Presentation of an Algorithm for the Automated Estimation and Gap-Filling of Missing Data in Daily Weather Records

J.S. Gosselin[a,*], R. Martel[a], C. Rivard[b]

[a]*Institut national de la recherche scientifique, Centre Eau Terre Environnement, 490 rue de la Couronne, Quebec City, Quebec, Canada*
[b]*Geological Survey of Canada, Quebec Division, 490 rue de la Couronne, Quebec City, Quebec, Canada*

Abstract

This paper presents a free and open-source algorithm, written in the Python programming language, that was developed to fill the gaps in daily weather datasets with an automated, robust, and efficient method. The missing data for a given weather station are estimated using a multiple linear regression model, generated using synchronous data from neighboring stations. The algorithm was tested for a network of 19 weather stations located in the Monteregie Est region, Quebec, Canada. The uncertainty of the estimates were assessed with a cross-validation procedure, which is included as part of the algorithm. The method gave consistent results for the daily mean, max, and min air temperature and daily total precipitation for all of the weather stations tested. Uncertainty of the results compared well with other studies that also used a similar approach. In addition, the algorithm can also be used with a Graphical User Interface (GUI) that is part of the free and Open Source software WHAT (Well Hydrograph Analysis Toolbox).

Keywords: `Weather Stations, Precipitation, Air Temperature, Daily Missing Data`

1. Introduction

Daily weather data are useful in several fields of Earth sciences, including hydrology, hydrogeology and agronomy. However, weather datasets are, most of the time, incomplete. This can represent a major hindrance in various applications, such as for the use of hydrological or hydrogeological simulators (e.g. SWAT, HELP, SHAW), which heavily depend on these data. Filling the gaps in weather datasets can quickly become a tedious task as the length of the data records and the number of stations increase. Moreover, it can also be quite complex when aspects such as time-efficiency of the method and accuracy of the estimated missing values are taken into account. This is particularly true for the estimation of missing daily precipitation data because of their high spatial and temporal variability (Simolo et al., 2010). Although various methods to estimate missing daily weather data are documented in technical papers (i.e. DeGaetano et al., 1995; Simolo et al., 2010), few published tools to perform this task efficiently and automatically are available.

*Corresponding authors
Email address: `jnsebgosselin@gmail.com` (J.S. Gosselin)

This paper addresses this issue by presenting an open source algorithm, written in the Python programming language, that was developed to fill the gaps in daily weather datasets with an automated, robust, and efficient method. The missing weather data in the records of a given weather station (hereafter called the target station) are estimated through a Multiple Linear Regression (MLR) model using synchronous measurements from neighboring stations. The algorithm also includes an option to asses the validity of the method and the uncertainty of the estimated missing values through a cross-validation resampling technique.

In addition, a Graphical User Interface (GUI) for the algorithm has been developed and is included in the WHAT software (Gosselin, 2015). The algorithm and the WHAT software are both available for free at this web address: `https://github.com/jnsebgosselin/WHAT`. An example of an application of the algorithm, using the GUI provided in WHAT, is also presented at the end of this paper for the Monteregie Est study area (southern Quebec, Canada).

2. Description of the Algorithm

The algorithm described in this paper is based on the classical MLR (Multiple Linear Regression) method presented in Eischeid et al. (2000). The MLR method is a robust and well known spatial interpolation technique that can indirectly account for local effects, such as topography, land cover, land use and surface water. While creating serially complete daily datasets of air temperature and total precipitation for the western U.S., Eischeid et al. (2000) found that the MLR method consistently outperformed the other classical methods tested (normal ratio, inverse distance, optimal interpolation, and single best estimator). The same result was also found by Xia et al. (1999) for a study in Bavaria, Germany. Moreover, in a study conducted in Iran for different climate conditions (dry to extra humid conditions), Kashani and Dinpashoh (2011) found that the estimation obtained with the MLR method compared well with those obtained with more recent methods, more specifically the artificial neural network (reference) and the genetic programming (references) techniques.

A flowchart of our gap-filling algorithm is shown in Fig. 1. The algorithm consists of two nested loops: the external 'Loop A' iterates over the weather variables of the dataset (min, max, and mean air temperature and total precipitation), while the inner 'Loop B' iterates over the missing values in each weather data series. Each missing value is estimated independently with a two-step procedure in 'Loop B'. The first step consists in the selection of the neighboring stations. The second step consists in building a MLR model, estimating the missing value, and filling the corresponding gap in the data series. The gap-filling algorithm of Fig. 1 is described in more details in the following sections of this paper.

2.1. Correlation Coefficients Calculations

Correlation coefficients are calculated between the available data of the target station and those of the neighboring stations for each weather variable individually. The coefficients are calculated using all the available data. However, neighboring stations that have less than 182 days (half a year) of synchronous data with the target station or that have

