INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **30**: 1564–1576 (2010)
Published online 5 August 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1992

RMetS
Royal Meteorological Society

Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach

C. Simolo,[a] M. Brunetti,[a]* M. Maugeri[a,b] and T. Nanni[a]
[a] *ISAC-CNR, via P. Gobetti 101, I-40129 Bologna, Italy*
[b] *Dipartimento di Fisica – Università degli Studi di Milano, via Celoria 16, I-20133 Milano, Italy*

ABSTRACT: This work presents a novel method for estimating missing values in daily precipitation series. It is aimed at identifying the event time location with good accuracy and reconstructing the correct amount of daily rainfall. In addition, the statistical properties of the time series, i.e. both probability distribution and long-term statistics, are preserved. The completion method is based on a two-step algorithm that uses information from a cluster of neighboring stations. First, wet and dry days are tagged, and subsequently, the full precipitation amount for wet-classified days is estimated by a modified multi-linear regression approach. This method avoids overestimation of the number of wet days and underestimation of intense precipitation events, which are typical side effects of common regression-based approaches. Copyright © 2009 Royal Meteorological Society

KEY WORDS daily precipitation; missing values; probability density function; interpolation

Received 2 December 2008; Revised 20 May 2009; Accepted 7 July 2009

1. Introduction

The handling of missing data in meteorological time series is a relevant issue to many climatologic analyses, such as studies of droughts and above/below-threshold events, such as those based on the well-known indexes defined within the joint World Meteorological Organization Commission for Climatology/World Climate Research Program project on Climate Variability and Predictability and Expert Team on Climate Change Detection, Monitoring and Indices (Alexander *et al*., 2006).

Almost all the instrumental time series are affected by a percentage of missing values. Lack of data concerns not only the early instrumental period, due to loss of yearbooks because of wars or fire accidents etc., but also the most recent period, due, for example, to occasional interruptions of automatic stations, instrument malfunctions and network reorganizations. A way out of this difficulty is to exclude periods with missing values from data analysis, or to ignore the problem if their amount is not very large. Such approaches, however, may disregard valuable information and can induce biases in many climate investigations.

To overcome the problem, a number of interpolation techniques have been developed over the decades, aimed at estimating missing observations in climatic time series, mainly on a monthly and seasonal basis. Methods for handling missing data with daily resolution, on the other hand, are scarce and show marked errors, even though such methods perform well at lower resolution time scales (e.g. DeGaetano *et al*., 1995; Xia *et al*., 2001). The situation becomes particularly complicated when dealing with precipitation, because of its large space and time variability; moreover, in this case, the problem is twofold, since both time location and rainfall amount of each single-day event must be reconstructed. Thus, performing accurate estimates of missing data in daily precipitation records remains a difficult task, even more so if long time series and coarse rain-gauge networks are considered.

Within-station methods for estimating missing observations in climate series are the simplest approaches. They are self-contained, in that they only use data from the series that is being filled, by replacing missing data, e.g. by the mean of values on previous and subsequent days, or by the series mean value (Kemp *et al*., 1983). In spite of their simplicity, such methods are suitable for variables with high autocorrelation and for calculating long-term averages, and are thus essentially useless as far as daily precipitation is concerned.

Traditional techniques for filling-in gaps in precipitation series, both on monthly and daily time scales, are mainly based on spatial interpolation, that is, imputed values at a target station are calculated by using synchronous observations from surrounding stations. The inverse distance weighting method (Cressman, 1959; Shepard, 1968) is one of the most commonly used distance-based schemes; basically, it amounts to computing a weighted average by using inverse squared distance between target and surrounding stations as weighting factors, the underlying assumption being the existence of positive spatial correlation between data from nearby observation sites. However, naturally, because distance

* Correspondence to: M. Brunetti, ISAC-CNR, Via P. Gobetti, 101, I-40129 Bologna, Italy. E-mail: m.brunetti@isac.cnr.it

alone is not enough to define a similarity criterion for precipitation time series, and the selection of surrounding stations is critically important to the accuracy of the results, several variants and adds-on to this method have been devised. For instance, inverse squared distance can be replaced by higher powers or by negative exponential functions of distance (e.g. Teegavarapu and Chandramouli, 2005; Garcia *et al*., 2008), or by topographical relationships that incorporate orographic effects (Daly *et al*., 1994; Lloyd, 2005). Alternatively, the use of correlation coefficients between data series as weights has been investigated on a daily basis (Teegavarapu and Chandramouli, 2005; Ahrens, 2006), and generally found to outperform distance-based schemes.

Data correlation between time series is also exploited in the revised normal ratio method (Young, 1992; Tang *et al*., 1996), originally developed by Paulhus and Kohler (1952); weight factors are related to daily (or monthly) correlation coefficients between target and surrounding stations and to the number of points the correlation coefficient is based on.

Recent conceptual improvements to inverse distance weighting method and the revised normal ratio method can be found, for example, in Teegavarapu and Chandramouli (2005) and Suhaila *et al*. (2008). Furthermore, simplified schemes, including the so-called 'closest station method' and 'single best estimator' (e.g. Wallis *et al*., 1991; Eischeid *et al*., 2000; Xia *et al*., 2001), are also currently used for estimating missing values in precipitation series.

It should be noted that, apart from the arbitrariness in choosing weights and defining a reliable measure of distance between observation sites, an obvious limitation of weighting approaches is the overestimation of the number of rainy days.

More sophisticated spatial interpolation methods to fill in gaps in precipitation series estimate the functional relationships between target and surrounding stations. These are, for example, spline-surface fitting (e.g. Hutchinson and Gessler, 1994), statistical methods, such as optimal interpolation and kriging (Creutin and Obled, 1982, and references therein), regression-based approaches with conventional least squares or least absolute deviations criterion (Tabios and Salas, 1985; Beauchamp, 1989; Eischeid *et al*., 1995, 2000; Bennis *et al*., 1997; Schneider, 2001). The efficacy of these methods for the interpolation of precipitation data has been investigated by a number of comparative studies, both on a monthly and daily basis (e.g. Ashraf *et al*., 1997; Teegavarapu and Chandramouli, 2005; Ramos *et al*., 2008), and generally found to be greater than that of weighting approaches. In particular, Eischeid *et al*. (2000) demonstrated that multi-linear regression (MLR) outperforms most of the commonly used techniques concerning missing data handling in daily resolution precipitation series.

However, regression-based methods, similarly to weighting methods, suffer from the overestimation of the number of rainy days; furthermore, the rainfall probability distribution is not preserved, in that heavy precipitation events are systematically underestimated.

Statistical properties of precipitation time series are exploited in the procedure devised by Karl *et al*. (1995) (Karl and Richard, 1998; Brunetti *et al*., 2001a, 2001b; Brunetti *et al*., 2004), which is based on the fit of the two-parameter Gamma distribution (Bradley *et al*., 1987) to each station's daily data; subsequently, to determine whether precipitation occurs on any missing day, a random number generator is used, with probability distribution given by the empirical probability of precipitation for that day. Then, the Gamma distribution and a random number generator are used once again to estimate rainfall amount in wet-classified days. Even though the average statistical properties of precipitation series (i.e. number of rainy days and total rainfall amount on monthly and yearly time scales) are preserved, such an approach is of little use on a daily time scale because it randomly locates precipitation events.

Finally, in addition to the above-discussed methods for imputation of missing precipitation values, data-driven approaches based on neural network algorithms should also be remembered (Gupta and Lam, 1996; Elshorbagy *et al*., 2000a, 2000b; Khalil *et al*., 2001), i.e. on learning relationships between data from target and surrounding stations. Recent developments along this line can be found, for example, in Teegavarapu and Chandramouli (2005), Boulanger *et al*. (2007) and Coulibaly and Evora (2007). In Coulibaly and Evora (2007), in particular, different architectures of artificial neural network have been investigated and their effectiveness at estimating missing daily precipitation values compared; the authors found quite accurate results for the best-performing models, even though observed data statistics are not systematically preserved.

