Stochastic Restoration Of Local-Scale Meteorological Records Under Urban Heat Island Signal

Paulo Lucio[1] and Ricardo Deus[2]

[1] CGE - Centre of Geophysics of Évora
Rua Romão Ramalho, 59, 7000-554, Évora – Portugal
(e-mail: pslucio@uevora.pt, ricardo.deus@meteo.pt)
[2] IM - Instituto Meteorologia
Rua C do Aeroporto 1749-077
Lisboa, PORTUGAL

Abstract. One of the biggest constraints to study meteorological fields is due to the fact that the ground-based meteorological network does not operate over a common time period of adequate length. In general, the biggest drawback is that recorded data available must be gap-filled and quality controlled (coherent and consistent) to provide a reliable continuous reference daily/monthly/yearly time series. Hence, this paper is addressed to procedures for reconstruction and evaluation of extremes air temperature time series obeying a sequential strategy divided in two moments: (1) the interpolation considering the cross-correlation and the autocorrelation time-memory; and (2) the spatial interpolation procedure based upon the "optimum distance" between stations. The latter is accomplished subdividing areas of a 2D region into triangles (simplex) to assess the interpolation structure that make use of the altitude of stations as a weight correction factor. Hence, an integrated model for the restoration of time series was developed, which conjugates small-scale space-time interaction between meteorological stations. To validate this work-algorithm, the diagnostic of extreme air temperatures was accomplished based on the analysis of daily time series (1941-2001) from eighteen meteorological stations placed in the Lisboa (Portugal) metropolitan region. As expected, this innovator and robust reconstruction method has good performance, since more information is introduced in the decision-making system.
Keywords: autocorrelation, bias, barycentric coordinates, statistical quality control, time series reconstruction.

1 Introduction

In general, the biggest drawback in climate time series research and diagnostic is that recorded data available must be gap-filled and quality controlled (coherent and consistent) to provide a reliable continuous reference daily $\Rightarrow$ monthly $\Rightarrow$ yearly time series (control or reference series). Hence, this manuscript is addressed to stochastic procedures for reconstruction and evaluation (quality control) of extremes air temperature time series. For this

reason, we have created a Time Series Reconstruction via Integrated (Interactive) Modelling algorithm: MIRS, an integrated model for the restitution of time series, which conjugates small-scale "space-time interaction" between meteorological stations, Fig.1. It is basically subdivided in two major steps: (1) The temporal linear interpolation, considering the time-memory; and (2) The spatial linear interpolation based upon the "optimum distance" between stations. The time series have been recovered and the empirical mean squared error (MSE) has been evaluated, taking into account the comparison between records in local neighbourhood (time and space small-scale) presenting the same climatic (seasonal) characteristics.

Throughout this work some predictors for spatiotemporal processes will be derived. It is realised on bases of linear techniques and assuming some conditions for the processes under study. First, it is important to formalise the notion of spatiotemporal process. Hence, consider a random function $\{Z(s,t) : s \in D \subseteq \Re^n; t = 0, \mp 1. \mp 2, ...\}$, realizations of a spatiotemporal stochastic process. Thus, for fixed t,$\{Z(s,t) : s \in D\} \equiv \{Z_t(s) : s \in D\}$ is a purely spatial processes and for a fixed location $s \in D$, $\{Z(s,t) : t = 0, \mp 1. \mp 2, ...\} \equiv \{Z_s(t) : t = 0, \mp 1. \mp 2, ...\}$is a time series. Hence, a spatiotemporal stochastic process is simply an infinite, possibly correlated, sequence of spatial processes in time or vice-versa. For our purpose we will make the distinction between space-partial trajectories $\{Z(t) : t = 0, \mp 1. \mp 2, ...\}$ (reconstruction on temporal domain) and time-partial trajectories $\{Z(s) : s \in D\}$ (reconstruction on spatial domain) of the spatiotemporal stochastic process [Kyriakidis and Journel, 1999].

Let $\hat{Z}(x_0)$ the predictor of a random function on partial trajectories based on the realisations $\{Z(x_1), Z(x_2), ..., Z(x_2)\}$; the prediction error associated is defined as $\varepsilon(x_0) = \hat{Z}(x_0) - Z(x_0)$ and the mean squared error, which is interconnected with the prediction's quality, is $MSE\left[\hat{Z}(x_0)\right] = E\left[\hat{Z}(x_0) - Z(x_0)\right]^2$. The best prediction function in terms of the minimum MSE [Graybill, 1976] is given by:

$$\psi_0\left[Z(x_1), Z(x_2), ..., Z(x_n)\right] = E\left[Z(x_0)|Z(x_1), Z(x_2), ..., Z(x_n)\right],$$

and the best linear prediction function is:

$$\psi_0^*\left[Z(x_1), Z(x_2), ..., Z(x_n)\right] = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i),\ \lambda_i \in \Re,\ i = 1, 2, ..., n,$$

where $\hat{Z}(x_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i)$. Moreover, it is well-known that the MSE can be written as follows: $MSE\left[\hat{Z}(x_0)\right] = Var\left[\hat{Z}(x_0) - Z(x_0)\right] + B^2\left[\hat{Z}(x_0) - Z(x_0)\right]$. An optimum predictor must be unbiased and ψ_0^* is unbiased $B\left[\hat{Z}(x_0) - Z(x_0)\right] = 0$, since

