INTERNATIONAL JOURNAL OF CLIMATOLOGY, VOL. 16, 213–227 (1996)

A METHOD FOR THE RECONSTRUCTION AND TEMPORAL EXTENSION OF CLIMATOLOGICAL TIME SERIES

F. VALERO, J. F. GONZALEZ, F. J. DOBLAS AND J. A. GARCÍA-MIGUEL

Universidad Complutense de Madrid, Departamento de Astrofisica y Física de la Atmósfera, Facultad de Ciencias Físicas, 28040 Madrid, España (Spain)

Received 6 June 1994
Accepted 3 May 1995

ABSTRACT

A method for the reconstruction and temporal extension of climatological time series is provided. This method was focused on a combination of methods, including harmonic analysis, seasonal weights, and the Durbin–Watson (DW) regression method. The DW method has been modified in this paper and is described in detail because it represents a novel use of the original DW method.

The method is applied to monthly means of daily wind-run data sets recorded in two historical observatories (M series and A series) within the Parque del Retiro in Madrid (Spain) and covering different time periods with an overlapping period (1901–1919). The aim of the present study is to fill up to and to construct a historical time series ranging from 1867 to 1992. The proposed model is developed for the 1906–1919 calibration period and validated over the 1901–1905 verification period, which includes the hypothesis of constant ratio of variances. The verification results are almost as good as those for the calibration period. Hence, the M series was extended back to 1867, which results in the longest climatological wind-run data-set in Spain. Also, the reconstruction is shown to be reliable.

KEY WORDS: reconstructing data; filling data; temporal extension; Durbin–Watson regression; time series; wind-run; Madrid

1. INTRODUCTION

When climatological time series are studied it is profitable to check data quality because it usually provides more reliable results. Unfortunately, we often find situations where data sets are too short to study long-term changes or where they show data gaps and so standard time series techniques cannot be used. The former problem may be addressed as a particular case of missing data (i.e. where data were lost or not recorded at the beginning or end of the data set). A number of methods have been proposed to deal with this concern. In general, they would be acceptable as far as the residuals could be considered as a white noise and their sum of squares (SSR) as small as possible.

This paper attempts to provide a simple method for filling and extending climatic data sets, different from those more commonly related with either the arithmetic mean or linear regression models, respectively. For the former, a method is proposed based on defining seasonal ratios (or weights) and a simple harmonic analysis. For the latter, a methodological approach for adjusting time series is specifically developed which makes a full use of station history information. Essentially, the Durbin–Watson (DW) regression (Durbin, 1953; Durbin and Watson, 1950, 1951, 1971) is used to reconstruct data. In this paper, this technique was modified. As an alternative to this method, there exist techniques focused on filtering dependent and independent variables once seasonal components have been filtered so as to obtain independent disturbances in the regression model; but this method involves acting on each of the variables (notice that it could also be applied in cases with more than one independent variable) with their respective coefficients, thereby complicating the procedure of obtaining the original predicted values once the regression is performed. We feel the proposed procedure is more compact and 'acts' directly on the residuals, this being the reason why this procedure has been chosen.

CCC0899-8418/96/020213-15

Monthly means of daily wind-run data were analysed in this paper. Even allowing for the fact that wind-run is a variable related directly to kinetic energy in the lower atmosphere, to pollution potential effects, and to other climatological processes, it still has not been used very much for climate studies. The wind-run time series is noisy and thus it is suitable for checking some methods for reconstructing historical time series.

Data used in this paper were those from the Meteorological (M) and Astronomical (A) Observatories, both within the Parque del Retiro in Madrid (Spain); they are about 500 m apart only, constituting the largest data sets in the Madrid area, ranging from 1901 until 1992 and from 1867 until 1919, respectively. Figure 1 shows an overlapping period (1901–1919). By using the data sets herein, the methodological approach developed is used for reconstructing and extending back to 1867.

2. THE RECONSTRUCTION METHOD

2.1 Filling data

Usually, data sets from observatories show a number of data gaps, not only for hourly or daily data series but also for monthly or yearly data series. Here, we have focused our attention on only monthly time series. Sometimes, gaps encompass randomly 'spaced' (isolated) missing data. At other times, gaps occur in single blocks. In this paper, gaps were classed into two categories: *class a*, for gaps consisting of monthly data missing for an entire year, and *class b*, for isolated missing monthly data.

The filling procedure requires an initial annual mean estimate, from which a decomposition of this estimated mean is 'projected' into its respective monthly values. Indeed, the estimate of annual means depends on whether missing data are grouped together forming relatively large data segments (class a) or whether they are isolated (class b). For the former, by using the annual mean series and by applying a harmonic analysis it becomes possible to extrapolate one time-step. For the latter, seasonal scores were

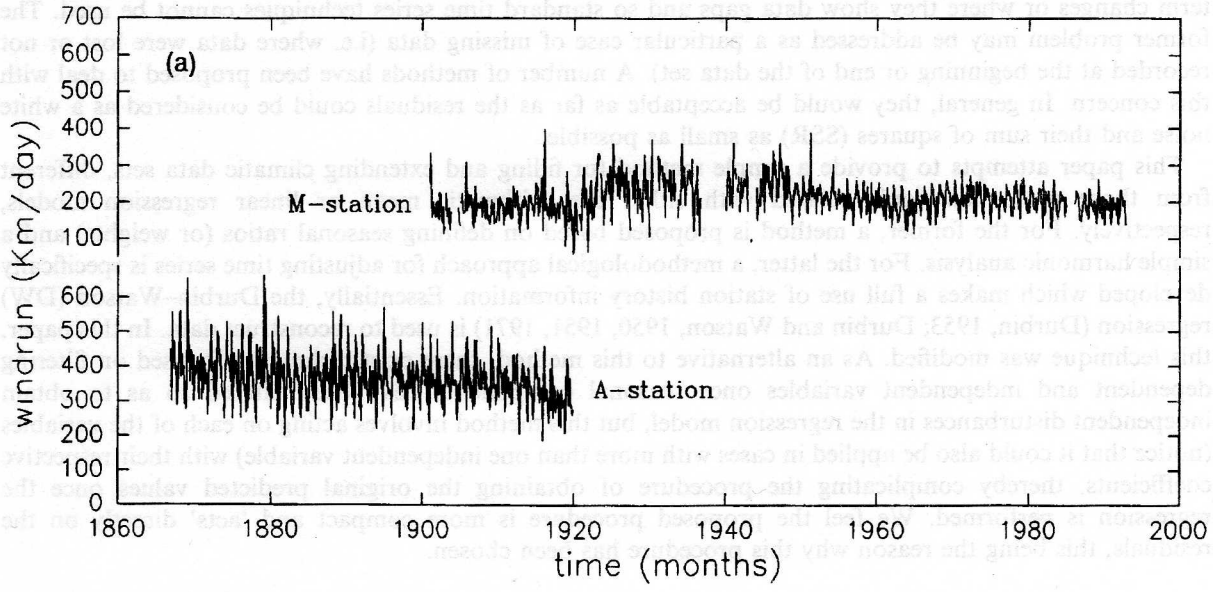

Figure 1. Composite raw time series of monthly mean of daily wind-run from the Meteorological (M) (1901–1992) and Astronomical (A) (1867–1919) Observatories

computed. Once the annual mean was obtained then it became feasible to fill monthly data through the seasonal score.