The purpose of this study is to present an objective automated procedure for imputing missing observations in daily precipitation series that enables the reconstruction of complete rainfall data sets. Unlike commonly used filling-in methods, the developed algorithm has the advantage of preserving both statistical properties of precipitation time series (probability distribution and long-term statistics) and single-day peculiarities (time location of precipitation events).

The method is based on a two-step algorithm, i.e. first rainfall occurrence (wet or dry day) is analyzed and subsequently rainfall amount for each wet-classified day is reconstructed. The case study area for the method validation is the Reno River basin in northern Italy. We considered a set of historical daily precipitation records from 1916 up to 2004, available at 36 rainfall-gauging stations. Details on the data set are provided in Section 2. Section 3 outlines the basics of the method. In Section 4, we extensively discuss the model performances, by first analyzing the determination of rainfall occurrence, and subsequently focusing on relevant aspects of rainfall amount reconstruction. Complete daily results and the

related uncertainty analysis are also provided. Finally, conclusions are drawn in Section 5.

2. Case study area

A data set of daily precipitation series at 36 Italian stations is used in this work as a baseline to investigate the effectiveness of the reconstruction procedure. These series were selected from a larger database because of their record length and completeness level. The rainfall-gauge network under consideration spans the area from about 44.0° to 44.7° north and from 10.8° to 12.0° east (Reno River basin in northern Italy), as illustrated in Figure 1. Elevation of the measuring sites varies from a few meters up to 900 m above sea level. Because of the orographic aspects, the region shows different precipitation regimes, with yearly total precipitation roughly ranging from 600 up to 2000 mm.

The data used, provided and validated by the Italian Autorita' di Bacino del Reno, cover the time period from 1916 up to 2004. The fraction of missing observations in each series does not exceed ∼8%. Table I summarizes each station name, record length, percentage of missing values, total yearly precipitation and number of wet days per year, both averaged across only complete years; additionally, mean and standard deviation (SD) of precipitation intensity are reported for each series. Figure 2 shows the data availability during the 1916–2004 period.

3. Basics of the method

Estimating missing values in daily precipitation series by a conventional MLR approach, as already noted, has the drawback of overestimating the number of rainy days and underestimating heavy precipitation as well. To circumvent the problem, we propose a two-step procedure that preserves both the correct event time location (wet/dry days) and the probability distribution function of daily precipitation. Indeed, first rainfall occurrence is estimated by a weighting-based method appropriately modified by the introduction of a wet/dry threshold; subsequently, rainfall amount of wet-classified days is estimated by a multivariate fit, and in turn, the generated values are rescaled to recover the original precipitation probability distribution on a daily time scale. The two steps of the reconstruction procedure are discussed in detail in sections 3.1 and 3.2, respectively.

The method is based on the fit to daily data of the Gamma distribution, which is believed to represent precipitation phenomena reliably (Bridges and Haan, 1972; Stern and Coe, 1984; Bradley *et al*., 1987; Wilks, 1995; Groisman *et al*., 1999; Nicholls and Murray, 1999; Dunn, 2004; Jones *et al*., 2004). To further support the use of the Gamma distribution, a Kolmogorov–Smirnov test has been applied to each station's daily data (a total of more than 13 000 daily Gamma distributions have been tested). In fact, we applied the Lilliefors test, a variant of the Kolmogorov–Smirnov test to be used when the distribution parameters have been fit to the same data used in the test (Wilks, 1995), as in our case. The significance level of 0.10 has been assumed, i.e. we reject the null hypothesis that the theoretical distribution is performing adequately in modeling the empirical values when the p-value is below 0.10. We found that in more than 70% of the analyzed cases, the Gamma distribution was not rejected.

The goodness-of-fit is also clear from the quantile-plot in Figure 3, where the theoretical values of cumulative distribution functions are plotted *versus* the empirical ones, for the 366 daily precipitation distributions of the station S25, chosen as an example.

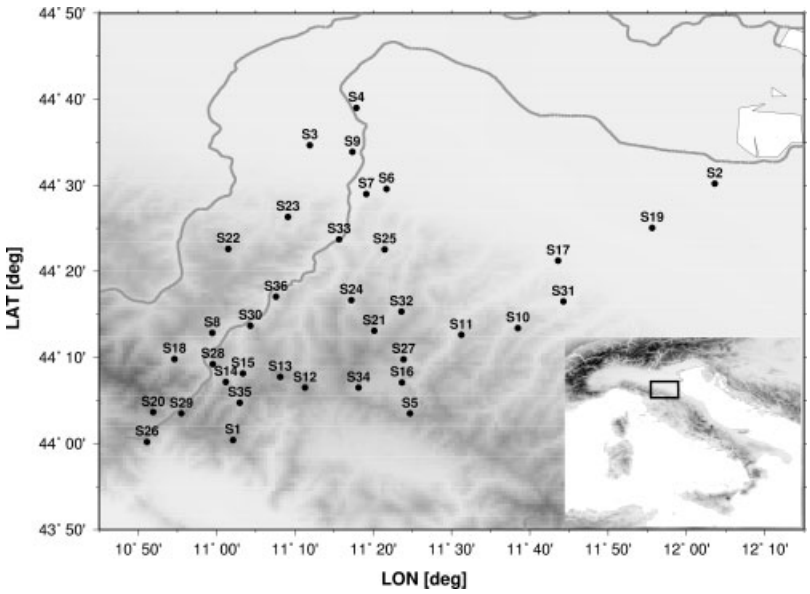

Figure 1. Location map of the study area. Stations are identified by a sequential code defined in Table I.

Table I. Summary of meteorological stations used for the method validation. Station codes and names are reported along with their record period and percentage of missing data. The fifth and sixth columns give the total yearly precipitation and the number of wet days per year, both averaged across only complete years; the last two columns give the mean intensity (i.e. total precipitation over the number of rainy days across the entire record period) and the associated standard deviation (SD).

Station code	Name	Period of record	Missing values (%)	Yearly total (mm/yr)	Wet days	Mean intensity (mm/day)	SD (mm/day)
S1	Acquerino	1929–2004	4.2	2002.7	117	17.1	19.2
S2	Alfonsine	1916–2004	0.6	684.4	76	9.0	10.5
S3	Anzola dell'Emilia	1935–2004	1.7	743.9	81	9.2	10.9
S4	Bagno di Piano	1918–2004	3.4	623.6	75	8.4	9.0
S5	Barco	1924–2004	6.5	1511.4	113	13.4	13.8
S6	Bologna Idrografico	1934–2004	0.7	767.8	80	9.6	10.9
S7	Bologna San Luca	1922–2004	5.0	809.9	81	10.0	11.6
S8	Bombiana	1924–2004	4.6	1151.4	103	11.2	12.3
S9	Calderara di Reno	1922–2004	2.3	707.1	79	9.0	10.1
S10	Casola Valsenio	1920–2004	4.4	907.0	89	10.1	11.8
S11	Castel del Rio	1920–2004	2.6	1061.6	93	11.4	12.9
S12	Cottede	1937–2004	3.8	1525.6	114	13.4	15.4
S13	Diga del Brasimone	1916–2004	1.5	1476.6	111	13.3	15.8
S14	Diga di Paviana	1947–2004	0.0	1448.4	105	13.7	15.9
S15	Diga di Suviana	1947–2004	0.1	1242.8	101	12.3	13.7
S16	Firenzuola	1920–2004	3.2	1285.9	106	12.2	14.2
S17	Imola	1919–2004	2.1	778.7	84	9.3	11.2
S18	Lizzano in Belvedere	1919–2004	5.8	1533.4	111	13.8	16.2
S19	Lugo di Romagna	1919–2004	2.5	757.2	77	9.9	11.5
S20	Maresca	1930–2004	4.0	1978.3	123	16.0	20.0
S21	Monghidoro	1920–2004	2.4	1113.7	99	11.3	13.0
S22	Monteombraro	1918–2004	0.6	959.1	92	10.5	12.2
S23	Monte San Pietro	1926–2004	2.0	854.4	86	9.9	11.6
S24	Monzuno	1920–2004	1.6	981.8	93	10.6	11.9
S25	Pianoro	1919–2004	2.4	894.6	87	10.3	12.0
S26	Piastre	1919–2004	1.5	2077.2	114	18.3	21.9
S27	Pietramala	1920–2004	2.0	1410.1	107	13.2	14.5
S28	Porretta Terme	1916–2004	0.7	1261.5	98	12.8	14.7
S29	Pracchia	1926–2004	0.1	1935.0	121	16.0	20.1
S30	Riola di Vergato	1920–2004	4.0	980.3	93	10.5	11.4
S31	Riolo Terme	1920–2004	7.1	825.2	85	9.7	11.7
S32	S. Benedetto del Querceto	1920–2004	7.6	995.3	94	10.5	11.9
S33	Sasso Marconi	1923–2004	5.3	873.2	87	10.0	11.4
S34	Traversa	1938–2004	2.4	1622.6	112	14.5	17.0
S35	Treppio	1920–2004	2.2	1768.2	109	16.2	19.6
S36	Vergato	1919–2004	5.2	854.1	91	9.4	10.4

