Proceedings of 8 ICSHMO, Foz do Iguaçu, Brazil, April 24-28, 2006, INPE, p. 29-36.

A BAYESIAN APPROACH FOR RECOVERING AND HOMOGENISING METEOROLOGICAL TIME SERIES.

P. S. Lucio[1,2]*; F. C. Conde[1]; A. M. Ramos[1]

[1]Centro de Geofísica de Évora (CGE) – Apartado 94, 7000-554 Évora – Portugal.
[2]Instituto Nacional de Meteorologia (INMET) – Brasília DF – Brazil.

pslucio@uevora.pt , fabconde@uevora.pt, andreara@uevora.pt,

**Corresponding author address:*
Instituto Nacional de Meteorologia (INMET)
Eixo Monumental Sul – Via S1 – Setor Sudoeste
70680-900 – Brasília DF - Brasil
pslucio@uevora.pt / paulo.lucio@inmet.gov.br

INTRODUCTION.

The biggest drawback in long-term meteorological time series analysis is that recorded data available must be gap-filled and quality controlled to provide a reliable continuous homogeneous reference series (where divergences are caused only by variations in weather and climate). A common problem in numerical climate characterization is the spatiotemporal processing (integration or interpolation) of data from different types and different origins or accuracy (the space-time change of support problem).

Data reconstruction is a methodology developed by climate scientists and meteorologists to remove inconsistencies in a time series due to factors unrelated weather, such as station location change, station environment change or change in instrumentation. A reconstructed time series behaves as if the station observed weather throughout its history using its current configuration. Objective: Availability, reliability and homogeneity of the historical series of meteorological data. The development of a continuous and complete daily dataset are useful in a variety of meteorological and hydrological research applications.

Time series is a special case of symbolic regression and can be done using the framework of mathematical modelling by an artificial intelligence network. The Artificial Neural Network (ANN) explores the dependence of meteorological attributes as a function of space & time on inputs to the computer simulations. The use of ANN has been recognized recently as a promising way of making predictions on time series, detecting irregular behaviour.

The Stochastic Artificial Neural Network approach using Empirical Bayesian Updating seems to be an important tool for the propagation of the related weather information to provide practical geostatistics solution of uncertainties associated with the interpolation, capturing the spatiotemporal structure of the data. The basic idea is to import the entire posterior distribution from other locations allowing prediction of unsampled weather parameters using spatial related sampled information. The temporal dependence of model parameters is evaluated in a Bayesian framework. A model is used to predict the process of interest Q at the time t. This Multivariate Stochastic procedure uses the available related weather data sets and climate proxies (monitoring and assembling network sites).

The Bayesian solution is the posterior predictive function. In practice, the posterior predictive distribution is obtaining by an updating of the a priori distribution of the Q(s,t) at local s. The spatiotemporal dependence can be explored by examining the distribution of nearest-neighbour distances. To validate the work-algorithm, the diagnostic of homogenised minimum-maximum temperature and total daily precipitation was accomplished.

The spatial distribution of temperatures and precipitation are summarised by the subjective descriptive four-moment measures: Mean, Variance, Skewness and Kurtosis, giving support to spatial pattern

recognition. A number of homogeneity tests with kinds to detect non homogeneities were employed (methods currently used Chow & Pettitt, SNHT Alexandersson, Range Buishand, Von Neumann ratio and Craddock tests) and the effect of natural variability is established taking into account ensembles of consecutive years. The significance of the adjustments was tested using the Robust Modified Wilcoxon Rank Sum Test. The Customised Kendall t Test is used as a non-parametric method to test the significance of trends. As expected, this robust reconstruction method has good performance; since more information is introduced in the decision-making system (the conclusion highlights the use of climate proxies response as potential weather predictor).

PROJECT APPROACH.

The biggest drawback in long-term meteorological time series analysis is that recorded data available must be gap-filled and quality controlled to provide a reliable continuous homogeneous reference series (where divergences are caused only by variations in weather and climate). Hence, spatiotemporal integration is required!

This work consists on the reconstruction of long-term weather time series for a number of Brazilian localities, consubstantiating the spatial consistency, apparent cycles and respective trends. We capture the intrinsic dynamics of atmospheric activities, producing good long-term forecasting for periods of at least a complete cycle of ENSO/PDO/NAO and Sunspots. It seems that the dynamics is essentially non-chaotic in this time scale, but perturbed by a fairly large amount of noise. Moreover, some meteorological variables over Brazil could be accurately predicted taking into account the model developed by artificial neural network. This approach recognizes very well the mutual dependence between spatiotemporal temperature and rainfall variability.