2.1.1. Estimating annual mean—class a. In analysing time series, we note first that the total variation or oscillation of a meteorological variable with time can be described as the sum of a number of oscillations. Therefore, the interpolating procedure of missing data may be well thought of as a smoothing procedure based on grouping the original data sequence just before and just after the missing time (BMT and AMT respectively) and on searching a Fourier description of the sequences, which will recover the missing value from each group.

Let us suppose that the BMT and AMT data groups are composed by all years around the missing datum (a particular year). We may describe the smoothed annual pattern of any climatological quantity for the two data segments by its finite sine–cosine Fourier expansions

$$X_t = \bar{X} + \sum_{i=1}^{m} \left(A_i \cos \frac{2\pi i}{N} t + B_i \sin \frac{2\pi i}{N} t \right)$$

where X_t is the t-year annual mean estimate, $\overline{X}$ is the mean of the series, A_i and B_i are the Fourier coefficients, N is the number of data points, i the order of the harmonic and m the truncation point. In practice, we retained only the first i harmonics, such that the variance explained by these i was > 80 per cent and the variance explained by $i-1$ was < 80 per cent. A more detailed treatment of harmonic analysis may be found in Panofsky and Brier (1968) or in Barry and Perry (1973). When this procedure is applied, a backward estimate is obtained for the AMT sequence and a forward one for the BMT sequence. The estimates obtained were then averaged to yield a unique estimation for the missing data point.

This method was also extended for estimating annual means corresponding to grouped missing data consisting of a few years. For both original BMT and AMT data sequences a forward and backward estimate, respectively, was obtained by enlarging the time domain for 1 year by applying harmonic analysis. This procedure is repeated until the missing period is entirely filled. The estimates for the central-point datum for those missing periods comprising an odd number of years were computed in the way described in the previous paragraph.

2.1.2. Estimating annual mean—Class b. Where only isolated monthly data within a year are found to be missing the method to be applied is substantially different. In this case, the aim of the procedure is to estimate the annual mean value through a weighted average by taking into account the monthly data available for the 'k-problem' year. Each monthly weight was obtained according to the following steps: (i) a 10-year subset (5 years BMT and 5 years AMT) was taken out; (ii) the 10-year monthly means $\bar{y}_{ik}$ and the 10-year annual mean $\overline{X}_k$ were computed; (iii) the monthly weights a_{ik} were computed according to the equation $a_{ik} = y_{ik}/\overline{X}_k$; (iv) for the k-problem year, the annual mean X_k is computed involving the available monthly data through the expression

$$X_k = \sum_{i=1}^{12-m} \left(\frac{y_i k}{a_i k} \right) / (12 - m)$$

where y_{ik} is the i-monthly datum for the year k, a_{ik} is the weight for the i-month, computed for the 10 years surrounding the k-year, and m is the number of missing monthly data for the k-year. The monthly weights account for the 'contribution' of present data into the year to the annual mean. As stated before, only two 5-year monthly data subsets were picked out from the raw data set so as to apply the method. This allows the capture of more recent information at both sides of the unknown data point. The length of the grouping is chosen subjectively in attempting to make the method adaptive.

2.1.3. Estimating missing monthly data. We are now curious about how to project the estimated annual mean X_k into each of missing month y_{ik}. By doing so, the monthly weights a_{ik} were again used to fill the gaps y_{ik} according to the expression.

$$y_{ik} = X_k a_{ik}$$

The BMT and AMT time segments were selected to compute the monthly weights via the above method. This method allows us to have a complete series without any gaps.

2.2. Lengthening data

Let us now suppose that we have two complete series of the same variable without gaps recorded at two very nearby observatories embedding different periods but with some overlap. We could attempt to extend one of the climatic data series from the other one. By doing so, we could reconstruct a new historical time series that includes both periods. It would have the advantage of being considered as a larger reference time series for the particular area.

To do this, we first found the interrelationships between series over the common periods in order to obtain some measure of the extent to which common data are related to each other. Essentially, a lengthening data procedure can be thought of as a particular case of checking on the homogeneity of the time series and subsequent suitable correction because different data sources have been used. In general, when standard deviations of two data sets could be considered the same, the method of differences is preferred and when their variation coefficients are statistically equal then the ratio method is recommended to be used. A simple regression model

$$y_t = \beta_0 + \beta_1 x_t + \xi_t \tag{1}$$

where y_t is the dependent variable, x_t the independent variable, and ξ_t the error term, accounts for the more general case when none of these two hypotheses can be accepted.

One of the standard assumptions in the regression model is that the error terms ξ_i and ξ_j associated with the ith and jth observations are uncorrelated. Too often, however, the regression residuals are autocorrelated and the simple linear regression is not strictly valid. Also, if the samples show autocorrelation, serious mistakes can be introduced in the regression model. If it does, the sample cross-correlation function at lag k ($k \neq 0$) is overestimated, and the standard deviation of the cross-correlation at lag 0 as well as the standard deviation of the slope of the regression model are poorly estimated (Katz, 1988). Autocorrelation may occur for several reasons, e.g. observations sampled from adjacent points tend to have residuals that are correlated because they are affected by similar external conditions (Chatterjee and Price, 1977). Correlation in the error terms suggests that there is additional explanatory information in the data that has not been exploited in the current model, which is the result of an independent variable having been omitted from the right-hand side of the regression equation. The presence of autocorrelation causes several negative effects when samples are used for climatic analysis (Johnston, 1967; Chatterjee and Price, 1977; Canavós, 1990), and therefore it becomes convenient to determine for each particular case study whether the common least-square techniques are reliable or alternative methods have to be used instead. The Durbin–Watson statistic provides a good reference to find out if residuals are autocorrelated. The test is based on the assumption that the errors constitute a first-order autoregressive (AR) series, namely

$$\xi_t = \rho \times \xi_{t-1} + \omega_t, \mid \rho \mid < 1 \tag{2}$$

where ω_t is a white noise process. In most situations, the error ξ_t may have a much more complex correlation structure. The first-order dependency structure given in equation (2) is taken as a simple approximation to the actual error structure. The statistic d is defined as

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} (e_t)^2} \tag{3}$$

where the residuals $e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_t)$, $\hat{\beta}_0, \hat{\beta}_1$ being the least-square estimates of β_0, β_1. The statistic d is used for testing the null hypothesis H_0 ($\rho = 0$) against an alternative H_1 ($\rho > 0$). Note that when $\rho = 0$ in equation (2), the ξs are uncorrelated. We may estimate the parameter ρ by $\hat{\rho}$, where

$$\hat{\rho} = \frac{\sum_2^n (e_{t-1} \times e_t)}{\sum_2^n e_t^2} \tag{4}$$

Durbin and Watson (1950, 1951) proposed a test of serial independence of the disturbances in the regression model (1) based on the distribution of the d statistic defined by equation (3). They showed that the distribution of d depends on the independent variable. However, they found a pair of bounding random x-independent variables d_L and d_U and tabulated their lower tail significance points. Tables for these values also can be found in many other test books (Chatterjee and Price, 1977; Canavós, 1990). Since their 1950 and 1951 papers were published, a number of alternative exact and approximate tests have been proposed, none of which showed skilled local invariance statistical properties, as was later demonstrated by Durbin and Watson (1971). The original test against positive serial correlation suggested that the hypothesis of independence H_0 ($\rho = 0$) should be rejected if d is less than the tabulated value of d and accepted if d is greater than the tabulated value of d. In the intermediate case, Durbin and Watson (1971) showed that the test is inconclusive, one could use the data and calculate the exact significance point numerically through an approximate procedure, which they called the bounds test, based on fitting a beta distribution. Alternatively, the inconclusive area can be treated as part of the rejection region considering the null hypothesis rejected ($\rho \neq 0$), which obviously simplifies the procedure. In this way, we can resolve if proceeding with a type (1) regression model is appropriate or if an alternative method should be used that takes account of the autocorrelation of residuals. If the null hypothesis should be rejected, a first-order AR structure should be accepted. If we insert equation (2) in (1) we have

$$y_t = \beta_0 + \beta_1 \times x_t + \rho \times \xi_{t-1} + \omega_t$$

If we now write equation (1) lagged one-time step, namely

$$y_{t-1} = \beta_0 + \beta_1 \times x_{t-1} + \xi_{t-1}$$

and by inserting this equation in the previous one we find

$$y_t - \rho \times y_{t-1} = \beta_0(1 - \rho) + \beta_1(x_t - \rho \times x_{t-1}) + \omega_t \tag{5}$$