3.1. Wet/dry event identification

Determination of a wet or dry event in the series considered (target series) is first based on a weighted average of synchronous precipitation data from surrounding stations (reference series). To deal with standardized values, daily precipitation records from the reference series are preliminary converted into probability values by using the two-parameter Gamma distribution.

Specifically, the shape and scale parameters of the Gamma distribution are estimated for each wet day of the reference series on the basis of the maximum likelihood. The data sample used to estimate Gamma parameters is provided by a running window of variable width centered on each Julian day and year examined; starting with a fixed number of days and years (typically, 31 days and 25 years), the window is gradually enlarged forward and backward in time until a minimum number of data (i.e. at least 150 data) is reached. In fact, rather than the probability density, we used the cumulative distribution, i.e. the incomplete Gamma function,

$$P(x, \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \int_0^{x/\beta} z^{\alpha-1} \mathrm{e}^{-z} \mathrm{d}z \qquad (1)$$

to ensure a one-to-one relation between precipitation amount and probability for each day. Thus, millimeter values in the reference series are replaced by the corresponding probabilities as given by Equation (1), whereas zero entries as well as missing values are left unchanged.

A complete replica of the target series is then generated day-by-day in terms of probability using a weighted average of synchronous (probability) values from the reference series. Weighting factors are given by Gaussian-like

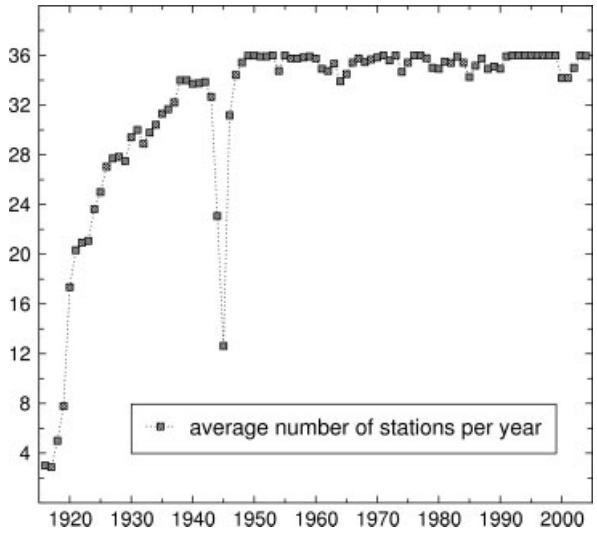

Figure 2. Mean data availability year-by-year over the period 1916–2004 for the 36 daily precipitation series used.

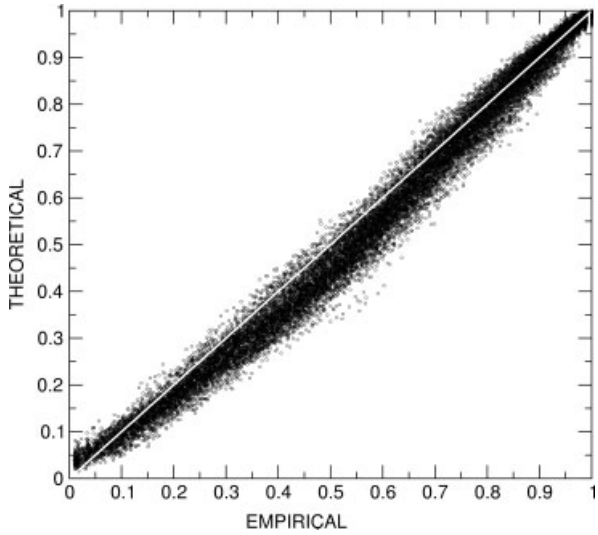

Figure 3. Quantile-plot: comparison between the theoretical and empirical values of the 366 daily cumulative distribution functions for the station S25 chosen as an example.

functions of local coordinates associated with the stations involved. Specifically, they are functions of distance, elevation difference and relative angular distribution with respect to the series under consideration,

$$w_i^d(x, y) = \exp\left(-\frac{d_i^2(x, y)}{c_d}\right),$$

$$w_i^h(x, y) = \exp\left(-\frac{\Delta h_i^2(x, y)}{c_h}\right)$$

and

$$w_{i\,(x,y)}^{\text{ang}} = 1 + \frac{\sum_{j\neq i} w_j^d(x, y) w_j^h(x, y)\left(1 - \cos\theta_{(x,y)}(j, i)\right)}{\sum_{j\neq i} w_j^d(x, y) w_j^h(x, y)} \quad (2)$$

respectively, where the indices i and j run over all the reference stations, and (x, y) stands for the target station location. $d_i(x, y)$ and $\Delta h_i(x, y)$ denote respectively the distance and the elevation difference between the target and i-th reference series; the coefficients c_d and c_h are defined as

$$c_d = \frac{\overline{d}^2}{\ln 2} \text{ and } c_h = \frac{\overline{h}^2}{\ln 2} \quad (3)$$

i.e. so that the related weighting factors reduce to 0.5 for fixed value of distance $\overline{d}$ and elevation difference $\overline{h}$. The last term in Equations (2) is used to correct for potential anisotropies of the information from the reference series (Shepard, 1968, 1984; Willmott *et al.*, 1985), $\theta_{(x,y)}(j, i)$ being the angular separation between the j-th and i-th reference series with respect to the target station location (x, y). The total weight associated with the i-th reference series is finally given by the product of the three terms in Equations (2), namely

$$w_i(x, y) = w_i^d(x, y) w_i^h(x, y) w_i^{\text{ang}}(x, y). \quad (4)$$

The exponential slowdown of weighting factors, controlled by the two parameters c_d and c_h, takes into account the poor spatial correlation of precipitation. Values of c_d and c_h must be fixed in advance with relation to the rain-gauge network density and complexity of terrain (Section 4), and act as objective selection criteria of surrounding stations for the target series reconstruction.

As stated, to obviate the systematic overestimation of rainy days, a key point here is the introduction of a time-dependent wet/dry threshold. Indeed, obviously, just a few (or even one) non-zero probability values from surrounding stations on a given day are sufficient to generate a rainy event in the target series on that day; in that case, the estimated probability value is typically very low, and the corresponding event is most likely to be a dry one. The threshold thus serves to discriminate between wrong estimates and actual precipitation events, recovering the correct number and time location of wet/dry days in the target series.

The threshold calculation is based on a data sample obtained by the same technique used above for the calculation of the Gamma parameters. Specifically, a running window of variable width is again considered around each day and each year of the target series recovered in terms of probabilities, requiring in this case at least 1000 values; the data included are only those corresponding to non-missing observations in the original (incomplete) series, and the elements are reordered according to decreasing values. The probability value associated with the n-th element plays the role of the wet/dry threshold, n being the number of rainy days within the same subset extracted from the original series.