Fig. 1. Map of Lisboa metropolitan area with the location of the automatic meteorological stations of the urban network. Identification of the eighteen meteorological stations and their respective altitudes: 1. Torres-Vedras/Dois-Portos (ID: #139) – 150m; 2. Salvaterra de Magos (ID: #141) – 5m; 3. Colares Sarrazola (ID: #148) – 55m; 4. Sintra (ID: #149) – 200m; 5. Cabo da Roca (ID: #150) – 142m; 6. Paiã/Escola-Agrícola (ID: #153) – 70m; 7. Sacavém (ID: #155) – 9m; 8. Cabo Ruivo (ID: #157) – 16m; 9. Sassoeiros (ID: #160) – 50m; 10 – Lisboa/Tapada-da-Ajuda (ID: #162) – 37m; 11. Lavradio (ID: #166) – 6m; 12. Sintra/Granja (ID: #532) – 134m; 13. Montijo/Base-Aérea (ID: #534) – 14m; 14. Lisboa/Geofísico (ID: #535) – 77m; 15. Lisboa/Portela (ID: #536) – 103m; 16. Alverca/Base-Aérea (ID: #537) – 2m; 17. Ota/Base-Aérea (ID: #539) – 40m; 18. Lisboa/Gago-Coutinho (ID: #579) – 104m.

$$E\left\{\psi_0\left[Z(x_1), Z(x_2), ..., Z(x_n)\right]\right\} = E\left\{E\left[Z(x_0)|Z(x_1), Z(x_2), ..., Z(x_n)\right]\right\} = E\left[Z(x_0)\right]$$

Then ψ_0^* is the best linear unbiased predictor (BLUP) of $Z(x_0)$. Therefore, minimise the MSE is reduce the variance of prediction: $Var\left[\varepsilon(x_0)\right] = Var\left[\hat{Z}(x_0) - Z(x_0)\right]$ and $\{\varepsilon(x_0)\}$, $\forall x_0$ unsampled points, determine series of uncorrelated random variables, supposed to be a zero mean and constant variance - white noise. Moreover, $\hat{Z}(x_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(x_i) + \varepsilon(x_0)$, where $\sum_{i=1}^{n} \lambda_i Z(x_i)$ is considered the large scale trend surface (1st order component in time or space domain; defined by the wide meaning neighbourhood influence zone) and $\varepsilon(x_0)$ the local component (2nd order factor) or residuals.

2 Daily Reconstruction

2.1 Temporal Domain

The reconstruction model is based on: (1) the use of the own series for filling records without information, considering the strong daily relationship of the internal variation between minimum and maximum air temperature, this

association can be verified performing the cross-correlation function analysis between both air temperature attributes for each station in time $t = t_0$, the antecedent t - 1 and the subsequent $t+1$ values (two-day time influence); (2) and (3) the use of the serial correlation, considering the strong connection between a record in time t, the antecedent t-2, t-1 and the subsequent $t+1$, $t+2$ values (four-day time influence). Hence, the autocorrelation and the partial autocorrelation functions are applied for each series of data, whenever an isolated missing value is found, taking into consideration a second order autoregressive (AR(2)) model [Box *et al.*, 1994].

The coefficient of correlation is used as a measure of the strength of linear association between both variables, a measure of the interdependence of two random variables that ranges in value from -1 to +1, indicating perfect negative correlation at -1, absence of correlation at zero, and perfect positive correlation at +1. The cross-correlation function is a standard method of estimating the degree to which two series are correlated.