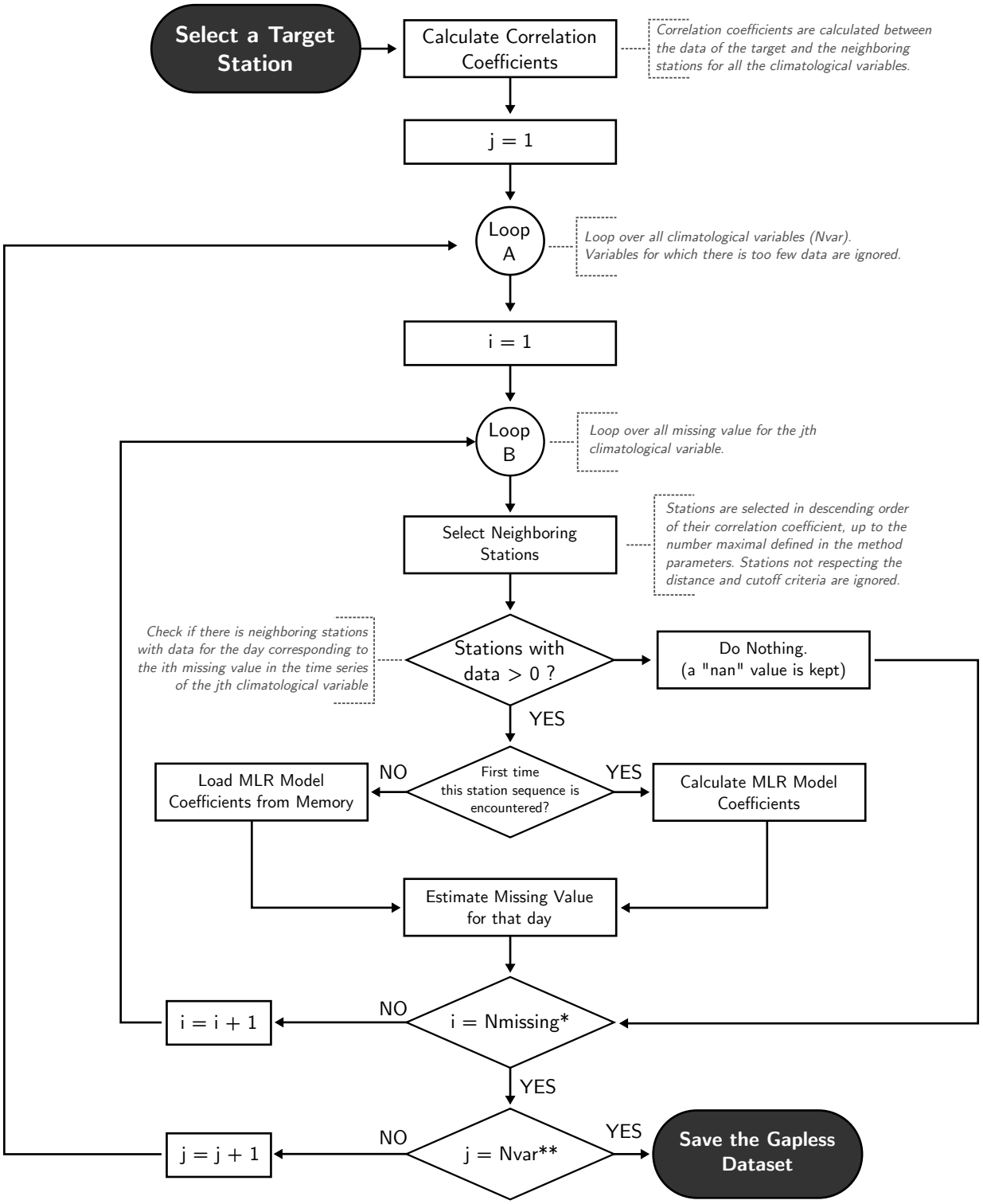


Figure 1: Flowchart of the gap-filling algorithm presented in this paper.

a correlation coefficient below a value of 0.35, for a given weather variable, are not used to fill the gaps in the data for that weather variable. The 0.35 threshold is based on the value used by Eischeid et al. (2000) in their application of the method.

Moreover, it is possible to discard completely from the gap-filling procedure neighboring stations that are located farther away from the target station than specified threshold values, either in the horizontal or the vertical direction. The default values are set to 100 km and 350 m for the horizontal and vertical distance respectively, based on the values found in the literature (Tronci et al., 1986; Xia et al., 1999; Simolo et al., 2010).

2.2. Selection of the Neighboring Stations

As stated by Eischeid et al. (2000), the selection of neighboring stations is critically important for the accurate estimation of missing weather data. Problems arise though because the list of neighboring stations with available data can vary from one day to the other. Therefore, the process of selecting the neighboring stations for the generation of a MLR model must be done for each missing value in the dataset of the target station separately.

For a given day with a missing value, the neighboring stations with available data on that day are selected in descending order of their correlation coefficient, up to a maximal number of stations that can be specified as a parameter in the algorithm. The default value for the maximal number of neighboring stations used for the generation of the MLR models is 4. Tests run by Eischeid et al. (2000) showed that using more then 4 neighboring stations did not significantly improve, and may even have degraded, the accuracy of the estimates. If for a given day with a missing value, no neighboring stations have a measured value, no calculation is done and a 'NaN' value is kept in the dataset.

2.3. Generation of the Multiple Linear Regression Model

Each time a MLR model is generated for a given sequence of neighboring stations, the result is stored into memory. Therefore, after a set of neighboring stations have been selected for a given day where a data is missing (Section 2.2), the program checks first if this sequence of stations has already been encountered before for the current weather variable. If so, the stored MLR parameters will be used directly to estimate the missing data for the current day. Otherwise, a new MLR model will be generated from the newly encountered set of neighboring stations to estimate the missing data. Since a MLR model is generated only once for a given sequence of neighboring stations, the algorithm becomes faster with time as the number of MLR models stored into memory increases.

The MLR models can be generated using either an Ordinary Least Square (OLS) or a Least Absolute Deviations (LAD) criteria. Since daily precipitation series are generally characterized by long-tailed and positively skewed distributions, the LAD method is more appropriate than the OLS method since it is more robust to outliers (Menke, 1989; Eischeid et al., 2000). The downside in using the LAD method is an increase in computation time by about a factor 10 compared to the OLS method. The resolution of the MLR models with the LAD method is achieved using an iterative reweighted least-squares method (reference).

2.4. Estimating Missing Daily Values

The value of the weather data for the target station is estimated from the synchronous measurements of the neighboring stations using the MLR model, such as:

$$Y_t = a_0 + \sum_{k=1}^{N} a_k \cdot X_k(t_i) \tag{1}$$

where $Y(t)$ is the value estimated for the target station at time t, $X_k(t)$ is the synchronous measured data for the k^{th} neighboring stations, a_k are the regression coefficients of the MLR model, and N is the total number of neighboring stations that were used for the regression. The intercept term, a_0, is estimated for the air temperature, but is set to zero for precipitation.

It is possible to have a negative regression coefficients for the less correlated neighboring stations used to generate the MLR model. If so, the MLR model will sometimes yield small negative values for daily precipitation. To correct that, negative values estimated for daily precipitation are always set to zero.

2.5. Uncertainty of the estimated values

The accuracy of each MLR model generated throughout the gap-filling procedure is estimated by computing the root mean of squared residuals of the regression. It is also possible to evaluate the accuracy of the whole method (instead of each MLR model individually) for the entire dataset of given weather station with a Leave One Out (LOO) cross-validation procedure. The procedure consists in estimating a value for each day of the data series of each weather variable for which a measured data is available in the dataset (in addition to the days with missing data). In other words, the loops A and B in the flowchart of Fig. 1 iterates over all the days and all the weather variables of the dataset instead of only iterating over days with a missing data. Before estimating a value for a given day, the corresponding measured data is temporarily discarded from the dataset of the target station to avoid self-influence of this observation on the generation of the MLR model. The accuracy of the method is then estimated by computing the Root-Mean-Square Error (RMSE), the Mean-Absolute Error (MAE), the Mean Error (ME) and the correlation coefficient (r) between the estimated weather data and the respective non-missing observations in the original dataset of the target station.

