Expedition Programming Challenge CASE III

Yu Luo

MS. Computer Science

Indiana University

Background

-DO datasets:

dissolved oxygen data collected by a sensor in reservoir- this sensor is deployed for month-long intervals and collects data every 15 minutes at 1m depth.

-FCR datasets:

5 temperature files collected by thermistors deployed at five depths (depth in the file name). These sensors were deployed in June 2013 and removed in March 2014, and collect data every five minutes.

Raw Data
(retrieved from data stream simulator)

DO Dataset sample

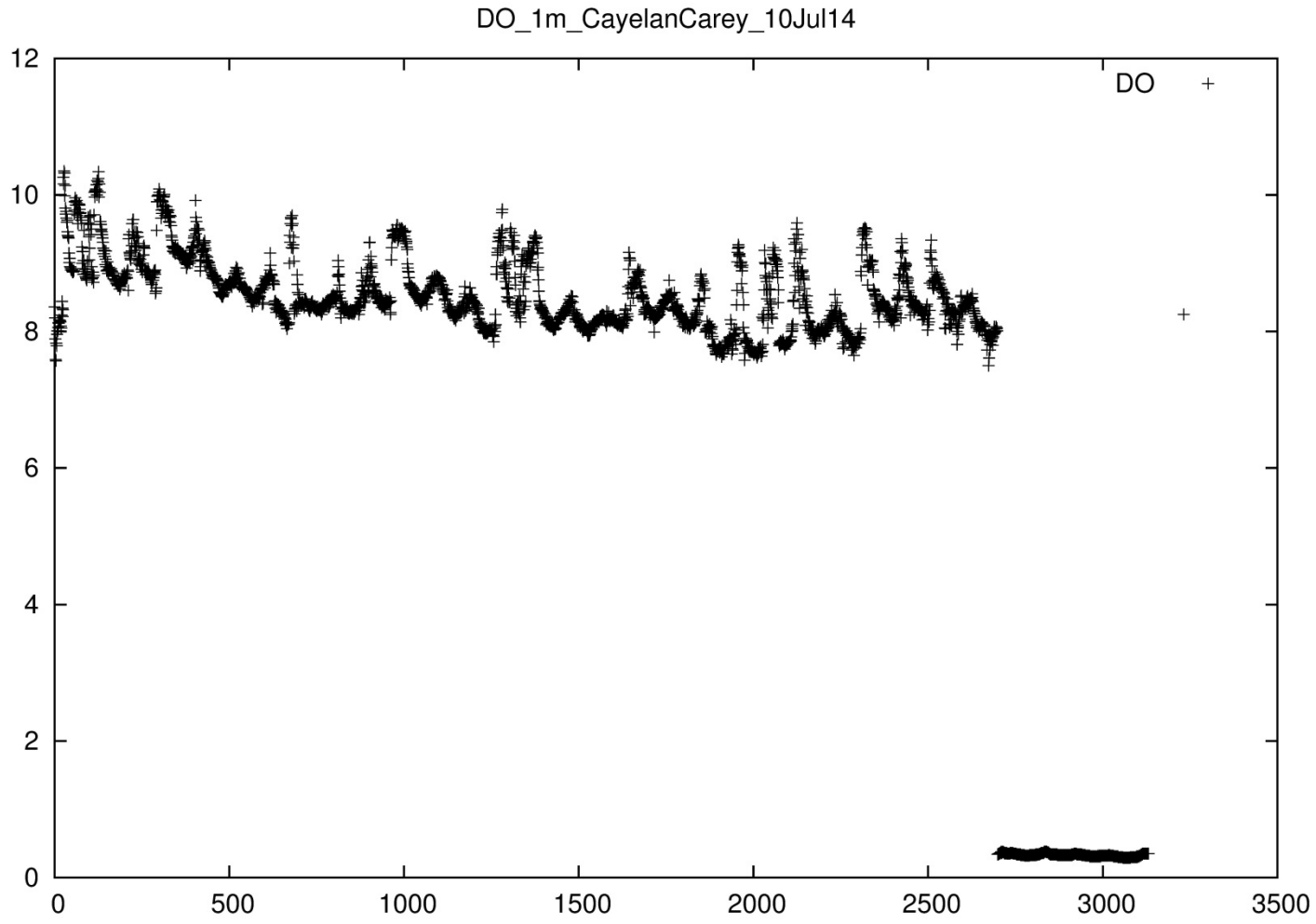
```
"Date","Time","","Temp","","SpCond","","LDO%","","LDO","","IBatt",""
"M/D/YYYY","HH:MM:SS","","◆C","","◆S/cm","","Sat","","mg/l","","%Left",""

5/29/2014,09:00:00,"","20.94","","0","","100.3","","8.36","","84",""
5/29/2014,09:15:00,"","23.09","","0","","102.5","","8.20","","94",""
5/29/2014,09:30:00,"","25.01","","0","","102.4","","7.89","","93",""
5/29/2014,09:45:00,"","26.18","","0","","100.5","","7.59","","91",""
5/29/2014,10:00:00,"","26.35","","0","","100.6","","7.57","","86",""
```

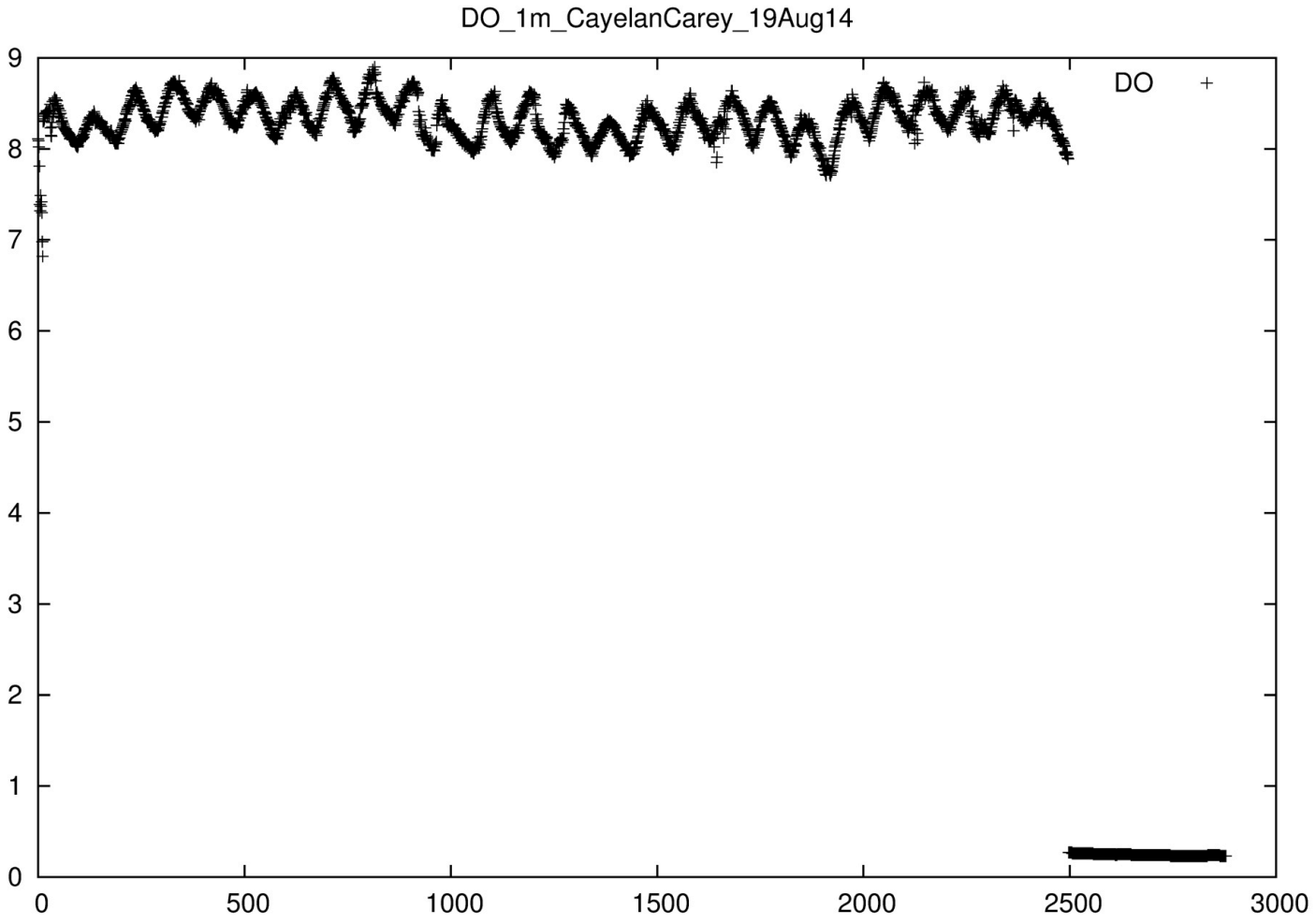
FCR Dataset sample

Date & Time	temp02
18-Jun-2013 10:19:13.000	23.5683599
18-Jun-2013 10:24:13.000	23.4423096
18-Jun-2013 10:29:13.000	23.0947846
18-Jun-2013 10:34:13.000	22.8760167
18-Jun-2013 10:39:13.000	22.5756361
18-Jun-2013 10:44:13.000	22.3987928
18-Jun-2013 10:49:13.000	22.2708988
18-Jun-2013 10:54:13.000	22.1756061

