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

Daily weather data are useful in several areas of Earth sciences, including hydrology, hydrogeology and agronomy. However, weather datasets are often incomplete. The estimation of missing data can be a complex and tedious task. This is particularly the case for daily precipitation because of their high spatial and temporal variability. A user friendly, menu-driven, and interactive computer program for rapid and automatic completion of daily climatological series has been developed. Missing data for a given weather station are estimated using a multiple linear regression model, generated using data from nearby stations. For daily precipitation, it is possible to activate an option that forces the algorithm to preserve the probability distribution of data. This is an advantage over conventional approaches that tend to overestimate the number of wet days and underestimate the high intensity precipitation events. The software also allows downloading and automatic formatting of raw data available on the Environment Canada website. The software is demonstrated for two weather station located in Monteregie Est region, southern Quebec. Cross-validation was used to check the method and to define the optimal parameters to minimize the error in estimating missing daily precipitation.

Keywords: `heat transport, recharge assessment, uncertainty analysis, subsurface temperature time series`

1. Introduction

Climate data are useful in several fields of Earth sciences, including hydrology, hydrogeology and agronomy. For this purpose, the Canadian Daily Climate Database (CDCD), owned by the Government of Canada, contains daily data for air temperature and precipitation dating back to 1840 to the present for about 8450 stations distributed across Canada. Data can be downloaded manually on the Government of Canada website (`www.climate.weather.gc.ca`) for each year individually and saved in a csv file. This process involves a lot of repetitive manipulations and is a time consuming task. Moreover, the re-organization of the individual data files, saved for each year separately, in a more

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

convenient format can also represent a tedious task when done manually. Alternatively, it is possible to order a DVD containing the entire database for a small fee. This option has the disadvantage of only providing an image in time as data cannot be updated.

Furthermore, climate 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 models that heavily depend on these data. Filling the gaps in weather datasets can quickly become a tedious task as the size 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 there are various methods to estimate missing daily weather data that are well covered in textbooks and technical papers, few tools to perform this task efficiently and automatically are available.

WHAT (Well Hydrograph Analysis Toolbox) is a computer program that addresses the aforementioned issues (Gosselin et al., 2015). Firstly, it 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 more convenient format. Secondly, the program also includes an automated, robust, and efficient method to quickly and easily fill the gaps in the daily weather datasets downloaded from the CDCD. WHAT also includes a cross-validation resampling algorithm to conveniently validate and assess the uncertainty of the estimated missing values.

This paper presents the algorithm that is used in the WHAT software to fill the gaps in daily weather datasets and to assess the uncertainty on the estimated values. An application of the method, using the WHAT software, is also presented for the Montérégie Est study area, located in southern Quebec, Canada. A guide for the operation of the software Gosselin (2015) is available for download at this web address: `https://github.com/jnsebgosselin/WHAT`.

2. Theory

The algorithm described in this paper is based on the implementation of 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.

Figure 1 presents a flowchart of the algorithm developed for filling the gaps in the daily weather dataset of a given target station using data from the neighboring stations. The algorithm consists of two nested loops: the external ‘Loop A’ iterates over the weather variables in the dataset of the target station (min, max, and mean air temperature and total precipitation), while the inner ‘Loop B’ iterates over the missing values in the data series of a given variable. Each missing value is estimated independently with a two-step procedure. The first step consists in the selection of the neighboring stations. The second step consists in building a MLR model and estimating the missing value in the dataset of the target station.

2.1. Loop A

2.1.1. Correlation Coefficients Calculations

The first step consists in calculating the correlation coefficients between the available data of the target station and those of the neighboring stations for the j^{th} weather variable of the dataset. A correlation coefficient is calculated for each neighboring station individually using all the available data. Neighboring stations that have less than 182 days (half a year) with synchronous data with the target station or with a correlation coefficient below a value of 0.35 are discarded and won’t be used for filling the gaps the time series of the target station for this weather variable. The 0.35 threshold for the correlation coefficient is based on the value used by Eischeid et al. (2000) in theirs application of the method.

