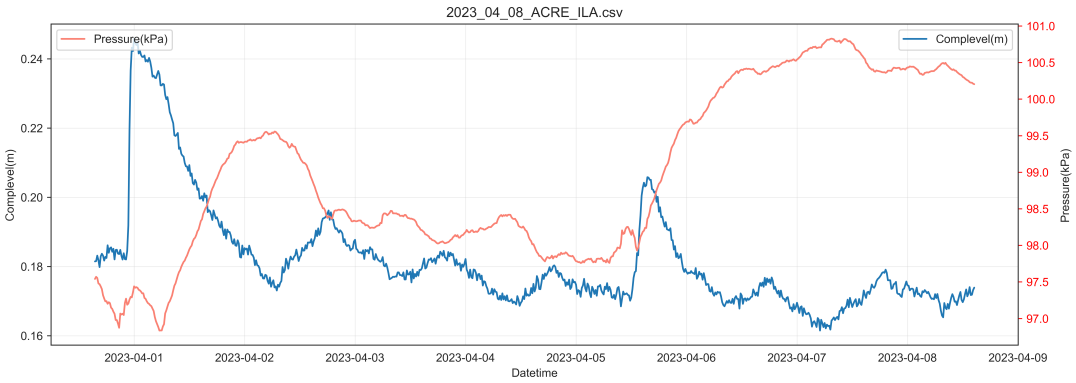


Figure 3. Graphical analysis for the compensation

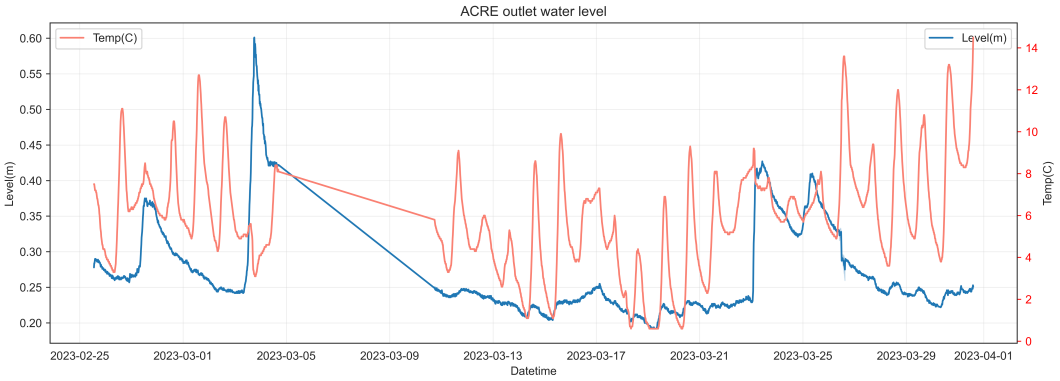


Figure 4. Graphical analysis for the timeseries check with absent of data
The data is absent from 25th Feb 2023 to 24th Mar 2023.

Data Quality Checking

For the waterlevel data quality checking (after compensation), this project is using boxplot graph to identify range of the outliers. From the concept of the boxplot, the project identifies the range of the IQR (Interquartile Range) and calculates the values exceeding whisker. However, if length of the data is too small to conduct the outlier, this project will remove minimum 10 points for removing outliers.

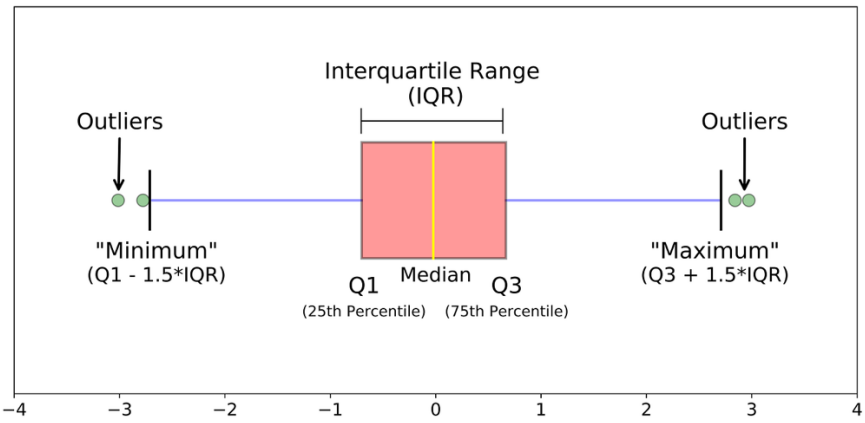


Figure 5. Outlier detection using boxplot (Agarwal, 2019)

Compensation

For the compensation of the waterlevel based on the pressure value, this study used equation from experience. The compensation requires waterlevel pressure and atmospheric pressure (barometric) values of the same day. And difference between these two values will be multiplied with the fraction to get the metric waterlevel value. Equation is following below:

$$Waterlevel_i = (WP_i - AP_i) \times F_C$$

Where Waterlevel is converted metric waterlevel, WP is Water pressure value, AP is atmospheric pressure (Barometric) value and F_C is Fraction of the conversion. In the case of the F_C, it is 10.1972 which means 10.1972 cm per 1 pressure difference.

Summary Statistics and Metrics

This project adopted statistics and metrics to interpret the behavior of the waterlevel after compensation.