Raw data visualization



More Raw Data Visualization



Discussion

Data Issues:

1. Outliers
2. Low Battery Screening
3. Drift

Solution:

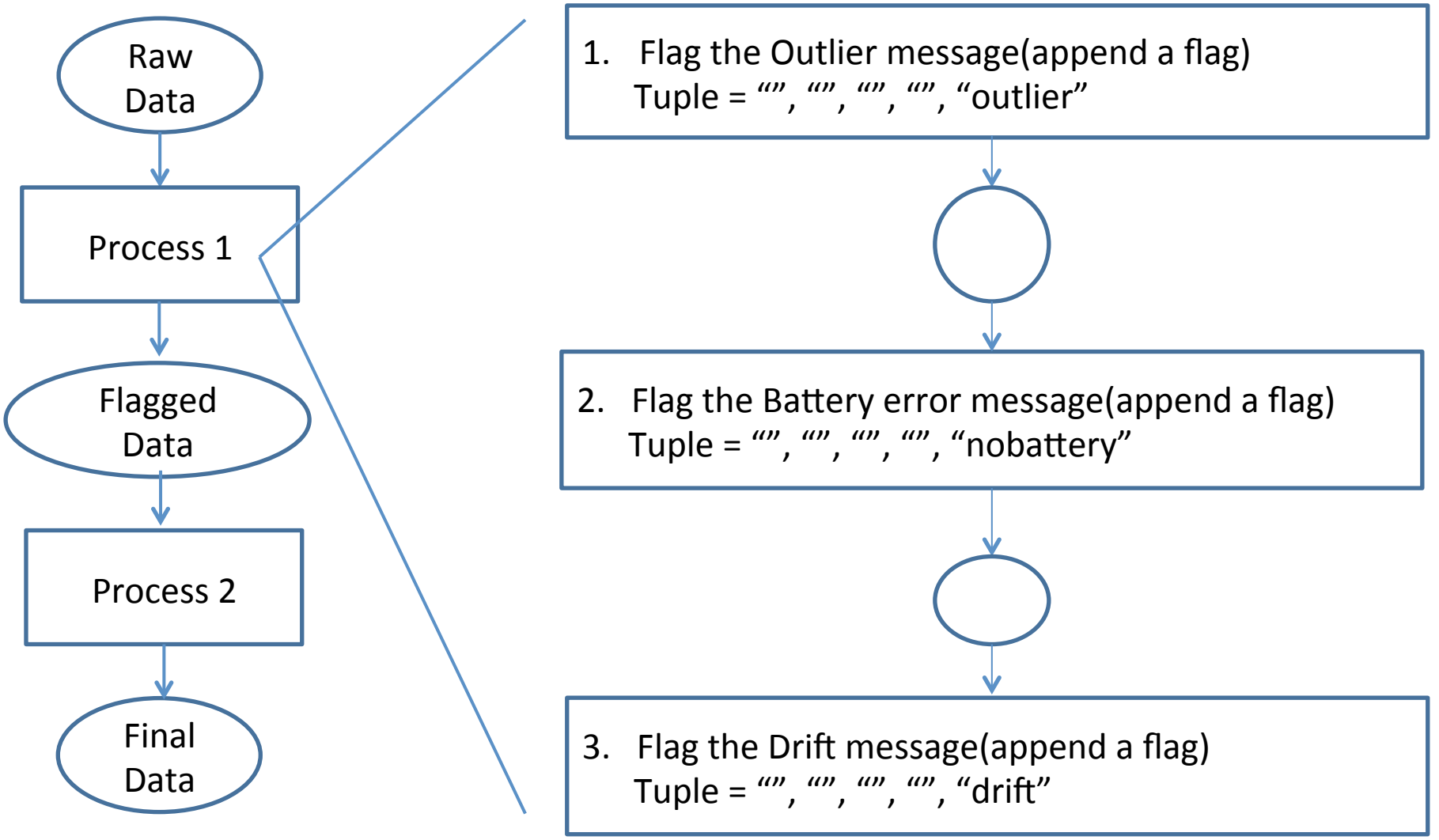
Data Cleaning.

Final Result:

Product two different .txt files

- .txt files that removes the outliers and sensor drift
- .txt files that leaves the raw data but adds flags

Data Cleaning Workflow

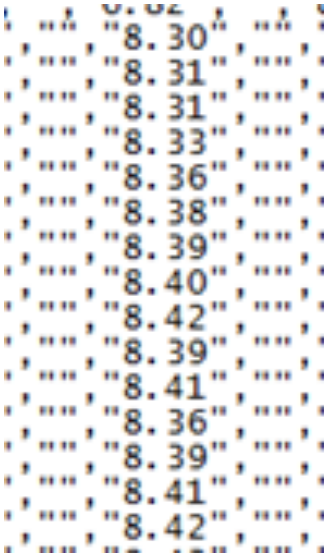


Outliers

What is outliers?

e.g., LDO > 15 mg/L or a negative number

LDO 1m depth



Outlier Filter

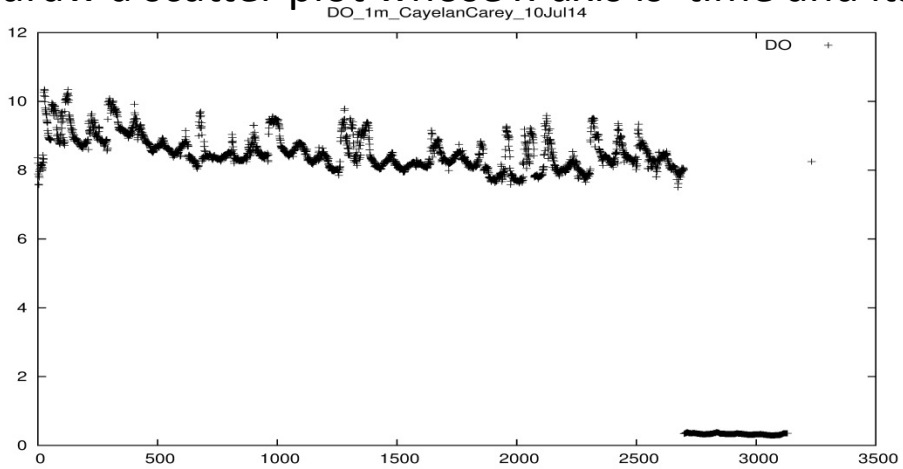
1. Read the raw data file line by line
   (delete the header lines and description lines)

2. For each line, check the attributes
   if “LDO” > 15 or “LDO” < 0,
   append flag “outlier” at the end of line;

3. Save all lines to new .text file

Low Battery Screening

After reading the data file, we draw a scatter plot whose x-axis is time and its y-axis is Temp or LDO.



From the above graphic, we could find that the some points are plotted at low degree.

Battery Dies:

In DO data sets, we could find battery information. When "IBatt" is 0, it means sensor is out of power.

In FCR data sets, when the Temp is null. It means the sensor is out of power.