Moreover, data correlation between two stations will generally decreases as the horizontal and vertical distances between them increase. It is therefore possible to specify a cutoff distance and a cutoff altitude difference for which neighboring stations that fall above these cutoff values are thereafter ignored in the gapfilling algorithm. The default values are set to 100 km and 350 m for the horizontal and vertical distance respectively based on the literature Tronci et al. (1986); Xia et al. (1999); Simolo et al. (2010).

2.2. Loop B

2.2.1. Selection of the neighboring stations

The selection of surrounding stations is critically important for the accurate estimation of missing weather data Eischeid et al. (2000). Problems arise though because the list of neighboring stations with available data can vary from one day to the other. Therefore, it is not possible to use a single MLR model to estimate the missing values in the dataset of the target station all at once. The selection of the neighboring stations and the generation of a MLR model must be done instead individually for each missing value in the dataset of the target stations.

Neighboring stations with available data are selected in descending order of their correlation coefficient, up the a maximal number of stations that was set in the method parameters. The default value for the maximal number of neighboring station used for the construction of the MLR models is set to four. Tests run by Eischeid et al. (2000) showed that using more then four stations did not significantly improve or degraded the accuracy of the estimate. If

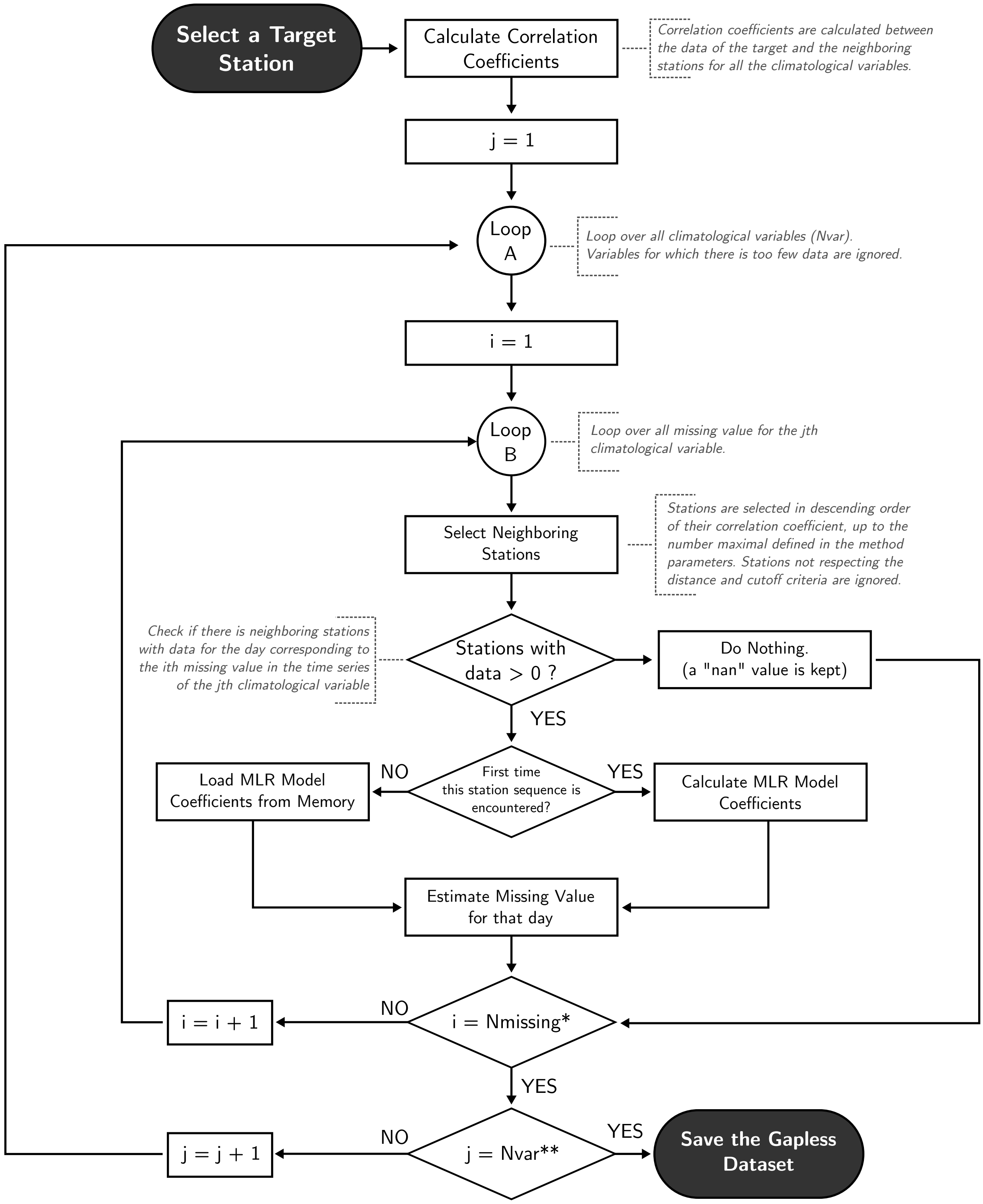
Select a Target Station
Calculate Correlation Coefficients
Correlation coefficients are calculated between the data of the target and the neighboring stations for all the climatological variables.
j = 1
Loop A
Loop over all climatological variables (Nvar). Variables for which there is too few data are ignored.
i = 1
Loop B
Loop over all missing value for the jth climatological variable.
Select Neighboring Stations
Stations are selected in descending order of their correlation coefficient, up to the number maximal defined in the method parameters. Stations not respecting the distance and cutoff criteria are ignored.
Check if there is neighboring stations with data for the day corresponding to the ith missing value in the time series of the jth climatological variable
Stations with data > 0 ?
Do Nothing. (a "nan" value is kept)
YES
Load MLR Model Coefficients from Memory
NO
First time this station sequence is encountered?
YES
Calculate MLR Model Coefficients
Estimate Missing Value for that day
i = i + 1
NO
i = Nmissing*
YES
j = j + 1
NO
j = Nvar**
YES
Save the Gapless Dataset