Defining two variables $x_{t'}, y_{t'}$

$$y'_t = y_t - \rho \times y_{t-1} \tag{6}$$

$$x'_t = x_t - \rho \times x_{t-1}$$

equation (5) becomes

$$y'_t = \beta'_0 + \beta'_t \times x'_t + \omega_t \tag{7}$$

where the error terms are now not correlated, thereby gaining better estimates of the coefficients. From equations (5) and (7)

$$\beta_0 - \beta'_0/(1 - \rho) \tag{8}$$

$$\beta_1 = \beta'_1$$

$\beta_0, \beta_1, \beta'_0, \beta'_1, \rho$ parameters are population parameters. We conclude, therefore, that the procedure to estimate them operates as follows.

(i) Compute the ordinary least-square estimates of β_0, β_1 using equation (1), obtaining $\hat{\beta}_0, \hat{\beta}_1$.
(ii) Calculate the residuals $e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 x_t)$ and from the residuals calculate $\hat{\rho}$ using equation (4).
(iii) With the variables defined in equation (6), fit equation (7) to obtain estimates of β'_0, β'_1 denoted by $\hat{\beta}'_0, \hat{\beta}'_1$, respectively.

(iv) According to equation (8), improved estimates $(\tilde{\beta}_0, \tilde{\beta}'_1)$, of the original parameters in equation (1) are given by $\tilde{\beta}_0, \tilde{\beta}'_1/(1-\hat{\rho})\ \tilde{\beta}_1, \tilde{\beta}'_1$,in terms of $\hat{\beta}'_0, \hat{\beta}'_1$ from stage (iii).

The usual statistical approach concentrates on obtaining new reliable estimates $(\tilde{\beta}_0, \tilde{\beta}_1)$ for the original (β_0, β_1) from the two least-square estimates $(\hat{\beta}'_0, \hat{\beta}'_1)$, which can be used in equation (1) to produce a valid regression model $(y_t = \tilde{\beta}_0 + \tilde{\beta}_1 \times x_t)$. Residuals from this last model would probably still show autocorrelation, nevertheless, the new estimated coefficients, $(\tilde{\beta}_0, \tilde{\beta}_1)$, are reliable because they have been estimated from the white noise disturbances (Canavós, 1990). Now, if we are interested in using the model for prediction purposes, as in our case, we can deal with equation (5) to produce a regression model in terms of the residuals autocorrelation.

$$\hat{y}_t = \hat{\rho} \times y_{t-1} + \hat{\beta}'_0 + \hat{\beta}'_1 \times (x_t - \hat{\rho} \times x_{t-1}) \tag{9}$$

Here, this equation provides extrapolated values of the variable y from the variable x value at the present time t and lagged time $t-1$ and from the y itself lagged one-time unit. However, as it was pointed out earlier, the first-order AR structure, given in equation (2), is taken as a simple approximation to the actual error structure, which can be much more complex. If, in particular, significant multiple autocorrelation coefficients are present we have found that better results are obtained by making a generalized s-order process assumption:

$$\hat{y}_t = \sum_{i=1}^{s} \hat{\rho}_i \times y_{t-i} + \hat{\beta}'_0 + \hat{\beta}'_1 \times \left(x_t - \sum_{i=1}^{s} \hat{\rho}_i \times x_{t-i} \right) \tag{10}$$

In such a case, the procedure to obtain the parameters is similar to that described previously for a first-order AR process. However, there are two aspects to discuss at this stage: one is the fact that the d test is no longer valid for lags different than one and the second point is that the coefficients for an s-order process would not be now the autocorrelation ones, but rather the autoregressive parameters that are related to them through the Yule–Walker equations (Box and Jenkins, 1976). The first point is easily circumvented by considering a different autocorrelation test to assess the significance of autocorrelation for lags different from the first one. To do this, we have plotted the autocorrelation coefficients and twice the standard error at lag k after Box and Jenkins (1976).

Regression model estimates usually show lower than that for real data variance. If we want to optimize, the regression estimates it may be advantageous to introduce a correction factor. We may assess this factor taking into account the following two hypotheses: the difference of station means as well as the ratio between station deviations remain constant throughout the period under study. These assumptions are consistent with the methods that have often been carried out to lengthen out a series based on homogeneity criteria. They do consider the source of variability for samples to be the same. This means that any possible forcing that causes climatic variation will affect both station data sets in a similar way, even if each series undergoes non-stationary changes.

3. A CASE STUDY

In order to apply the methods above described for filling and extending climatic data series we have chosen two monthly mean series of daily wind-run (wind speed multiplied by duration in km day^{-1}) recorded by two Robinson's cup anemometers in two very close observatories in Madrid city (Spain). Raw data from the Meteorological (M) and Astronomical (A) Observatory stations were supplied by the Spanish National Institute of Meteorology (INM) and the Spanish National Geographic Institute (IGN), respectively. The two stations are within the Parque del Retiro and they are only about 500 m apart. Both are the largest data sets in the Madrid area but they cover different time periods, ranging from 1901 to 1992 for the M station and from 1867 to 1919 for the A station. In Table I(a), means and variances for different time periods are displayed, from which it is apparent that an overlapping period (1901–1919) occurs. It might well be feasible to construct a continuous data set extending back to 1867. In order to apply the methodology developed in section 2 and to minimize the effects of predictable seasonal variations of the variables, raw

Table I(a) Estimates of mean (km/day), variance (km^2/day) and ratio of variances of monthly means of daily wind-run of M series and A series for different periods·

	1867–1900	1867–1919	1901–1992	1901–1905		1906–1919		1901–1919	
	A	A	M	A	M	A	M	A	M
Mean	373·3	362·6	208·9	353·5	180·9	340·0	186·7	343·5	185·2
Variance	4858·2	4539·8	1824·1	3332·5	1464·3	3421·3	1328·0	3418·6	1364·1
Mean difference	—	—	—	172·6		153·3		158·3	
Ratio variance	—	—	—	2·27		2·57		2·50	

Table I(b) Means (km/day) and variances (km^2/day) of monthly means of daily wind-run of the M series and A series for different months and periods