Project: "Spatiotemporal Stochastic Characterization of the Brazilian Climatology." A research program with the aim of better investigating the impact of data quality and homogeneity issues (filtering "anthropogenic" factors) on the detection of Brazilian weather attributes: temperatures, precipitation, insolation, etc., "trends", "bias" and "shifts", in the last century. The series must be free of discontinuities, gaps and spurious values. The series must be homogeneous. **Design:** "Determination of Good Quality Reference Climatology."

Background: In JON K. EISCHEID et al, 2000 (Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States, Journal of Applied Meteorology, AMS), six different methods of spatial interpolation were used to create the serially complete dataset for the western United States (all states west of the Mississippi River) includes 2034 minimum-maximum temperature stations and 2962 total daily precipitation locations. The methods were: (1) the normal ratio method (NR); (2) simple inverse distance weighting (IDW); (3) optimal interpolation (OI); (4) multiple regression using the least absolute deviation criterion (MLAD); (5) the single best estimator; and (6) the median (MED) of the previous five methods (Eischeid et al. 1995). Results: The interpolation schemes were evaluated by monthly integration method. The cross-validation of the results indicated a distinct seasonality to the efficiency of the estimates, although no systematic bias in the estimation procedures was found. Statistical summaries were generated using cross correlations between observed daily values and those estimated for each of the six different methods described. The six techniques respond to variations in season and geography, and the best estimation method is selected based on the efficiency of the estimate over time. The cross correlations were used to measure the efficiency of each method, and the method that exhibits the highest correlation relative to the other methods is utilized to replace missing values.

Additional investigations performed by the Northeast Regional Climate Center (DeGaetano et al., 1993) have shown that regression based methods of data estimation tend to be more accurate than within-station methods. Additional work (Huth and Nemesova, 1995) has shown that other weather elements, such as relative humidity, wind speed, and cloudiness, contribute very little to

regression-based methods and that temperature at neighboring stations has by far the highest spatial correlations. DeGaetano goes on to mention that "while such methods are useful over limited areas, they are computationally intensive and therefore not feasible when data estimates are needed for a large number of stations over a long period of time" (DeGaetano et al., 1993). These limitations have been partially overcome with the use of new high-speed workstations and large mass storage capabilities that now provide the horsepower required to perform these intensive calculations in a reasonable time period.

Because estimates are required for each day separately over a variety of terrain with a differing number of available surrounding observations, we have chosen a different method for filling meteorological gaps.

ARTIFICIAL NEURAL NETWORK (ANN) - EMPIRICAL BAYESIAN UPDATING.

Lemma 1: The prior information at time t can be modelled by a temporal prior function:

$$Y(\Theta(t)) = f(\Theta(t) \mid Z(t1),\ldots, Z(tn)).$$

The efficiency of the temporal prediction process depends on some considerations, as the decision of stationarity.

Lemma 2: The model of temporal dependence allows an Empirical Bayesian Updating of any prior $Y(\Theta(t))$ by "neighbouring" related data.

The basic idea is to interpret the prior distribution $Y(\theta \in \Theta(t))$ as realisations of the corresponding temporal random function $Y(\Theta(t))$, which allows an updating of the prior distribution, $f(\Theta(t))$. The implementation is via Gibbs sampler, where the degree of belief (credibility) is assessed to exploit the uncertainties associated with the interpolation process!

The spatiotemporal dependence can be explored by examining the distribution of nearest-neighbour distances. The bandwidth determines the amount of smoothing of the point pattern (using the weighted Euclidean distance). This study estimated bandwidth as the space-time average (k, φ, θ) nearest-neighbour "distance" among points. The value of (k, φ, θ) is chosen by the analyst to specify the desired degree of smoothing of the data. Small k (φ, θ) values result in a small spatial (temporal) bandwidth, producing a spiky map with little smoothing (good time dependence). Larger k (φ, θ) values result in a larger spatial (temporal) bandwidth and smoother density map (fake time dependence).

EXPERIMENTAL DATASET.

The 27 (26+1DF) states reflect a wide variety of terrain and a diversity of climatic regimes, which allows a means for testing the efficacy of daily estimates for regional and seasonal differences. In addition, with few exceptions, the geographic distribution across the Brazilian states is non-uniform, which provides a non-stable estimation environment.

Practical Aspects: In any spatial interpolation scheme the selection and quantity of surrounding stations are critically important to the results of the interpolations. Problems arise when using climatological data because of missing values and the varying availability of stations through time. In order to determine which stations are to be used, surrounding stations are pre-selected based on their relationship with the target station. The closest stations are identified for each target station and are ranked by the value of the correlation coefficient between the candidate station and its neighbours.

