



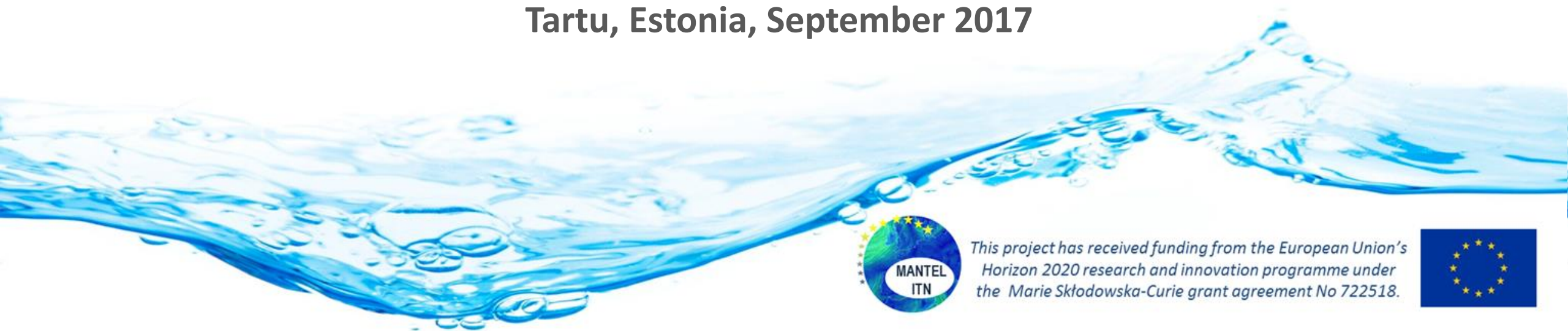
# Data Quality Control and Management

**Don Pierson**

**Uppsala University**

**MANTEL ITN Workshop**

**Tartu, Estonia, September 2017**



*This project has received funding from the European Union's  
Horizon 2020 research and innovation programme under  
the Marie Skłodowska-Curie grant agreement No 722518.*



# Changes in Data Practice That Will Affect Your Work as a Scientist

- It is now expected that all data collected by scientists will be archived and available to others on request.
  - Monitoring data
  - Data generated by analysis and simulation
  - This is now required by funding agencies.
  - But there is very little financial support for this
- It is now expected that all data associated with a peer reviewed paper be deposited in an archive that is publicly available.
- Computer programs scripts and other data files (spreadsheets etc.) should also be available with data files so that work can be replicated

# Challenges of Data Intensive Science

- Large amounts of data to be archived. Easily in the 100 MB-GB range
- Many different programs and software is used
  - Open Source
  - Proprietary
- Complex Workflows
  - Can be difficult to repeat or remember
  - May change over time
- Programing
  - Many languages – always changing
  - Inevitable bugs and reanalysis
- SAVING ALL OF THIS SO THAT OTHERS CAN USE AND UNDERSTAND IT TAKES TIME!
  - However this also means that you will be able to remember and understand what you did several years ago!

# What to Do?

- A Better More Rational Way – Organize by Project!
- Realize from the beginning that you are going to need to produce a data archive, and also document all programs developed workflows etc
- Make data archiving and documentation part of day to day work so it is essentially done by the time your paper is written.

# Caveats and Qualifiers

- This is not necessarily what I do myself, but it is what I think we probably should be moving towards.
  - New Students don't have to break as many old habits
- Much of what I describe is related to documenting simulations and workflows associated with them. Other types of data may require other strategies.
- There is no fixed way to best organize your data. What I present is meant to stimulate discussion

# Some Principle for Project Based Organization

## (With High Ambitions)

- All data from initial measurements to final manuscript are in a single directory.
  - Initial measurements (automated HF data and lab data)
  - Data analysis (statistical, custom programs, model simulations visualizations)
  - Correspondence and documentation files
  - Manuscript (all versions and revisions)
- All programs input files, output files, and control scripts are in this directory.  
Work can be reproduced easily
  - Allows other to check your work
  - Allows you to rerun your analysis when you discover inevitable errors
  - To some extent the workflow can be considered the documentation
- Allows easy development of data archives supplementary material etc.
- Project directory is always backed up on a regular basis

# Some Examples of How My Work has Slowly Evolved Towards Project Based Organization

# NYC DEP Modeling Group – Four Commandments of Model Data Management

- Thou shall save and archive data in ASCII format
- Thou shall save date and time in ISO standard format in the first column of the data file
- Thou shall have informative file headers
- Thou shall create a brief but informative text file in simulation directory that describes input files output files and the programs that created the output.