	Years		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
M	1901 to 1992	$\bar{x}$	189·5	215·5	231·5	240·6	217·2	215·1	217·0	212·9	194·7	191·6	189·4	189·0
		σ^2	1722·5	1697·4	1310·4	1528·8	1281·6	1129·0	894·0	1428·8	1056·2	1149·2	1466·9	1632·2
M	1901 to 1919	$\bar{x}$	147·8	203·1	210·6	209·6	187·3	186·2	199·0	178·9	177·6	169·6	179·8	165·2
		σ^2	2611·2	2180·9	1310·4	556·9	620·1	246·5	302·8	510·8	445·2	739·8	1436·4	1176·5
A	1867 to 1919	$\bar{x}$	328·0	375·6	421·1	419·4	382·1	377·1	375·2	355·4	336·1	331·2	323·9	326·6
		σ^2	5944·4	7779·2	5852·3	5055·2	2520·0	1436·4	948·6	1310·4	1459·2	2227·8	2926·8	4435·6
A	1901 to 1919	$\bar{x}$	304·1	381·0	390·2	376·7	346·6	358·1	355·3	338·2	314·2	313·1	313·7	331·5
		σ^2	3249·0	7656·2	3831·6	2450·2	2601·0	1075·8	789·6	1303·2	1108·9	1883·6	2672·9	4515·8

Table I(c) Means (km/day) and variances (km/day^{-2}) of estimated monthly means of daily wind-run of the M series by the different methods used both in the calibration (1906–1919) and verification (1901–1905) periods

Period		Real	Estimated			
			Simple	DW(1)	DW(1,6)	DW(1,6) + VC
1901 to 1905	$\bar{x}$	180·9	—	191·0	186·2	186·2
	σ^2	1464·3	557·8	—	787·3	1293·6
1906 to 1919	$\bar{x}$	186·7	186·7	185·4	188·6	—
	σ^2	1328·0	538·8	752·0	934·7	—

data were first transformed by subtracting the long-term monthly mean from each monthly value. Table I(b) shows the monthly climatological means and the standard deviations for both observatories and for both common and complete periods. In order to obtain good estimates of monthly climatological means and standard deviations, each of the monthly time series were previously winsorized to manage outliers; for a detailed description of winsorization see Hoaglin *et al.* (1983).

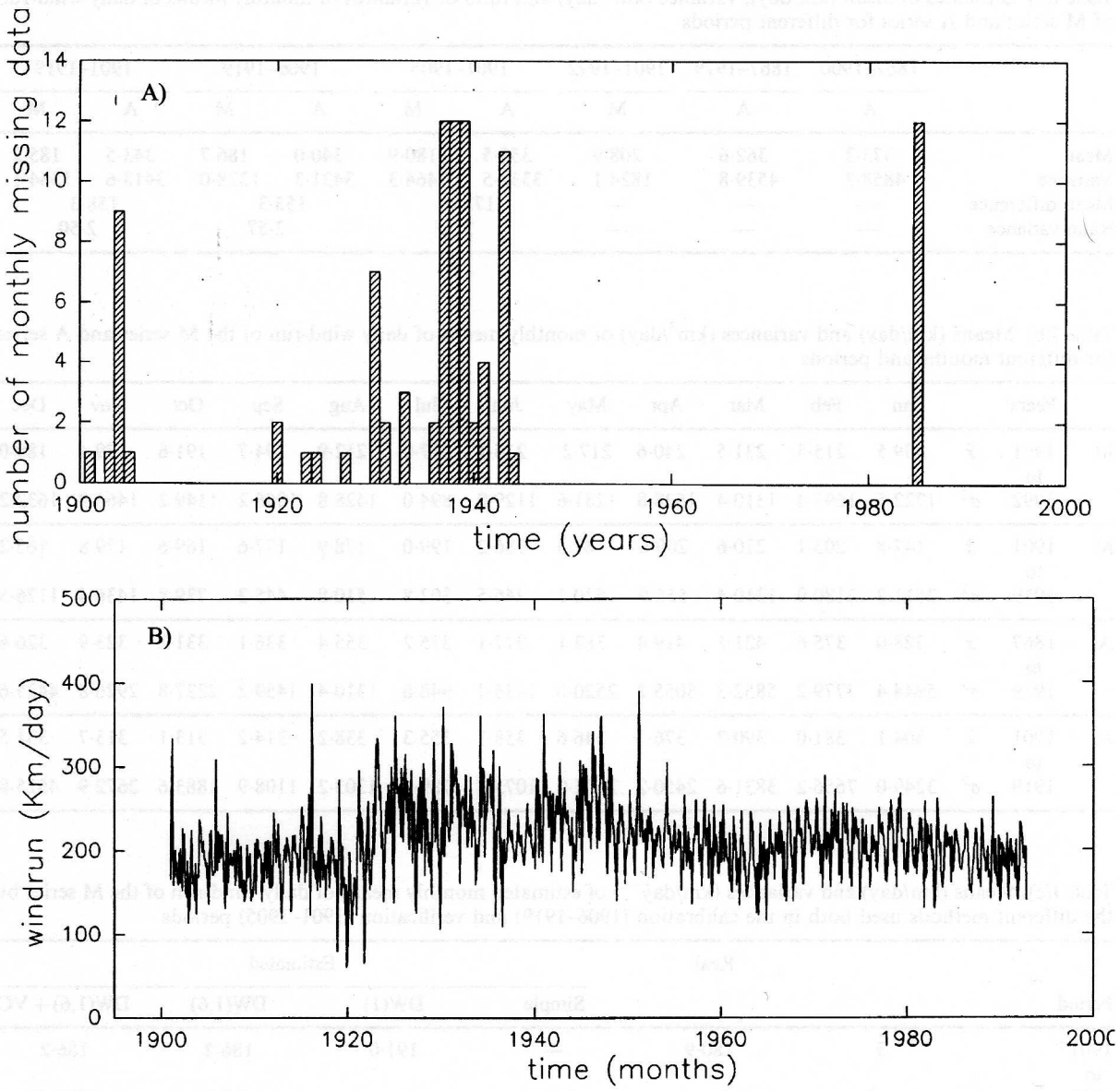

Figure 2. (A) Number of missing monthly data for the M series. (B) The reconstructed M series after filling stage

3.1. Filling data

Prior to applying the methods described in section 2, verification is advisable. There are a number of ways to validate the method. A way to ascertain the goodness of filling data is to do it over a chosen segment with no data gaps. For the selected period, from 1956 to 1976, the IMSL routine RNSRI was run to generate pseudorandom numbers from a discrete uniform distribution so as to create a subset (or arrangement) of 30 gaps. Next, the filling procedure was applied specifically to derive data-point estimates for the gaps previously generated. This procedure has been repeated 100 times. The comparison of actual versus estimated data has been carried out by a two-sample analysis. This procedure estimates and tests the

means and variances of two independent samples. The means difference test and the variances ratio test (Snedecor and Cochran, 1972) have been used. Both actual and estimated data were found to be statistically equal to a 5 per cent significance level. The agreement was regarded as sufficiently good so as to allow the simulated data to be used to fill up the gaps.

Finally, the method was used only for the M series (because the A series has no gaps) to be reconstructed. The total number of estimated missing data is 98, of which 50 are class b. Figure 2(A) displays the missing data arrangement for the whole period. The series without gaps is displayed in Figure 2(B).

3.2. Extending data

To produce a climatological time series extending from 1867 to the present time requires us to lengthen the previously reconstructed M series from A series information. Our purpose is to derive an optimum statistical model where no residual autocorrelation actually exists. The model was obtained by splitting the common period (1901–1919) into two subperiods. The 1901–1905 subperiod (I) was viewed as a *verification* period and the 1906–1919 subperiod (II) as a *calibration* period. As mentioned above, the Durbin–Watson linear regression was used to fit the model.