Since a new MLR model must be generated for each day independently when doing the cross-validation procedure, the computation time of the gap-filling procedure is thus significantly increased. This is especially true if the MLR model is generated with the least absolute deviation regression method. For this reason, the cross-validation procedure is by default not activated in the algorithm.

3. Operation of the Algorithm

It is possible to use the algorithm with the Graphical User Interface (GUI) that is included in the free and open source software WHAT (Well Hydrograph Analysis Toolbox). A detailed description on the use of the algorithm with

WHAT is provided in the user guide of the software (Gosselin, 2015).

Alternately, the gap-filling algorithm can be run directly from the command line in a Python interpreter version 2.7 or 3.4 or later. This section will cover this approach. The external libraries *NumPy*, *Matplotlib*, *xlrd*, *PySide* and *Statsmodels* are required for the program to run. A minimal working example of an application is documented at the end of the python file. Some data samples to run the example are also provided with the algorithm.

The present section of this paper covers the format of the input data that is required for running the algorithm, the parameters of the method, and the various outputs that are generated after a gap-less weather dataset has been successfully produced with the algorithm. Additional information about the input and output of the gap-filling algorithm is also provided in the user guide of the WHAT software.

3.1. Input Data

It is possible to use weather data from any sources with the gap-filling algorithm, as long as the data are saved in tab-separated values file with the '.csv' extension. Also, the labels in the first column of the file must be faithfully observed, since the algorithm is reading these to know where to retrieve the station information and the weather data within the file. It is recommended to use a copy of one of the sample files that are provided with the algorithm and fill-in directly the station information and the weather data. A "NaN" value must be entered where data are missing. The daily data must also be in chronological order, but do not need to be continuous over time. That is, missing blocks of data (e.g., several days, months or years) can be completely omitted from the time-series.

All the input weather data files must be saved in one single location that must be specified to the gap-filling algorithm. The algorithm will automatically scan this location for valid weather data files and will store the data in memory for future analysis.

3.2. Parameters

The gap-filling algorithm is written as a Python class object, with the method parameters defined as class attributes. When using the gap-filling algorithm directly in a Python interpreter (without the GUI), the method parameters are specified by directly defining the value of their corresponding class attribute. An example is given at the end of the python file and each parameter is documented in the help section of the algorithm class, within the code. Additional information is also provided in the user guide of the WHAT software. A list of the different method parameters for the current version of the algorithm is presented in Table A.3 in Appendix A.

3.3. Output

All the outputs that are produced after a gapless weather dataset has been produced successfully are saved in a sub-folder that is named after the name of the target station, in a directory that must be specified to the gap-filling algorithm. A list of the different output files that are produced with the current version of the algorithm is presented in Table A.4 in Appendix A a brief overview is given below.

The gap-less weather datasets are saved in a tab-separated values file with a '.out' extension. Detailed information about the estimated values that were used to fill the gaps in the data series are also saved in an accompanying '.log' file. An histogram showing the yearly and monthly weather normals, calculated from the gap-less data series previously generated with the algorithm, is also produced and saved in a pdf format.

The results from the cross-validation procedure, if the option is enabled in the algorithm, are saved in a tab-separated values file with a '.err' extension. A figure comparing the probability density function of the original and the estimated daily precipitation series is also produced and saved in a pdf format. Scatter plots comparing the estimated and measured daily weather data are also produced for each variable of the dataset and saved in a pdf format. Example of the figure produced with the gap-filling algorithm are shown in Section 4 where a study case is presented.

4. Application: Monteregie Est Case Study

4.1. Materials and Method

4.1.1. Study Area

The algorithm was tested using data from 32 land-based Canadian weather stations in and around the Monteregie Est region, located in southern Quebec, Canada. This region covers a total area of more than 9000 km^2, from the St. Lawrence River at its northern limit to the border of the United States (states of New York and Vermont) at its southern limit (see Fig. 2). It is characterized by strongly variable topography and land cover conditions, as well as warm summers and cold winters (Carrier et al., 2013).

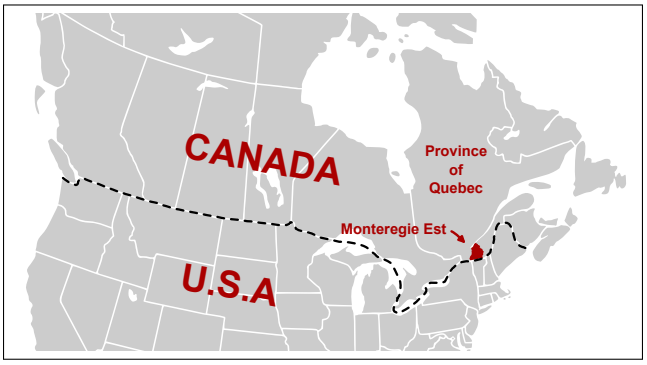


Figure 2: Location of the Monteregie Est area in North America.

4.1.2. Weather Stations

A total of 32 weather stations were selected from the Canadian Daily Climate Database (CDCD) based on the availability and continuity of the measured weather data between 1980 and 2009 inclusively. The data were downloaded and formatted using the software WHAT (Gosselin et al., 2015). WHAT provides a graphical interface to the online CDCD that allows to search for stations interactively using location coordinates, download the available data

for the selected weather stations, and automatically organize the data in a format that is compatible with the gap-filling algorithm presented in this paper.

Table 1 presents the list of the stations used in this study with their corresponding location coordinates (latitude and longitude), altitude, percentage of days with missing data, and yearly averages for each weather variable. Most of the information presented in Table 1 are generated automatically when loading data into the gap-filling routine and are saved in a file named 'weather_datasets_summary.log'. The geographical disposition of the weather stations is also presented in the map of Fig. 3.

On average for all the stations, the mean annual total precipitation is 1100 mm/y. The highest total annual precipitation are observed at the *Sutton* station (~1300.4 mm/y), while the lowest at the *Nicolet* station (~924.3 mm/year). Mean annual air temperature in the study area is 5.9 °C, ranging from 4.6 to 7.1 °C. The highest temperatures are observed at the *Philipsburg* station and the lowest at *Bonsecours* station.

