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. 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 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

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

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.

This paper presents an open source algorithm, written in the Python programming language, that can be used to automatically 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.

for filling the gaps in the daily weather datasets of a given weather station (hereafter called the target station) using data from the neighboring stations.

Can also be used to compute daily potential evapotranspiration.

The algorithm is available for free at : . In addition, it is included as part of the WHAT software, which blablabla.

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.

In addition the algorithm can also b

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 shows a flowchart of the gap-filling algorithm presented in this paper. The algorithm consists of two nested loops: the external 'Loop A' iterates over the weather variables contained 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 the current weather variable in 'Loop A'. 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, while the second step consists in building a MLR model, estimating the missing value, and filling the corresponding gap in the data

series.

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, using all the available data. Neighboring stations that have less than 182 days (half a year) of synchronous data with the target station or that have 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 defined for the correlation coefficient 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 further of the target station than specified thresholds, 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 selection of the neighboring stations and the generation of a MLR model must be done individually for each day with a 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 is specified as a parameter of the algorithm. The default value for the maximal number of neighboring station 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 estimate. 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 resulting model parameters are stored into memory. Therefore, after the neighboring stations have been selected for a given day with a missing data (section 2.2), the program checks if this sequence of selected 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, 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, the algorithm becomes faster with time.

The MLR model can be generated using either an Ordinary Least Square (OLS) or a Least Absolute Deviations (LAD) criteria. For the OLS criteria, the MLR model is obtained by solving the linear matrix equation $\mathbf{Xa} = \mathbf{Y}$ by

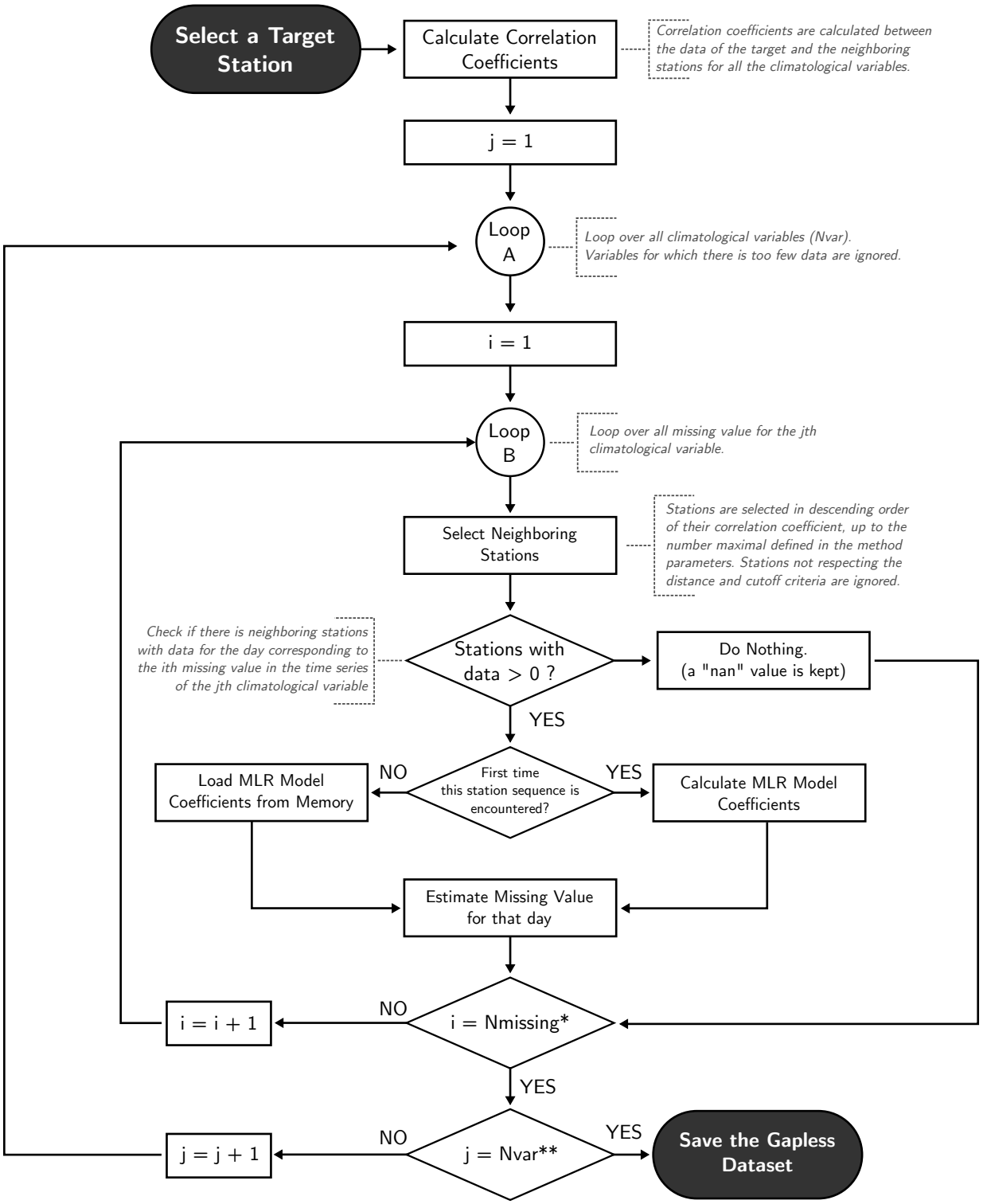
Select a Target Station
Calculate Correlation Coefficients
Correlation coefficients are calculated between the data of the target and the neighboring stations for all the climatological variables.
j = 1
Loop A
Loop over all climatological variables (Nvar). Variables for which there is too few data are 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