3.2.1. Calibration. Regression was performed fitting subset II. The estimated simple regression parameter values were $\hat{\beta}_0 = 7{\cdot}9(\pm 2{\cdot}9)$ km day^{-1} and $\hat{\beta}_1 = 0{\cdot}32(\pm 0{\cdot}04)$; $\hat{\beta}_0, \hat{\beta}_1$ being the intercept and the slope of linear regression, respectively. Values between parentheses are plus or minus one standard deviation. The sample autocorrelation function (SACF) of the residuals is plotted in Figure 3(A). Note that the autocorrelation shows coefficients, at lags 1 and 6, significantly different from 0. Next, following the methods described in section 2.2, the critical values, d_L and d_U, for the Durbin–Watson statistic d were derived. For 168 data and two variables, d_L and d_U proved to be 1·78 and 1·82, respectively. The value of d ($= 1{\cdot}19$) is calculated from equation (3). Obviously, this value is lower than the critical values (d_L and d_U). Accordingly, the null hypothesis H_0 is rejected, and so the use of the Durbin–Watson method is recommended to fit a new model, the residuals of which are free of correlation in time. Given that the value of $\hat{\rho}$ in equation (4) equals 0·37, the autoregressive model for the residuals can be described by equation (2) or in terms of the estimates by $e_t = \hat{\rho} \times e_{t-1}$. Following the operational procedure described in section 2.2 for this first-order AR process, estimated values for the model parameters are obtained (Table II). Notice that the new values estimated for d ($= 1{\cdot}96$) and for the regression correlation coefficient, r ($= 0{\cdot}62$), using the residuals of equation (7) have increased compared with same parameters in simple regression, whereas SSR (sum of squares of residuals) has decreased substantially. Furthermore, the estimated variance of the estimates (once the annual cycle has been added) has also improved, even though its real value is still underestimated (see Table I(c). Nevertheless, the residual's SACF (Figure 3(B) still show significant autocorrelation at lag 6. Here the residual's model could be improved by assuming an autoregressive model described by $e_t = \hat{\rho}_1 \times e_{t-1} + \hat{\rho}_6 \times e_{t-6}$, thereby deriving the estimated autoregressive parameters at lags 1 and 6 ($\hat{\rho}_1 = 0{\cdot}32; \hat{\rho} = 0{\cdot}20$). By virtue of $\hat{\rho}_1$ and $\hat{\rho}_6, x'_t$ and y'_t can be written as

$$y'_t = y_t - \hat{\rho}_1 y_{t-1} - \hat{\rho}_6 \times y_{t-6}$$

$$x'_t = x_t - \hat{\rho}_1 x_{t-1} - \hat{\rho}_6 \times x_{t-6}$$

The regression between these variables allows the estimators $(\tilde{\beta}_0, \tilde{\beta}_1)$ for (β_0, β_1) to be obtained from the regression coefficients $(\hat{\beta}'_0, \hat{\beta}'_1)$ between the transformed variables (see Table II). Thus:

$$\tilde{\beta}_0 = \hat{\beta}'_0(1 - \hat{\rho}_1 - \hat{\rho}_6)^{-1} = 10{\cdot}6 \pm 3{\cdot}7$$

$$\tilde{\beta}_1 = \hat{\beta}'_1 = 0{\cdot}39 \pm 0{\cdot}03$$

These estimated parameters $(\tilde{\beta}_0, \tilde{\beta}_1)$ are better than those for the initial simple regression $(\hat{\beta}_0, \hat{\beta}_1)$, and also their standard deviations are not underestimated, which from a statistical standpoint turns out to

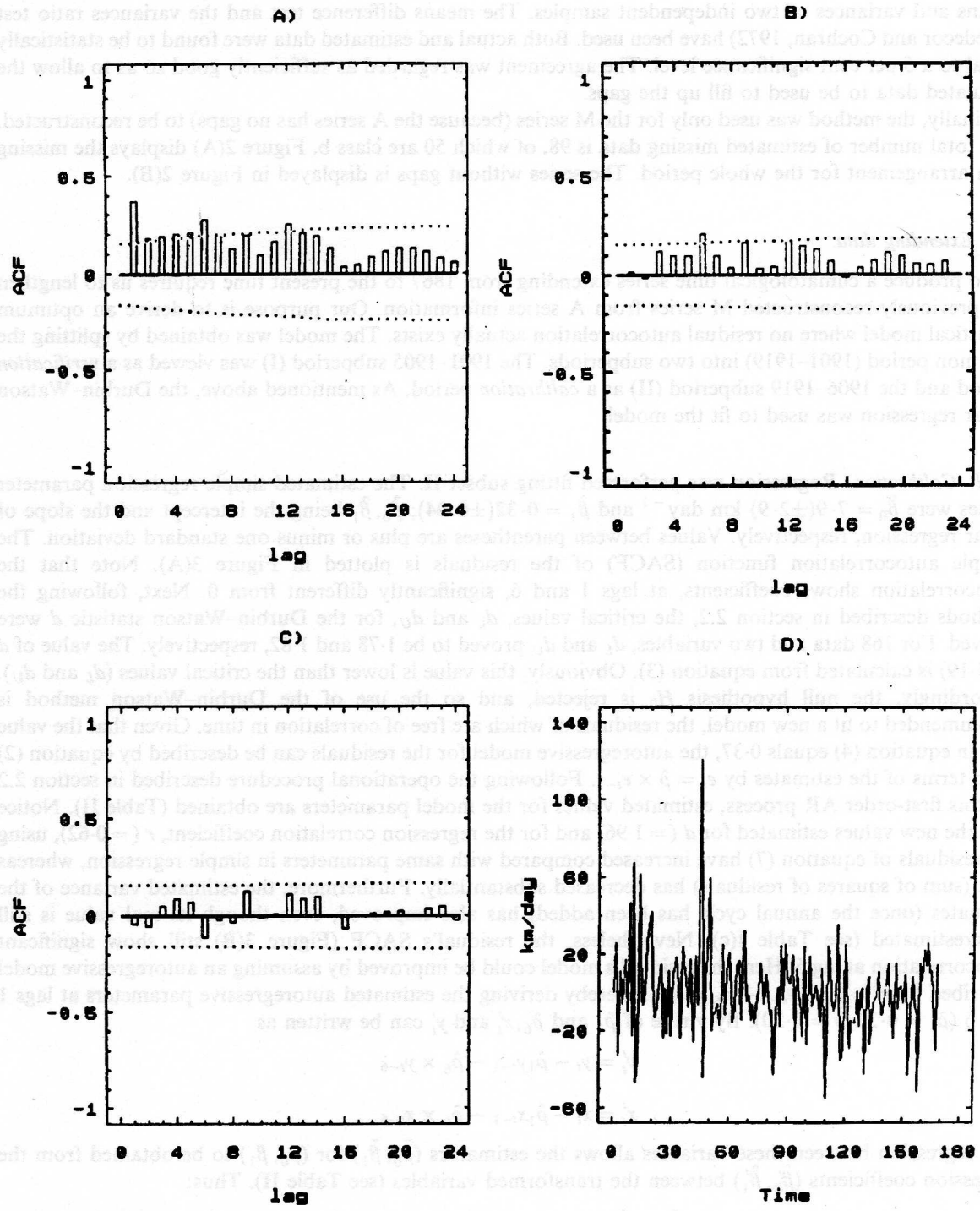

Figure 3. (A) The sample autocorrelation function (SACF) for the residuals from fitting simple regression to the 1906–1919 calibration period data. (B) As (A) but using DW regression with an AR(1) hypothesis on the residuals. (C) As (A) but using DW regression with an AR(1,6) hypothesis on the residuals. (D) composite wind-run difference between original data and reconstructed data by DW regression model for the calibration period