# Thou shall save and archive data in ASCII format

- Why ASCII?
  - Still only other universal file format (at least that I know of)
  - Will ALWAYS be readable by others. Not linked to proprietary software.
  - Easy to compare different versions of files
- Other open source advanced file formats (good but more complicated to work with) Support available Phyton MatLab IDL etc)
  - HDF – mainly spatial data
  - Net CDF – mainly climate and met data
  - Viewers are available
  - Key outputs can still be saved as ASCII
- Proprietary software (Excel, JMP, SAS, Vensim, Stella etc) Powerful software for interactive analysis. Easy learning curve.
  - Input and output can usually be stored in ASCII format.
  - None read HDF or NetCDF

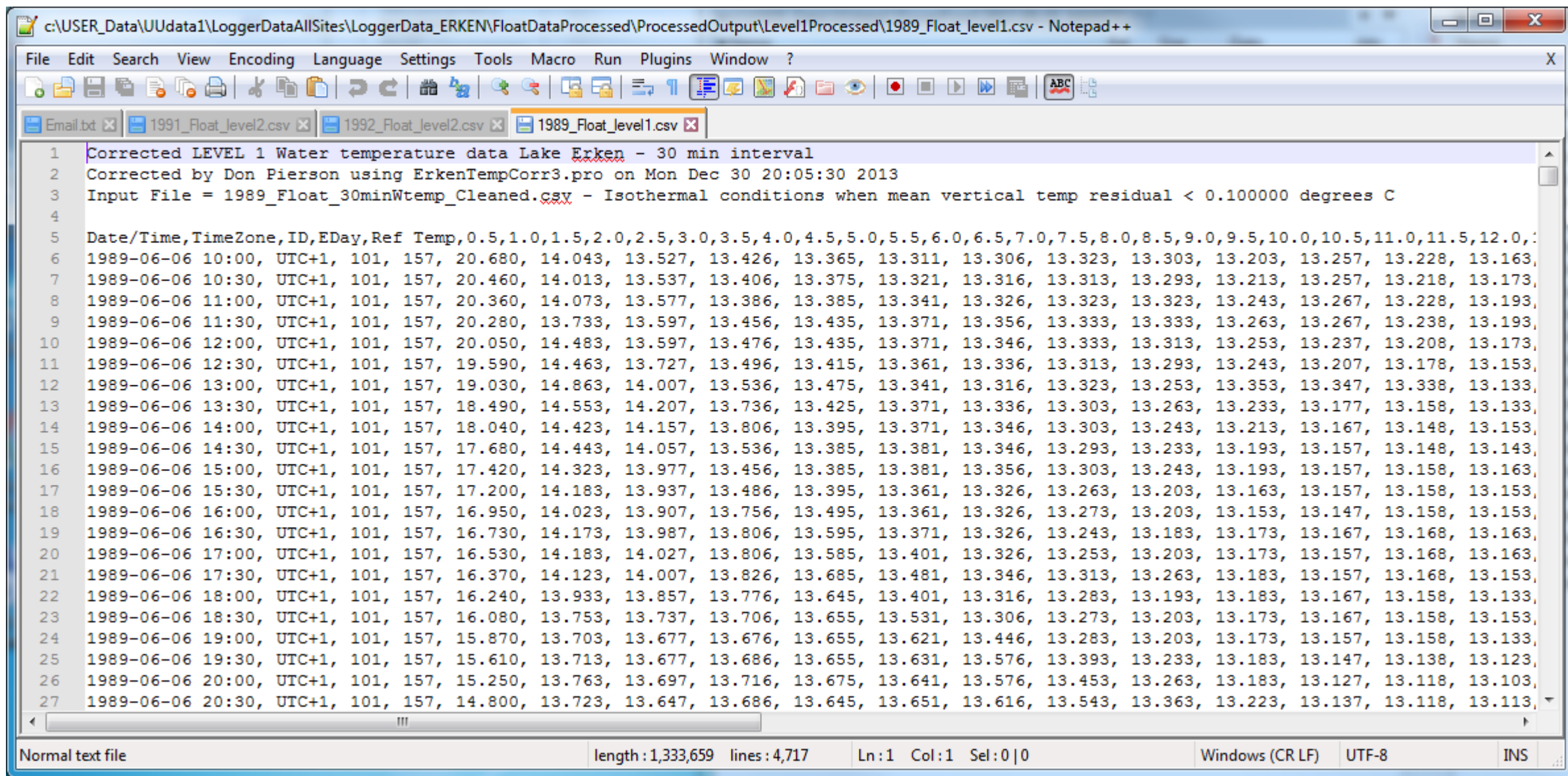
# Thou shall save date and time in ISO 8061 standard format in the first column of the data file