computing the $N \times 1$ parameter vector $\mathbf{a}$ that minimizes the Euclidean L2-norm $\|\mathbf{Y} - \mathbf{Xa}\|_2$, where $\mathbf{Y}$ is a $M \times 1$ vector containing the M daily data of the target station and $\mathbf{X}$ is a $M \times N$ matrix containing the M synchronous daily data of the N selected neighboring stations. (REFERENCE: Numpy Documentation)

However, daily precipitation series are generally characterized by long-tailed, positively skewed, distributions. In this case, the generation of the MLR model using the more robust LAD method is a more appropriate approach that will yield more reliable results than using the OLS method as described above Menke (1989); Eischeid et al. (2000). Resolution of the MLR model with the LAD method is achieved in the gap-filling algorithm using an iterative reweighted least-squares method. The downside in using the LAD method compared to using the OLS is an increase in computation time by about a factor 10.

2.4. Estimating Missing Daily Values

Once the parameters of the MLR model are known, the missing value for the corresponding day at time t_i can be estimated as follows:

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

where $Y_p(t_i)$ is the value estimated at time t_i for the $j_t h$ weather variable in the dataset of the target station , $X_k(t_i)$ is the synchronous available data of the $k_t h$ neighboring stations, a_k are the regression coefficients, and N is the total number of selected neighboring stations used for the regression.

When all the missing values in the dataset of target station have been estimated and filled, the resulting gapless time series are saved in a file with a '.out' extension. Moreover, detailed information about the estimated values are also saved in an accompanying '.log' file. The outputs of the gap-filling algorithm are discussed in more details in section 3.3.

2.5. Uncertainty of the estimated values

Each time a new MLR model is generated from a sequence of neighboring stations, an estimation of the accuracy of the model is made for the weather variable for which missing data are being estimated. This is done by first using the model to estimate the values for the days in the data series of the target station for which there exists a measured value. The accuracy of the MLR model is then approximated by computing a Root-Mean-Square Error (RMSE) between the estimated values and the respective measured values. The RMSE thus calculated is saved, along with the estimated value, in the '.log' file.

The algorithm also includes a cross-validation re-sampling procedure to estimate the accuracy of the method to fill the gaps in the dataset with a more rigorous approach. The procedure to enable this functionality in the algorithm is presented in section 3.1.