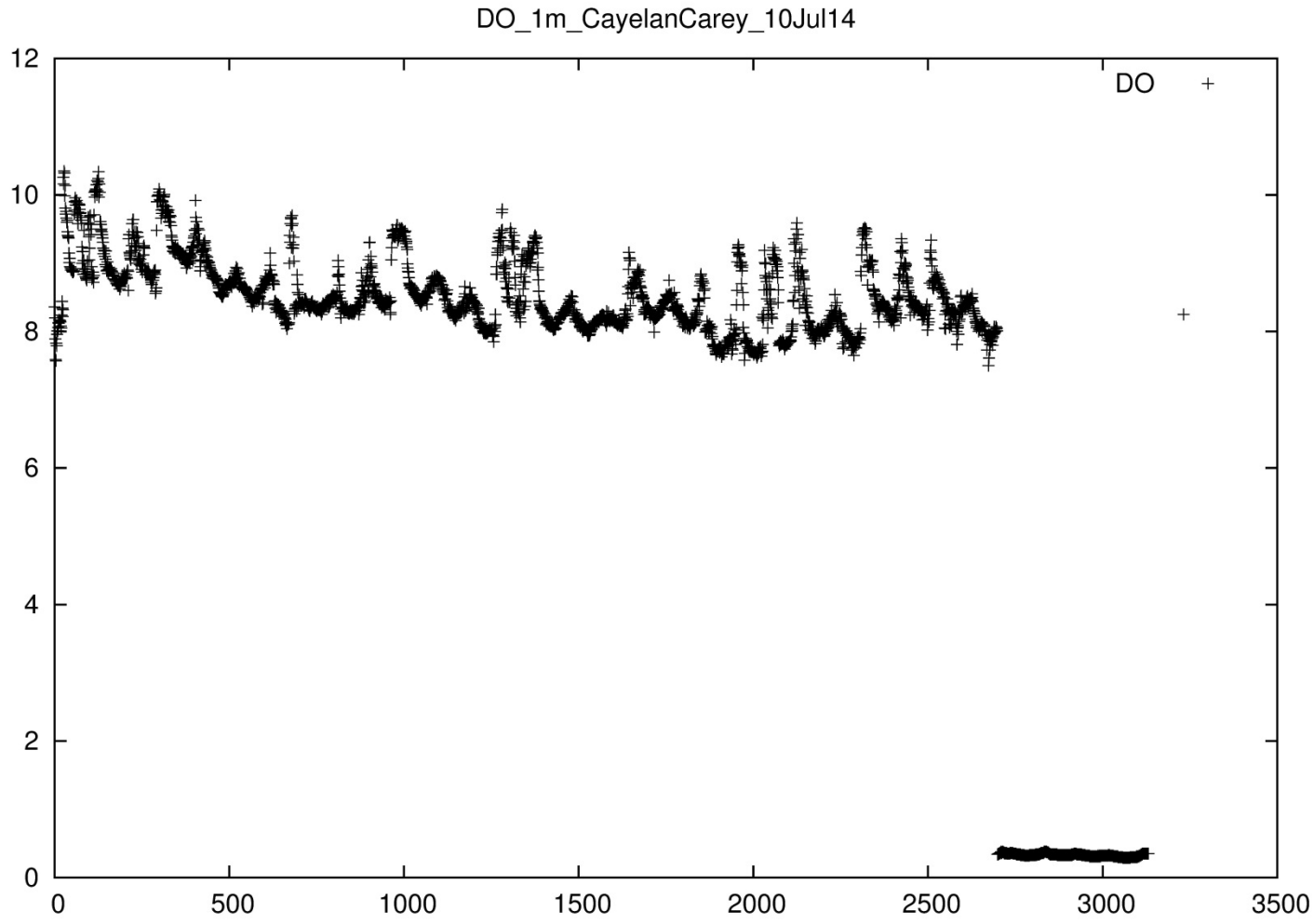
Battery Filter

1. Read the raw data file line by line
   (delete the header lines and distribution lines)

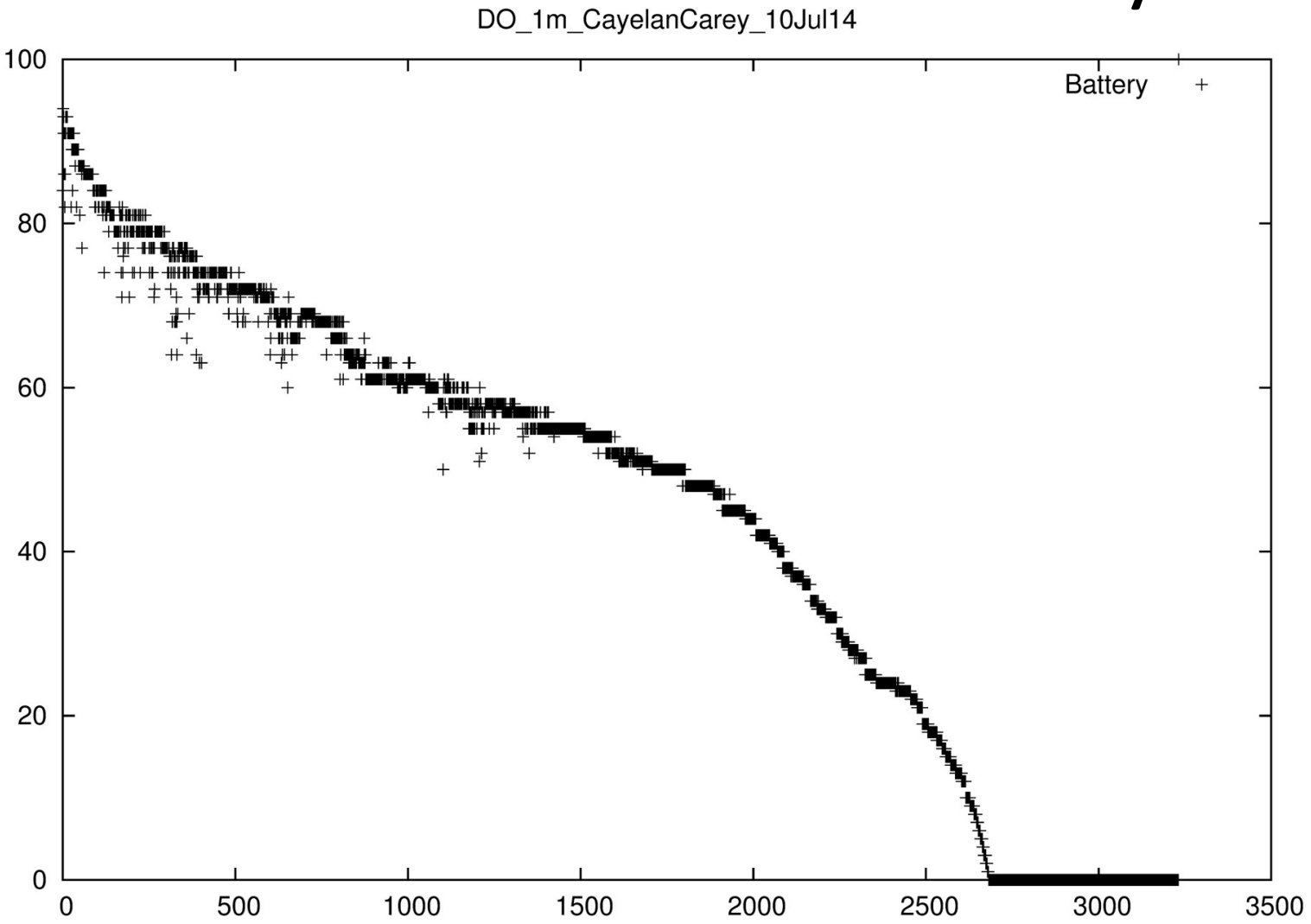
2. For each line, check the attributes
   if “LDO” > 15 or “LDO” < 0,
   append flag “outlier” at the end of line;
   if “IBatt” is 0,
   append flag “nobattery” at the end of line;
   if “Temp” is null,
   append flag “nobattery” at the end of line;