Table II. Estimated values for the coefficients and parameters used in the methods selected

Linear regression				Test Durbin–Watson		
$\hat{\beta}_0$	$\hat{\beta}_1$	r	SSR	d_U	d_L	d
7·9 ± 2·9	0·32 ± 0·04	0·5 ± 0·1	120212	1·82	1·78	1·19

Durbin–Watson AR(1)							
$\hat{\beta}'_0$	$\hat{\beta}'_1$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\hat{\rho}$	r	d	SSR
6·03 ± 1·92	0·35 ± 0·3	9·57 ± 3·04	0·35 ± 0·03	0·37	0·62 ± 0·1	1·96	95012

Durbin–Watson AR(1,6)							
$\hat{\beta}'_0$	$\hat{\beta}'_1$	$\tilde{\beta}_1$	$\tilde{\beta}_1$	$\hat{\rho}_1$	$\hat{\rho}_6$	r	SSR
5·1 ± 1·8	0·39 ± 0·03	10·6 ± 3·7	0·39 ± 0·03	0·32	0·2	0·7 ± 0·1	87687

be advantageous. Finally, according to equation (10), the fitted model would be given by

$$\hat{y}_t = \hat{\rho}_1 \times y_{t-1} + \hat{\rho}_6 \times y_{t-6} + \hat{\beta}'_0 + \hat{\beta}'_1(x_t - \hat{\rho}_1 \times x_{t-1} - \hat{\rho}_6 \times x_{t-6}) \tag{11}$$

Let us now consider the differences between observed and estimated data in period II. These differences or residuals do not show autocorrelation, as can be observed in the autocorrelogram in Figure 3(C). If the observed and estimated means (186·7 and 188·6, respectively) and variances (1328 and 935, respectively) are compared, once the annual cycle has been added, it is remarkable that both subsamples are statistically equal ($\alpha = 0{\cdot}05$). If we set equation (11) model variance estimates against those of equations (1) and (9), a significant improvement is found, as can be noted in Table 1(C). A composite wind-run difference between original and reconstructed data can now be performed for the calibration period, as displayed in figure 3(D). Non-significant bias is revealed and so the equation (11) model may well be thought of as an unbiased model to extend the original M series from the A series, preserving the more common features of the two series under study.

The means and variances of the estimated values by each one of the models (after monthly climatological means have been added) are summarized in Table I(C). Also, the regression parameters and statistics are registered in Table II.

3.2.2. Verification. Just as with any regression method, the reliability of the regression model specifications cannot be determined solely from the calibration period performance. Although the model developed in this study is highly significant statistically, based on a conventional test, the only way to assess the true value of the reconstructions is to verify the model over an independent period for which both stations windrun data are available. This independent testing is referred to commonly as *verification* (Jones *et al.*, 1987). In general, regression models can be assessed (for either the calibration or verification periods, or both) in a number of ways. For instance, overall performances can be judged from the total explained variance. This criterion is used in the verification period.

The autocorrelogram of the residuals and their normal probability plot when model (11) is applied to the verification period are presented in Figure 4(A and B). Note that significant autocorrelation coefficients are not apparent (Figure 4(A)), which means that statistical tests are applicable, because independent data are assumed and the data set is reasonably well fitted to a normal distribution (Figure 4(B)). Both issues allow us to infer that the estimates of the model are non-contaminated with biases. Nevertheless, based on the results from period I, for both observed and estimated (adding up the annual cycle) data sets, is the relatively striking, substantially reduced estimated variance ($S_e^2 = 787$) with respect to the observed variance ($S_0^2 - 1464$), which means that the estimates explain only 54 per cent of the sample variance. Therefore, a correction factor was derived on a monthly basis so as to obtain improved estimates. For this

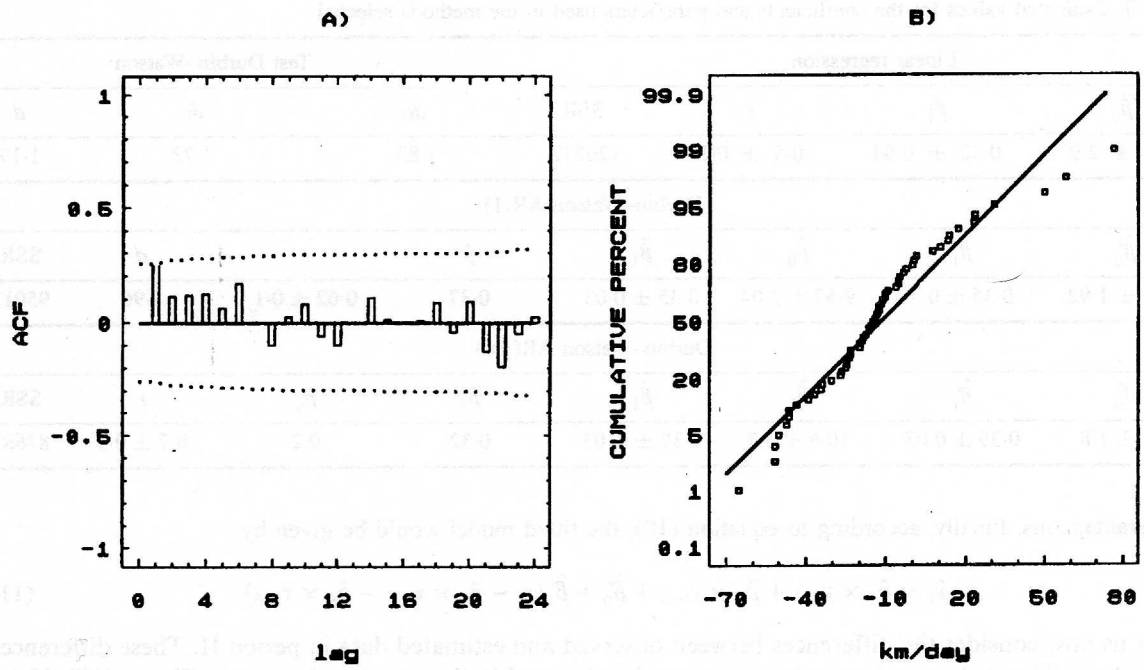

Figure 4. (A) As Figure 3(D), but for the verification period. (B) Normal probability plot of the residuals from fitting DW regression with an AR(1,6) to the verification period data

purpose, both differences of means and ratios of variances were checked in order to assess whether they are maintained or not with time. We may notice (see Table 1(a)) that neither hypothesis can strictly be assumed. In any case, the correction of the shift in means is achieved through applying the regression and therefore it is reasonable to derive a correction factor by making use of the ratios of variances only rather than differences of means. The aim of this correction factor is merely to adjust the shift in variances between the target station and the supplementary station, but it does not provide further information about natural variability, which is accounted for by the regression and the Durbin–Watson method. A suitable and simple value to be used as a correction factor is the ratio of M- and A-series variances within a period where data for both observations are available. For instance, if in particular we want to correct the extended data through years 1867 to 1900 we would use the value obtained from the period 1901–1919. This value is more convenient as a scale factor than values obtained from individual years or shorter time intervals. In fact, the year-to-year change of this ratio can differ notably. The authors have tested this by calculating series of ratios of variances for the M and the A series. Each value was obtained by dividing the M-series variance by the A-series variance for a certain time segment and displacing it in time in a similar way to the centred moving average procedure. A variance ratio series was derived for different segments, ranging from 1 year to the complete 19-year period (1901–1919). These series showed great variability, which increased as the interval range decreased. Their mean values were relatively close to the 19-year period mean (Table 1(a)). Therefore, and for simplicity, the value of 2·5 was chosen as a correction factor.