Particularly, in this first reconstruction phase, three lags (days) are considered, taking into account the presumed strong linear association between a record in timet, the previous t - 1 and the subsequent $t+1$ observed values:

$$\begin{aligned}
\beta_1(t) &= \lambda_1 \cdot TMAX(t-1) + \lambda_2 \cdot TMAX(t) + \lambda_3 \cdot TMAX(t+1) \\
\alpha_1(t) &= \lambda_1 \cdot TMIN(t-1) + \lambda_2 \cdot TMAX(t) + \lambda_3 \cdot TMIN(t+1), \\
\hat{X}(t) &= TMIN(t) = [\alpha_1(t) + \beta_1(t)] + \varepsilon(t) \qquad (1) \\
\beta_2(t) &= \lambda_1 \cdot TMIN(t-1) + \lambda_2 \cdot TMIN(t) + \lambda_3 \cdot TMIN(t+1), \\
\alpha_2(t) &= \lambda_1 \cdot TMAX(t-1) + \lambda_2 \cdot TMIN(t) + \lambda_3 \cdot TMAX(t+1), \\
\hat{Y}(t) &= TMAX(t) = [\alpha_2(t) + \beta_2(t)] + \varepsilon(t) \qquad (2)
\end{aligned}$$

where $\lambda_1 = \frac{\hat{\rho}(-1)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$, $\lambda_2 = \frac{\hat{\rho}(0)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$, $\lambda_3 = \frac{\hat{\rho}(1)}{(\hat{\rho}(-1)+\hat{\rho}(0)+\hat{\rho}(1))}$ and $\varepsilon(t)$ denote the empirical series of uncorrelated random variables (residues), whose the ensemble is supposed to be a white noise. It is well known that this prediction method is not optimum at all, since it considers that the attributes are correlated when a linear change in one variable is associated with a change in another one - an unrealistic assumption for daily temperature extremes. The serial correlation is the correlation of a variable with itself over successive time intervals. In climatology we use serial correlation to determine how well the past climate could predicts the future climate and impacts. When the correlation is calculated between a series and a lagged version of itself it is called autocorrelation. The autocorrelation is a correlation coefficient. However, instead of correlation between two different variables, the correlation is between two values of the same variable at times. A high correlation is likely to indicate a periodicity in the signal of the corresponding time duration. The partial autocorrelations, like autocorrelations,

are correlations between sets of ordered data pairs of a time series; partial autocorrelations measure the strength of relationship with other terms being accounted for. In practice, for daily data, only two lags are necessary to be considered, taking into account the presumed strong association between a record in time t, the previous (t - 2, t - 1) and the subsequent (t + 1, t + 2) observed values (to get supplementary available information – backward and forward second order autoregressive AR(2) predictor):

$$\alpha(t) = \frac{\hat{\varphi}(-2).X(t-2) + \hat{\varphi}(-1).X(t-1) + \hat{\varphi}(1).X(t+1) + \hat{\varphi}(2).X(t+2)}{\hat{\varphi}(-2) + \hat{\varphi}(-1) + \hat{\varphi}(1) + \hat{\varphi}(2)}$$

$$\Downarrow$$

$$\alpha(t) = \lambda_1.(X(t-2) + X(t+2)) + \lambda_2.(X(t-1) + X(t+1)),$$

$$\hat{X}(t) = \alpha(t) + \varepsilon(t), \tag{3}$$

where$\lambda_1 = \frac{\hat{\varphi}(2)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$, $\lambda_2 = \frac{\hat{\varphi}(1)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$ and the ensemble $\varepsilon(t)$ is supposed to be a white noise process. The partial autocorrelation at a lag κ is the correlation between residuals at time t from an autoregressive model and observations at lag κ with terms for all intervening lags present in the autoregressive model. The PACF associated to a stochastic process is defined as a sequence of $\hat{\varphi}(\kappa)$'s obtained by the resolution of the Yule-Walker equations for κ = 1,2,3,...:

$$\alpha(t) = \frac{\hat{\varphi}(-2).X(t-2) + \hat{\varphi}(-1).X(t-1) + \hat{\varphi}(1).X(t+1) + \hat{\varphi}(2).X(t+2)}{\hat{\varphi}(-2) + \hat{\varphi}(-1) + \hat{\varphi}(1) + \hat{\varphi}(2)}$$

$$\Downarrow$$

$$\alpha(t) = \lambda_1.(X(t-2) + X(t+2)) + \lambda_2.(X(t-1) + X(t+1)),$$

$$\hat{X}(t) = \alpha(t) + \varepsilon(t), \tag{4}$$

where$\lambda_1 = \frac{\hat{\varphi}(2)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$, $\lambda_2 = \frac{\hat{\varphi}(1)}{2\hat{\varphi}(1)+2\hat{\varphi}(2)}$ and the ensemble $\varepsilon(t)$ is supposed to be a white noise process. It is furthermore well known that these of prediction linear methods (2) and (3) are not favourable at all, while they consider that the extremes attributes are correlated when a linear change in one day is associated with a change in the adjacent two days - an improbable postulation, mainly considering severe events. However, we also believe that is worthwhile to make an effort in this direction for regular time series reconstruction.