On this basis, a complete wet/dry series associated with the target series is generated in terms of binary entries, i.e. 1 if the reconstructed probability turns out to be above or equal to the threshold, otherwise 0.

3.2. Estimating daily precipitation amount

The binary series of wet/dry days obtained in the previous step is used as input for the reconstruction of the rainfall amount in wet-classified days via a MLR with ordinary least squares. This implies the minimization of the sum of squared residuals for each non-zero entry of the input series, namely

$$\chi^2 = \sum_i (x_i^{\mathrm{obs}} - x_i^{\mathrm{mlr}})^2, \;\; x_i^{\mathrm{mlr}} = \beta_0 + \sum_{j=1}^{N} Y_{ij}\beta_j. \qquad (5)$$

In Equation (5), x^{obs} denotes the vector of observations extracted from the original series, whereas x^{mlr} is the vector of values calculated by the model; the columns of matrix Y are given by the observations of the N explanatory variables, i.e. the reference series. Index i runs over the data sample used for the estimation of the $N+1$ time-dependent regression coefficients β_0 and β_j. These data are provided as a rule by a running window centered on the day that is being reconstructed, and including only non-missing values in both the target (incomplete) and reference series.

Actually, to improve the data sample for each day, not all the available series are included in the fit; instead, an upper bound of 20 series with the largest weighting factors as given by Equation (4) is imposed. The best-fit coefficients are then used for calculating precipitation amount of each wet-classified day, according to the second of Equations (5).

As already pointed out, the time-dependent probability distribution is not preserved by the MLR approach in that, even though the number of events and their time location have been constrained by the previous step, heavy precipitation events are typically underestimated. To correct for the bias induced by the fit, the generated values are forced to satisfy the daily probability distribution associated with the original series. Specifically, the cumulative probability defined by Equation (1) $\hat{P} = P(x^{\mathrm{mlr}}; \alpha^{\mathrm{mlr}}, \beta^{\mathrm{mlr}})$ is calculated for every positive value x^{mlr} of the MLR-series, and the equation

$$P(z; \alpha^{\mathrm{obs}}, \beta^{\mathrm{obs}}) = \hat{P} \qquad (6)$$

is solved for z, where $P(z; \alpha^{\mathrm{obs}}, \beta^{\mathrm{obs}})$ denotes the time-dependent cumulative probability function derived from the original series. Precipitation value x^{mlr} reconstructed by MLR is finally replaced by the rescaled one z as given by Equation (6) and denoted by x^{rmlr} in what follows.

The correct distribution function for the rescaled MRL-series (RMLR-series hereinafter) is thus recovered, and, as a result, determination of heavy precipitation events is improved, as demonstrated in Section 4.2.

4. Results and discussion

The method validation was performed with reference to the rain-gauge network presented in Section 1, by using a jackknife-like procedure, that involves the removal of subsets of data from the target series before reconstruction is carried out. This avoids 'self-influence' of the observations that are being estimated. Specifically, one year of the target series at a time was fully discarded, together with a n-year long window centered on that year, for fixed n, and subsequently reconstructed. Imputed data were finally compared with the original (removed) ones to assess the accuracy of the results.

In the following, the cutoff distance and elevation difference in Equation (3) have been fixed as a rule to $\overline{d} = 50$ km and $\overline{h} = 350$ m, respectively. The performances of the method concerning time location of rainfall occurrence are discussed in detail in Section 4.1, whereas Section 4.2 focuses on the impact of probability rescaling on intense precipitation events. In Section 4.3, detailed results from the full reconstruction algorithm are then illustrated, for the sake of brevity, with reference to one representative series. The related uncertainty analysis is provided, for comparative purposes, by using conventional error measures both on a daily and monthly basis.

4.1. Accuracy of rainfall occurrence estimation

To begin with, the stability of the method with respect to the gap width in the target series was studied by varying the gap from 1 to 15 consecutive years centered on the year being reconstructed. The present analysis was restricted to 1951–2000 since, as it is clear from Figure 2, data availability is at its maximum and roughly constant within this time period. This choice allows us to disentangle the dependence of the results on the gap width in the target series from that on the number of reference series used.

The absolute number of days incorrectly reproduced over the total number of reconstructed days was computed across 1951–2000 for each one of the 36 daily precipitation series considered. The absolute percentage error is shown in Figure 4 as a function of the gap width. The most striking aspect is the robustness of the algorithm if the amount of missing data in the target series is significantly increased. Indeed, for each series, the related error rises very slowly with the gap width, i.e. by less than 1% from a 1-year up to a 15-year gap, and, on average across all the series, it varies from a mean of 10.2% (ranging from 7.2 to 14.1% for the different stations) to a mean of 10.4% (ranging from 7.3 to 14.4%).

In addition, no systematic tendency toward overestimation or underestimation of the number of rainy days was observed, as it is evident from Table II, where the total number of estimated and observed wet and dry days over the entire data set are given for a gap width of 5 years. In Table II, also false alarm and hit rate are both reported for wet, dry and total days. Moreover, it should be stressed that if a tolerance of one day is accepted on the wet/dry event location, the error drastically decreases from an average value across all the series of 10.2 down to 2.3% (for a 5-year gap).

Actually, a 1-day tolerance turns out to be a reasonable criterion for the error estimate, because the record time is

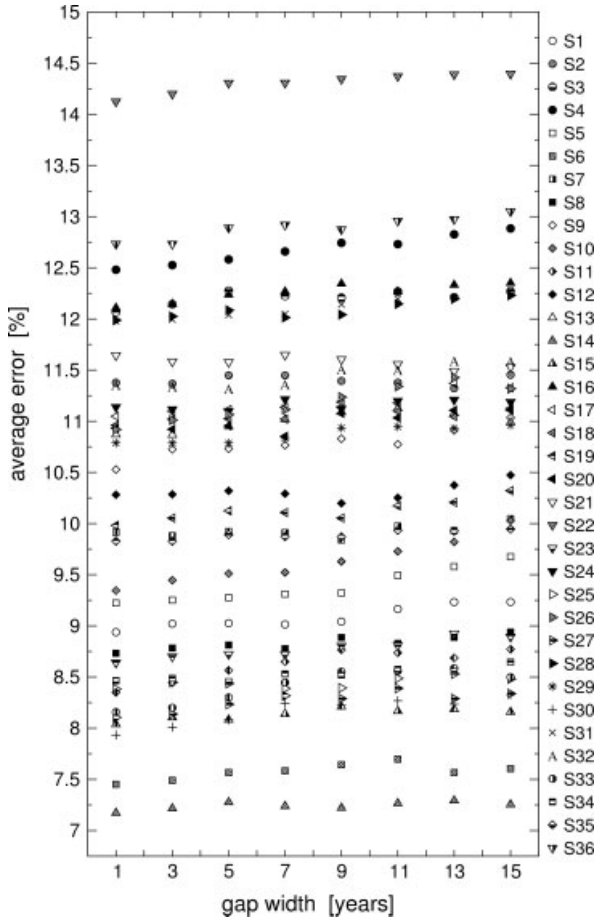

Figure 4. Percentage error in the wet/dry day determination as a function of the gap width in the target station. The uncertainty is computed as the absolute number of days incorrectly identified over the total number of daily estimates across 1951–2000, for each series.

itself uncertain (such information is not always reported) and possibly varies among the stations: The measures may refer to the 24 h before midnight, before 9 a.m., or before 9 p.m., and this may cause ±1 day error in the time location of precipitation events.

In addition, being a common practice to consider a day as wet only if the precipitation amount exceeds 1 mm/day (Nicholls and Murray, 1999), we also evaluated the error in the wet/dry day identification setting all input data below 1 mm equal to zero. In this case, the mean error (ME) decreases down to 7.6 and 1.8%, respectively without and with a tolerance of ±1 day in the event location. Nonetheless, since the threshold value is somewhat arbitrary, to maintain the discussion at a general level, we will not consider any threshold on the daily precipitation amount in what follows.