Graphs showing the yearly and monthly averages for total precipitation and max, min, and mean air temperature for the *Sutton*, *Nicolet*, *Philipsburg*, and *Bonsecours* weather stations are presented in Fig. 4. From these graphs, it can be seen that the climate of the region is characterized by significant seasonal differences in temperature, resulting in warm summers and cold winters. The minimum monthly temperatures are observed in January while the maximum monthly temperatures are observed in July. Total precipitation, as rain or snow, are distributed rather evenly throughout the year. Precipitation as rain also occurs frequently in the winter season due to mild spells.

4.1.3. Validation With Cross-Validation Procedure

In order to validate the procedure for the Monteregie Est region and assess the uncertainty of the estimates, the cross-validation procedure described in Section 2.5 was run for the 19 weather stations that are located within the study area (stations that are shown in red in Table 1). Data from the stations bordering the limits of the study area, but outside of it, were used to improve the spatial distribution of the weather data.

The method was tested with the parameter values that are set by default in the algorithm, as shown in Table A.3 in Appendix A. That is, the maximum number of neighboring stations was set to 4, the horizontal and vertical distance thresholds were kept at 100 km and 350 km values respectively. However, the OLS method was chosen for the regression instead of the LAD to save in computation time.

4.2. Results and Discussion

Table 2 presents the results of the cross-validation procedure. The Root-Mean-Square Error (RMSE), the Mean-Absolute Error (MAE), the Mean Error (ME) and the correlation coefficient (r), are given for each of the 19 weather stations, and each of the four weather variables tested (T_{max}, T_{min}, T_{mean}, and P_{tot}). The mean, max, and min values for each estimator (RMSE, MAE, ME, and r) are also provided at the bottom of the table.

Table 1: List of selected weather stations in and around the study area and related information about location coordinate, altitude, missing data and yearly averages for the 1980-2009 period. The 19 stations that are located within the study area are shown in red while the bordering stations that are outside of it are in black.

					% of days with missing data				Yearly Averages			
#	Station name	Lat. °N	Lon. °W	Alt. m	T_{max} %	T_{min} %	T_{mean} %	P_{tot} %	T_{max} °C	T_{min} °C	T_{mean} °C	P_{tot} mm
1	Auteuil	45.65	73.73	53.0	20.8	19.1	22.8	18.4	11.3	1.4	6.4	989.7
2	Bonsecours	45.40	72.27	297.2	4.7	5.1	6.3	3.2	10.0	−0.8	4.6	1226.2
3	Brome	45.18	72.57	205.7	2.6	2.3	3.0	2.3	11.0	−0.4	5.3	1296.7
4	Bromptonville	45.48	71.95	130.0	3.9	4.2	6.0	1.8	11.1	−0.1	5.5	1137.9
5	Danville	45.82	71.98	190.0	30.8	31.1	33.1	30.2	10.6	0.4	5.5	1074.5
6	Drummondville	45.88	72.48	82.3	2.5	2.4	3.4	1.8	11.0	1.5	6.2	1122.1
7	Farnham	45.30	72.90	68.0	5.2	4.7	6.1	3.7	11.5	1.1	6.3	1131.9
8	Fleury	45.80	73.00	30.5	1.1	1.2	1.6	1.1	10.8	0.8	5.8	1097.3
9	Georgeville	45.13	72.23	266.7	26.5	26.2	27.6	5.7	10.7	−0.3	5.2	1253.3
10	Granby	45.38	72.72	175.0	1.3	1.3	2.1	0.5	10.7	1.6	6.2	1219.2
11	Hemmingford	45.07	73.72	61.0	6.0	6.0	6.8	4.6	11.9	1.2	6.5	945.7
12	Iberville	45.33	73.25	30.5	7.3	7.5	9.7	4.2	11.5	1.6	6.6	1098.4
13	Magog	45.27	72.12	274.0	4.2	4.2	5.3	4.6	10.3	0.7	5.5	1145.7
14	Marieville	45.40	73.13	38.0	10.2	10.4	11.2	9.8	11.5	1.4	6.4	1102.4
15	Nicolet	46.20	72.62	30.4	4.1	4.2	5.2	3.6	10.3	0.1	5.2	924.3
16	Philipsburg	45.03	73.08	53.3	4.9	5.1	6.9	3.2	11.9	2.2	7.1	1066.1
17	Pierreville	46.08	72.83	15.2	6.3	5.4	6.9	4.9	10.7	0.8	5.8	979.4
18	Richmond	45.63	72.13	123.1	3.3	3.3	3.8	4.0	10.9	0.1	5.5	1166.4
19	Riviere des Prairies	45.70	73.50	9.0	2.4	4.0	4.8	1.4	11.5	1.3	6.4	986.6
20	Sabrevois	45.22	73.20	38.1	25.5	26.0	27.1	5.2	11.6	1.4	6.5	1020.5
21	Sorel	46.03	73.12	14.6	5.7	5.9	6.2	4.5	11.1	1.2	6.2	999.5
22	St. Amable	45.67	73.30	41.1	8.6	10.4	11.8	7.7	11.4	1.0	6.2	1007.0
23	St. Bernard	45.08	73.38	49.3	9.6	9.7	10.5	9.0	11.7	1.7	6.7	979.6
24	St. Guillaume	45.88	72.77	43.9	4.3	4.6	5.7	3.0	11.0	0.4	5.7	1022.6
25	St.Hyacinthe 2	45.57	72.92	33.0	6.7	6.9	7.5	6.6	11.3	1.3	6.3	1060.8
26	St. Jacques	45.95	73.58	69.0	11.7	11.9	14.0	10.8	11.0	−0.1	5.5	1010.5
27	St. Janvier	45.73	73.88	61.0	40.5	40.7	41.6	21.9	10.9	−0.1	5.4	1036.3
28	St. Nazaire	45.73	72.62	68.6	3.8	3.8	5.4	2.7	11.0	0.4	5.7	1087.8
29	Ste. Madeleine	45.62	73.13	30.0	5.8	6.4	7.0	5.1	11.4	1.1	6.3	1035.1
30	Ste. Martine	45.22	73.85	38.1	6.9	6.6	7.8	6.4	11.6	1.8	6.7	985.0
31	Sutton	45.07	72.68	243.8	0.4	0.6	0.7	0.5	11.1	1.0	6.1	1300.4
32	Vercheres	45.77	73.37	21.0	3.5	3.5	4.7	2.2	11.3	1.8	6.5	973.8