We considered a "bivariate linear interpolation", for reconstructing daily extreme air temperatures, since for each t - 1 ,t and t + 1 three values (2 of TMIN and 1 of TMAX or vice versa) were available; once verified the strong correlation (in phase - same day) between TMIN and TMAX and a less strong or even weak one (out of phase). If we had used an in phase model we would have a colinearity problem due to the great dependence between

the meteorological variables, implying a certain redundancy. So the 1-day lag would be strongly satisfactory for our "bivariate linear interpolation". It was done in the practice, and the equations 3 and 4 are reliable interpolations scheme to reconstruct these kind of meteorological variables taking into account the coupled phenomena.

When all these linear interpolation approaches are applicable and appropriate (the weighted correlations are statistically significant) the decision-making criterion is based on the minimum empirical MSE among the ensembles. These temporal stochastic reconstructions were achieved for overall meteorological stations data series. In Fig.2 we cover the residuals graphical summary that includes: histogram with an overlaid normal curve, box-plot, 95% confidence intervals for the means and 95%confidence intervals for the median. The graphical summary also displays a table of descriptive statistics. No more than 0.3% of the missing values (Tab.1) records were recovered considering the temporal domain! Hence, let us move now to the spatial approach.

2.2 Spatial Domain

The reconstruction model is based on two influence factors: (1) the Euclidean and angular distances, defining a triangle-based network and the areas of each elementary cell (2D simplex); and (2) the Euclidian distance between stations and the respective difference between altitudes (heights). The objective of the first scheme is to construct elementary cells (simplex structure) by triangulation of the convex sub-region S $\subseteq \Re^2$ and the interpolating tool is adopted using the areas of a region subdivided in triangles. The Voronoi region of an object is the region of space closer to the given object than to any other object of the sample. The set of Voronoi triangulations for a set of spatial objects, called a Voronoi diagram (also known as a Dirichlet tessellation or Thiessen polygons), provides a partition of a point-pattern according to its spatial structure. Features of this kind can also be used for analysis of the underlying point process. In practice, the triangulated network of a sub-region of the two-dimensional convex envelope must be determined. Then, let $m \geq 3$ events of a random sample over a sub-region S $\subseteq \Re^2$ and assume that the two-dimensional convex hull of this sub-region has area |A| and that the partition produces (M = 2m - ν - 2) triangles, where ν is the number of extreme points of the two-dimensional partition of the unity$A$, with areas$|A_1|, |A_2|, ..., |A_{2m-\nu-2}|$, respectively. Hence, for this schematic prediction we employ the following topological concepts. The tessellation and oriented areas - A closed ensemble K of the $n-$dimensional space $\Re^n$ is convex if for any $x \in K$, $y \in K$ and $0 \leq \lambda \leq 1$, the linear combination$\{\lambda x + (1 - \lambda)\, y \in K\}$. A point $\varpi \in K$ is an extreme point of K if it may not be written as a convex combination of κ different elements of ϖ. A two-dimensional convex envelope of a finite set $C = \{p_1, p_2, ..., p_m\}$ of m

events of $\Re^n$ is defined as the set of all the convex combinations of elements:

$$conv(C) = \{\sum_{i=1}^{m} \lambda_i p_i | \lambda_i \geq 0 \text{ and } \sum_{i=1}^{n} \lambda_i = 1\}. \quad (5)$$

The polygon $C = \{p_1, p_2, ..., p_m\}$ is convex if, and only if, each internal angle is convex, *i.e.*, if each triangle p_{i-1}, p_i, p_{i+1} has the same polygon orientation. A triangle (the 2D simplex structure) defines a coordinate system in the plane (Farin, 1993). Let us consider p_{i-1}, p_i, p_{i+1} non-collinear events onto a triangle $\Delta \subset \Re^2$. Each point $p \in \Delta \subset \Re^2$ can be written as a unique linear combination satisfying

$$\{\sum_{i=1}^{3} \lambda_i p_i : \lambda_i \geq 0 \text{ and} \sum_{i=1}^{3} \lambda_i = 1\}. \quad (6)$$

The parameters $(\lambda_1, \lambda_2, \lambda_3)$ are the barycentric coordinates of p (the relative centroids) in relation to(p_{i-1}, p_i, p_{i+1}). For $(p, p_{i-1}, p_i, p_{i+1})$ with $p = (x, y)$ and$p_j = (x_{ji}, y_{ji}), j = i-1, i, i+1$, the parameters $(\lambda_{i-1}, \lambda_i, \lambda_{i+1})$ satisfying some initial conditions are solutions of the system represented below:

$$\begin{cases} \lambda_{i-1}x_{i-1} + \lambda_i x_i \ + \lambda_{i+1}x_{i+1} = x \\ \lambda_{ii-1}y_{i-1} + \lambda_i y_i + \lambda_{i+1}y_{i+1} = y \\ \lambda_{i-1} + \lambda_i + \lambda_{i+1} \ = 1 \end{cases}. \quad (7)$$