Finally, the robustness of the algorithm with relation to the number of surrounding stations used for the target series reconstruction was also evaluated, since in practice the number of stations might not be the same over time due to missing observations in the reference series as well. Specifically, the number of incorrect daily estimates over the total number of reconstructed days was computed across 1951–2000, by varying the number of reference series from 35 down to 2, for a fixed gap width of 5 years. The absolute percentage errors are shown in Figure 5 for all the series. As can be seen, decreasing the number of reference series does not critically affect the results, as long as the number of input stations is still greater or equal to 5; on the other hand, the level of accuracy is not significantly enhanced if the number of reference series exceeds 20.

4.2. Rescaling time-dependent probability: heavy precipitation

To substantiate the whole methodology outlined in Section 3, the reconstruction of rainfall occurrence and precipitation amount was carried out in turn for each series over its own record period. As a rule, a 5-year gap was applied to the target series, around each year being reconstructed. This choice, which is a rather pessimistic one as far as daily precipitation is concerned, was made to test the model effectiveness at reconstructing highly incomplete series.

In Figure 6, the Gamma probability densities associated with both the MLR- and RMLR-series (dashed and dot-dashed lines, respectively) are compared with that derived from the original series (solid line); the parameters refer to series S25 (Table I), here used as a benchmark, and represent average values across the series record period, i.e. 1919–2004. As can be seen, the probability density associated with the MLR-series deviates from that of the original series mostly for heavy precipitation, whereas the correct (i.e. original) distribution function is recovered on readjusting MLR-values according to Equation (6). As a result, the bias induced by the MLR estimation turns out to be significantly reduced, in particular as far as heavy precipitation is concerned, as

Table II. Hit and false event rate for the wet, dry and total reconstructed events (estimated over the entire data set for a gap width of 5 years).

		Observed	Reconstructed	Hit	Hit ±1 day	False	False ±1 day
Wet	Days	220 953	220 388	187 516	210 869	33 437	5 566
	%	34.1	34.0	84.9	95.4	15.1	2.5
Dry	Days	427 355	427 920	394 483	422 354	32 872	9 519
	%	65.9	66.0	92.3	98.8	7.7	2.2
Total	Days	648 308	648 308	581 999	633 223	66 309	15 085
	%	100	100	89.8	97.7	10.2	2.3

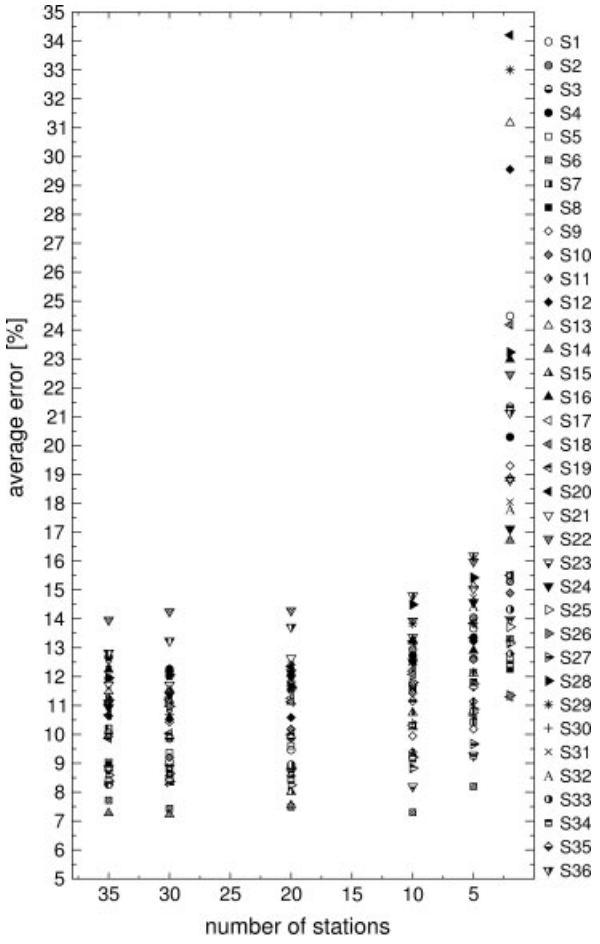

Figure 5. Percentage error in the wet/dry day determination as a function of the number of stations used for the target station reconstruction. For each station, the error is given as in Figure 4 by the absolute number of days incorrectly identified over the total number of daily estimates across 1951–2000.

can be explicitly verified by comparing the uncertainties related to MLR- and RMLR-values.

Indeed, for each series, precipitation events above the 90th percentile (estimated for each Julian day) were extracted, and the ME

$$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i, \quad \varepsilon_i = x_i^{\mathrm{est}} - x_i^{\mathrm{obs}} \tag{7}$$

was calculated for both MLR- and RMLR-values as a measure of the systematic uncertainty. Residuals ε_i in Equation (7) are given by the difference between the estimated daily value (the MLR- and RMLR-value in the two cases, i.e. x_i^{mlr} and x_i^{rmlr}, respectively) and the corresponding observation, and the sum runs over all non-missing precipitation values exceeding the 90th percentile. As expected, an overall reduction of the ME absolute value was observed after probability rescaling. Specifically, for each series, the ME associated with the RMLR-values turns out to be always smaller in absolute value than that associated with the MLR-values, with respective averages of 0.6 mm/day and −2.9 mm/day across all the series. If these deviations are normalized to

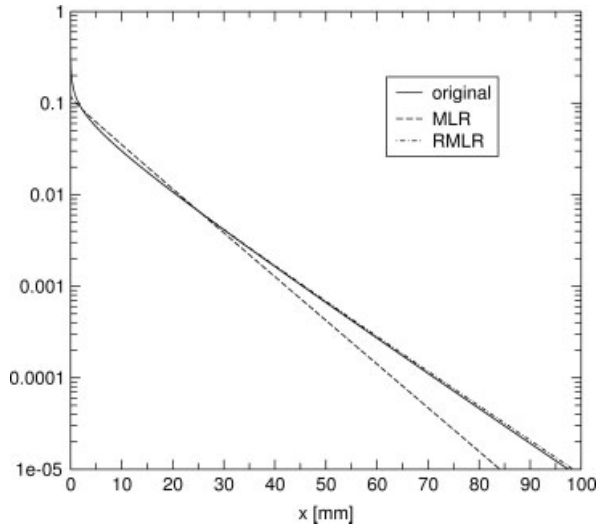

Figure 6. Comparison between the Gamma probability density function associated with the original series (solid line) and those associated with the MLR- and RMLR-series (dashed and dot-dashed lines, respectively). The shape and scale parameters refer to series S25 (Table I) and are computed in each case by averaging across the entire record period (1919–2004).

the mean precipitation per event above the 90th percentile for each series, the bias is on average 1.2 and −6.3% in the two cases, respectively, and is thus significantly reduced in absolute value after probability rescaling.

Finally, to give some further insights into the reconstruction of precipitation amount, we consider also the mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}(\varepsilon_i) \tag{8}$$

where residuals ε_i denote differences between estimated (RMLR) and actual values. The calculation is restricted only to true wet days over the entire data set, clustered into ten categories defined by the percentile thresholds from the 10th up to the 90th. The results are given in Table III for each category together with the associated mean precipitation amount. As can be seen, the MAE increases very slowly from the first to the last decile, so that the percentage error decreases with increasing the mean precipitation amount; this is a positive aspect of the reconstruction method, in particular from the hydrological point of view, since high relative errors on low precipitation events and low ones on heavy events do not significantly affect the annual/seasonal/monthly precipitation budget. In the following section, a comprehensive estimate of the algorithm's performances, including also the uncertainty on the dry days reconstruction, is provided.

4.3. Assessment of uncertainty

The complete results of the reconstruction algorithm are shown as an example for the series S25 in Figure 7. Here, imputed values are directly compared with actual observations on a day-to-day basis. By way of illustration,

Table III. Mean absolute error (MAE) evaluated on all the original wet days clustered into deciles. The mean precipitation per event of each decile is also indicated in the first column.