- ISO format yyyy-mm-dd hh:mm
- <https://www.iso.org/iso-8601-date-and-time-format.html>
- Now understood by Excel
- Save lots of time and errors
- Easily parsed in to components
- Allow sub-programs or procedures to simplify file reading
  - Format known
  - Position known
  - Even better if you fill in all missing rows

# Thou shall have informative file headers

- Date that file was created
- Name of program that created the file
- Input files to the program that created the file
- Name of Analyst
- Blank line
- Column headers
- Column units

# Example of File That Meets Commandments 1-3



The screenshot shows a Notepad++ window with the title bar "c:\USER\_Data\Udata1\LoggerDataAllSites\LoggerData\_ERKEN\FloatDataProcessed\ProcessedOutput\Level1Processed\1989\_Float\_level1.csv - Notepad++". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Tools, Macro, Run, Plugins, Window, and ?. The toolbar contains various icons for file operations and editing. The tab bar shows four open files: Email.txt, 1991\_Float\_level2.csv, 1992\_Float\_level2.csv, and 1989\_Float\_level1.csv. The main text area displays the following content:

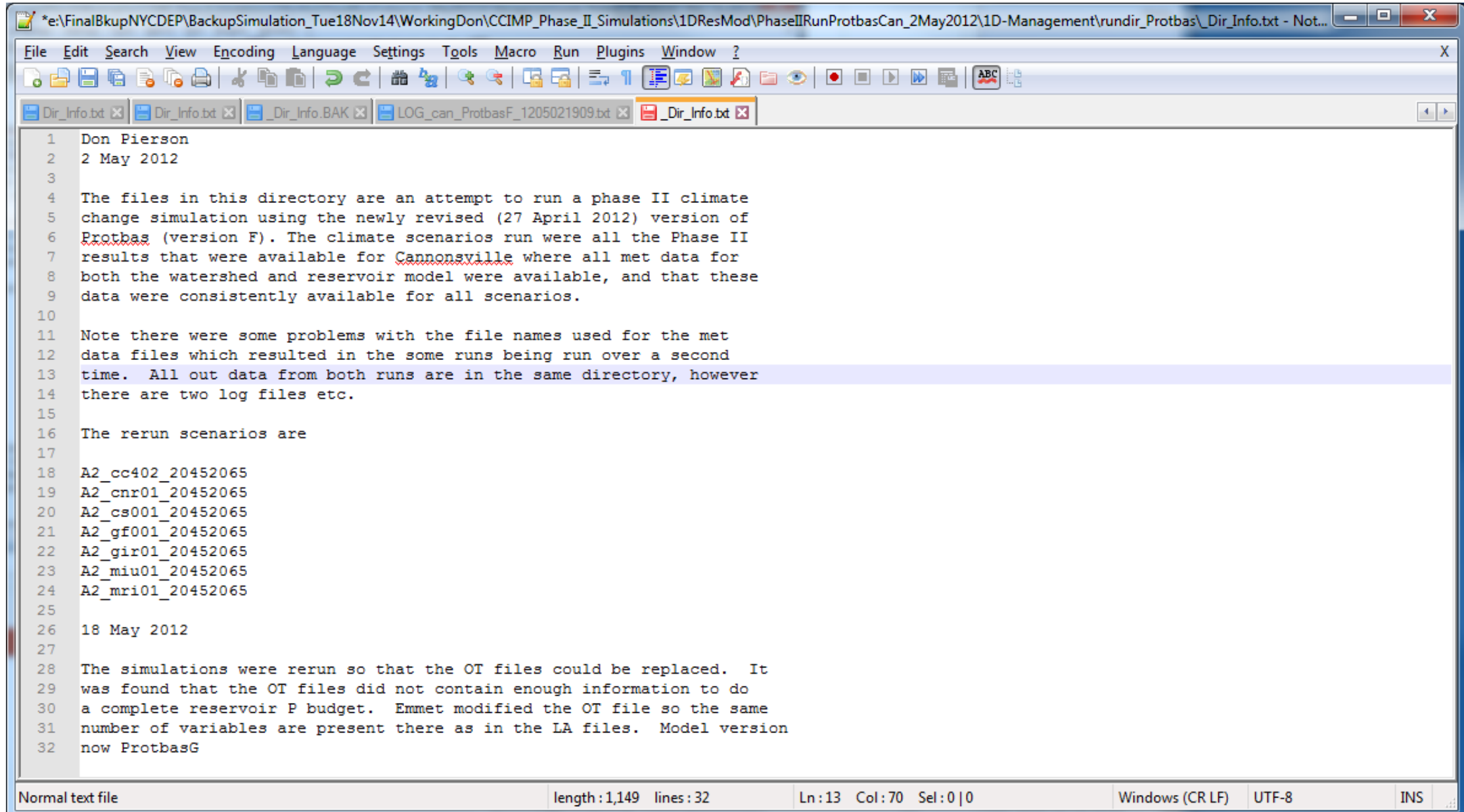
```
1 Corrected LEVEL 1 Water temperature data Lake Erken - 30 min interval
2 Corrected by Don Pierson using ErkenTempCorr3.pro on Mon Dec 30 20:05:30 2013
3 Input File = 1989_Float_30minWtemp_Cleaned.csv - Isothermal conditions when mean vertical temp residual < 0.100000 degrees C
4
5 Date/Time,TimeZone,ID,EDay,Ref Temp,0.5,1.0,1.5,2.0,2.5,3.0,3.5,4.0,4.5,5.0,5.5,6.0,6.5,7.0,7.5,8.0,8.5,9.0,9.5,10.0,10.5,11.0,11.5,12.0,:
6 1989-06-06 10:00, UTC+1, 101, 157, 20.680, 14.043, 13.527, 13.426, 13.365, 13.311, 13.306, 13.323, 13.303, 13.203, 13.257, 13.228, 13.163,
7 1989-06-06 10:30, UTC+1, 101, 157, 20.460, 14.013, 13.537, 13.406, 13.375, 13.321, 13.316, 13.313, 13.293, 13.213, 13.257, 13.218, 13.173,
8 1989-06-06 11:00, UTC+1, 101, 157, 20.360, 14.073, 13.577, 13.386, 13.385, 13.341, 13.326, 13.323, 13.323, 13.243, 13.267, 13.228, 13.193,
9 1989-06-06 11:30, UTC+1, 101, 157, 20.280, 13.733, 13.597, 13.456, 13.435, 13.371, 13.356, 13.333, 13.333, 13.263, 13.267, 13.238, 13.193,
10 1989-06-06 12:00, UTC+1, 101, 157, 20.050, 14.483, 13.597, 13.476, 13.435, 13.371, 13.346, 13.333, 13.313, 13.253, 13.237, 13.208, 13.173,