The determinant (Δ) of the solution matrix of the system above is the scalar 2S,

$$\Delta = \begin{vmatrix} x_{i-1} & x_i & x_{i+1} \\ y_{i-1} & y_i & y_{i+1} \\ 1 & 1 & 1 \end{vmatrix} = 2S, \quad (8)$$

where S is the area of the triangle (p_{i-1}, p_i, p_{i+1}). The values of each one of the elements $(\lambda_{i-1}, \lambda_i, \lambda_{i+1})$ can be determined by Cramer's rule:

$$\lambda_{i-1} = \frac{S_{p\ p_i\ p_{i+1}}}{S_{p_{i-1}\ p_i\ p_{i+1}}} = \frac{S_{i-1}}{S}, \lambda_i = \frac{S_{p_{i-1}\ p\ p_{i+1}}}{S_{p_{i-1}\ p_i\ p_{i+1}}} = \frac{S_i}{S}, \lambda_{i+1} = \frac{S_{p_{\prime}{}_{-1}\ p_i\ p}}{S_{p_{i-1}\ p_i p_{i+1}}} = \frac{S_{i+1}}{S}. \quad (9)$$

The system (10) determines oriented areas. The weights $\lambda_j, j = i-1, i, i+1$ are positive if and only if $S_j, j = i-1, i, i+1$, and S has the same orientation (signal). The barycentric linear interpolation can be used to determine, for continuous phenomena, the unknown values at unsampled points in the spatial point pattern. The barycentric interpolation - Neighbour relationships can also be weighted. Weights based on barycentric coordinates are the subject of this section. Given an ensemble $C = \{x_1, x_2, ..., x_m\}$ of events and real values$f(x_i), i = 1, ..., m$, a piecewise linear function F(x) , defined inside an adequate domain D, such as$F(x) = \sum f(x_i),\ i = 1, ..., m$, can be obtained. The natural choice for this

domain D is a $conv(C)$. However, given a point$x \in conv(C)$, the calculation of F(x) is not obvious. The basic idea is to write$x \in conv(C)$ as a disjoint union of an ensemble of triangles (the simplex). Based on the construction of this triangle network, given $x \in conv(C)$, it can be determined whether x belongs to a particular triangle p_{i-1}, p_i, p_{i+1}, and then F(x) can be computed using equations (7), (8) and (9). The fundamental step in this approach consists of solving the related problem of the triangulation of an ensemble $C = \{x_1, x_2, ..., x_m\}$ based on the construction of $conv(C)$. Notice that, on a two-dimensional space, the triangulation does not exhibit the property of unicity (*i.e.* there are several ways to triangulate a convex network). However, it is possible to determine the optimal number of triangles on the $conv(C)$. Hence, a robust first-order scheme based on barycentric coordinates is used to interpolate the observations at elementary cell vertices on a denser grid. For each unsampled location, the values are evaluated and updated by linear interpolation using the values at the vertices of the triangle. Notice that the precision of the linear interpolation can be estimated with the same properties as the kriging methodology; and without loss of generality the variogram model can be considered linear.

When this approach is legitimate (the weights can be determined, *i.e.*, we can define a triangular network do interpolate the unknown inner point) the stochastic reconstruction is achieved for a particular meteorological station (Fig.3), presenting missing values, represented in the barycentric coordinates system [Lucio and Brito, 2004]. In effect, the spatial linear interpolation is the representation of the data as a parametric model plus a random process function of space$\hat{Z}(s) = \mu(s) + \varepsilon(s)$. The parametric model $\mu(s) = \sum_{i=1}^{n} \lambda_i Z(s_i)$, representing the smooth variation and $\varepsilon(s)$ the deviations from$\mu(s)$.

To clarify the interpolation scheme, consider an ensemble and suppose that we would like to know an attribute value over the point $P = (303.8825\ ,\ -141.2425)$, this point is undoubtedly inside the triangle with vertices $P_{i-1} = (-2436.0075\ ,\ 3814.9875)$, $P_i = (1959.3325,\ 348.9375)$, $P_{i+1} = (172.7925,\ -4022.4825)$, with total area equal to $12,703,497$. The barycentric coordinates satisfy the systems (10), and the output of the program (algorithm in S-Plus) gives the solution: $\lambda_1 = \frac{3,180,636}{12,703,497} = 0.2503748$, $\lambda_2 = \frac{3,946,190}{12,703,497} = 0.3106381$, $\lambda_3 = \frac{5,576,670}{12,703,497} = 0.438987$, with $\sum_{j=i-1}^{i+1} \lambda_j = 1$. Taking an associated continuous function into account, *e.g.* temperature, we obtain an estimate for each point using barycentric interpolation. This estimate value is given by the expression $\hat{Z}(s) = \sum_{j=i-1}^{i+1} \lambda_j Z(s_j)$, where $Z(s_j)$ is the attribute observed on P_j. Hence, in our illustration, let $Z(P_{i-1}) = 18^oC$, $Z(P_i) = 20^oC$ and $Z(P_{i+1}) = 25^oC$:

$$\hat{Z(P)} = 0.2503748 \times 18^oC + 0.3106381 \times 25^oC + 0.438987 \times 20^oC = 21.05244^oC.$$