The procedure consists in estimating a value for each day of the dataset of the target station, even for days for which data are not missing. In other words, when this option is enabled, the loop B in the flowchart of fig. 1 will iterate over all the days 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 in the dataset of the target station is temporarily discarded to avoid self-influence of this observation on the generation of the MLR model. Therefore, a new MLR model must also be generated for each day independently. Since model parameters cannot be recalled from previous models, the computation time of the gap filling procedure is significantly increased, especially if the least absolute deviation regression method is selected.

After a value has been estimated for a given day, the corresponding observed data is put back in the dataset of the selected station. When a value for every day of the dataset has thus been estimated, the estimated values are saved in a file with the extension '.err', along with the '.log' and '.out' files described in section 2.4. The accuracy of the method can then be estimated by computing the RMSE between the estimated weather data and the respective non-missing observations in the original dataset of the selected station. Though costly in computation time, enabling this option can provide interesting insights on the performance of the procedure for the specific datasets used for a given project.

Moreover, the graphs that are presented in Section blabla are also produced upon the completion of the gap-filling routine for a given station blablabla.

3. Operation

The code is compliant with either Python 3.4 or 2.7.9 or later. It requires numpy , xlrd and PySide or any later version of these librairies. The algorithm is organized as a base class of the Qt GUI Framework using the PySide binding. Signals are also emitted at various stade in the gap-filling routine. This has been done to facilitate the addition of a Graphical User Interface on top of the algorithm with the Qt GUI Development framework. There is an mininimal working example of application that is documented at the end of the file with the algorithm at the end of the file.

The algorithm is also implemented in the free and open source software WHAT which provides a user friendly and convenient interface. Detailed information about the use of the algorithm with the interface of WHAT are provided in the user guide of WHAT.

3.1. Parameters

3.2. Input Data

It is possible to use weather data from any sources in WHAT, given the right format is used, either to fill the gaps in the weather time series and/or to interpret water level time series. For this purpose, it is recommended to use a copy of one of the sample files that are provided in the project example (distributed with the software) and fill the information and the data directly in it. The file must be kept in a text format using tab-separated values either with the extension ".csv" or ".out", depending if you want to fill the gaps in the weather time series or interpret water level time series. This can be achieved with any standard spreadsheet application such as Microsoft Excel or LibreOffice Calc. The format of the header must be faithfully observed for those files. In addition, "NaN" values must be entered where data

are missing. 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 in the time-series. These missing blocks of data will be filled during the gap filling procedure or will be ignored for the plotting of the hydrograph.

3.3. Output

4. Application: Monteregie Est Case Study

4.1. Materials and Method

4.1.1. Study Area

The method was tested using data from land-based Canadian weather stations located in and around the Monteregie Est region, which is located in southern Quebec, Canada. This region 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). It is characterized by strongly variable topography and land cover conditions. The climate of this region is characterized by significant seasonal differences in temperature, resulting in warm summers and cold winters. Total precipitation, as rain or snow, are distributed rather evenly throughout the year.

4.1.2. Weather Dataset

Among all the weather stations for which data were available in and around the study area in the Canadian Daily Climate Database (CDCD), a total of 32 was selected based on the availability and continuity of the measured weather data between 1980 and 2014. Table 2 presents the list of these selected stations with their corresponding climate ID, location coordinates (latitude and longitude), altitude, time periods for which data were available, mean annual cumulative precipitation, and mean annual air temperature. Most of the information presented in table 2 are generated automatically when loading data into the gap-filling routine and saved in a file named 'weather_datasets_summary.log' within the previously defined output folder. The geographical disposition of the weather stations is also presented in the map of fig. 3.

The data were downloaded and formatted to the format described in section X using the WHAT (Well Hydrograph Analysis Toolbox) a computer program (Gosselin et al., 2015). In addition to offering a set of tools to assist in the interpretation of water level time series, WHAT also 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. As mentioned previously, it also included an interface to easily use the gap-filling algorithm presented in this paper.