11 1989-06-06 12:30, UTC+1, 101, 157, 19.590, 14.463, 13.727, 13.496, 13.415, 13.361, 13.336, 13.313, 13.293, 13.243, 13.207, 13.178, 13.153,
12 1989-06-06 13:00, UTC+1, 101, 157, 19.030, 14.863, 14.007, 13.536, 13.475, 13.341, 13.316, 13.323, 13.253, 13.353, 13.347, 13.338, 13.133,
13 1989-06-06 13:30, UTC+1, 101, 157, 18.490, 14.553, 14.207, 13.736, 13.425, 13.371, 13.336, 13.303, 13.263, 13.233, 13.177, 13.158, 13.133,
14 1989-06-06 14:00, UTC+1, 101, 157, 18.040, 14.423, 14.157, 13.806, 13.395, 13.371, 13.346, 13.303, 13.243, 13.213, 13.167, 13.148, 13.153,
15 1989-06-06 14:30, UTC+1, 101, 157, 17.680, 14.443, 14.057, 13.536, 13.385, 13.381, 13.346, 13.293, 13.233, 13.193, 13.157, 13.148, 13.143,
16 1989-06-06 15:00, UTC+1, 101, 157, 17.420, 14.323, 13.977, 13.456, 13.385, 13.381, 13.356, 13.303, 13.243, 13.193, 13.157, 13.158, 13.163,
17 1989-06-06 15:30, UTC+1, 101, 157, 17.200, 14.183, 13.937, 13.486, 13.395, 13.361, 13.326, 13.263, 13.203, 13.163, 13.157, 13.158, 13.153,
18 1989-06-06 16:00, UTC+1, 101, 157, 16.950, 14.023, 13.907, 13.756, 13.495, 13.361, 13.326, 13.273, 13.203, 13.153, 13.147, 13.158, 13.153,
19 1989-06-06 16:30, UTC+1, 101, 157, 16.730, 14.173, 13.987, 13.806, 13.595, 13.371, 13.326, 13.243, 13.183, 13.173, 13.167, 13.168, 13.163,
20 1989-06-06 17:00, UTC+1, 101, 157, 16.530, 14.183, 14.027, 13.806, 13.585, 13.401, 13.326, 13.253, 13.203, 13.173, 13.157, 13.168, 13.163,
21 1989-06-06 17:30, UTC+1, 101, 157, 16.370, 14.123, 14.007, 13.826, 13.685, 13.481, 13.346, 13.313, 13.263, 13.183, 13.157, 13.168, 13.153,
22 1989-06-06 18:00, UTC+1, 101, 157, 16.240, 13.933, 13.857, 13.776, 13.645, 13.401, 13.316, 13.283, 13.193, 13.183, 13.167, 13.158, 13.133,
23 1989-06-06 18:30, UTC+1, 101, 157, 16.080, 13.753, 13.737, 13.706, 13.655, 13.531, 13.306, 13.273, 13.203, 13.173, 13.167, 13.158, 13.153,
24 1989-06-06 19:00, UTC+1, 101, 157, 15.870, 13.703, 13.677, 13.676, 13.655, 13.621, 13.446, 13.283, 13.203, 13.173, 13.157, 13.158, 13.133,
25 1989-06-06 19:30, UTC+1, 101, 157, 15.610, 13.713, 13.677, 13.686, 13.655, 13.631, 13.576, 13.393, 13.233, 13.183, 13.147, 13.138, 13.123,
26 1989-06-06 20:00, UTC+1, 101, 157, 15.250, 13.763, 13.697, 13.716, 13.675, 13.641, 13.576, 13.453, 13.263, 13.183, 13.127, 13.118, 13.103,
27 1989-06-06 20:30, UTC+1, 101, 157, 14.800, 13.723, 13.647, 13.686, 13.645, 13.651, 13.616, 13.543, 13.363, 13.223, 13.137, 13.118, 13.113,
```