3. Save all lines to new .text file

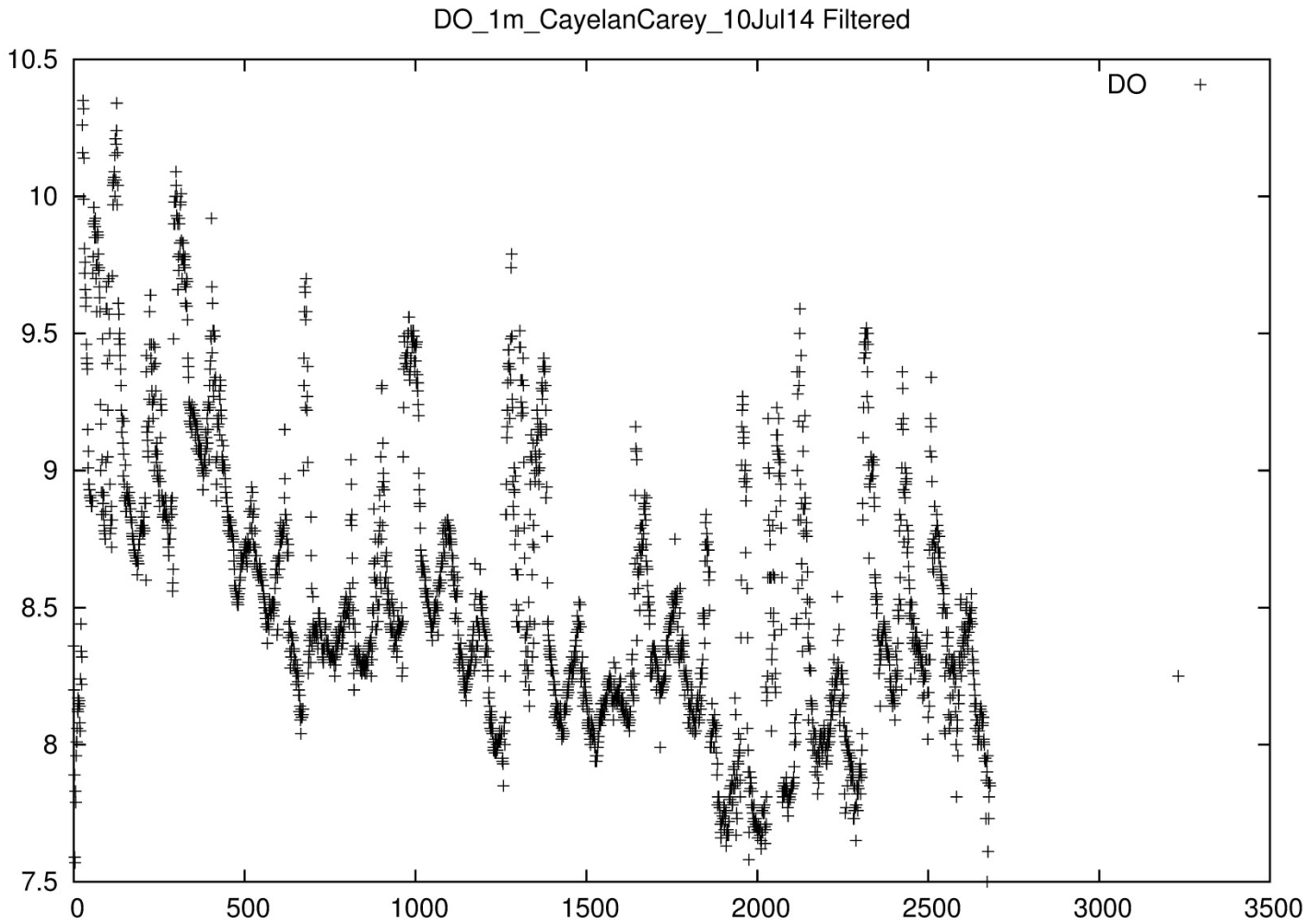
Before Outlier Filter and Battery Filter



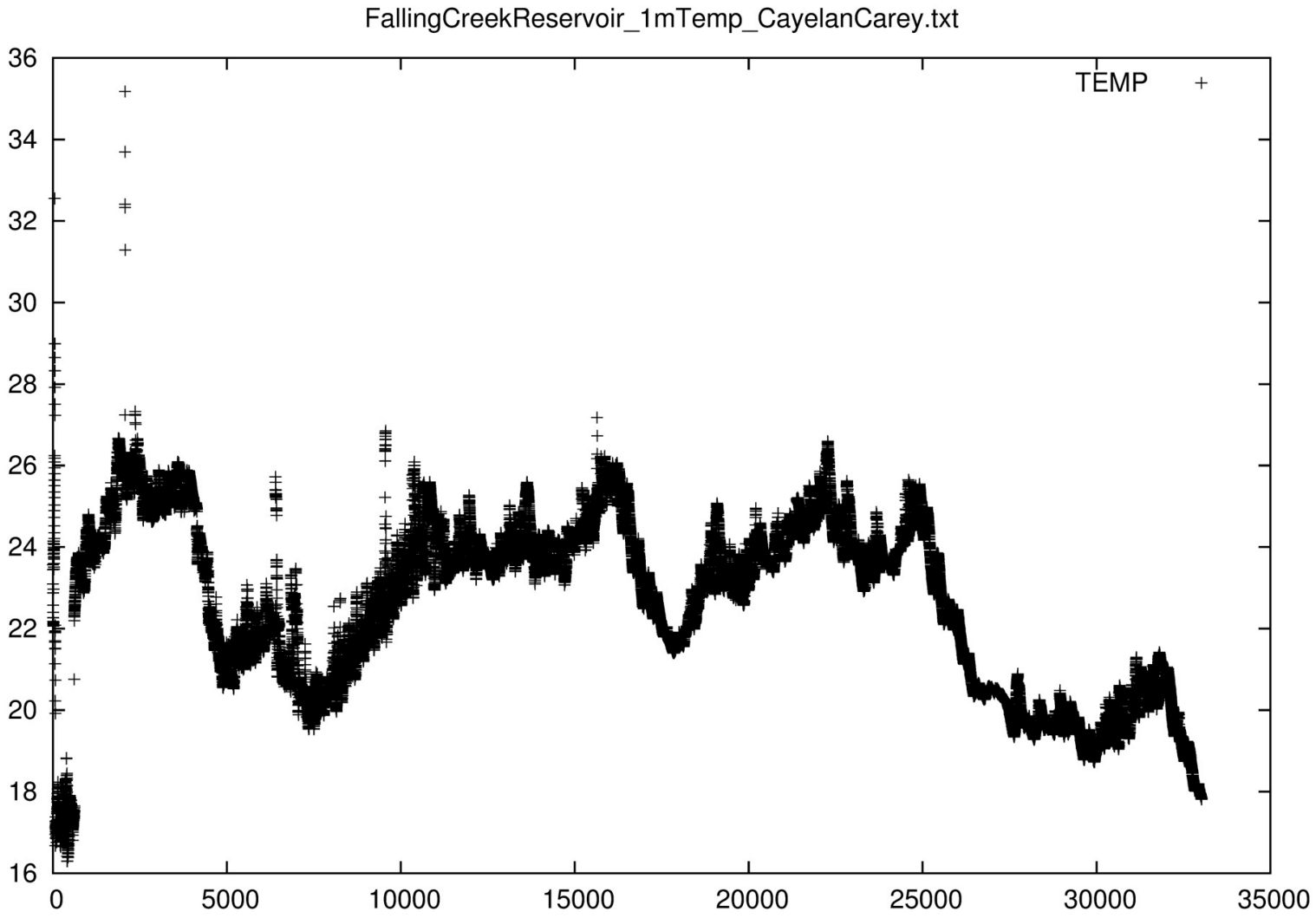
Before Outlier Filter and Battery Filter