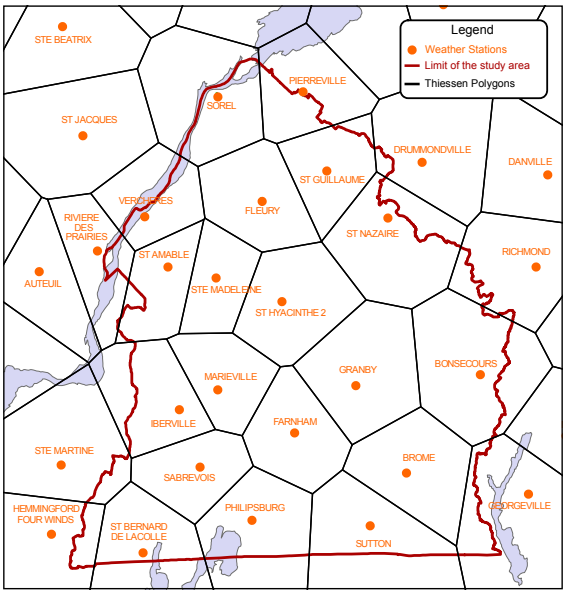
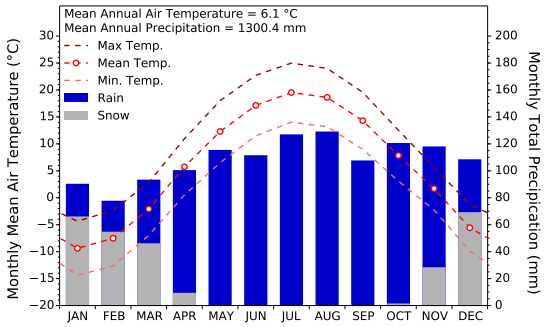
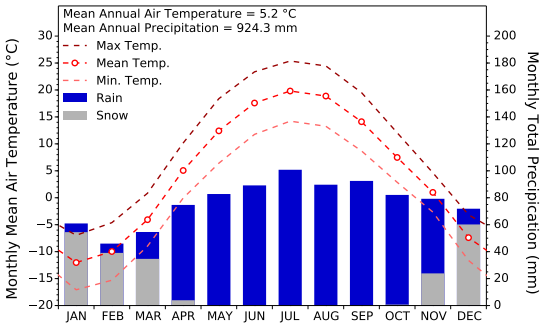


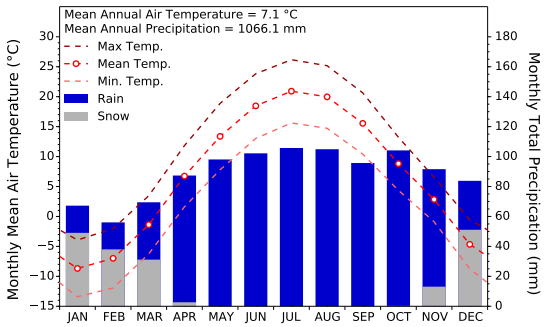
Figure 3: Spatial distribution of the weather stations in and around the Monteregie Est area, Quebec, Canada.



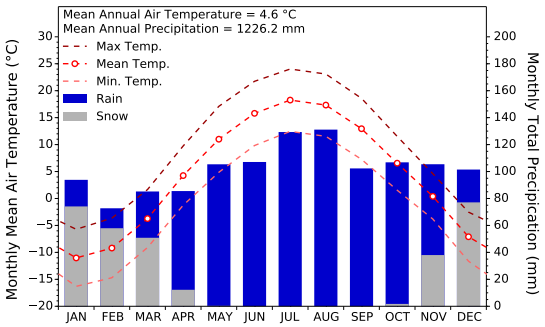
(a) Sutton Weather Station



(b) Nicolet Weather Station



(c) Philipsburg Weather Station



(d) Bonsecours Weather Station

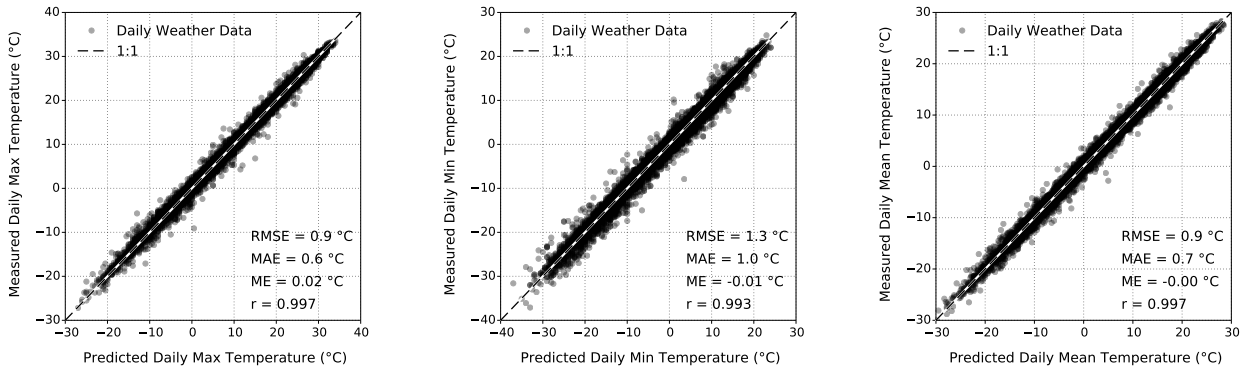
Figure 4: Yearly and monthly weather normals for the weather stations with the highest (Sutton) and lowest (Nicolet) annual total precipitation and the warmer (Philipsburg) and colder (Bonsecours) air temperature.

Table 2: Results of the cross-validation procedure for the 19 weather stations located within the Monteregie Est study area for the 1980-2009 period. The Root-Mean-Square Error (RMSE), the Mean-Absolute Error (MAE), the Mean Error (ME) and the correlation coefficient (r), are given for each weather variables. The mean, max, and min values for all the station, for each estimator (RMSE, MAE, ME, and r), are also provided at the bottom of the table.