The status bar at the bottom shows "Normal text file", "length : 1,333,659 lines : 4,717", "Ln : 1 Col : 1 Sel : 0 | 0", "Windows (CR LF)", "UTF-8", and "INS".

# Thou shall create a brief but informative text file in simulation directory

- Date
- Analyst name
- Purpose of simulation
- Files in the directory
  - Input
  - Output
  - Other
- Informative description
- But concise enough that it is actually read

# Dir\_Info.txt File That Meets Commandment 4



```
*e:\FinalBkupNYCDEP\BackupSimulation_Tue18Nov14\WorkingDon\CCIMP_Phase_II_Simulations\1DResMod\PhaseIIRunProtbasCan_2May2012\1D-Management\rundir_Protbas\_Dir_Info.txt - Not...
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
Dir_Info.txt Dir_Info.txt _Dir_Info.BAK LOG_can_ProtbasF_1205021909.txt _Dir_Info.txt
1 Don Pierson
2 2 May 2012
3
4 The files in this directory are an attempt to run a phase II climate
5 change simulation using the newly revised (27 April 2012) version of
6 Protbas (version F). The climate scenarios run were all the Phase II
7 results that were available for Cannonsville where all met data for
8 both the watershed and reservoir model were available, and that these
9 data were consistently available for all scenarios.
10
11 Note there were some problems with the file names used for the met
12 data files which resulted in the some runs being run over a second
13 time. All out data from both runs are in the same directory, however
14 there are two log files etc.
15
16 The rerun scenarios are
17
18 A2_cc402_20452065
19 A2_cnr01_20452065
20 A2_cs001_20452065
21 A2_gf001_20452065
22 A2_gir01_20452065
23 A2_miu01_20452065
24 A2_mri01_20452065
25
26 18 May 2012
27
28 The simulations were rerun so that the OT files could be replaced. It
29 was found that the OT files did not contain enough information to do
30 a complete reservoir P budget. Emmet modified the OT file so the same
31 number of variables are present there as in the LA files. Model version
32 now ProtbasG
Normal text file length: 1,149 lines: 32 Ln: 13 Col: 70 Sel: 0 | 0 Windows (CR LF) UTF-8 INS
```