Total annual precipitation for all of these stations, the mean annual precipitation is 1100 mm/y. The highest total precipitation are observed at the Brome station (~1280 mm/y), while the lowest at the Sorel station (~960 mm/year). Mean annual air temperature in the study area is 5.9 °C, ranging from 4.3 to 6.7 °C while average monthly temperatures fluctuate between −12 to 21 °C. The minimum monthly temperatures are observed in January −17.1 to −13.6 °C) while

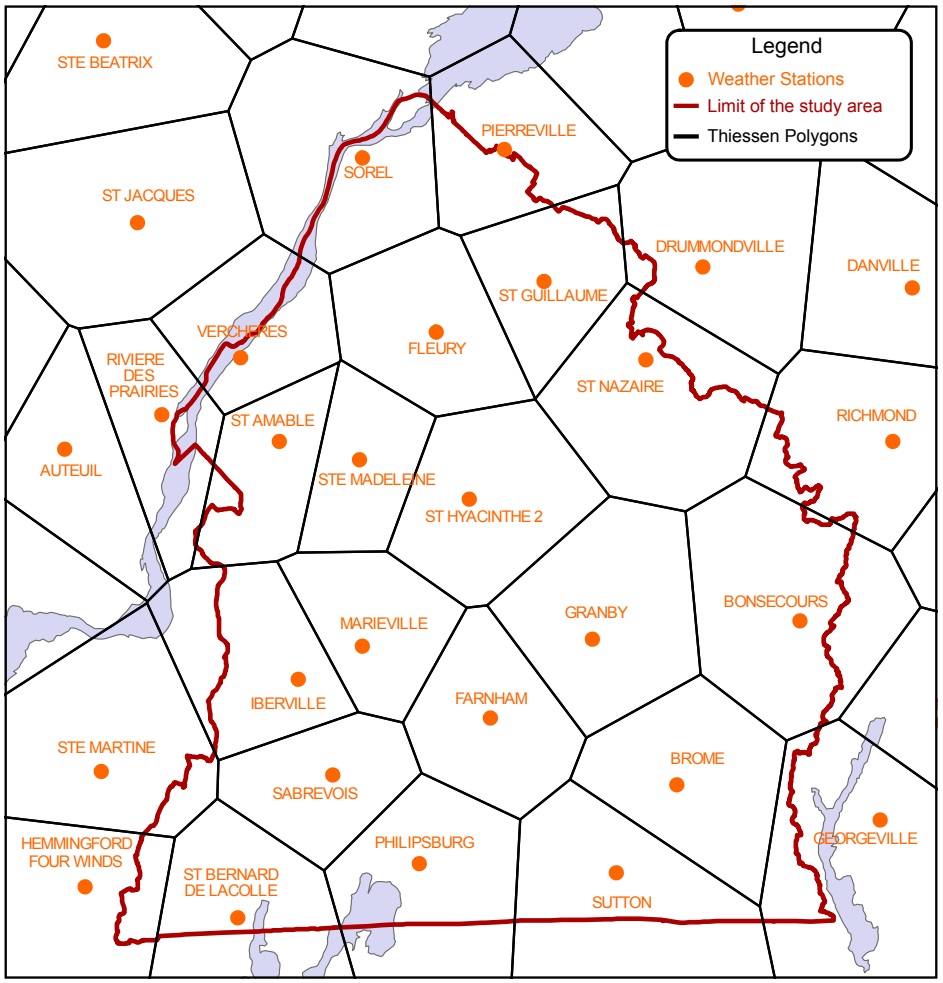

Figure 2: Locations of the weather stations in the Monteregie Est area.

the maximum monthly temperatures are observed in July (24 to 26.7 °C). The highest temperatures are observed at the Philipsburg et Saint-Bernard stations and the lowest at Bonsecours station.

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. 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.

4.1.3. Gap-Filling the Data

The algorithm described in section 2 of this paper was used to estimate the missing weather data and fill the gaps in the weather dataset of the station located within the study area. Stations bordering the limits but outside the study area were used to fill the data of the stations within the area, but their data were not filled. Data from these stations were use only to improve the spatial distribution of the neighboring stations for the weather station in the study area located near the limits.