	Mean decile intensity (mm)	MAE (mm)
Decile 1	1.2	1.6
Decile 2	2.1	2.0
Decile 3	3.7	2.5
Decile 4	5.4	3.1
Decile 5	7.4	3.7
Decile 6	9.8	4.3
Decile 7	12.8	5.1
Decile 8	16.8	6.0
Decile 9	22.9	7.4
Decile 10	42.5	11.5

5 equal-spaced years, from the early up to the latest period, are displayed as representative of the entire series. A glance at Figure 7 reveals that the method allows a faithful reconstruction of both rainfall occurrence and precipitation amount. Similar results are also found for all the series considered.

To determine the accuracy of the daily estimates, we consider, along with the ME (7) and the MAE (8), the root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2} \qquad (9)$$

where, as in Equation (8), residuals ε_i denote deviations of RMLR-values from observed ones. In the evaluation of all these errors, the sums now run over all non-missing data in the target series, that is, over both dry and precipitation events. The ME, MAE and RMSE displayed in Figure 8 (denoted by squares, circles and diamonds, respectively) refer again to the benchmark series S25 and were computed year-by-year. As can be seen by Figure 8, essentially no systematic bias can be traced, the ME being always well below 1 mm/day in absolute value. Moreover, the MAE turns out to be quite moderate even in the early period, where the completeness level of the data set is rather low as can be seen by Figure 2; indeed, the MAE stays always below 2 mm/day, with an average value of 1.1 mm/day across the entire record period (1919–2004). The RMSE shows greater variability from year to year, due to its greater sensitivity to the larger discrepancies mostly related to heavy precipitation events. Then, also non-dimensional forms of the errors (7), (8) and (9) were estimated, by normalizing the uncertainties to the range of observed data (i.e. the maximum occurred value) for each year of the series under consideration. These are seen in Figure 9 as before for series S25.

Similar results were found also for the other series. Daily errors (7), (8) and (9), averaged across the record period of each series, are shown altogether in Figure 10. Indeed, the ME turns out to be generally negligible, with an average of −0.1 mm/day across all the series. Apart from very few exceptions, the MAEs gather around 1.4 mm/day, with the lowest error given by 0.7 mm/day and the highest one not exceeding 2.5 mm/day. The larger discrepancies, emphasized by the RMSE, concern

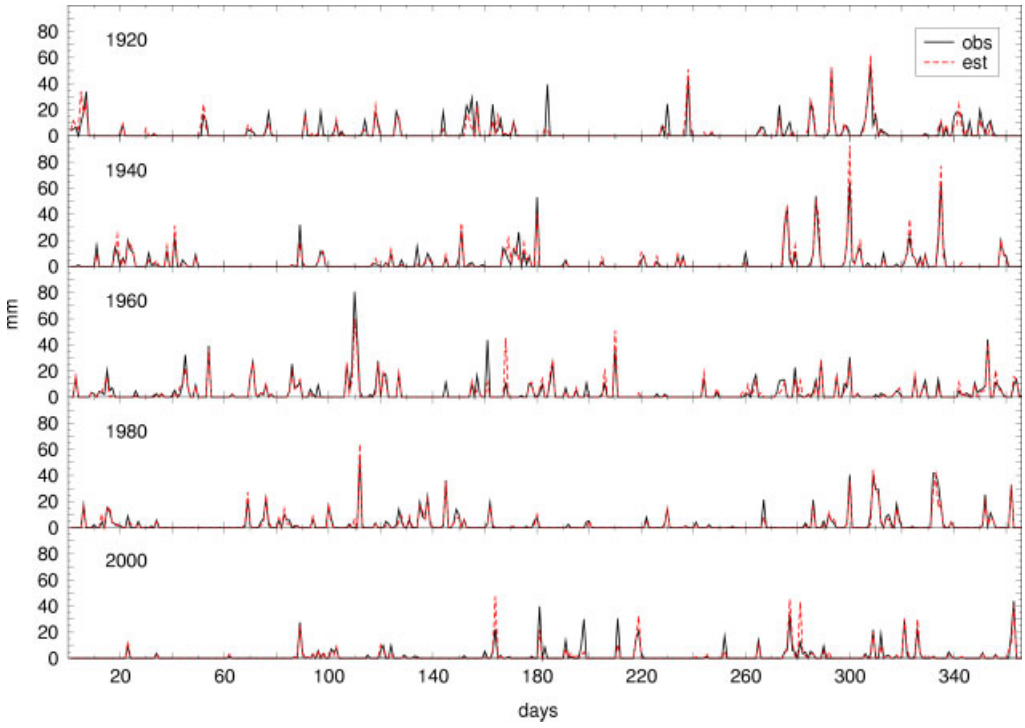

Figure 7. Direct comparison between daily estimates (rescaled multi-linear regression-values) and observed data (red dashed and black solid lines, respectively) for five representative years of series S25, from the early up to the latest period. This figure is available in colour online at www.interscience.wiley.com/ijoc

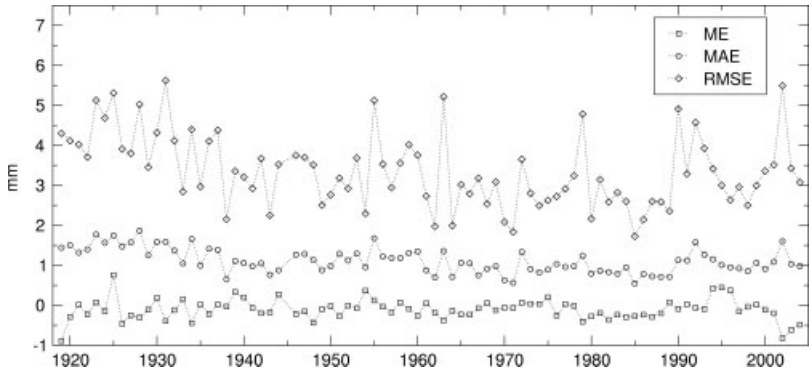

Figure 8. Representation of mean daily uncertainties *versus* years for series S25. The mean error, mean absolute error and root mean squared error (squares, circles and diamonds, respectively), as given by Equations (7), (8) and (9), are calculated by averaging across each year and are expressed in mm/day.

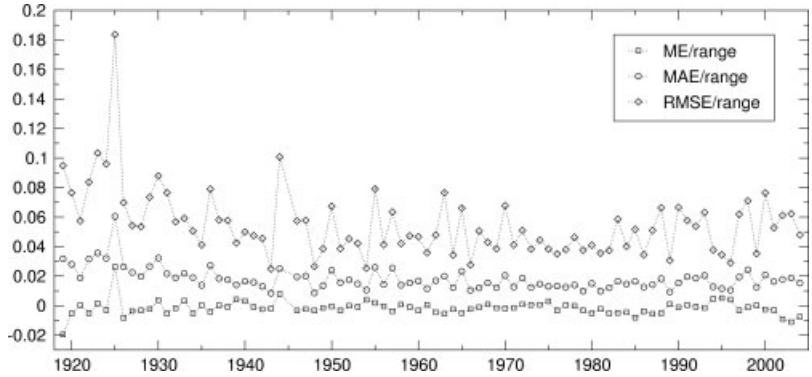

Figure 9. Year-by-year representation of the mean error, mean absolute error and root mean squared error (squares, circles and diamonds, respectively) for series S25, as in Figure 7, but normalized here to the range of data, i.e. to the largest value observed in each year.

in particular four time series (S1, S5, S26 and S35) from higher elevation stations, located at the boundaries of the area under consideration, as can be verified by Figure 1 and Table I.

Note that the uncertainty in the record time of stations was not taken into account in the error estimates discussed above; thereby, if 1-day tolerance was applied, the actual errors might be even less severe than those obtained here, as is the case for the time location of wet/dry events (Section 4.1).