# Workflow Log File Provides Self Documentation

```
e:\FinalBkupNYCDEP\BackupSimulation_Tue18Nov14\WorkingDon\CCIMP_Phase_II_Simulations\1DResMod\PhaseIIRunProtbasCan_2May2012\1D-Management\rundir_Protbas\LOG_can_ProtbasF...
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
1991_Float_level2.csv 1992_Float_level2.csv 1989_Float_level1.csv _Dir_Info.txt Dir_Info.txt Dir_Info.txt Dir_Info.txt _Dir_Info.BAK LOG_can_ProtbasF_1205021909.txt
1 ***** Reservoir Model Log*****
2
3
4 1D reservoir model simulation info linked to all simulations taged 1205021909
5 Created using the 1D reservoir model - ProtbasF on Wed May 2 19:09:33 EDT 2012
6
7 -----
8
9 ProtbasF run for can Run completed Wed May 2 19:27:04 EDT 2012
10
11 Input files
12 Elev_ca-bin25_A1B_cc402_20452065_1108121918.tab
13 Qin_ca-bin25_A1B_cc402_20452065_1108121918.tab
14 Qout_ca-bin25_A1B_cc402_20452065_1108121918.tab
15 Met_ca-bin25_A1B_cc402_20452065_1106081706.tab
16 NutLoad_ca-bin25_A1B_cc402_20452065_1108121918.tab
17 TribT_ca-bin25_A1B_cc402_20452065_1108121918.tab
18 Output files
19 Daily averaged Layer Files
20 LA_Alga_can_A1B_cc402_20452065_1205021909.tab LA_Nutr_can_A1B_cc402_20452065_1205021909.tab
21 Daily averaged Reservoir Outflow Files
22 OT_Alga_can_A1B_cc402_20452065_1205021909.tab OT_Nutr_can_A1B_cc402_20452065_1205021909.tab
23 Daily averaged Profile Files
24 PR_Alga_can_A1B_cc402_20452065_1205021909.tab PR_Nutr_can_A1B_cc402_20452065_1205021909.tab
25
26 -----
27
28 ProtbasF run for can Run completed Wed May 2 19:44:57 EDT 2012
29
30 Input files
31 Elev_ca-bin25_A1B_cc402_20802100_1108121918.tab
32 Qin_ca-bin25_A1B_cc402_20802100_1108121918.tab
33 Qout_ca-bin25_A1B_cc402_20802100_1108121918.tab
Normal text file length: 37,489 lines: 830 Ln: 1 Col: 1 Sel: 0 | 0 Windows (CR LF) UTF-8 INS
```

# Some Examples of How I Try to Use Project Based Organization



# Example of Project Based Organization Met data Quality Control

The screenshot displays the Total Commander file manager interface, showing a hierarchical project-based organization of met data. The left pane shows the root directory structure, and the right pane shows a sub-directory containing various data files and programs.

**Left Pane (Root Directory):**

Name	Ext	Size	Date	Attr
[..]	<DIR>		2017-09-02 14:25	----
[2016-11-10_MeanDailyErkenIslandData]	<DIR>		2017-02-20 14:55	----
[2016-12-02_ErkenProcessed_1986-1996]	<DIR>		2017-09-02 14:25	----
[2017-02-06_GaltenCumHour]	<DIR>		2017-02-06 11:00	----
[2017-02-06_WinterSpringDataRequest]	<DIR>		2017-08-29 13:59	----
[2017-02-09_ProcessSLUMetData]	<DIR>		2017-02-09 12:57	----
[2017-02-09_SWRcompare]	<DIR>		2017-02-09 16:33	----
[2017-02-26_CalculateMeanDailyFloatTemp]	<DIR>		2017-03-15 11:35	----
[2017-03-14_ProcessGOTM_TempData]	<DIR>		2017-03-16 11:52	----
[2017-05-11_CalcO2Sat]	<DIR>		2017-05-30 20:02	----
[2017-05-31_CalibErkenTempSys]	<DIR>		2017-06-01 15:16	----
[2017-05-31_ReFormatYSI_Erken]	<DIR>		2017-05-31 13:03	----
[2017-06-11_ErkenInFlowChem]	<DIR>		2017-06-14 14:35	----
[2017-07-21_Island_WTemp_QC]	<DIR>		2017-08-05 20:55	----