The second phase considers the Euclidian distance between stations: $\Delta\delta^2$ and the respective difference between altitudes (heights): Δz^2, calculated for each pair of stations ($\Delta z_{j,k} = z_k - z_j$,$\Delta\delta_{j,k} = \sqrt{(x_k - x_j)^2 + (y_k - y_j)^2}$,$j, k = 1, ..., 18$) to construct the matrix $\Omega = \frac{1}{\Delta h^2}$ (the inverse of the hypotenuse: $\Delta h^2 = \Delta\delta^2 + \Delta z^2$) that is used as an issue to recover the missing values and factor corrector for the

barycentric interpolation. The value to be predicted for a station meteorological ($\kappa \in K$, where K is a closed ensemble), which has no record in a day t is based on the weighted sum ($\frac{\omega_{[i,j]}}{\sum\limits_{j} \omega_{[j,\kappa]}}$) of the values of the records of all other available stations ($i \neq j = 1, ..., m$) at the same instant t using the idea of "nearest neighbourhood" and the symmetric weight matrix is given by:

$$\Omega = \begin{bmatrix} 0 & \omega_{[2,1]} & \cdots & \omega_{[m,1]} \\ \omega_{[1,2]} & 0 & \cdots & \omega_{[m,2]} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{[1,m]} & \omega_{[2,m]} & \cdots & 0 \end{bmatrix}. \tag{10}$$

In fact, we consider that height is a dependent variable of longitude ($x \in K$) and latitude ($y \in K$) in terrain surface (this idea is widespread used in mapping) and apply the information $\lambda_{\kappa,j} = \frac{\omega_{[i,j]}}{\sum\limits_{j} \omega_{[j,\kappa]}}$as an improvement factor after 2D linear interpolation based on barycentric coordinates [Farin, 1993]. Consequently,

$$\hat{Z}(\kappa) = \hat{Z}_0(\kappa) + \sum_{j} \lambda_{\kappa,j} Z(j) + \varepsilon(\kappa). \tag{11}$$

where $\hat{Z}_0(\kappa)$ is the result of the barycentric interpolation (when applicable, otherwise it is zero) and $Z(j)$ are the contributors (with valuable data) meteorological stations.

Moreover, our method allows us to determine an interpolation criterion, similar to kriging methodology [Kyriakidis and Journel, 1999] since the variogram has to be linear, based on the barycentric coordinates in influence zones. The MSE can be considered independent (uncorrelated) and approximately zero-centred. In addition, they give us an idea about the spatial interpolation misfit based on the variance of prediction. This stochastic reconstruction was achieved for overall meteorological stations data series. In practice, the available records of a station are used to predict the extreme air temperature attributes of the missing value records, considering the neighbourhood and the own station historical records.

As a result of these applications all the series were recovered, except for the first six months of 1997 (181 days without available observations), observed data do not exist in any station or at least it is presumed to not exist. So, we now consider the monthly model identification and characterisation of extreme time series making use of an appropriated forecast model, which might be extrapolated to high levels of the climatological process: the autoregressive integrated moving average (ARIMA) modelling (cf. [Box and Jenkins, 1976]).

THE SCHEME:

DAILY DATA $\Rightarrow$ MONTHLY DATA

TMIN $\Rightarrow$ Min {TMIN}, Mean {TMIN} and Max {TMIN},

TMAX $\Rightarrow$ Min {TMAX}, Mean {TMAX} and Max {TMAX}.

In this work, no more than the extremes was analysed. This stochastic reconstruction was achieved for overall meteorological stations data series and the reference period of validation was 1992-1996.

3 Conclusions

These time series recoverable approach is a very simple way to offer high efficiency results for a low computational cost. Furthermore, this alternative method allows barycentric interpolation of the unsampled points into a two-dimensional simplex (triangular) framework. Moreover, our method allows us to determine an interpolation criterion, similar to kriging methodology since the variogram has to be linear, based on the barycentric coordinates in influence zones. Nevertheless, we can identify two main sources of uncertainties:

i) The induced error when applying the autocorrelation function in the reconstruction of the daily series varies between -3°C and 3°C;

ii) The error generated when estimating values of the air temperature for missing values record considering the spatial reconstruction depends on certain expected conditions. When few stations contribute for filling the gaps, the associate error presents fail values, e.g. for the last 5 years only two stations present records (Lisboa/Geofísico - urban and Lisboa/Gago-Coutinho - suburban); the result under the "heat island effect" may overestimate (contaminate) the calculated values for the other stations;

iii) The integration of new parametrization on the spatial interpolation procedure, like land declination, ocean/river distance can reduce the error associated with this step.

