An algorithm for automated estimation of missing daily climate data

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 convenient format can also represent a tedious task when done manually. Alternatively, it is possible to order a DVD

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

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 exist 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 conveniently 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. In addition to the handling of missing data, WHAT includes a cross-validation resampling technique to conveniently validate and assess the uncertainty of the estimated missing values.

This paper presents the algorithm that was developed as part of the WHAT software. The operation of the user interface of the software is provided in Gosselin (2015), available for download at this address: `https://github.com/jnsebgosselin/WHAT`.

2. Theory

The algorithm described in this paper for filling the gaps in daily air temperature and total precipitation datasets is based on the implementation of the classical MLR method presented in Eischeid et al. (2000). The MLR method is a robust, efficient, accurate, and well known method 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.

The algorithm that was developed in this work for filling the gaps in weather datasets is presented in the flowchart of fig. 1. It consists of two nested loops: the external 'Loop A' iterates over the time series of four weather variables

(min, max, and mean air temperature and total precipitation) for the target station while the inner 'Loop B' iterates over every missing value in each data series. The estimation of a single missing value is achieved with a two-step procedure. The first step consists in selecting the data series with the best correlation coefficient, which also respect a certain number of conditions. The second step consists in building a MLR model and estimating the missing values. This is described in more details below.

2.1. Loop A

2.1.1. Quality Control

Prior to the analysis of weather time series, it is important to apply quality control constraints to ensure that the data do not violate obvious constraints associated with minimum, maximum, and average daily air temperature and daily cumulative precipitation.

The program will identify irregularities or inconsistencies to insure that maximum, minimum and average daily temperatures are coherent for a given day and that all daily precipitation values are positive. Erroneous values are replaced by nan values in the dataset. These values will subsequently be estimated by the program from neighboring stations.

2.1.2. Station Correlation Assessment

The first step consists in calculating the correlation coefficients between data of the target station and those of the neighboring stations for each of the four weather variables: minimum, maximum and average daily temperatures and daily cumulative precipitation. These coefficients are calculated for the entire time-series for each neighboring station individually. If there are less than 182 synchronous values between the data of the target station and those of a neighboring station for a given variable, the correlation is not computed and a "NaN" value is kept instead.

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., 1995). Problems arise though because of synchronized missing values in the target and neighboring weather station datasets that varies trough time. This is illustrated in Table 8.1, where theoretical time-series of air temperature data with a realistic distribution of missing values are presented.

In table 1, there are missing values in the target station dataset for days 2, 4, and 5. The missing value on day 2 will then be estimated with the data of the neighboring stations Y1, Y3, and Y4 since station Y2 is also missing a value on this day. All neighboring stations will be used for the estimation of the missing value on day 4, while only stations Y1 and Y2 have data available for the estimation of the missing value on day 5.

Data correlation between two stations will generally decreases as the horizontal and vertical distances increase. It is possible to specify a cutoff distance and a cutoff altitude difference for which neighboring stations that fall above