Table 1. Metrics in this study

Index	Explanation
TQ-Mean	TQ-Mean is the fraction of time that daily waterlevel exceeds mean streamflow for each year. This value is based on the duration rather than the volume of value.
RB-Index	RB-index is used to analyze flashiness of the hydrological component such as streamflow. Equation is following below: $$RB - Index = \frac{\sum_{i=1}^{n} \lvert Q_i - Q_{i-1} \rvert}{\sum_{i=1}^{n} Q_i}$$ Q_i is streamflow of day i
7Q Values	7Q Values is the minimum value of the 7-days average waterlevel.
3X Median	3X Median value is the days exceeding the values three times of the median of the waterlevel data.
These functions could be changed, added, or deleted depends on the update.	

```
================== Metric Analysis ==================
   TQ-Mean    |   RB-Index    |   7Q Values   |  3X Median
    0.340     |     0.005     |     0.000     |      0
=====================================================
```

Figure 6. Example of Metrics in this project (ILA 2023-01-25 – 2023-02-12)

Conclusions

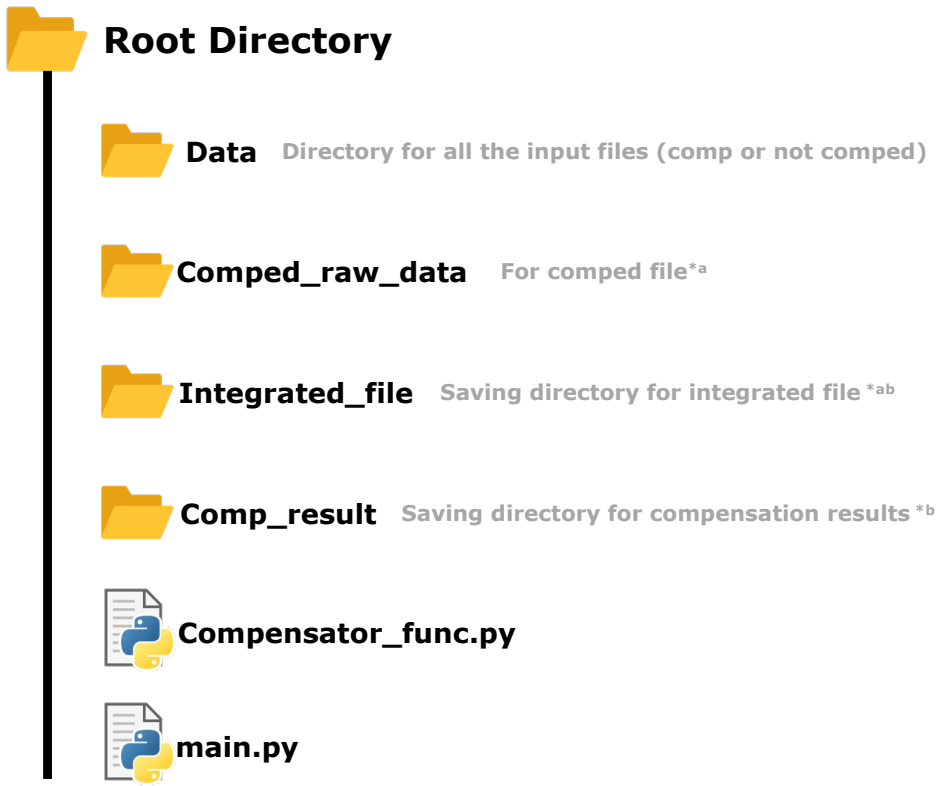
- Using the developed program, this study conducted sample data process using the field data from Jan 2023 to Apr 2023.
- It was found that this program is doing great job for compensating the data.

- Outlier graph and timeseries graphs are helpful to examine the data quality checking. But some of outliers are ignored during the process.
  - From the data quality checking process, it was able to find sudden drop of the value (monitoring time)
  - It was easier to find absent of data by graphical analysis.
- It was found that all stations have data absent for 28 days.
- Some stations had unit problem since not all the data loggers are set to the same unit.

Check list	Stations	Reason	Soultion
Absent of the data	All station	Barometric error	Download Baro data from different source
Missing outliers	Some stations	Low water level	Adapt new method for outlier removal
Metric analysis	Some stations	Format error	Unify all the dtypes through out the functions

- The problems found in this study will be considered for the next update of the program.
- For outlier removal process, it happened when the waterlevel was low so the pressure is not different from outside the observation point. Therefore, other outlier removal process will be considered in the near future.
- The unit conversion criteria will be 500. For example, if the pressure value is over 500, the unit will be cm and below that, it will be m.
- Since this study is designed for the certain site, there is no flexibility for additional observation point or different date. Therefore, the recognition of the station or timeseries method will be updated in the near future. For this process, the input file ‘comp_set’ will get more complicated to embrace information about the observation points.

Appendix A. Directory management



*a : Saved or moved to this directory automatically during process
*b : includes .csv file and results

Appendix B. File name method

File name format:

[Samped date]_[Study area]_[Station_code].csv	: if not compensated (raw data)
[Samped date]_[Study area]_[Station_code]_COMP.csv	: if compensated
[Station code]_integrated.csv	: integrated data file

Sample date : yyyy_mm_dd
Study area : text
Station code : text

Appendix C. setting.comp (input file for program)

This program requires one file for setting. All the lines starting with '#' will be ignored while reading the variables in the file. For example, from the sample file, only line 9,10 will be read in the program. The variables will be separated using comma.

LINE#	CONTENTS
1	# this is comment line
2	#
3	#Height variables(location of the loggers)
4	#first line = name of the stations
5	#second line = offset depth of the loggers for each stations
6	#third line = Latitude of the station (UTM WGS 1984 16N)
7	#Fourth line = Longitude of the station (UTM WGS 1984 16N)
8	#
9	OUT,ILA,ILB,CEN,F63,BARO
10	0.5,0.8,0.45,0.37,0.5,0.6