Metadata

Author: Tao Huang (huan1441@purdue.edu), Purdue University

Created: March 20, 2020 (Version 1.0, used in Python 3.6)

Name of the script: program_09.py

Purpose: Script to 1) read a meteorological data file (DataQualityChecking.txt) containing daily precipitation, daily maximum and minimum air temperature, and daily wind speed from 2015-01-01 to 2016-12-31, 2) conduct four basic data quality checking on the data, 3) plot each dataset before and after correction has been made, 4) output the data that has passed the quality check into a new file (Checked_Data.txt), and 5) output a summary of the failed checks to a separate Tab delimited file (Checked_Results.txt).

Description of Data Quality Check:

(1) Remove No Data values
This check is conducted to replace all values of -999, which means a no data value, in this file with the NumPy NaN values. Meanwhile, it records the number of values replaced for each data type in the dataframe “ReplacedValuesDF” with the index “1. No Data”.

(2) Check for Gross Errors
This check is conducted to replace values outside this range with NaN, the error thresholds of which are $0 \leq P \leq 25$ mm, $-25\ ℃ \leq T \leq 35\ ℃$, and $0 \leq WS \leq 10$ m/s. Meanwhile, it records the number of values replaced for each data type in the dataframe “ReplacedValuesDF” with the index “2. Gross Error”.

(3) Check for Relation between Max Temp and Min Temp
This check is conducted to swap the value of Max Temp is less then that of Min Temp for the current day's observations. Meanwhile, it records the number of values replaced for each data type in the dataframe “ReplacedValuesDF” with the index “3. Swapped”.

(4) Check for Daily Temperature Range Exceedance
This check is conducted to identify days with temperature range (Max Temp minus Min Temp) greater than 25°C, and replace both Tmax and Tmin of these days with NaN. Meanwhile, it records of the number of values replaced for each data type in the dataframe “ReplacedValuesDF” with the index “4. Range”.

Summary Table of Failed Checks

The table below presents the number of each type of error in each dataset.

	Precip	Max Temp	Min Temp	Wind Speed
1. No Data	2.0	2.0	2.0	0.0
2. Gross Error	15.0	14.0	2.0	2.0
3. Swapped	0.0	4.0	4.0	0.0
4. Range Fail	0.0	5.0	5.0	0.0

The comparison plot for each dataset before and after the data quality check are shown in the following four figures (.jpeg).

(1) Comparison for Precipitation Dataset

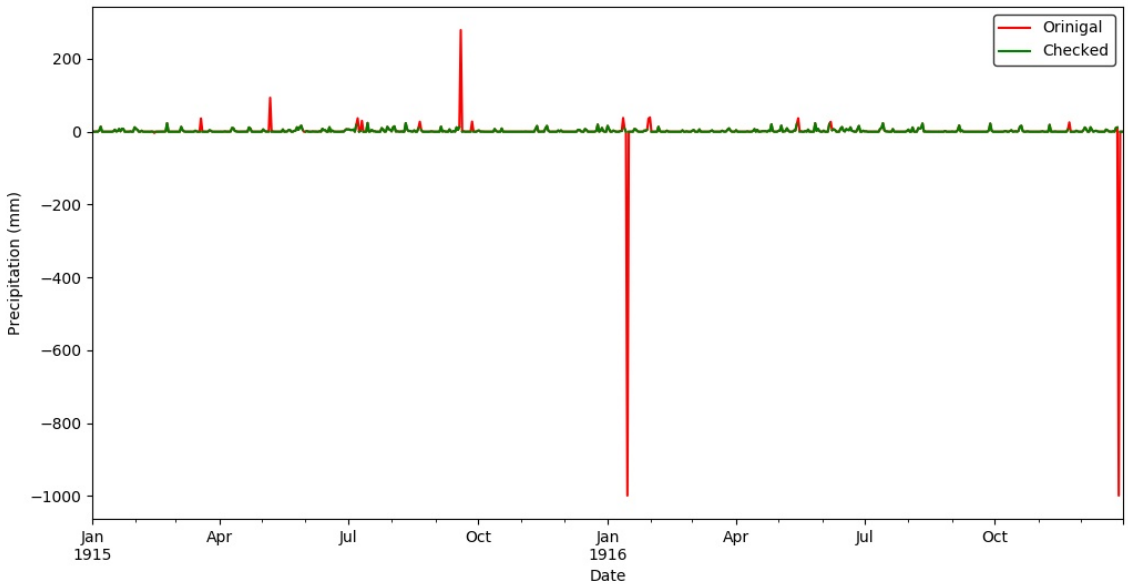

Figure 1. Comparison for precipitation dataset before and after check.