Figure 1

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.

2.2.2. *Generation of the Multiple Linear Regression Model*

Each time a MLR model is generated for a given sequence of neighboring stations, the program stores the model parameters into memory. Therefore, after the selection process of the neighboring stations is done for a given day with a missing data (section 2.2.1), the program checks if the sequence of selected stations has already been encountered before for the current weather variable. If so, the stored MLR parameters will be use directly to estimate the missing data for the current day. Otherwise, the model will generate a new MLR model and will store the results into memory. Since a MLR model is generated only one time for a given sequence of neighboring stations, this means that the algorithm becomes faster with time.

using either an Ordinary Least Square (OLS) or a Least Absolute Deviations (LAD) criteria (both options are available). Since daily precipitation series generally represented by long-tailed, positively skewed, distributions, the LAD criterion is typically a better option than the OLS criteria for handling this kind of distribution because it is more robust to outliers (Eischeid et al., 2000, 1995). The downside is an increase in computation time. The MLR using a LAD criterion is computed in WHAT with an iterative reweighted least-squares method (Schlossmacher, 1973; (Eischeid et al., 1995).

2.2.3. *Estimating Missing Daily Values*

Once the parameters of the MLR model for a given day with a missing value are known, the missing value in the target time series is estimated as:

$$X(t) = a_0 + \sum_{i=1}^{N} a_i \cdot Y_i(t) \tag{1}$$

where Y(t) is the missing value of the target station estimated at time t, Xi are the synchronous values of the neighboring stations, ai are the regression coefficients and N is the total number of neighboring stations used for the regression, up to the maximal value defined in the method (default is four).

When all missing values in the target station dataset have been estimated, the resulting gapless time series is saved in a ".out" file. Detailed information about the estimated values are saved in an accompanying ".log" file (see Section 8.3.1).

2.3. *Step 3: Validation*

WHAT also includes an option to perform a validation of the method used for a particular weather dataset with a jackknife procedure. This option is an advanced feature that can be activated by changing the value of the field "Full Error Analysis" from 0 to 1 in the file named "WHAT.pref" (see Section 8.3.1).