In order to illustrate how this correction performs it has been tested in the verification period to re-scale the variance of the Durbin–Watson regression estimates. As M-series data for the verification period are not supposed to exist, only the period II data are used to develop the correction factor. The estimates of the model (11) are then modified according to the formula

$$\hat{\hat{y}}_t = \left(\frac{\hat{y}_t - \bar{y}}{S}\right) \times S_{M,I} + \bar{y}$$

where $\hat{\hat{y}}_t$ are the *corrected estimates*, $\hat{y}_t$ the model (11) estimates, $\bar{y}$ the mean of model (11) estimates, S the

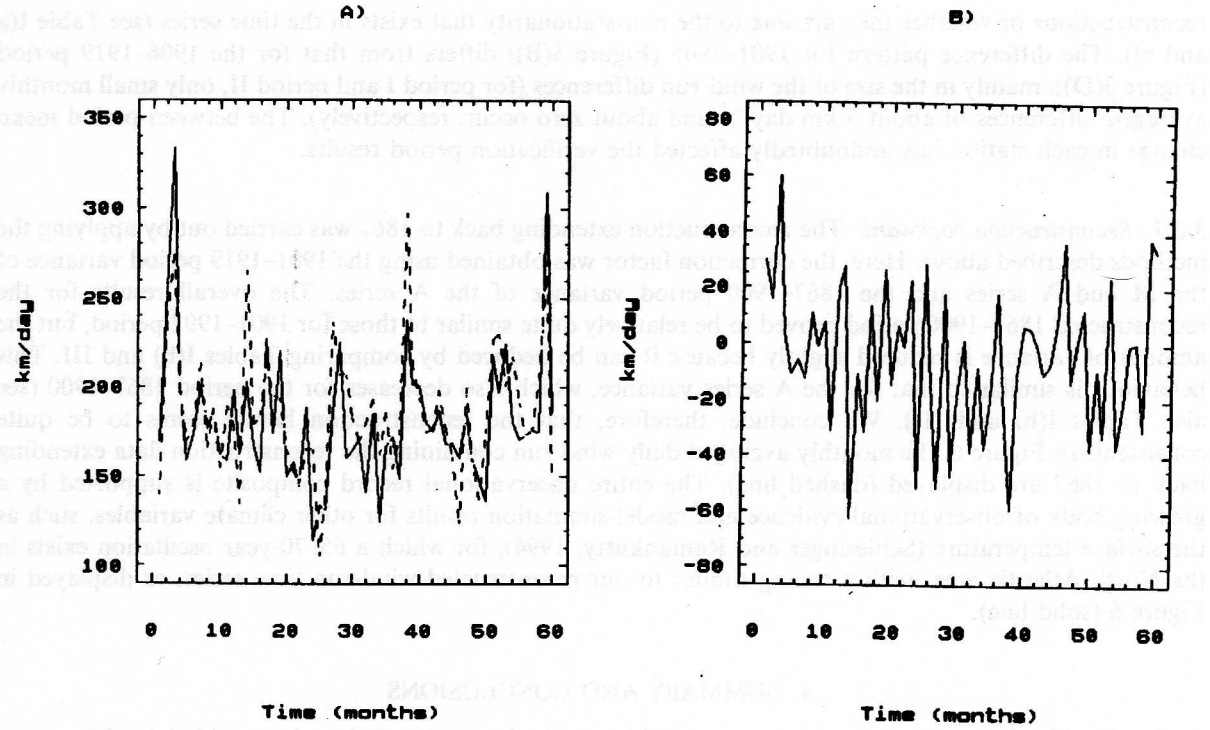

Figure 5 (A) Time series of true data (solid line) and corrected estimates (dashed line) for verification period. (B) Composite wind-run difference between original data and corrected estimates for verification period

standard deviation of model (11) estimates, and $S_{M,I}$ the standard deviation for period I after correction

$$S_{M,I} = (S_{M,II}/S_{A,II})S_{A,I}$$

where $S_{M,II}$ and $S_{A,II}$ are the period-II standard deviations of the M and A series respectively, and $S_{A,J}$ is the period-I standard deviation of the A series. The corrected estimates are then added up to the annual cycle amplitude to give rise to a reconstructed data set for period I. If we now compare the reconstructed and observed data sets we may find that the time data variability is quite well reproduced in reconstructed data, i.e. the reconstructed variance accounts for about 88 per cent of the sample variance. It should be noticed also that both means and variances for reconstructed ($\bar{y}' = 193{\cdot}4, S_e'^2 = 1294$) and real ($\bar{y} = 180{\cdot}9, S_0^2 = 1464$) data subsets are statistically equal ($\alpha = 0{\cdot}05$).

The results are encouraging because they perform reasonably well over an independent 1901–1905 verification period (Figure 5(A)). The verification results are almost as good as those obtained in calibration. It is not immediately clear whether the departures are due to errors in the model-based

Table III. Monthly means (km/day) and variances (km^2/day) of the A series and estimated M-series wind-run values during 1867–1900

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
M	$\bar{x}$	177·8	209·1	237·4	231·1	207·2	200·3	213·2	191·9	192·6	185·4	192·9	166·3
	σ^2	2631·7	3003·1	2323·2	1857·6	529·0	501·8	275·6	408·0	466·6	800·9	1239·0	1656·5
A	$\bar{x}$	341·4	372·5	438·4	443·3	401·9	387·6	386·4	365·0	348·3	341·3	329·6	323·9
	σ^2	7072·8	8064·0	6272·6	4984·4	1413·8	1354·2	712·9	1089·0	1267·4	2190·2	3058·1	4489·0

reconstructions or whether they are due to the non-stationarity that exists in the time series (see Table I(a and c)). The difference pattern for 1901–1905 (Figure 5(B)) differs from that for the 1906–1919 period (Figure 3(D)), mainly in the size of the wind-run differences (for period I and period II, only small monthly averaged differences of about 5 km day^{-1} and about zero occur, respectively). The between-period mean change in each station has undoubtedly affected the verification period results.

3.2.3. Reconstruction backward. The reconstruction extending back to 1867 was carried out by applying the methods described above. Here, the correction factor was obtained using the 1901–1919 period variance of the M and A series and the 1867–1900 period variance of the A series. The overall results for the reconstructed 1867–1900 period proved to be relatively quite similar to those for 1901–1992 period, but the amount of variance is reduced slightly because it can be deduced by comparing Tables I(b) and III. This behaviour is similar to that for the A series variance, which also decreases for the period 1867–1900 (see also Tables I(b) and III). We conclude, therefore, that the reconstruction herein seems to be quite consistent. In Figure 6, the monthly averaged daily wind-run containing the reconstruction data extending back to 1867 are displayed (dashed line). The entire observational record composite is supported by a growing body of observational evidence and model-simulation results for other climate variables, such as the surface temperature (Schlesinger and Ramankutty, 1994), for which a 65–70-year oscillation exists in the North Atlantic area, with a timing similar to our reconstructed wind-run time series, as displayed in Figure 6 (solid line).