As a cross-check of our results, we compared average climatological quantities extracted from the original and reconstructed series, namely, total monthly precipitation, number of rainy days per month and monthly precipitation intensity. Figure 11 shows, for example, the correlation coefficients between the reconstructed and the original series concerning total monthly precipitation. Correlation was calculated over the record period of each of the 36 series at hand, and by including only complete months in the original series. As can be seen, correlation coefficients turn out to be generally pretty near to the unit, thereby further substantiating the good agreement between the model predictions and observed data. However, the correlation has a pronounced seasonal cycle indicating that the method performs better during winter and transition seasons. Similar effects were also found by Eischeid *et al*. (2000). These results were also confirmed by the analysis of the number of rainy days per month and monthly precipitation intensity, and may be explained by the greater coherence of the atmospheric circulation during winter months.

Furthermore, since the daily probability distribution is preserved by construction, long-term statistics is faithfully reproduced. In particular, mean and SD of precipitation intensity (i.e. total precipitation over the number of rainy days across the entire record period) for the reconstructed series agree well with those calculated on the original series (Table I); discrepancies are found to vary from 0.09 to 5.06% for the mean and from 0.01 to 8.07% for the SD, with respective average values of 1.82 and 2.62% across all the series.

In the current literature, as already noted, little is devoted to the reconstruction of missing values in daily precipitation series. In addition, comparison between different investigations is hardly feasible, in that results are sometimes given on a monthly basis and conventional performance indexes are not always reported. Nonetheless, a rather exhaustive investigation of commonly adopted approaches to filling in gaps in daily precipitation can be found, for example, in Teegavarapu and Chandramouli (2005), where comparisons are made on the basis of standard error measures (e.g. MAE and RMSE); the authors concluded that correlation-based weighting methods and the neural network model adopted yield

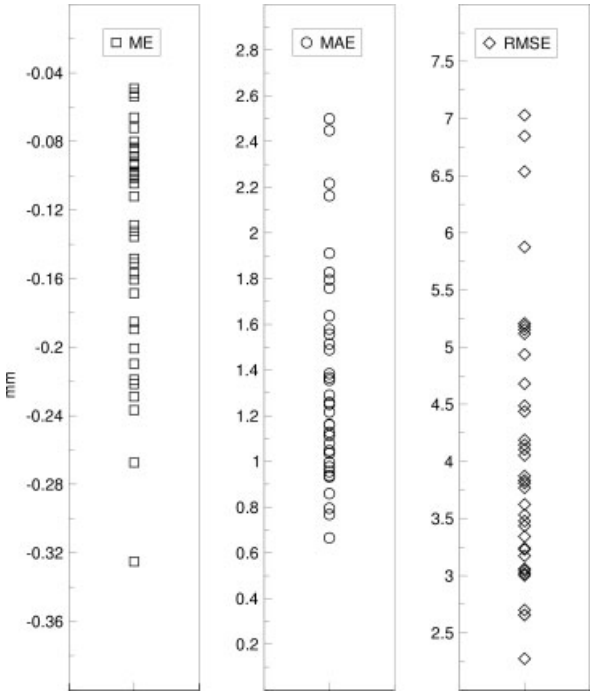

Figure 10. Representation of the mean error, mean absolute error, root mean squared error (squares, circles and diamonds, respectively) averaged across the record period for each series.

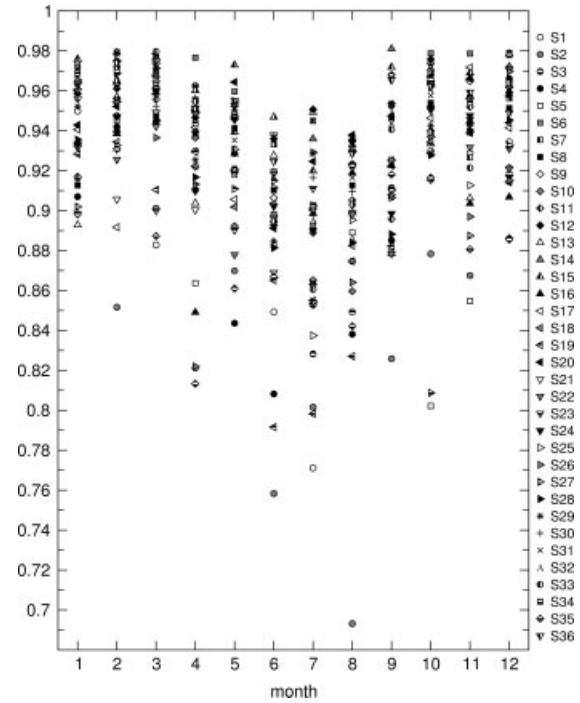

Figure 11. Correlation coefficients between estimated and actual monthly precipitation for all the series considered. Only complete months in the original series are included.

as a whole the best performances, with, e.g. rather low MAEs (about 2–3 mm/day). Additional insights into the performances of different neural network models can be found, for example, in Coulibaly and Evora (2007), where also MAEs are provided for the best-performing models (about 1–2 mm/day). We refer to the quoted literature for further details and results. Actually, it should be highlighted that accuracy of the results is naturally related to a number of factors, such as complexity of terrain, variety of meteorological regimes, network density and seasonality, which also make not straightforward comparisons with related works. Thus, individual model performances should be evaluated on a case-by-case basis, by taking into considerations all the aforementioned conditions.

5. Conclusions

This article discusses a novel method for estimating missing data in daily precipitation series, which is effective in both reconstructing single-day events and preserving the statistical properties on a daily time scale. The main focus here is to correct for common biases, such as the overestimation of the number of rainy days and the underestimation of intense rainfall, which affect traditional filling-in models.

The method performances were thoroughly investigated step by step, with reference to a rain-gauge network in northern Italy (Reno River basin), by using a jackknife procedure. A faithful estimation of uncertainty was also provided on the basis of conventional error indexes.

First, our results show that the method allows a quite accurate determination of rainfall occurrence, the overall uncertainty being around 10.8%, and even smaller (3.6%) if 1-day tolerance is accepted in the event time location. Furthermore, the algorithm proved to be quite robust with increasing gap width in the target series as well as with decreasing the number of reference series used. This issue is of importance when dealing with coarse rain-gauge networks and daily precipitation series affected by long gaps of many years.

Second, imposing the correct time-dependent probability distribution in wet-classified days significantly improves the reconstruction of intense precipitation events, thus reducing their systematic underestimation induced by MLR. Furthermore, unlike many currently used interpolation methods, basic data statistics, such as mean precipitation intensity and the associated SD for each series, were systematically preserved by our technique.

Thus, as our analysis has clearly shown, recovering the correct probability density function of daily precipitation series is a fundamental issue, even though this aspect is sometimes hidden, or completely ignored, in the current literature on this subject.

Finally, daily estimates as a whole show pretty small MAEs, with an averaged value of 1.4 mm/day across all the series; also, because the MEs are essentially negligible, no marked systematic bias can be traced. As already noted, however, geographical factors and rain-gauge network density associated with the area under consideration may influence the accuracy of the results and also complicate direct comparison with related works.

Bearing these limitations in mind, we conclude that the technique proposed in this work proved to be very well-performing for reconstructing missing values in daily precipitation series, in that it is reliable with regard to both event time location and precipitation amount estimation, and at the same time preserves all the statistical properties of time series. In addition, being a low time-consuming technique, it can be usefully exploited for creating large complete data sets, which are the basis of many climatologic investigations.

Acknowledgements

This study was carried out in the framework of the Italy–USA co-operation on Science and Technology of climatic change (2006–2008) funded by Centro Euro - Mediterraneo per i Cambiamenti Climatici (CMCC). The authors wish to thanks EU-COST-ACTION ES0601 'Advances in homogeneization methods of climate series: an integrated approach (HOME)'. Data were kindly provided by the Autorità di Bacino del Reno (http://www.regione.emilia-romagna.it/bacinoreno/). We also want to thank the two anonymous reviewers for making helpful comments and suggestions.

References

Ahrens B. 2006. Distance in spatial interpolation of daily rain gauge data. *Hydrology and Earth System Sciences* **10**: 197–208.