**Right Pane (Sub-directory: 2017-02-06\_GaltenCumHour):**

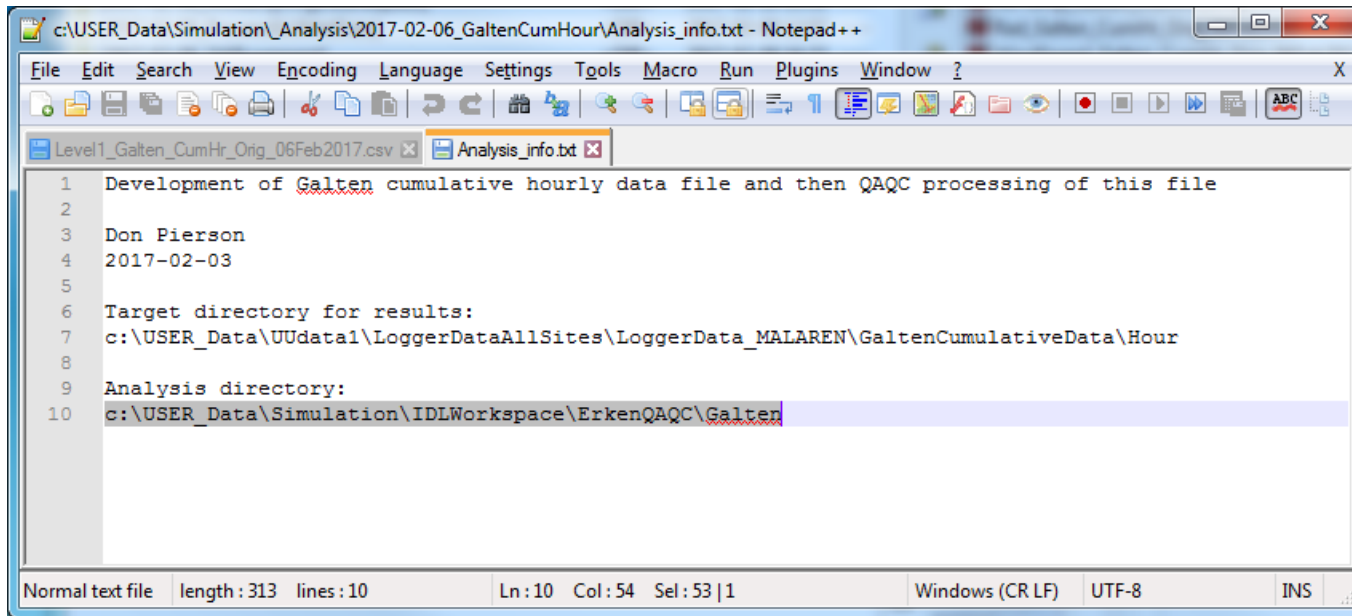
Name	Ext	Size	Date	Attr
[..]	<DIR>		2017-02-06 11:00	----
Galten_CumHr_Orig	csv	13,553,562	2017-02-06 09:04	-a--
Level0_Galten_CumHr_Orig_06Feb2017	csv	16,568,954	2017-02-06 09:51	-a--
Level1_Galten_CumHr_Orig_06Feb2017	csv	16,568,986	2017-02-06 09:51	-a--
AirTemp_Galten_CumHr_Orig_06Feb2017	jpg	1,569,678	2017-02-06 09:51	-a--
Rad_Galten_CumHr_Orig_06Feb2017	jpg	1,087,915	2017-02-06 09:51	-a--
WindSpeed_Galten_CumHr_Orig_06Feb2017	jpg	1,711,388	2017-02-06 09:51	-a--
Wtemp_Galten_CumHr_Orig_06Feb2017	jpg	1,125,007	2017-02-06 09:51	-a--
GaltenHrQC_1	pro	21,506	2017-02-06 09:51	-a--
AirTemp1_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	81,452	2017-02-06 09:51	-a--
AirTemp2_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	1,316	2017-02-06 09:51	-a--
TotalRad_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	734,720	2017-02-06 09:51	-a--
WindDir_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	1,361	2017-02-06 09:51	-a--
WindSpeed1_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	1,314	2017-02-06 09:51	-a--
WindSpeed2_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	1,232	2017-02-06 09:51	-a--
Wtemp1_QC_Loc_Galten_CumHr_Orig_06Feb2017	tab	1,906,673	2017-02-06 09:51	-a--
Analysis_info	txt	313	2017-02-06 10:59	-a--
FileHeaderInfo	xlsx	10,038	2017-02-06 16:52	-a--
Galten_CumHr_Orig	xlsx	18,536,744	2017-02-06 09:03	-a--

**Annotations:**

- Input Files:** Points to the `Galten_CumHr_Orig` file.
- Graphic Output:** Points to the `AirTemp_Galten_CumHr_Orig_06Feb2017` file.
- Computer Program:** Points to the `GaltenHrQC_1` file.
- Output Files:** Points to the `Analysis_info` file.
- Description File:** Points to the `FileHeaderInfo` file.