4. SUMMARY AND CONCLUSIONS

A simple objective method for reconstructing historical observational data is provided in this paper. Basically, it makes use of harmonic analysis and of the Durbin–Watson regression. In order to check its worthiness it was applied to two monthly wind-run data records.

The reconstruction approach can be partitioned into two parts: (i) filling missing data and (ii) extending time series backwards. The procedure can be summarized as follows. For the former, from an initial annual mean estimate its monthly values were projected on the basis of seasonal scores (weights) previously

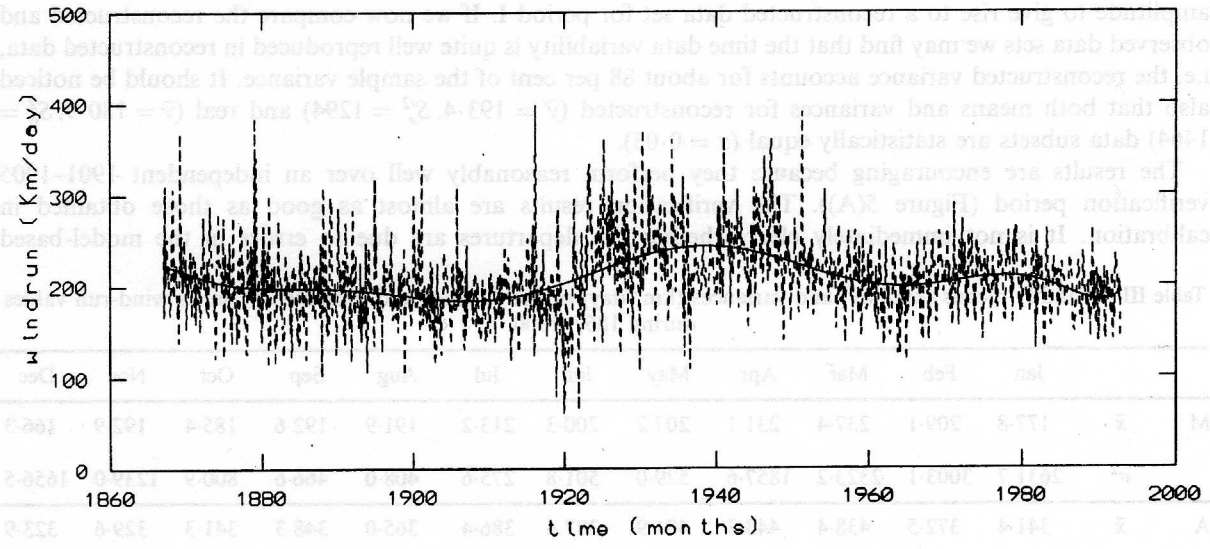

Figure 6. Reconstructed time series of monthly averaged daily wind-run ranging from 1867 to 1992 after applying the method described in this paper (dashed line). Solid line displays a polynomial fitting to the series

assessed. For the latter, the Durbin–Watson regression model was used to remove the autocorrelation in the residuals from a simple regression model. As the estimate of variance from a regression model proves to be lower than the real variance, the DW estimates were transformed by introducing a scale variance-based factor.

The model developed was validated over the independent verification period 1901–1905. The results are encouraging because the verification results are almost as good as those for the calibration period.

The reconstruction was extended back to 1867, which results in the longest climatological wind-run data set in Spain. The reconstructed climatological series is available from the authors.

ACKNOWLEDGEMENTS

We are grateful to the referees for all their valuable comments, corrections and suggestions and also to Dr V. Quesada and Dr P. Cuesta for their useful comments and technical support. Financial support for this work was provided partially by the Comunidad Autónoma de Madrid (Spain), under Contract CAM 286/92 and Grants Orden 316/92. The authors also thank the data supplied by the Spanish National Institute of meteorology (INM) and the Spanish National Geographic Institute (IGN).

REFERENCES

Alexandersson, H. 1986. 'A homogeneity test applied to precipitation data', *J. Climatol.*, **6**, 661–675.

Barry, R. G. and Perry, A. H. 1973. *Synoptic Climatology*, Methuen, London, 555 pp.

Bloomfield, P. 1976. *Fourier Analysis of Time Series: an Introduction*, John Wiley, New York, 258 pp.

Box, G. E. P. and Jenkins, G. M. 1976. *Time Series Analysis: Forecasting and Control*, (revised), Holden-Day, San Francisco, 575 pp.

Canavós, G. C. 1990. *Probabilidad y Estadistica. Aplicaciones y Métodos*, McGraw Hill, México, 651 pp.

Chatterjee, S. and Price, B. 1977. *Regression Analysis by Example*, John Wiley, New York, 228 pp.

Durbin, J. 1953. 'A note on regression when there is extraneous information about one of the coefficients', *J. Am. Statis. Assoc.*, **48**, 799–808.

Durbin, J. and Watson, G. S. 1950. 'Testing for serial correlation in least squares regression, I', *Biometrika*, **37**, 409–428.

Durbin, J. and Watson, G. S. 1951. 'Testing for serial correlation in least squares regression, II', *Biometrika*, **38**, 159–178.

Durbin, J. and Watson, G. S. 1971. 'Testing for serial correlation in least squares regression, III', *Biometrika*, **58**, 1–19.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. 1983. *Understanding Robust and Exploratory Data Analysis*, Wiley, New York, 447 pp.

Johnston, J. 1972. *Econometric Methods*, McGraw Hill, New York, 464 pp.

Jones, P. D., Wigley, T. M. L. and Briffa, K. R. 1987. *Monthly Mean Pressure Reconstructions for Europe (back to 1780) and North America (to 1958)*, DoE Technical Report No. TR037, US Department of Energy, Carbon Dioxide Research Division, Washington, DC.

Katz, R. W. 1988. 'Use of cross correlations in the search for teleconnections', *J. Climatol*, **8**, 241–253.

Makridakis, S., Wheelwright, S. C. and McGee, V. E. 1983. *Forecasting: Methods and Applications*, 2nd edn, Wiley, New York, 926 pp.

Mitchell J. M. Jr., (chairman), Dzerdzeevskii, B., Flohn, H., Hofmeyr, W. L., Lamb, H. H., Rao, K. N. and Wallén, C. C. 1966. *Climate Change*, Technical Note No. 79, World Meteorological Organization No. 195 TP100, Geneva.

Panofsky, H. A. and Brier, G. W. 1968. *Some Applications of Statistics to Meteorology*, Pennsylvania State University Press, Pennsylvania, 224 pp.

Schlesinger, M. E. and Ramankutty, N. 1994. 'An oscillation in the global climate system of period 65–70 years', *Nature*, **367**, 723–726.

Shimshoni, M. 1971. 'On Fisher's test of significance in harmonic analysis', *Geophys. J. R. Astron. Soc.*, **23**, 373–377.

Snedecor, G. W. and Cochran, W. G. 1972. *Statistical Methods*, Iowa State University Press, Iowa, 593 pp.

Solow, A. R. 1987. 'Testing for climate change: an application of the two-phase regression model', *J. Clim. Appl. Meteorol.*, **26**, 1401–1405.