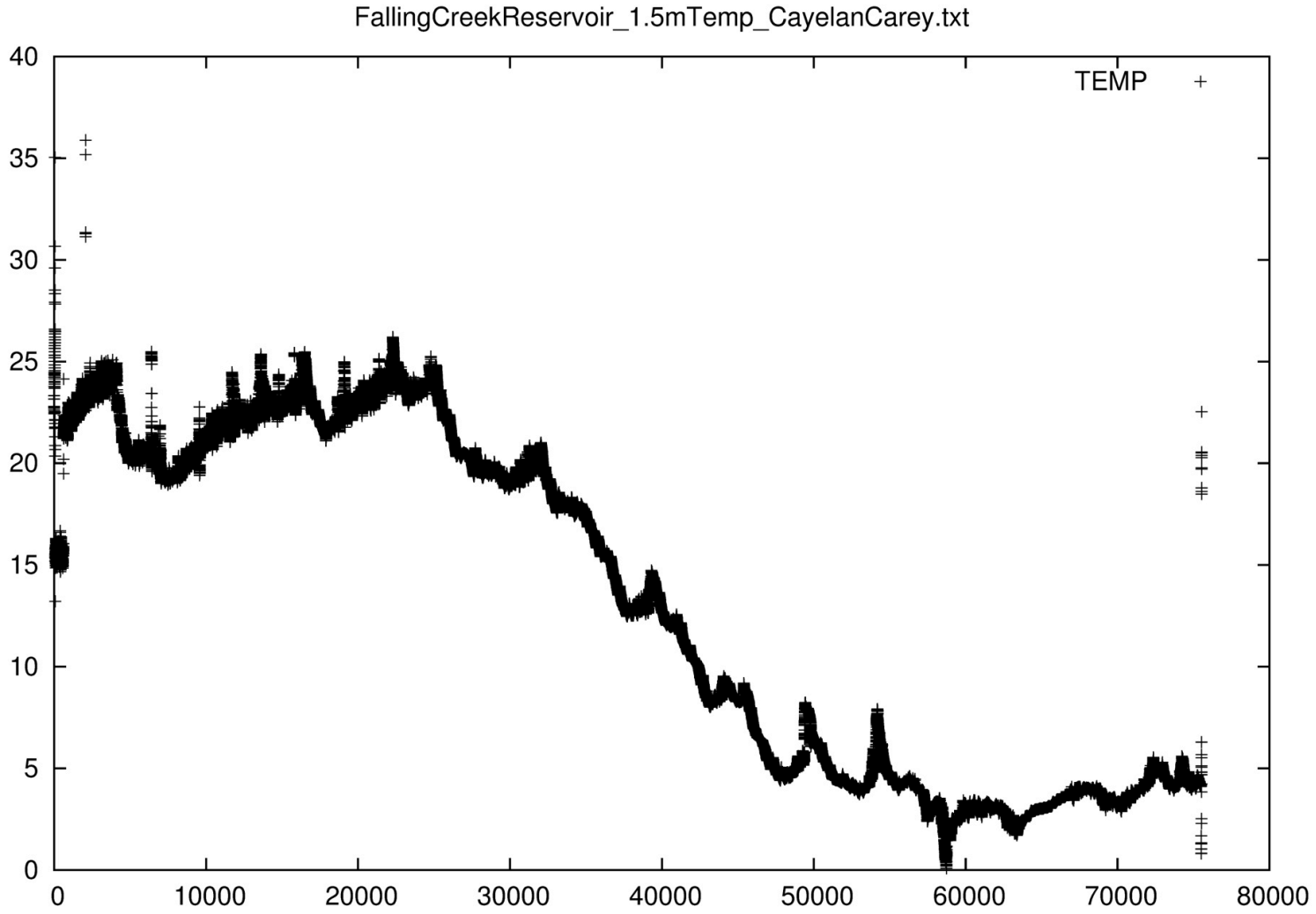
After Outlier Filter and Battery Filter



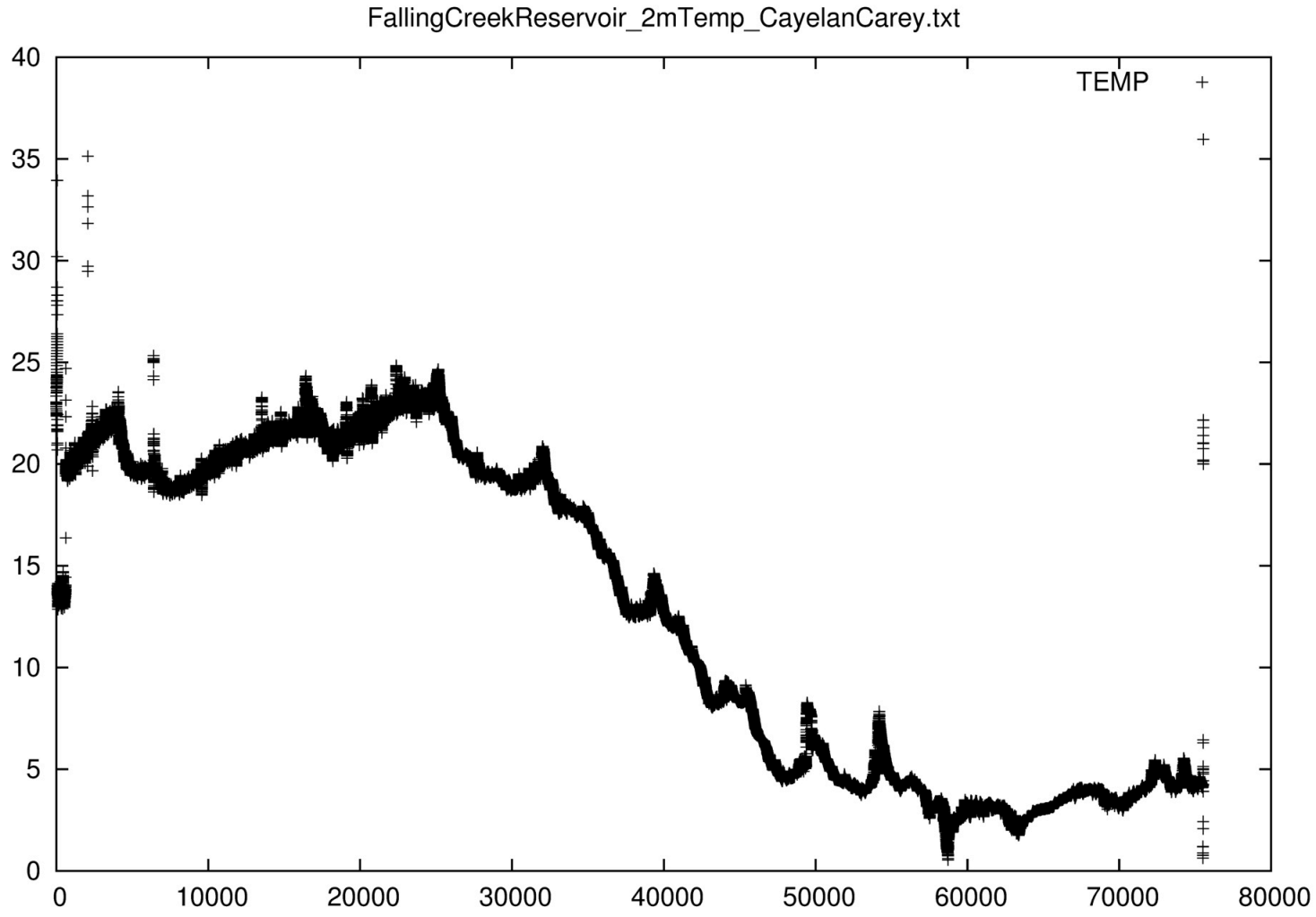
After Outlier Filter and Battery Filter



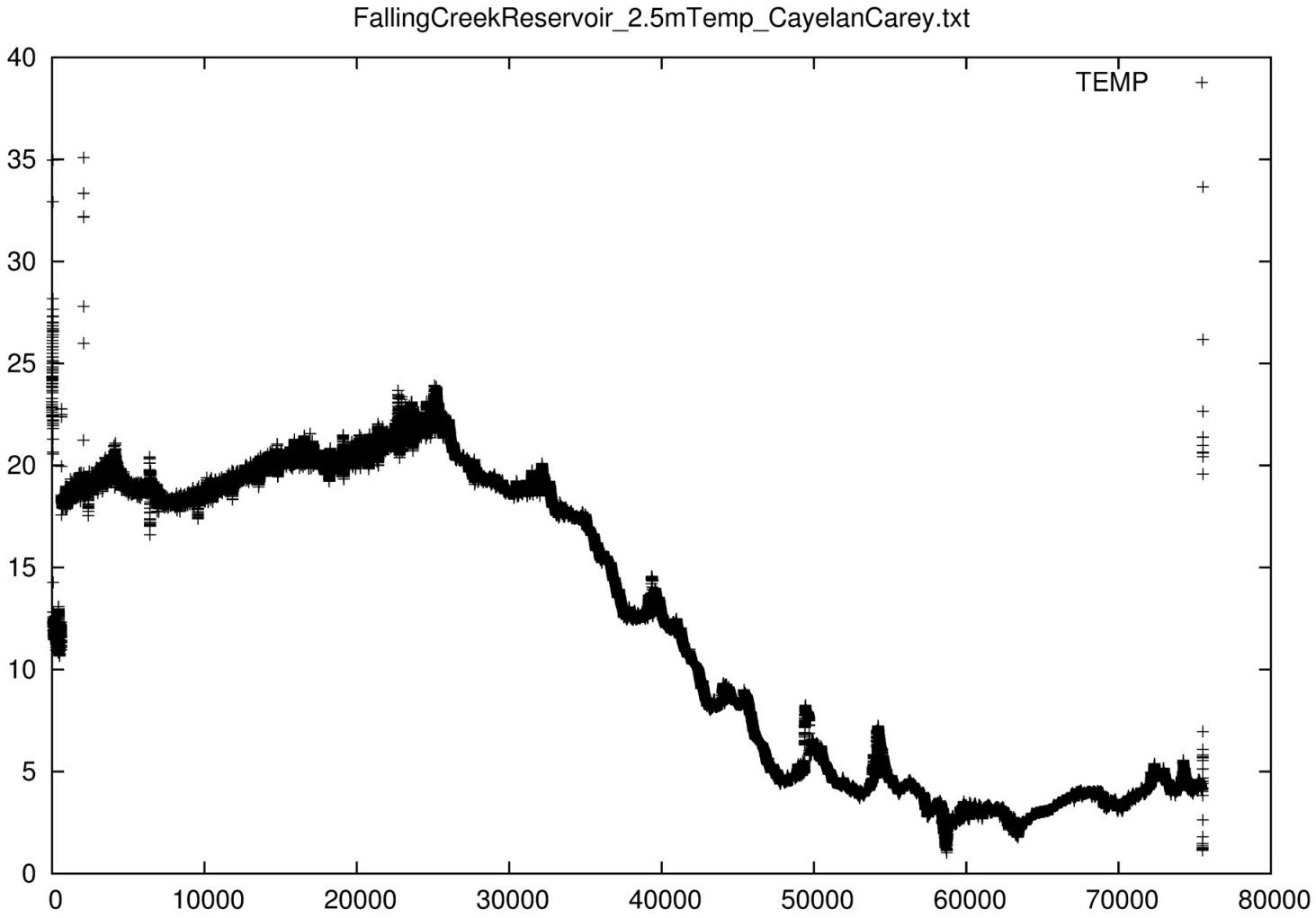
After Outlier Filter and Battery Filter



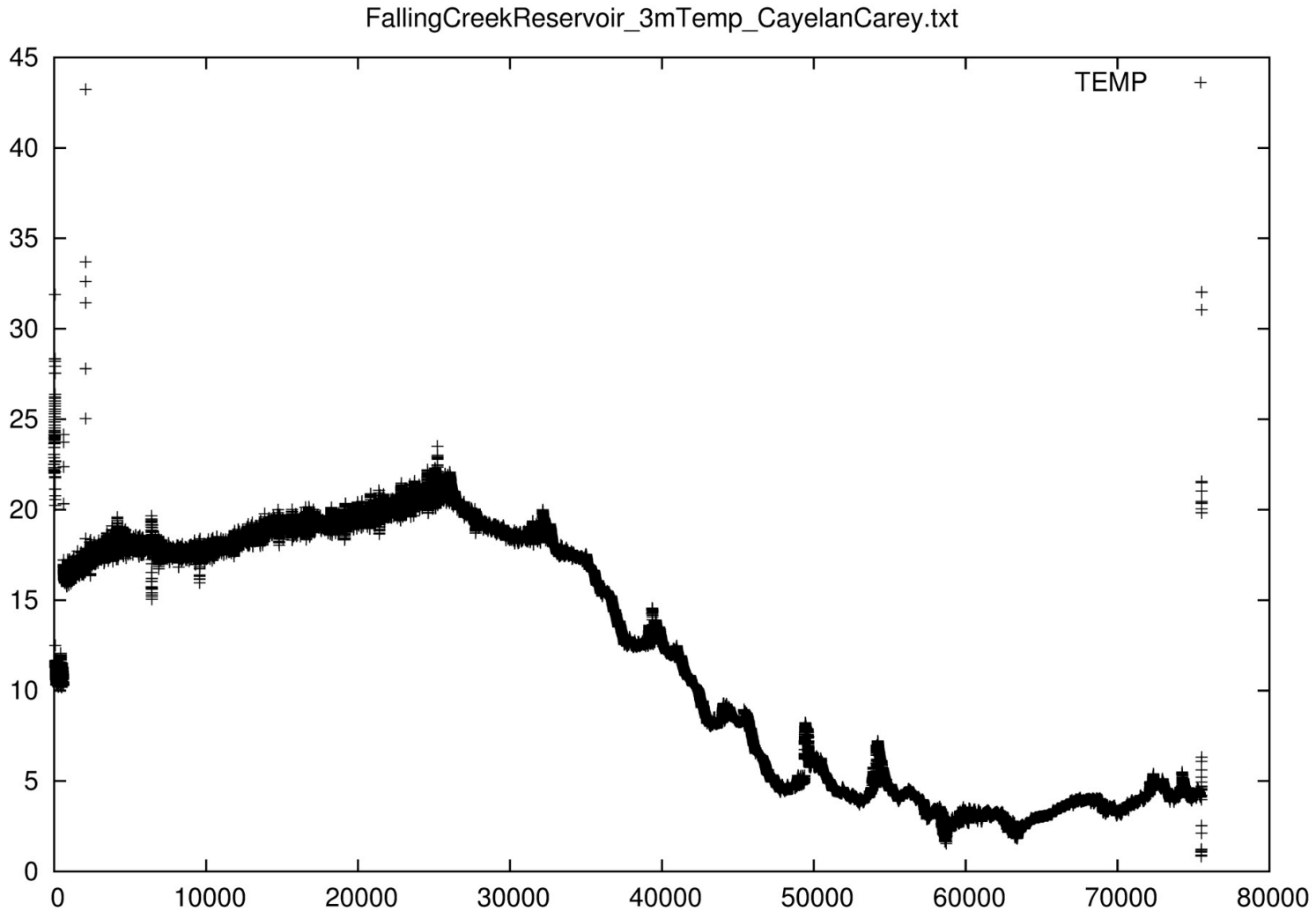
After Outlier Filter and Battery Filter

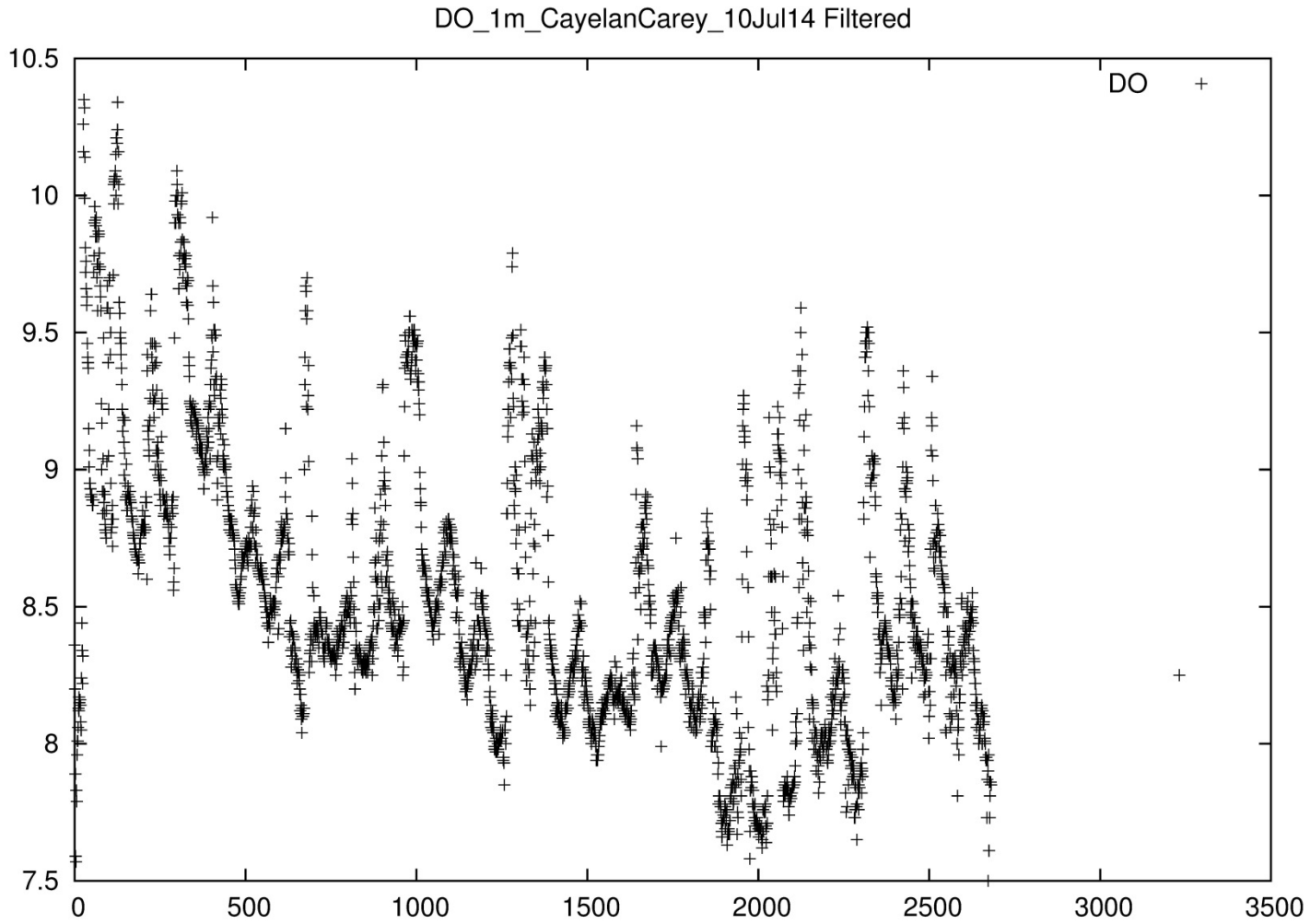