Alexander LV, Zhang X, Peterson TC, Caesar J, Gleason B, Klein Tank AMG, Haylock M, Collins D, Trewin B, Rahimzadeh F, Tagipour A, Rupa Kumar K, Revadekar J, Griffiths G, Vincent L, Stephenson DB, Burn J, Anguilar E, Brunet M, Taylor M, New M, Zhai P, Rusticucci M, Vazquez-Aguirre JL. 2006. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research* **111**: D05109. DOI:10.1029/2005JD006290.

Ashraf M, Loftis JC, Hubbard KG. 1997. Application of geostatisticals to evaluate partial weather station network. *Agricultural Forest Meteorology* **84**: 255–271.

Beauchamp JJ. 1989. Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin* **25**: 961–975.

Bennis S, Berrada F, Kang N. 1997. Improving single-variable and multivariable techniques for estimating missing hydrological data. *Journal of Hydrology* **191**: 87–105.

Boulanger JP, Martinez F, Penalba O, Segura EC. 2007. Neural Network based daily precipitation generator (NNGEN-P). *Climate Dynamics* **28**: 307–324.

Bradley RS, Diaz HF, Eischeid JK, Jones PD, Kelly PM, Goodess CM. 1987. Precipitation fluctuations over northern-hemisphere land areas since the mid-19th century. *Science* **237**: 171–175.

Bridges TC, Haan CT. 1972. Reliability of precipitation probabilities estimated from the gamma distribution. *Monthly Weather Review* **100**: 607–611.

Brunetti M, Maugeri M, Nanni T. 2001a. Changes in total precipitation, rainy days and extreme events in northeastern Italy. *International Journal of Climatology* **21**: 861–871.

Brunetti M, Colancino M, Maugeri M, Nanni T. 2001b. Trends in the daily intensity of precipitation in Italy from 1951 to 1996. *International Journal of Climatology* **21**: 299–316.

Brunetti M, Maugeri M, Monti F, Nanni T. 2004. Changes in daily precipitation frequency and distribution in Italy over the last 120 years. *Journal of Geophysical Research – Atmosphere* **109**: D05102. DOI:10.1029/2003JD004296.

Coulibaly P, Evora ND. 2007. Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology* **341**: 27–41.

Cressman GP. 1959. An operational objective analysis system. *Monthly Weather Review* **87**: 367–374.

Creutin JD, Obled C. 1982. Objective analysis and mapping techniques for rainfall fields: an objective comparison. *Water Resources Research* **18**: 413–431.

Daly C, Neilson R, Phillips D. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology* **33**: 140–158.

DeGaetano AT, Eggleston KL, Knapp WW. 1995. A method to estimate missing maximum and minimum temperature observations. *Journal of Applied Meteorology* **34**: 371–380.

Dunn PK. 2004. Occurrence and quantity of precipitation can be modeled simultaneously. *International Journal of Climatology* **24**: 1231–1239.

Eischeid JK, Baker CB, Karl TR, Diaz HF. 1995. The quality control of long-term climatological data using objective data analysis. *Journal of Applied Meteorology* **34**: 2787–2795.

Eischeid JK, Pasteris PA, Diaz HF, Plantico MS, Lott NJ. 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.

Elshorbagy AA, Panu US, Simonovic SP. 2000a. Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrological Sciences Journal (Journal Des Sciences Hydrologiques)* **45**: 849–866.

Elshorbagy AA, Panu US, Simonovic SP. 2000b. Group-based estimation of missing hydrological data: II. Application to streamflows. *Hydrological Sciences Journal (Journal Des Sciences Hydrologiques)* **45**: 867–880.

Garcia M, Peters-Lidard CD, Goodrich DC. 2008. Spatial interpolation in a dense gauge network for monsoon storm events in the southwestern United States. *Water Resources Research* **44**: W05S13, DOI: 10.1029/2006WR005788.

Groisman PY, Karl TS, Easterling DR, Knight RW, Jamason PF, Hennessy KJ, Suppiah R, Page CM, Wibig J, Fortuniak K, Razuvaev VN, Douglas A, Forland E, Zhai Pan-Mao. 1999. Changes in the probability of heavy precipitation: important indicators of climate change. *Climatic Change* **42**: 243–283.

Gupta A, Lam MS. 1996. Estimating missing values using neural networks. *Journal of the Operational Research Society* **47**: 229–238.

Hutchinson MF, Gessler PE. 1994. Splines more than just a smooth interpolator. *Geoderma* **62**: 45–67.

Jones C, Waliser DE, Lau KM, Stern W. 2004. Global occurrences of extreme precipitation events and the Madden-Julian Oscillation: observations and predictability. *Journal of Climate* **17**: 4575–4589.

Karl TR, Knight RW, Plummer N. 1995. Trends in high-frequency climate variability in the twentieth century. *Nature* **377**: 217–220.

Karl TR, Richard WK. 1998. Secular trends of precipitation amount, frequency, and intensity in the United States. *Bulletin of the American Meteorological Society* **79**: 231–241.

Kemp WPD, Burnell DG, Everson DO, Thomson AJ. 1983. Estimating missing daily maximum and minimum temperatures. *Journal of Climate and Applied Meteorology* **22**: 1587–1593.

Khalil M, Panu US, Lennox WC. 2001. Groups and neural networks based streamflow data infilling procedures. *Journal of Hydrology* **241**: 153–176.

Lloyd CD. 2005. Assessing the effect of integrating elevation data into the estimation of monthly precipitation in Great Britain. *Journal of Hydrology* **308**: 128–150.

Nicholls N, Murray W. 1999. Workshop on indices and indicators for climate extremes, Asheville, NC, USA, 3–6 June 1997. Breakout group B: precipitation. *Climatic Change* **42**: 23–29.

Paulhus JLH, Kohler MA. 1952. Interpolation of missing precipitation records. *Weather Review* **80**: 129–133.

Ramos-Calzado P, Gomez-Camacho J, Perez-Bernal F, Pita-Lopez MF. 2008. A novel approach to precipitation series completion in climatological datasets: application to Andalusia. *International Journal of Climatology* **28**: 1525–1534.

Schneider T. 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate* **14**: 853–871.

Shepard D. 1968. A two-dimensional interpolation function for irregularly spaced data. *Proceeding of the Twenty-Third National Conference of the Association for Computing Machinery*: Washington, DC; 517–524.

Shepard D. 1984. In *Computer Mapping: The SYMAP Interpolation Algorithm, in Spatial Statistics and Models*, Gaile GL, Willmott CJ (ed). Springer: New York, 133145.

Stern RD, Coe R. 1984. A model fitting analysis of daily rainfall data (with discussion). *Journal of the Royal Statistical Society, Series A: Statistics in Society* **147**: 134.

Suhaila J, Sayang MD, Jemain AA. 2008. Revised spatial weighting methods for estimation of missing rainfall data. *Asia-Pacific Journal of Atmospheric Sciences* **44**: 93–104.

Tabios GQ, Salas JD. 1985. A comparative-analysis of techniques for spatial interpolation of precipitation. *Water Resources Bulletin* **21**: 365–380.

Tang WY, Kassim AHM, Abubakar SH. 1996. Comparative studies of various missing data treatment methods – Malaysian experience. *Atmospheric Research* **42**: 247–262.

Teegavarapu RSV, Chandramouli V. 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology* **312**: 191–206.

Xia Y, Fabian P, Winterhalter M, Zhao M. 2001. Forest climatology: estimation and use of daily climatological data for Bavaria, Germany. *Agricultural and Forest Meteorology* **106**: 87–103.

Young KC. 1992. A three-way model for interpolating for monthly precipitation values. *Monthly Weather Review* **120**: 2562–2569.

Wallis JR, Letten-Mayer DP, Wood EF. 1991. A daily hydroclimatological data set for the continental United States. *Water Resources Research* **27**: 1657–1663.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press: New York; 464.

Willmott CJ, Rowe CM, Philpot WD. 1985. Small-scale climate maps: a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *The American Cartographer* **12**(1): 5–16.