Figure 1 shows the comparison between the original data (red line) and checked data (green line) of the daily precipitation. Most pairs of the data overlap very well. However, two distinct spikes reaching the value of -999 represent the two no data values in this dataset, and the other shorter spikes represent the fifteen gross errors, which are located outside the range, $0 \leq P \leq 25$ mm.

(2) Comparison for Max Air Temperature Dataset

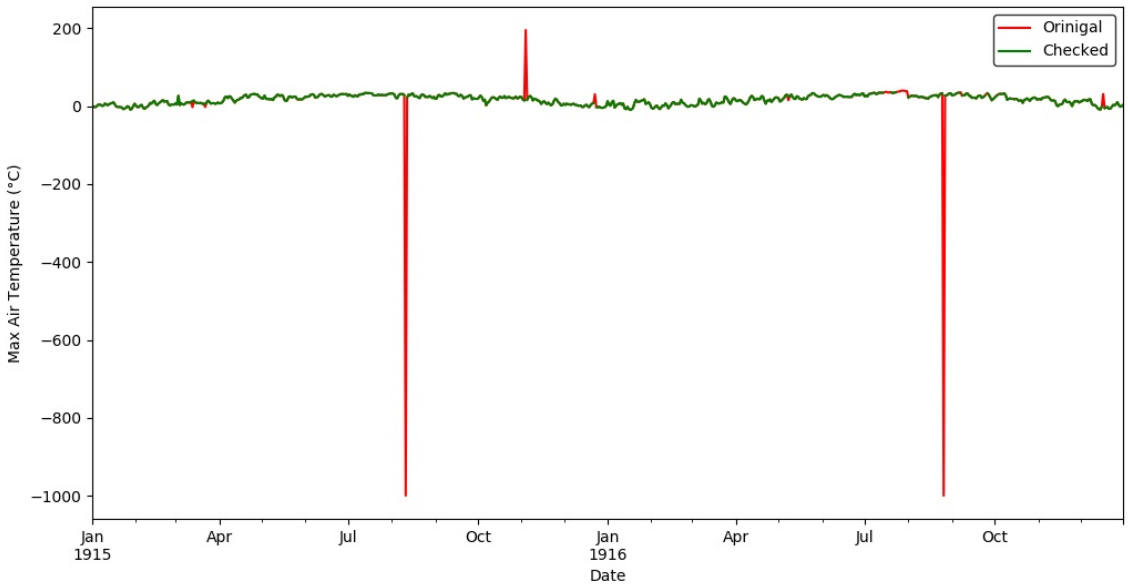

Figure 2. Comparison for max air temperature dataset before and after check.

Figure 2 shows the comparison between the original data (red line) and checked data (green line) of the daily max air temperature. Most pairs of the data overlap very well. However, two distinct spikes reaching the value of -999 represent the two no data values in this dataset, and the other shorter spikes represent the fourteen gross errors (values located outside the range, -25 ℃≤ T ≤ 25 ℃), four swapped errors ($T_{max} < T_{min}$) and five range exceedance errors ($T_{max} - T_{min} > 25$ ℃).