After Outlier Filter and Battery Filter



After Outlier Filter and Battery Filter




DO_1m_CayelanCarey_10Jul14 Filtered
DO
10.5
10
9.5
9
8.5
8
7.5
0
500
1000
1500
2000
2500
3000
3500

Drift Filter(1)

Whatever Temp and LDO, they increase and decrease by a constant slope.
When there are drifts on the sensor, the slope will change.
e.g.,
for LDO, we calculate slopes between different points.(x1 is before x2)

Slope of x1= (x2.ldo –x1.ldo)/15

Flag and remove the x2 whose slope is larger than a limited slope (n).

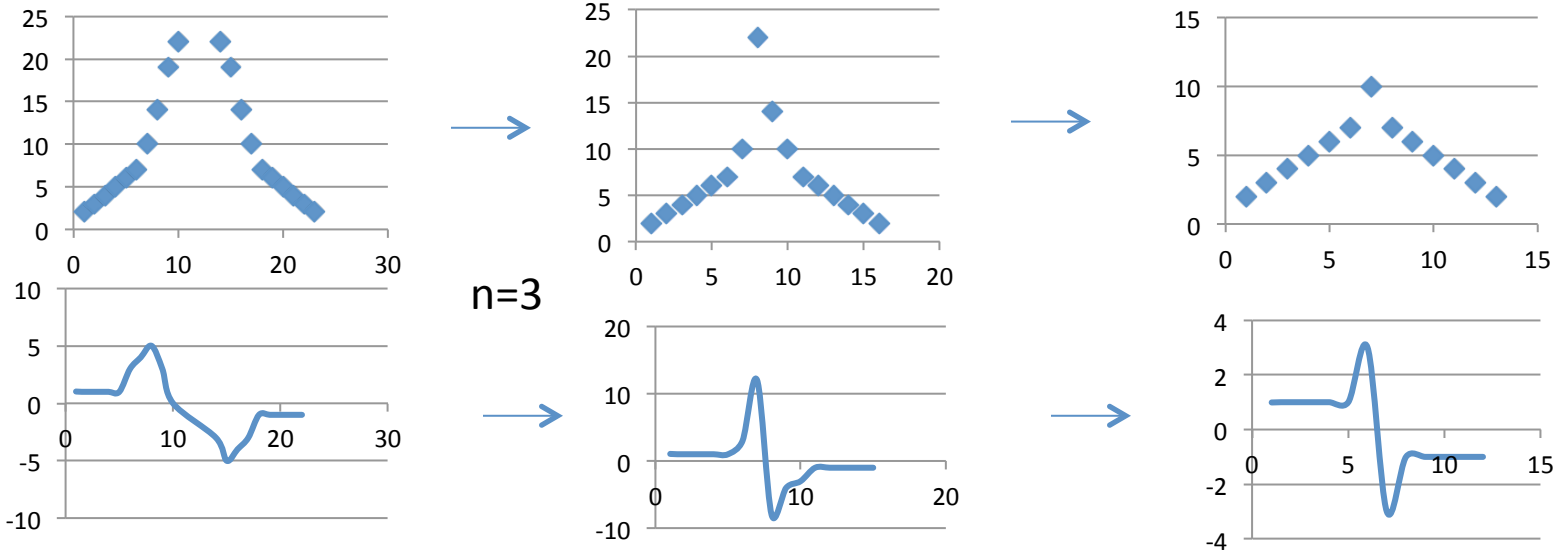
Flag and remove the x1 whose slope is smaller than a limited slope (–n).

-> we reduce the range of points which are affected by drifts. (not efficient)