The ANN estimation technique based on spatiotemporal objective analysis scheme is used to estimate daily values, with the "best" estimate chosen as a missing value replacement for the development of regional daily minimum-maximum temperatures and total precipitation time series over Brazil.

MISSING DATA ESTIMATION.

The replacement of missing daily values for temperatures and total precipitation includes the use of nearby simultaneous values to calculate an estimated value at the target station over the period of time for which adequate data are available. The efficiency, or accuracy, of the estimates over a long period of time provides the information used to assess the quality of

estimated daily values. Estimated daily values are used in lieu of missing values as a means of making a particular station serially complete. There are numerous spatial interpolation methods available for point estimation with irregularly spaced data. Typically, the choice of methodology is dependent on several factors: the meteorological variable under consideration, the geographical area, the spatial distribution of surrounding observations, and the day–month–season for which the target station is to be estimated

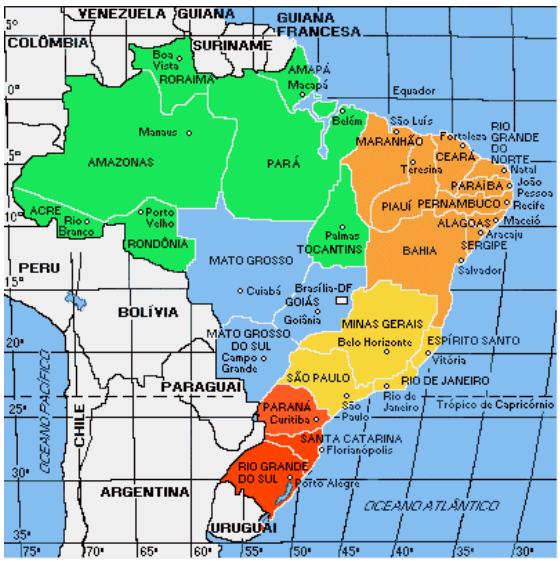


Fig. 1: Brazil Political Map.

In this research work we propose an interpolation method based on the representation of the data as a parametric model plus a random process (like Kriging):

$$RG_{TARGET}(s,t) = \mu_{TARGET}(s,t) + \varepsilon_{TARGET}(s,t),$$

where (s,t) are the discrete location and the time coordinates, respectively:

$$\mu_{TARGET}(s,t) = \beta_1.RG_{NN1}(t) + \ldots + \beta_K.RG_{NNK}(t) + \alpha_{1,-\varphi}.RG_{NN1}(t-\varphi) + \ldots + \alpha_{1,\varphi}.RG_{NN1}(t+\varphi) + \ldots + \alpha_{K,-\varphi}.RG_{NNK}(t-\varphi) + \ldots + \alpha_{K,\varphi}.RG_{NNK}(t-\varphi) + \lambda_{-\theta}.\mu_{TARGET}(s,t-\theta) + \ldots + \lambda_{\theta}.\mu_{TARGET}(s,t+\theta);$$

with NN1, …, NNK the k-nearest-neighbours from φ and θ (the model order) prior time lags.

$$RG(s+\kappa,t) = \mu(s+\kappa) + \delta(s+\kappa,t) + IH(s+\kappa,t),$$

where $\mu(s+\kappa)$ is RG's climatological normal value, $\delta(s+\kappa,t)$ is the anomaly related to the instant t and $IH(s+\kappa,t)$ is the possible inhomogeneity lying in the measured value $RG(s+\kappa,t)$. By using an analogous notation, a reference series which is constituted, for example, by the data of a neighbouring station can be written as follows:

$$RG(s,t) = \mu(s) + \delta(s,t) + IH(s,t).$$

If the two series belong to the same climatic area, it can be assumed that $\mu(s,t) = \mu(s+\kappa,t)$ for each value of t. If the reference series is homogeneous: $IH(s+\kappa,t) \rightarrow 0$. Therefore, the series of the differences and ratios between stations of the same area will be:

$$\varepsilon(t) = RG(s,t) - RG(s+\kappa,t)$$
$$= (\delta(s,t) - \delta(s+\kappa,t)) + IH(s,t) \rightarrow \gamma \text{ (constant)};$$

$$\pi(t) = RG(s,t) / RG(s+\kappa,t)$$
$$= [\mu(s+\kappa) + \delta(s+\kappa,t) + IH(s+\kappa,t)] /$$
$$[\mu(s) + \delta(s+\kappa,t) + IH(s,t)] \rightarrow \eta \text{ (constant)}.$$

The number of neighboring stations meeting the criteria is not fixed in time. It varies depending on available station data for the year/month/ day in question. As such, the interpolation model may also change in time. Moreover, the surrounding stations that may be optimal for a particular calendar month may not be optimal for a different month. Thus, the station selection procedures are computed for each calendar month separately.