	T_{max}				T_{min}				T_{mean}				P_{tot}			
Station	RMSE	MAE	ME	r	RMSE	MAE	ME	r	RMSE	MAE	ME	r	RMSE	MAE	ME	r
name	°C	°C	°C	-	°C	°C	°C	-	°C	°C	°C	-	mm	mm	mm	-
Bonsecours	1.2	0.8	0.00	0.995	1.7	1.3	0.15	0.989	1.1	0.8	0.02	0.996	3.1	1.4	−0.10	0.890
Brome	1.1	0.7	−0.07	0.996	1.8	1.3	−0.11	0.989	1.1	0.8	−0.08	0.996	3.0	1.3	−0.17	0.902
Farnham	1.0	0.7	−0.02	0.997	1.3	1.0	−0.03	0.994	0.9	0.6	−0.02	0.997	2.7	1.1	−0.09	0.912
Fleury	0.9	0.6	0.00	0.997	1.2	0.9	0.00	0.995	0.8	0.6	0.00	0.998	2.5	1.0	−0.04	0.928
Granby	0.9	0.6	0.02	0.997	1.3	1.0	−0.01	0.993	0.9	0.7	0.00	0.997	2.8	1.2	−0.03	0.912
Iberville	0.9	0.6	0.00	0.997	1.2	0.9	0.06	0.994	0.8	0.6	0.03	0.998	2.6	1.1	−0.12	0.915
Marieville	0.9	0.6	0.00	0.997	1.2	0.8	0.03	0.995	0.8	0.5	0.01	0.998	2.8	1.1	0.00	0.908
Philipsburg	1.2	0.8	0.02	0.995	1.4	1.0	0.00	0.993	0.9	0.7	0.01	0.997	2.8	1.1	0.00	0.908
Pierreville	1.0	0.7	0.00	0.997	1.3	1.0	−0.02	0.994	0.9	0.7	−0.01	0.997	2.6	1.0	−0.01	0.903
Sabrevois	1.1	0.7	0.00	0.996	1.2	0.8	0.01	0.995	0.8	0.6	0.01	0.998	3.1	1.2	−0.2	0.875
Sorel	1.2	0.8	0.01	0.996	2.0	1.4	0.13	0.987	1.1	0.8	0.07	0.996	2.9	1.1	−0.06	0.886
St. Amable	1.5	1.0	−0.01	0.993	1.8	1.3	0.02	0.988	1.2	0.9	0.00	0.995	2.9	1.2	−0.04	0.890
St. Bernard	1.4	1.0	−0.03	0.993	1.6	1.1	−0.06	0.990	1.1	0.8	−0.04	0.995	3.4	1.3	0.03	0.847
St. Guillaume	1.0	0.7	0.00	0.997	1.2	0.8	−0.01	0.995	0.8	0.6	0.00	0.998	2.3	1.0	−0.01	0.927
St.Hyacinthe 2	0.9	0.6	−0.01	0.997	1.2	0.9	−0.01	0.995	0.8	0.6	−0.02	0.998	2.5	1.0	−0.06	0.921
St. Nazaire	1.0	0.7	0.01	0.997	1.4	1.0	0.02	0.993	0.9	0.6	0.01	0.997	2.6	1.1	−0.08	0.911
Ste. Madeleine	1.0	0.7	0.02	0.996	1.2	0.8	−0.01	0.995	0.8	0.6	0.00	0.998	2.6	1.0	0.01	0.919
Sutton	0.9	0.6	0.00	0.997	1.2	0.9	0.00	0.994	0.8	0.6	0.00	0.998	2.9	1.3	−0.07	0.912
Vercheres	1.1	0.7	0.03	0.996	1.3	1.0	0.05	0.993	0.9	0.6	0.04	0.997	2.6	1.1	−0.10	0.895
Mean	1.1	0.7	0.00	0.996	1.4	1.0	0.01	0.993	0.9	0.7	0.00	0.997	2.8	1.1	−0.06	0.901
Max	1.5	1.0	0.03	0.997	2.0	1.4	0.15	0.995	1.2	0.9	0.07	0.998	3.4	1.4	0.03	0.928
Min	0.9	0.6	−0.07	0.993	1.2	0.8	−0.11	0.987	0.8	0.5	−0.08	0.995	2.3	1.0	−0.17	0.847

4.2.1. Air Temperature

The method gave consistent results for each of the three temperature-related weather variables, for all the 19 weather stations. The RMSE and MAE are both below 2.0 and 1.4 °C for max ,min, and mean daily air temperature. There is also no bias in the estimations for any of the temperature-based variable with a ME that is, on average for all the stations, less than 0.01 °C for the max, min, and mean temperature time series. The correlation coefficient between the estimated and measured time series is also above 0.987 for all the weather stations and all the temperature-based variables. The goodness of fit between the estimated and observed values for the max, min, and mean daily temperature are presented for the weather station Granby, located in the center of the study area, in the three graphs of Fig. 5. These graphs are generated automatically at the end of the gap-filling procedure when the cross-validation option is set to *True*.



(a) (b) (c)

Figure 5: Scatter plots comparing the predicted versus the observed daily values for (a) max air temperature, (b) min air temperature, and (c) mean air temperature for the *Granby* weather station.

4.2.2. Total Precipitation

Unlike air temperature, daily precipitation are characterized by high spatial and temporal variability and is thus a weather variable that is more difficult to estimate accurately. Nevertheless, the method gave consistent results for the Monteregie Est region, with a RMSE and MAE that are less than 3.4 and 1.4 mm for all the stations. These values are comparable to the results of Xia et al. (1999) and Eischeid et al. (2000) who also used the MLR method to estimate daily precipitation values. The correlation between the estimated and observed daily precipitation is good with values above 0.847 for all the stations. The goodness of fit between the estimated and observed values for the daily total precipitation is presented for the weather station Granby in the graph of Fig. 6

Moreover, from the results of Table 2, it can be seen that the differences between the RMSE and MAE are larger for precipitation than for temperature. Moreover, the ME are slightly negatives for 15 of the 19 stations tested. This is due to the fact that regression-based techniques, like the MLR method used in this paper, tend to systematically underestimate heavy precipitation events. This is well demonstrated in Fig. 7, where are shown gamma probability density