The default value for the method parameters were kept in the algorithm. That is, the maximum number of neighboring station was set to 4, the horizontal and vertical distance thresholds were kept at 100 and 350 values respectively. The data were filled using the OLR and LRM method to see if it makes any real difference. Also, during the entire process, the option for the cross-validation was activated.

The accuracy of the method was also assessed with different value imposed on the maximum number of neighboring stations to see the impact on the accuracy of the method.

4.2. Results and Discussion

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

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.

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.

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.

Table 1: List of selected weather stations in the study area and related information about missing data for the 1980-2014 period

#	Station	Climate ID	Lat.	Lon.	Alt.	% of days with missing data			
						T_{max}	T_{min}	T_{mean}	P_{tot}
1	Auteuil	7020392	45.65	−73.73	53.0	20.8	19.1	22.8	18.4
2	Bonsecours	7020828	45.40	−72.27	297.2	4.7	5.1	6.3	3.2
3	Brome	7020840	45.18	−72.57	205.7	2.6	2.3	3.0	2.3
4	Bromptonville	7020860	45.48	−71.95	130.0	3.9	4.2	6.0	1.8
5	Danville	7021954	45.82	−71.98	190.0	30.8	31.1	33.1	30.2
6	Drummondville	7022160	45.88	−72.48	82.3	2.5	2.4	3.4	1.8
7	Farnham	7022320	45.30	−72.90	68.0	5.2	4.7	6.1	3.7
8	Fleury	7022375	45.80	−73.00	30.5	1.1	1.2	1.6	1.1
9	Georgeville	7022720	45.13	−72.23	266.7	26.5	26.2	27.6	5.7
10	Granby	7022800	45.38	−72.72	175.0	1.3	1.3	2.1	0.5
11	Hemmingford	7023075	45.07	−73.72	61.0	6.0	6.0	6.8	4.6
12	Iberville	7023270	45.33	−73.25	30.5	7.3	7.5	9.7	4.2
13	Magog	7024440	45.27	−72.12	274.0	4.2	4.2	5.3	4.6
14	Marieville	7024627	45.40	−73.13	38.0	10.2	10.4	11.2	9.8
15	Nicolet	7025440	46.20	−72.62	30.4	4.1	4.2	5.2	3.6
16	Philipsburg	7026040	45.03	−73.08	53.3	4.9	5.1	6.9	3.2
17	Pierreville	7026043	46.08	−72.83	15.2	6.3	5.4	6.9	4.9
18	Richmond	7026465	45.63	−72.13	123.1	3.3	3.3	3.8	4.0
19	Riviere des Prairies	7026612	45.70	−73.50	9.0	2.4	4.0	4.8	1.4
20	Sabrevois	7026734	45.22	−73.20	38.1	25.5	26.0	27.1	5.2
21	Sorel	7028200	46.03	−73.12	14.6	5.7	5.9	6.2	4.5
22	St. Amable	7026818	45.67	−73.30	41.1	8.6	10.4	11.8	7.7
23	St. Bernard	7026916	45.08	−73.38	49.3	9.6	9.7	10.5	9.0
24	St. Guillaume	7027302	45.88	−72.77	43.9	4.3	4.6	5.7	3.0
25	St.Hyacinthe 2	7027361	45.57	−72.92	33.0	6.7	6.9	7.5	6.6
26	St. Jacques	7017380	45.95	−73.58	69.0	11.7	11.9	14.0	10.8
27	St. Janvier	7017386	45.73	−73.88	61.0	40.5	40.7	41.6	21.9
28	St. Nazaire	7027588	45.73	−72.62	68.6	3.8	3.8	5.4	2.7
29	Ste. Madeleine	7027517	45.62	−73.13	30.0	5.8	6.4	7.0	5.1
30	Ste. Martine	7027540	45.22	−73.85	38.1	6.9	6.6	7.8	6.4
31	Sutton	7028292	45.07	−72.68	243.8	0.4	0.6	0.7	0.5
32	Vercheres	7028700	45.77	−73.37	21.0	3.5	3.5	4.7	2.2