More specifically, when this option is activated, WHAT will estimate a value for the target station for every day of the time series. In other words, loop B in the flowchart of Figure 8.1 will run over all the days of the dataset and not only over days for which there is a missing data. In addition, the memory feature will be deactivated and a MLR model will be estimated for each day independently. If a measured value is present for the current day being estimated, this value will be temporarily discarded from the data series to avoid self-influence of the observation on the estimation procedure

Consequently, the activation of this feature will significantly raise the computation time for filling the gaps in weather time series and should be used only when a detailed analysis of the estimation errors is required. The default information provided in the ".log" should contain sufficient material to fill the needs of a large number of projects. Thus, this process leads to the production of a weather time-series for which every value has been estimated in WHAT. The results are saved in a tsv (tab-separated values) text file with the extension ".err" that is named after the station name and ID similarly to the ".log" and ".out" files. The accuracy of the estimation technique can then be assessed by comparing the estimated weather data with the respective non-missing observations in the original weather data file. There is currently no tool provided in WHAT to directly analyze the results from the Jackknife procedure. However, all the source code that has been written? for the production of the figures of Section 8.4 can be downloaded freely on GitHub at (https://github.com/jnsebgosselin/WHAT).

3. Operation

3.1. Input Data

3.2. Method Parameters Input File Format

4. Application: Monteregie Est Case Study

The Monteregie Est region is located in southern Quebec, Canada, on the south shore of the St. Lawrence River. It covers a total area of 9032 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 Figure X).

This region has been the subject of an extensive characterization project within the "Programme d'acquisition de connaissances sur les eaux souterraines du Québec" (PACES) whose main objective was to prepare a realistic and concrete picture of the groundwater resources for the region (**?**).

4.1. Study Area

The climate is characterized by significant seasonal differences in temperature, resulting in warm summers and cold winters. Precipitation, as rain or snow, are distributed rather evenly throughout the year.

Among all the weather stations for which data were available in and around the study area, a total of 32 was selected based on the availability and continuity of the weather data between 1980 and 2014. A list of these selected stations is presented in Table X with their coordinates, altitude, total time periods for which data were available, mean annual cumulative precipitation, and mean annual air temperature. Most of these information are generated automatically by WHAT in the file “weather_datasets_summary.log” (see Section **??**).

4.2. Materials and Method

Pour l’ensemble de ces stations, les précipitations totales annuelles sont d’environ 1100 mm/y en moyenne. Les précipitations totales les plus élevées sont observées à la station de Brome (~1280 mm/y et les plus faibles à la station de Sorel (~960 mm/year). La température annuelle moyenne dans la région d’étude est de 5.9 °C, variant de 4.3 to 6.7 °C tandis que les températures mensuelles moyennes fluctuent entre −12 to 21 °C. Les températures mensuelles minimales sont observées en janvier (−17.1 to −13.6 °C) tandis que les températures mensuelles maximales sont observées en juillet (24 to 26.7 °C). Les températures les plus élevées sont généralement observées aux stations de Philipsburg et Saint-Bernard-de-Lacolle (température annuelle moyenne de 6.7 °C) et les plus faibles à la station de Bonsecours (température annuelle moyenne de 4.3 °C)

Les figures 1.3 et 1.4 illustrent respectivement les variations spatiales des précipitations totales annuelles et de la température moyenne annuelle pour la période de 1970-2000. Les valeurs présentées sur ces figures ont été interpolées par krigeage ordinaire sur une grille de 250 x 250 m, à partir des valeurs rapportées pour les 16 stations actives mentionnées ci-haut.Dans la région d‘étude, la tendance générale indique que les précipitations annuelles totales diminuent du sud-sud-est vers le nord-nord-ouest et que les températures annuelles moyennes diminuent du sud-ouest vers le nord-est. Outre l‘influence de la latitude, la température de la région est également influencée par la présence des Appalaches au sud-est et du fleuve Saint-Laurent au nord-ouest.

The weather network of the Monteregie Est region, located in the province of Quebec, Canada, has been used to test the method. This region feature strongly variable topography and land cover conditions. The network is presented in figure X. Also, stations from bordering states were extracted to improve the spatial distribution of sites surrounding target stations located near state borders.

Daily weather data for 32 weather stations in and around the Monteregie Est area were also retrieved from the Canadian Daily Climate Database (CDCD) with the software WHAT for the years 2000 to 2012. Missing values in the weather time series were also estimated with WHAT to produce gapless meteorological records of daily air temperature and precipitation.

Tests have shown that inclusion of more than four stations does not significantly improve the interpolation and may in fact degrade the estimate.