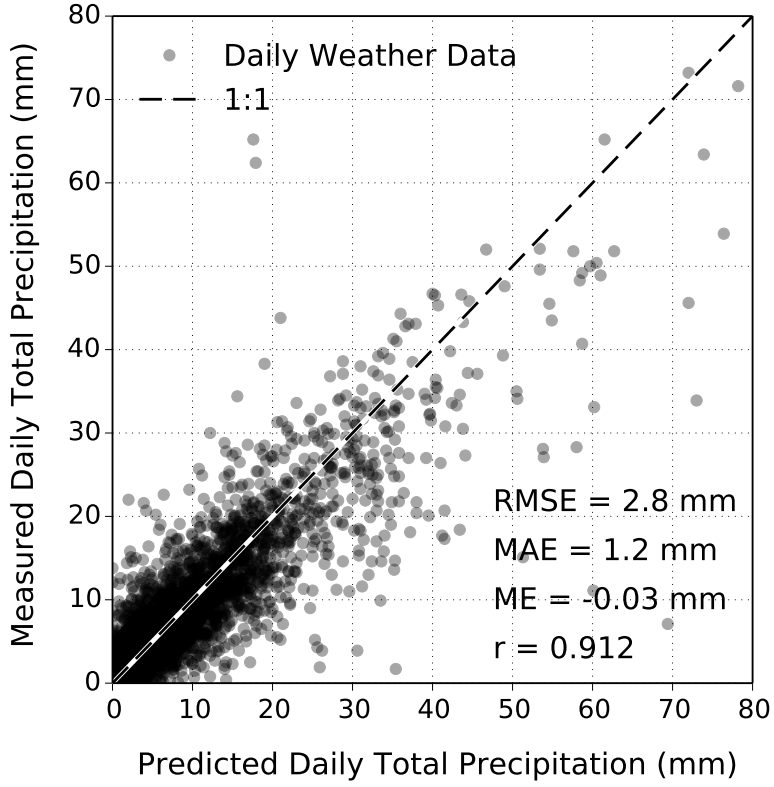


Figure 6: Scatter plots comparing the predicted versus the observed daily total precipitation for the *Granby* weather station

functions that were estimated from the estimated (dashed red line) and observed (solid blue line) daily precipitation time series. The histogram of the distribution of the observed daily precipitation events is also shown on the same figure in light blue. As can be seen, the occurrence of heavy precipitation events is reduced for the estimated time series compared to the measured one. Furthermore, it can also be seen on the graph of Fig. 7 that the occurrence of zero and light precipitation events is overestimated by the method. This resulted in an overestimation of the number of wet days by 30 % on average for all the weather stations tested, with a maximal value of 53 % for the *Vercheres* stations and a minimum value of 13 % for the *Sutton* station.

The rainfall probability distribution is thus not preserved when using the MLR method. This is a known issue that has been discussed by Simolo et al. (2010). They also propose a two-step procedure that modifies the MLR method to address both the overestimation of the number of wet days and the underestimation of the heavy precipitation events. Their approach may be incorporated in a future version of the gap-filling algorithm presented in this paper.

5. Conclusion

This article presented the main capabilities of an open source algorithm, written in the Python programming language, for filling the gaps in daily weather data with an automated, robust and efficient method. The method has

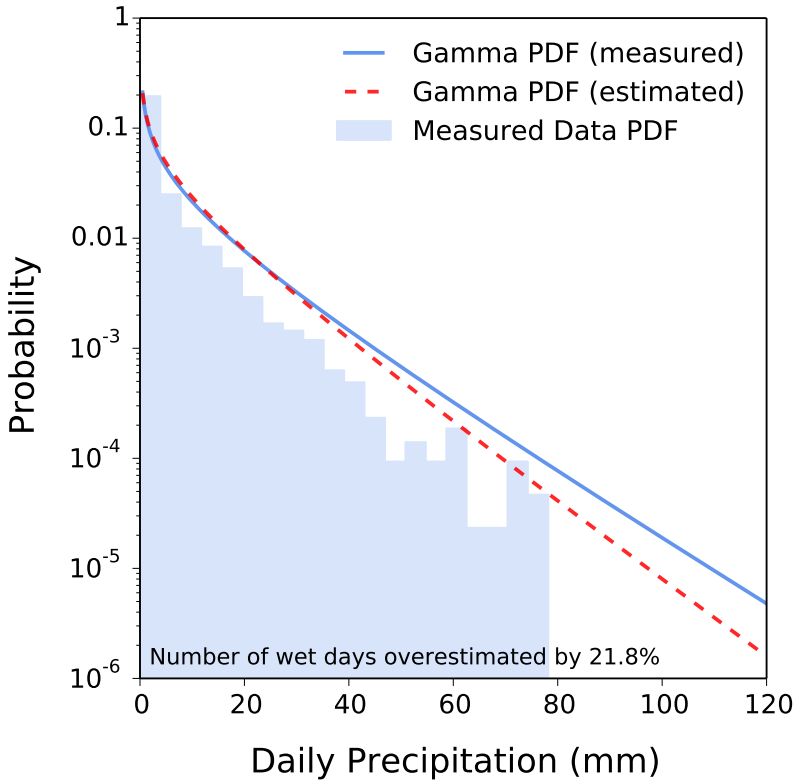


Figure 7: Gamma probability density functions that were estimated from the estimated (dashed red line) and observed (solid blue line) daily precipitation time series for the *Granby* weather station. The histogram of the distribution of the observed daily precipitation events is also shown in light blue.

also been validated against a set of data from 19 weather stations located in the Monteregie Est region, in Quebec, Canada. For this purpose, the cross-validation procedure, that is included with the algorithm, was used to conveniently asses the uncertainty of the method for each of the 19 weather stations. The method yielded consistent and reliable estimates for all the weather station tested. The RMSE and MAE are both below 2.0 and 1.4 °C for max, min, and mean daily air temperature, while it less than 3.4 and 1.4 °C, respectively, for precipitation. These results compare well with other published studies that used a similar method that the one used in this study.

In addition, the algorithm can also be used with a Graphical User Interface (GUI) that is part of the free and Open Source software WHAT (Well Hydrograph Analysis Toolbox). WHAT has been developed to estimate groundwater recharge by combining water level measurements with daily weather data time series. It also includes a graphical interface to easily search for weather stations in the online Canadian Daily Climate Database (CDCD) and download and format automatically the available data. The gap-filling algorithm that was presented in this paper can represent a powerful tool that could save a lot of time in any project requiring complete daily weather data time-series. Development of the algorithm is still in progress and new features might be added in the future.