ACKNOWLEDGEMENTS The authors wish to thank the contributions of the Portuguese Meteorological Institute (IM – Portugal). P. S. Lucio was sponsored by a grant from FCT (Portugal). Grateful thanks to the referees.

References

[Box and Jenkins, 1976]G.E.P. Box and G.M. Jenkins. *Time series analysis: forecasting and control.* Holden-Day, San Francisco, 1976.

[Box *et al.*, 1994]G.E.P. Box, G.M. Jenkins, and G.C. Reinsel. *Time series analysis: forecasting and control.* Prentice Hall, 1994.

[Farin, 1993]G. E. Farin. *Curves and surfaces for computed aided geometrical design.* Academic Press, London, 1993.

[Graybill, 1976]F. A. Graybill. *Theory and application of the linear model.* Duxbury Press, Massachusetts, 1976.

[Kyriakidis and Journel, 1999]P. C. Kyriakidis and A. G. Journel. Geostatistical space-time models: a review. *Mathematical Geology*, pages 651–684, 1999.

[Lucio and Brito, 2004]P. S. Lucio and N. L. Brito. Detecting spatial randomness: A stat-geometrical alternative. *Mathematical Geology*, pages 79–99, 2004.

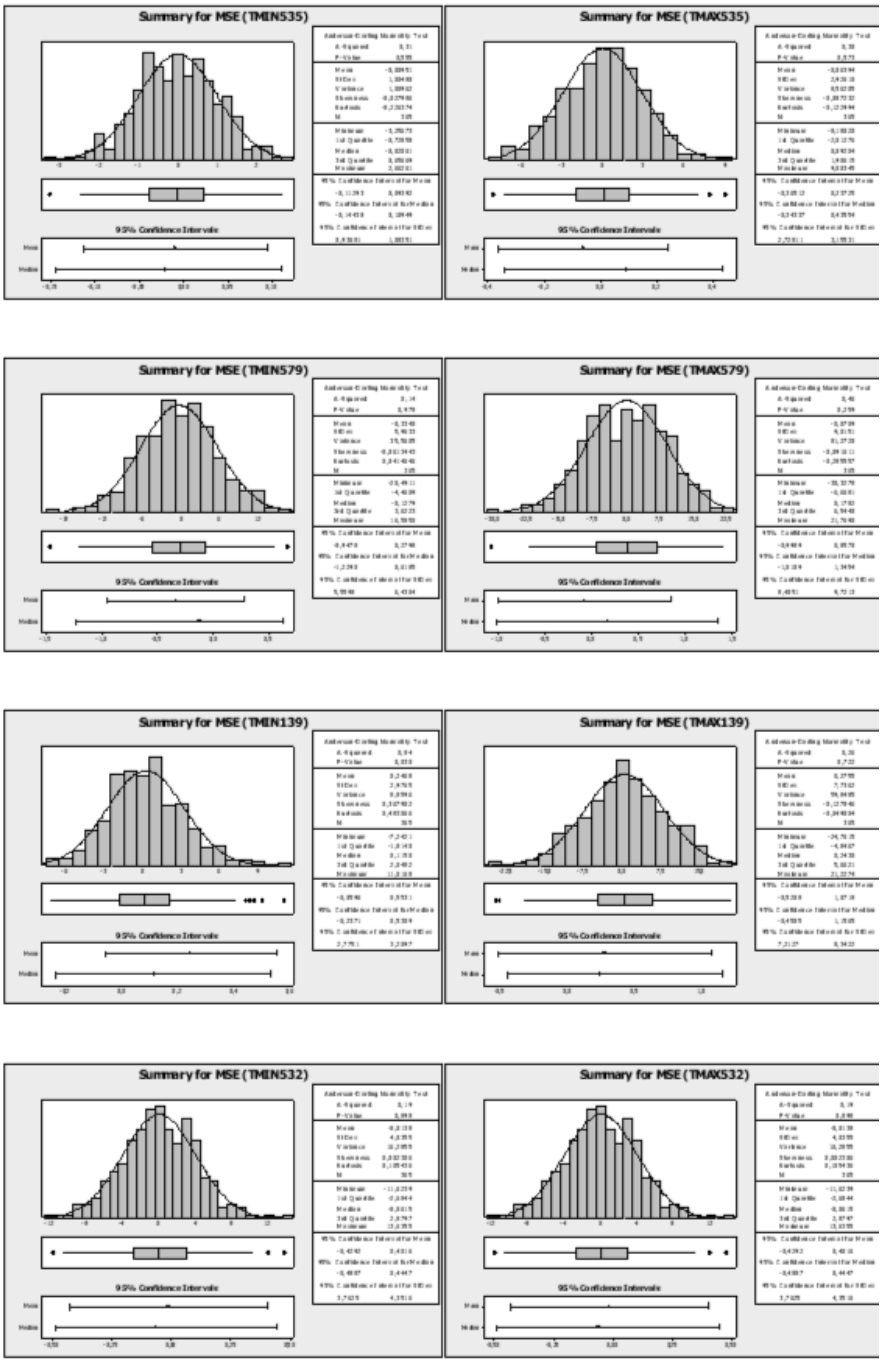

Fig. 2. The residuals graphical summary for TMIN (left) and TMAX (right): Lisboa/Geofísico (535), Lisboa/Gago-Coutinho (579), Torres-Vedras/Dois-Portos (139) and Sintra/Granja (532).

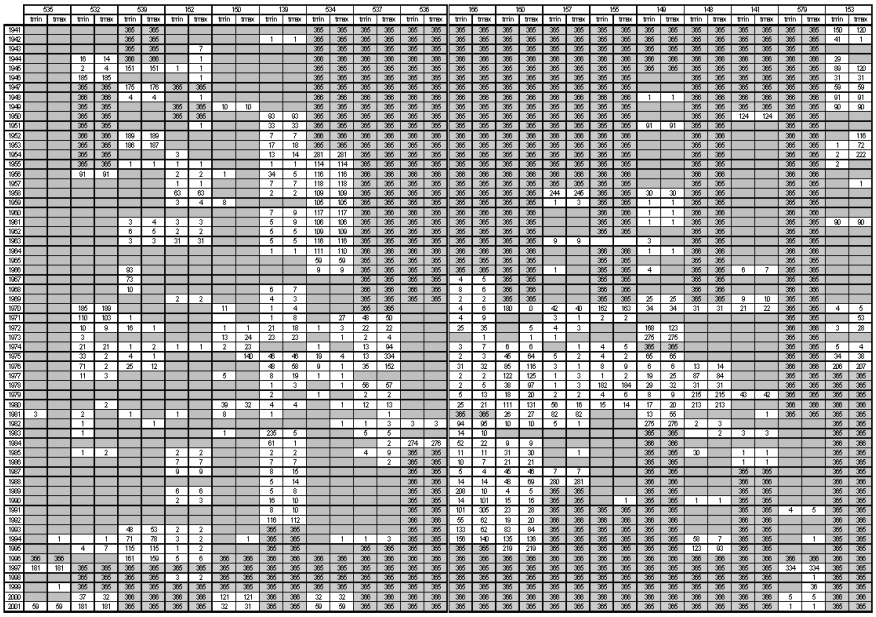

Fig. 3. Original meteorological data time series size. In darkish empty cells – complete year of records with good-quality information for TMIN and/or TMAX; in white and darkish filled cells – number of records without information in a year and "complete" year without records, respectively. Note that the data series from Lisbon/Gago-Coutinho (579) has no data for the period that precedes 1982.

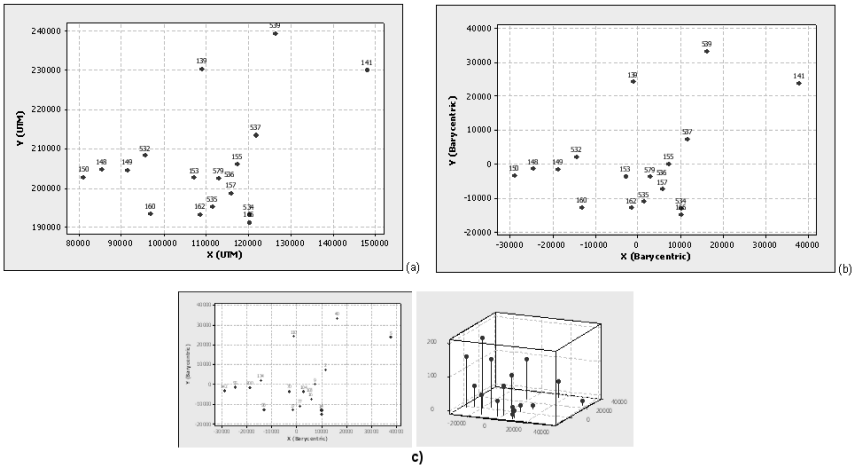

Fig. 4. The geographical position of the eighteen meteorological stations (the data labels) in (a) the UTM transform system and (b) the barycentric coordinates system. The altitude (data label on the left side) representation of the network superimposed over the barycentric coordinates system (c).