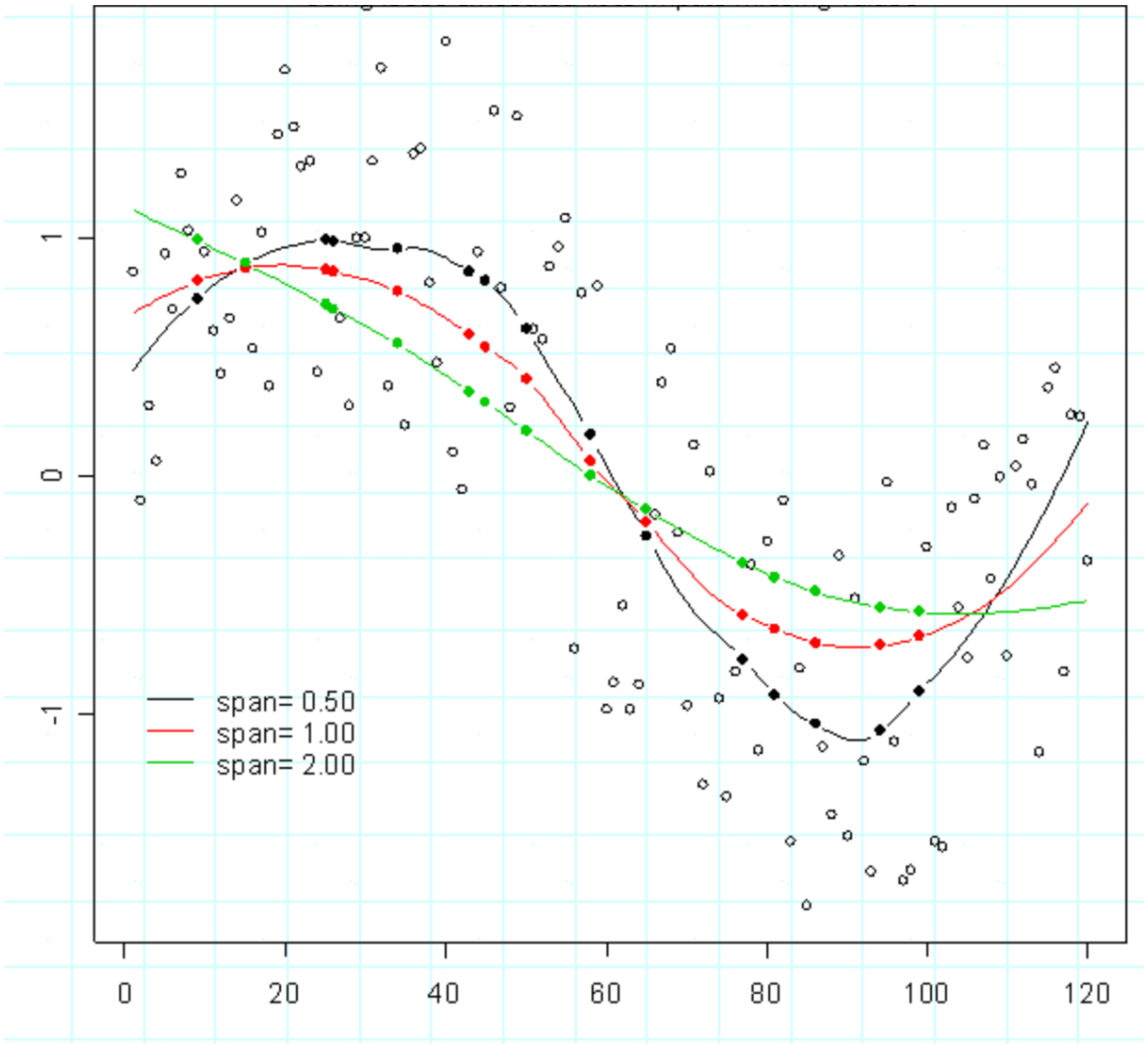
point

slope

n=3



Loess Curve

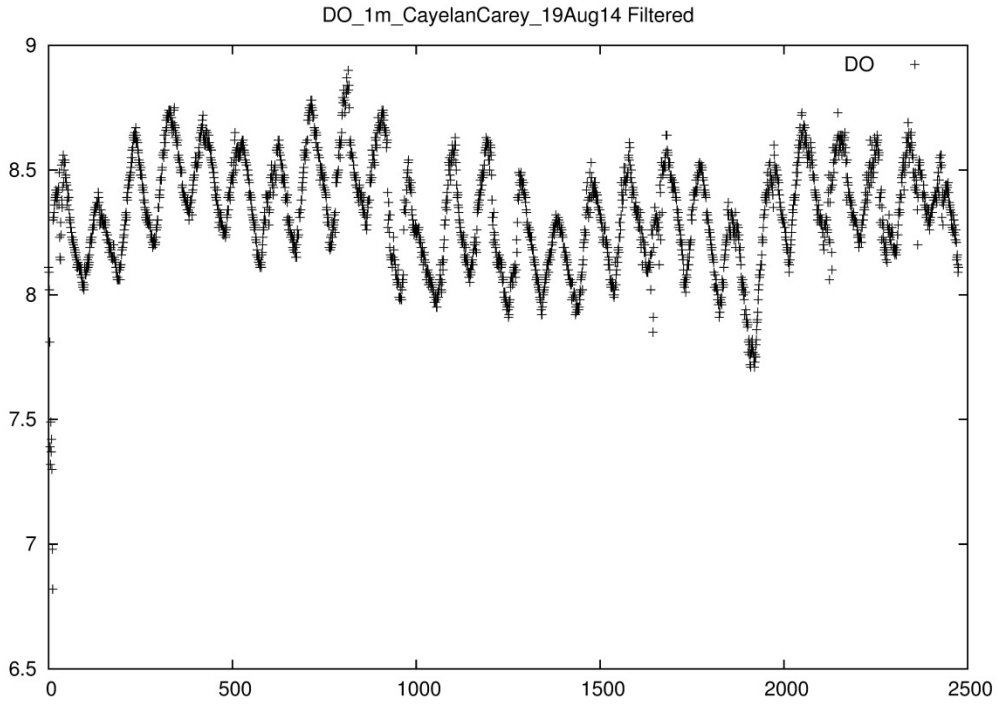


If Span is lower, we get more curved line.
Then Loess Curve will be more accurate

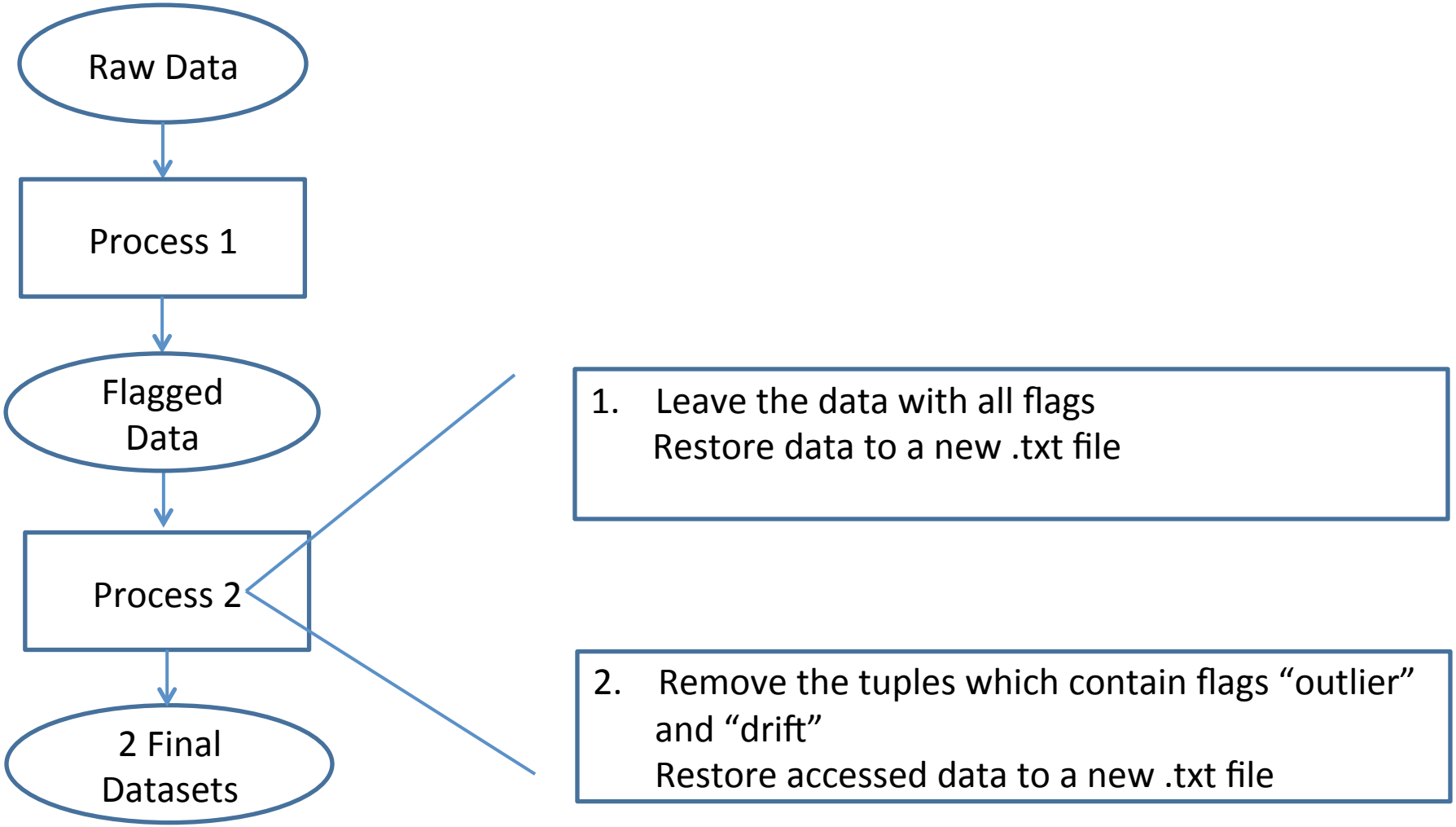
Drift Filter(2)

The scatter plot shows Temp and LDO are affected by time and each of them has a high correlation with it.
The Loess Curve is a good trend curve to describe the data changes with time passing.
-> The point which is further from Loess Curve is more likely to be the Drift point.

There is a online sample of Loess Curve on Java. It could help me make a efficient algorithm as Drift Filter.
But that's the future work.



Data Cleaning Workflow (cont'd)



Conclusion and Future Work

- Retrieve raw data
- Visualization raw data
- Data cleaning
  - Outlier
  - Out of battery
  - Drift (ongoing work)
- Visualization on cleaned data
- Loess Curve(Local Regression)
- QC/QA

Thank you!