References

Carrier, M.A., Lefebvre, R., Rivard, C., Parent, M., Ballard, J.M., Benoît, N., Vigneault, H., Beaudry, C., Malet, X., Laurencelle, M., Gosselin, J.S., Ladevèze, P., Thériault, R., Beaudin, I., Michaud, A., Pugin, A., Morin, R., Crow, H., Gloaguen, E., Bleser, J., Martin, A., Lavoie, D., 2013. Portrait des ressources en eau souterraine en Montérégie Est, Québec, Canada. Projet réalisé conjointement par l'INRS, la CGC, l'OBV Yamaska et l'IRDA dans le cadre du Programme d'acquisition de connaissances sur les eaux souterraines. Technical Report Research Report R-1433. Institut national de la recherche scientifique, Centre Eau Terre Environnement. Quebec City, Quebec, Canada. URL: http://espace.inrs.ca/1639/1/R001433.pdf.

DeGaetano, A.T., Eggleston, K.L., Knapp, W.W., 1995. A Method to Estimate Missing Daily Maximum and Minimum Temperature Observations. Journal of Applied Meteorology 34, 371–380. URL: http://journals.ametsoc.org/doi/abs/10.1175/1520-0450-34.2.371, doi:10.1175/1520-0450-34.2.371. d270.

Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S., Lott, N.J., 2000. Creating a serially complete, national daily time series of temperature and precipitation for the Western United States. Journal of Applied Meteorology 39, 1580–1591. doi:1520-0450(2000)039<1580:CASCND>2.0.CO;2. d181.

Gosselin, J.S., 2015. WHAT (Well Hydrograph Analysis Toolbox). URL: https://github.com/jnsebgosselin/WHAT.

Gosselin, J.S., Rivard, C., Martel, R., 2015. User Manual for WHAT. Document written for software version 4.1.7-beta. Technical Report. INRS-ETE, Quebec City, Qc, Can. URL: https://github.com/jnsebgosselin/WHAT/raw/master/WHATMANUAL/WHATMANUAL.pdf.

Kashani, M.H., Dinpashoh, Y., 2011. Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment 26, 59–71. URL: http://link.springer.com/article/10.1007/s00477-011-0536-y, doi:10.1007/s00477-011-0536-y. d265.

Menke, W., 1989. Geophysical Data Analysis: Discrete Inverse Theory. 1 edition ed., Academic Press, San Diego.

Simolo, C., Brunetti, M., Maugeri, M., Nanni, T., 2010. Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. International Journal of Climatology 30, 1564–1576. doi:10.1002/joc.1992. d184.

Tronci, N., Molteni, F., Bozzini, M., 1986. A comparison of local approximation methods for the analysis of meteorological data. Archives for Meteorology, Geophysics, and Bioclimatology, Series B 36, 189–211. URL: http://link.springer.com/article/10.1007/BF02278328, doi:10.1007/BF02278328. d218.

Xia, Y., Fabian, P., Stohl, A., Winterhalter, M., 1999. Forest climatology: estimation of missing values for Bavaria, Germany. Agricultural and Forest Meteorology 96, 131–144. doi:S0168-1923(99)00056-8. d182.

Appendix A.

Table A.3: List of method parameters for version 1.0 of the gapfilling algorithm.

Parameter name	Default value	Description
Nbr_Sta_max	4	Set the maximal number of neighboring stations that is used for the generation of the MLR models to estimate the missing daily weather data.
limitDist	100 km	Neighboring stations that are farther away from the target station than the specified value are completely excluded from the gap-filling procedure.
limitAlt	350 m	Neighboring stations with an absolute elevation difference with the target station that is higher than the specified value are completely excluded from the gap-filling procedure.
regression_mode	LAD	Define the optimization criteria that is used for the regression in the generation of the MLR model as described in Section 2.3. The two options available are *OLS* (Ordinary Least Squares) or *LAD* (Least Absolute Deviations).
full_error_analysis	False	When set to *True*, the accuracy of the method, for the dataset of the target station, will be estimated with the cross-validation procedure described in Section 2.5.
add_ETP	False	When set to *True*, daily potential evapotranspiration will be estimated from the daily temperature data series and will be saved in the '.out' file, along with the gapless data series produced with the gapfill algorithm.
inputDir	-	Directory where the algorithm search for valid weather data file.
outputDir	-	Directory where are saved all the outputs data files and figures when the gap-filling process for a station is completed successfully. Files associated with a given station are saved in a sub-folder named after the weather station.
time_start	-	Time in the weather dataset from which the gap-filling procedure will start.
time_end	-	Time in the weather dataset to which the gap-filling procedure will be completed.

Table A.4: List of outouts for version 1.0 of the gapfilling algorithm. The name of the file are given for the weather station BROME.

File name	File type	Description
BROME (7020840)_1980-2009.out	data	Tab-separated values file containing the gap-less weather dataset.
BROME (7020840)_1980-2009.log	data	Tab-separated values file containing detailed information about each daily weather value estimated to fill the gaps in the original weather data series.
BROME (7020840)_1980-2009.err	data	Tab-separated values file containing the results of the cross-validation procedure.
weather_datasets_summary.log	data	Tab-separated values file listing all the weather station for which a data file was available in the 'inputDir'. For each station, information about the location coordinates (latitude and longitude), elevation, years for which data are available, and proportion of missing data are also provided.
weather_normals.pdf	figure	Graphs presenting the yearly and monthly weather normals for precipitation and max, min, and mean air temperature, calculated from the gap-less weather dataset. Examples of these graphs are shown in Fig. 4.
precip_PDF.pdf	figure	Graph showing the gamma probability density functions that were estimated from the estimated and observed daily precipitation time series. The histogram of the distribution of the observed daily precipitation events is also shown. An example is provided in Fig. 7.
Max Temp (deg C).pdf	figure	Scatter plot comparing the goodness of fit between the observed and estimated daily max temperature series. An example is presented in Fig. 5a.
Min Temp (deg C).pdf	figure	Scatter plot comparing the goodness of fit between the observed and estimated daily min temperature series. An example is presented in Fig. 5b.
Mean Temp (deg C).pdf	figure	Scatter plot comparing the goodness of fit between the observed and estimated daily mean temperature series. An example is presented in Fig. 5c.
Total Precip (mm).pdf	figure	Scatter plot comparing the goodness of fit between the observed and estimated daily total precipitation series. An example is presented in Fig. 6.