4.3. Results and Discussion

The quality of the estimates is strongly affected by seasonality. Stations at higher elevations are difficult to estimate accurately, in large part because of the topographical diversity of the surrounding stations leading to degradation of spatial coherence among stations.

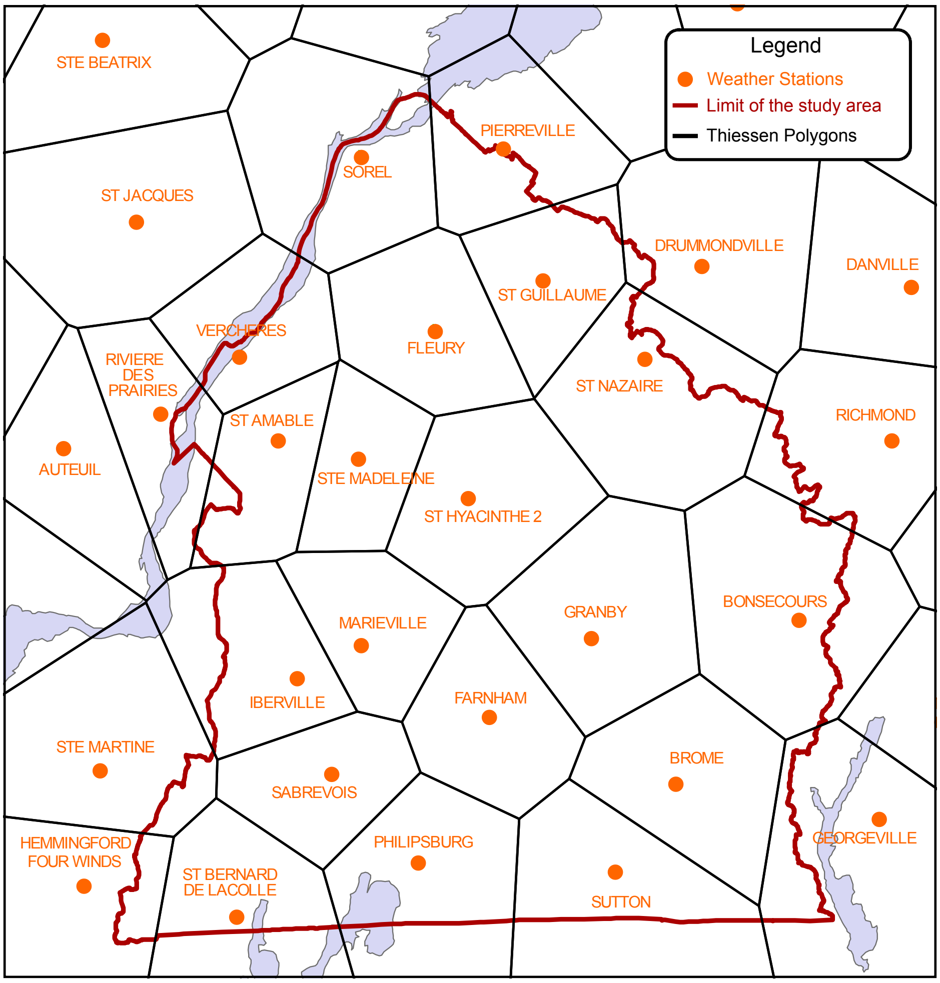

Figure 2: Locations of the weather stations in the Monteregie Est area.

The tendency for all of the methods to have a negative bias is indicative of the nature of precipitation distributions to be positively skewed (interpolated values will tend to cluster about the median error rather than the mean).

According to Xia et al. (1999), the two most important factors in climatology are the inter-correlations in the station network, and the seasonal variations in the relations between the stations.

5. Discussion

However, weighing and regression-based techniques, including the MLR method, all tend to overestimate the number of rainy days, while heavy precipitation events are systematically underestimated. Therefore, the rainfall probability distribution is usually not preserved with these techniques). However, Simolo et al. (2010) have proposed a two-step procedure to modify the MLR method to address these issues.

An alternative approach would have been to calculate the correlation coefficient with a subset of data from the target series centered around the missing value, as it was done in Simolo et al. (2010) for instance. This approach allows for a better representation of the seasonal variations in the relationships between the stations. The downsides include a more complex algorithm to implement and a reduction of the method robustness and efficiency.

References

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.

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.