**Analysis organized by date and Description**

# Documentation in Project Directory

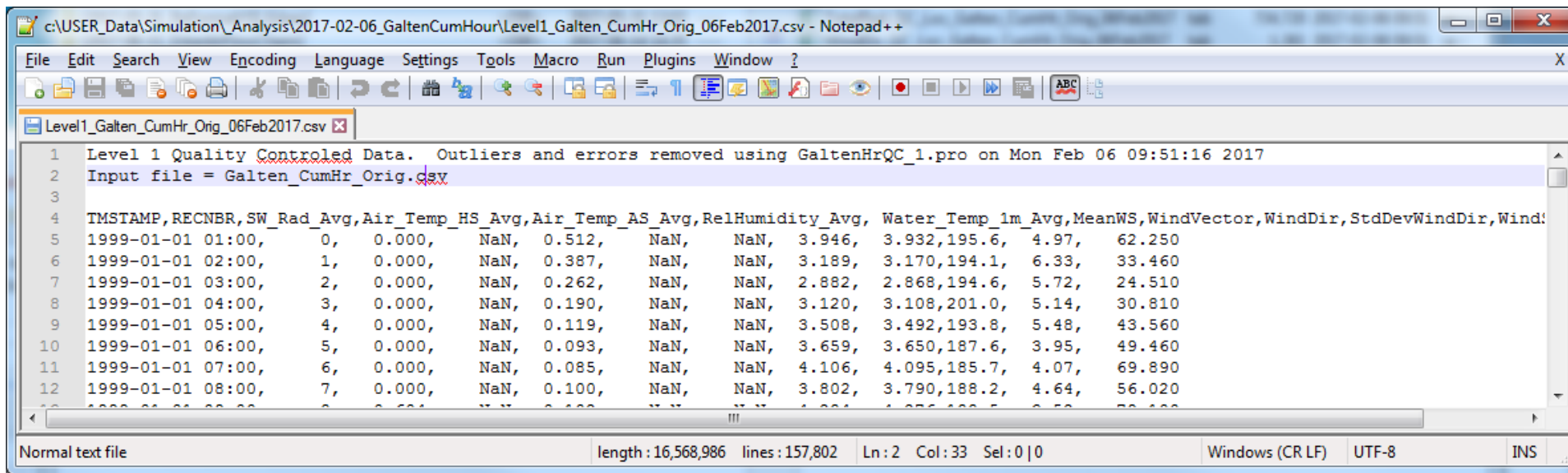


The screenshot shows a Notepad++ window titled "c:\USER\_Data\Simulation\Analysis\2017-02-06\_GaltenCumHour\Analysis\_info.txt - Notepad++". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Tools, Macro, Run, Plugins, and Window. The toolbar contains various icons for file operations. The text area shows the following content:

```
1 Development of Galten cumulative hourly data file and then QAQC processing of this file
2
3 Don Pierson
4 2017-02-03
5
6 Target directory for results:
7 c:\USER_Data\Udata1\LoggerDataAllSites\LoggerData_MALAREN\GaltenCumulativeData\Hour
8
9 Analysis directory:
10 c:\USER_Data\Simulation\IDLWorkspace\ErkenQAQC\Galten
```

The status bar at the bottom indicates "Normal text file", "length: 313 lines: 10", "Ln: 10 Col: 54 Sel: 53 | 1", "Windows (CR LF)", "UTF-8", and "INS".

Descriptive Text File in Directory



The screenshot shows a Notepad++ window titled "c:\USER\_Data\Simulation\Analysis\2017-02-06\_GaltenCumHour\Level1\_Galten\_CumHr\_Orig\_06Feb2017.csv - Notepad++". The menu bar and toolbar are the same as the previous screenshot. The text area shows the following content:

```
1 Level 1 Quality Controlled Data. Outliers and errors removed using GaltenHrQC_1.pro on Mon Feb 06 09:51:16 2017
2 Input file = Galten_CumHr_Orig.csv
3
4 TMSAMP,RECNR,SW_Rad_Avg,Air_Temp_HS_Avg,Air_Temp_AS_Avg,RelHumidity_Avg, Water_Temp_1m_Avg,MeanWS,WindVector,WindDir,StdDevWindDir,Wind!
5 1999-01-01 01:00, 0, 0.000, NaN, 0.512, NaN, NaN, 3.946, 3.932,195.6, 4.97, 62.250
6 1999-01-01 02:00, 1, 0.000, NaN, 0.387, NaN, NaN, 3.189, 3.170,194.1, 6.33, 33.460
7 1999-01-01 03:00, 2, 0.000, NaN, 0.262, NaN, NaN, 2.882, 2.868,194.6, 5.72, 24.510
8 1999-01-01 04:00, 3, 0.000, NaN, 0.190, NaN, NaN, 3.120, 3.108,201.0, 5.14, 30.810
9 1999-01-01 05:00, 4, 0.000, NaN, 0.119, NaN, NaN, 3.508, 3.492,193.8, 5.48, 43.560
10 1999-01-01 06:00, 5, 0.000, NaN, 0.093, NaN, NaN, 3.659, 3.650,187.6, 3.95, 49.460
11 1999-01-01 07:00, 6, 0.000, NaN, 0.085, NaN, NaN, 4.106, 4.095,185.7, 4.07, 69.890
12 1999-01-01 08:00, 7, 0.000, NaN, 0.100, NaN, NaN, 3.802, 3.790,188.2, 4.64, 56.020
```

The status bar at the bottom indicates "Normal text file", "length: 16,568,986 lines: 157,802", "Ln: 2 Col: 33 Sel: 0 | 0", "Windows (CR LF)", "UTF-8", and "INS".

Good header  
practice – self  
documentation

ISO time format