(3) Comparison for Min Air Temperature Dataset

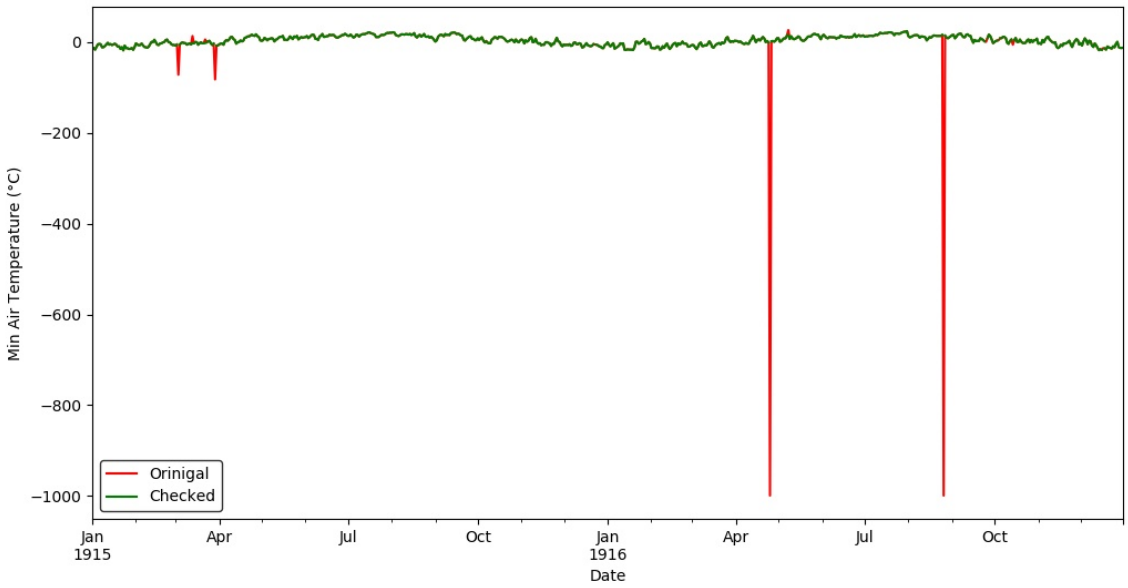

Figure 3. Comparison for min air temperature dataset before and after check.

Figure 3 shows the comparison between the original data (red line) and checked data (green line) of the daily min air temperature. Most pairs of the data overlap very well. However, two distinct spikes reaching the value of -999 represent the two no data values in this dataset, and the other shorter spikes represent the two gross errors (values located outside the range, -25 ℃≤ T ≤ 25 ℃), four swapped errors ($T_{max} < T_{min}$) and five range exceedance errors ($T_{max} - T_{min} > 25$ ℃).

(4) Comparison for Wind Speed Dataset

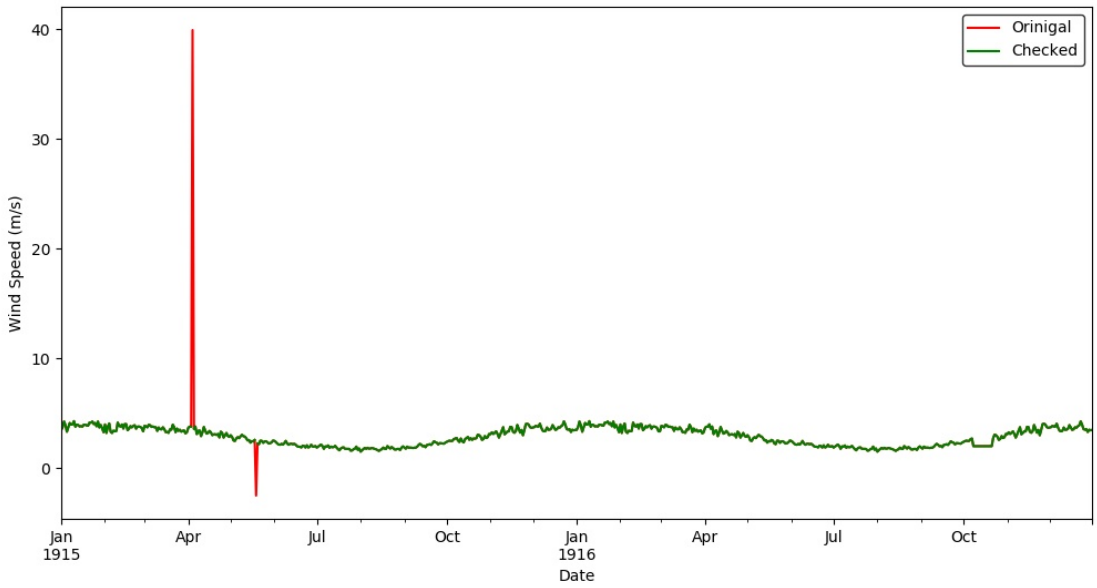

Figure 4. Comparison for wind speed dataset before and after check.

Figure 4 shows the comparison between the original data (red line) and checked data (green line) of the daily wind speed. Most pairs of the data overlap very well. However, two distinct spikes reaching the value of 39.9 and -2.5 represent the two gross errors (values located outside the range, $0 \leq WS \leq 10$ m/s), and it is the only type of error exists in this dataset.