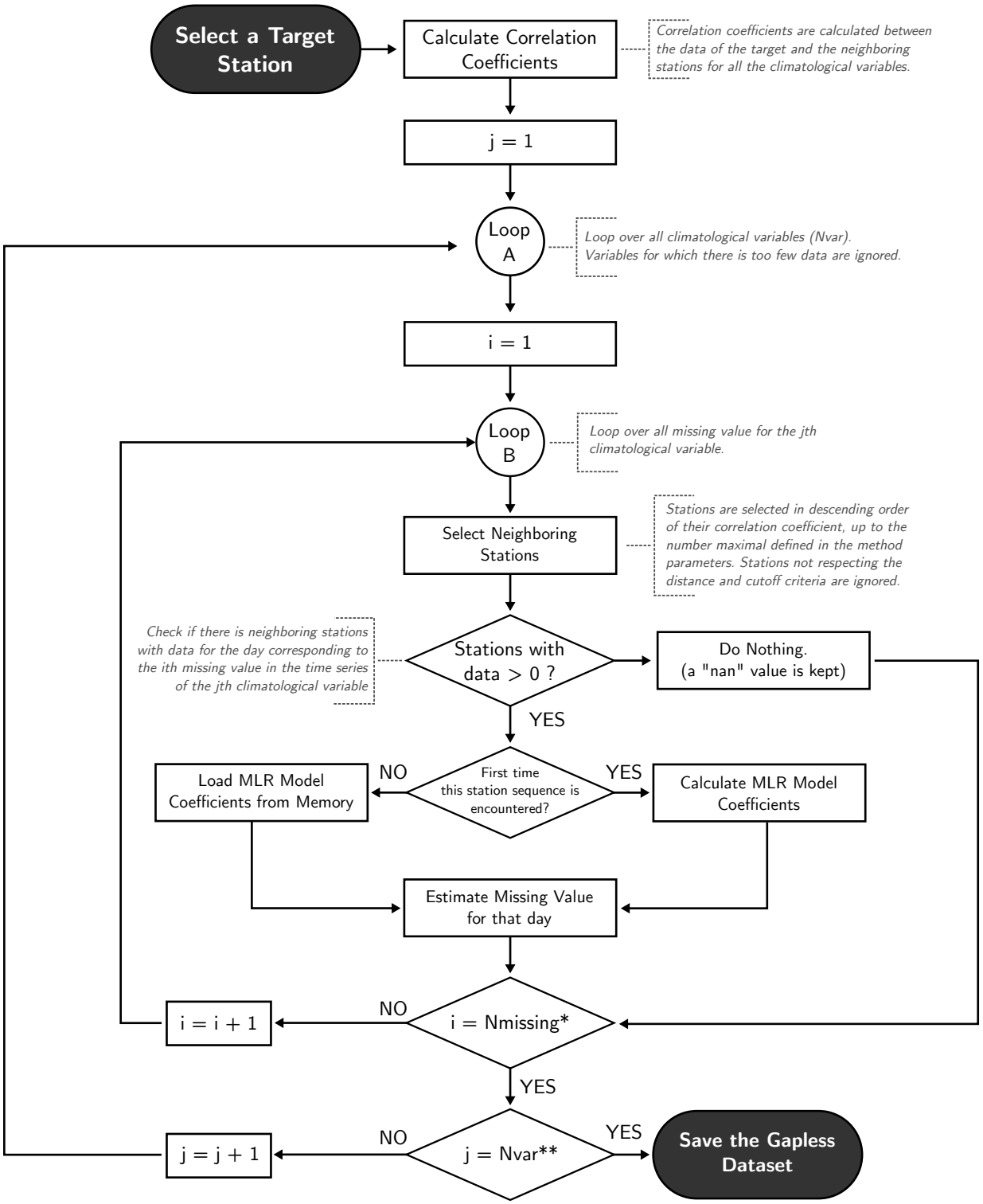
Select a Target Station
Calculate Correlation Coefficients
Correlation coefficients are calculated between the data of the target and the neighboring stations for all the climatological variables.
j = 1
Loop A
Loop over all climatological variables (Nvar). Variables for which there is too few data are 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
NO
First time this station sequence is encountered?
YES
Load MLR Model Coefficients from Memory
Calculate MLR Model Coefficients
Estimate Missing Value for that day
NO
i = Nmissing*
i = i + 1
YES
NO
j = Nvar**
YES
j = j + 1
Save the Gapless Dataset

Figure 1

these cutoff values are ignored by the program. The default values are set to 100 km and 350 m for the horizontal and vertical distance respectively based on the literature (Simolo et al., 2010; Tronci et al., 1986; Xia et al., 1999).

Table 1: This table shows some data

	Target	Neighbors			
Day	Y	X1	X2	X3	X4
1	11.0	12.0	12.0	12.5	10.0
2	**nan**	12.0	**nan**	13.0	12.2
3	7.5	8.5	8.5	8.0	8.9
4	**nan**	6.0	4.5	5.0	4.4
5	**nan**	8.0	8.5	**nan**	**nan**

Since the number of neighboring stations with available data is not fixed in time, it is not possible to use a single MLR model to fill all the missing values for the target station all at once. For each missing value in the target station dataset, the program keeps only the datasets of the neighboring stations that also have data at this particular time. Data series of stations that do not respect the cutoff criteria for distance and elevation differences are also ignored. Data from neighboring stations are selected in descending order of their correlation coefficient with the target station, up to a maximal number of stations defined in the method parameters. The default value for the maximal number of neighboring station was set to four, based on the literature (Eischeid et al., 1995; Xia et al., 1999).

If for a given day, no neighboring stations have a measured value to fill a gap in the target station dataset, no calculation is done and a “NaN” value is kept in the series instead and the program pass to the next missing value in the target series.

3. 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.

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.