LARGE-SCALE TELECONNECTIONS.

The El Niño Southern Oscillation (ENSO) phenomenon is the major cause of year-to-year variations in climate over lower latitudes and one of the most significant causes of global climate change on this timescale. The ENSO is associated with disruption to tropical climates in many regions. The Southern Oscillation Index (SOI) is a pronounced disturbance of the atmospheric circulation over lower latitudes of the Pacific sector.

The Trans-Niño Index (TNI), which is given by the difference in normalised (1950-79) anomalies of SST between Niño1+2 and Niño4 regions, is used as an

optimal description of the character and evolution of El Niño or La Niña.

The Pacific Decadal Oscillation (PDO) is a leading index associated to the ENSO phenomenon by taking into account the monthly Sea Surface Temperature (SST) anomalies in the North Pacific Ocean. In effect, to characterize the nature of the ENSO, SST anomalies in different regions of the Pacific is used.
The North Atlantic Oscillation (NAO) is a major disturbance of the atmospheric circulation and climate of the North Atlantic region, linked to a waxing and waning of the dominant middle-latitude westerly wind flow during winter.

The NAO exerts a strong influence on year-to-year climate variability and there is evidence of long-term trends in variability of this phenomenon. It is related to the shorter-term shift between zonal and meridional circulation types that occurs on a day-to-day time scale and is know as the index-cycle.

ARTIFICIAL NEURAL NETWORK.
The use of ANN has been recognized recently as a promising way of making predictions on time series, detecting irregular behaviour.

In practice, one determines the embedding dimension (number of past observations) of the time series attractor (delay time that determine how data are processed) and uses these number to define the network's architecture.

Physically, the attractor is the object to which the time series in a phase space (space in which each point describes the state of a dynamical system as a function of the non-constant parameters of the system) is attracted to.

Meteorological attributes can be accurately predicted by the spatiotemporal ANN model architecture: designing, training, validation and testing. The best generalization of new data is obtained when the mapping represents the systematic aspects of the data, rather capturing the specific details (e.g. noise contribution) of the particular training set. The evaluation of the error function.

ANN TRAINING ALGORITHM – BAYESIAN UPDATING.
Mathematical techniques for minimizing the discrepancy between a parameterized function and a set of pairs of inputs and "correct" outputs, where the overall function is partitioned into layers of vector functions.

Back Propagation: Back propagation is the best-known training algorithm for multi-layer neural networks. It defines rules of propagating the network error back from network output to network input units and adjusting network weights along with this back propagation. It requires lower memory resources than most learning algorithms and usually gets an acceptable result, although it can be too slow to reach the error minimum and sometimes finds not the best solution.

Quick Propagation: Quick propagation is a heuristic modification of the back propagation algorithm. This training algorithm treats the weights as if they were quasi-independent and attempts to use a simple quadratic model to approximate the error surface. In spite the fact that the algorithm hasn't theoretical foundation, it's proved to be much faster than standard back-propagation for many problems. Although sometimes the quick propagation algorithm may be instable and inclined to stuck in local minima.

MODELLING UNCERTAINTIES.
After estimating daily temperatures and precipitatino, a series of internal consistency checks were performed to ensure that estimates did not violate obvious constraints associated with recording weather attributes.

Modelling the uncertainty associated to the selected meteorological time series. The fuzzy set theory models uncertainty related to stochastic evaluation procedure based on expert knowledge through subjective modelling.

The fuzzy set theory models uncertainty based on expert knowledge through subjective modelling. Fuzzy theory is a method that facilitates systems' uncertainty analysis where uncertainty arises due to vagueness or "fuzziness" rather than due to randomness alone.

Fuzzification is a methodology to generalize any specific theory from a discrete (crispy) to a continuous (fuzzy) form. The fuzzy logic can be a generalization of the classical set theory, the statements are described in terms of membership functions that are continuous and have a range [0,1]. In practice:

1. Define variables to be used in determining Sensitivity and Adaptive Capacity.
2. Apply a multi-criteria model to develop a Sensitivity index and an Adaptive Capacity index:

$$I_j = \sum_{i=1}^{n} w_{ij} c_{ij}$$

ωij is obtained through ANN, which determine weights (e.g., importance) of each variable; cij is obtained through value functions, which transform the natural scales of all variables or criteria into a scale of [0,1].

3. Aggregate the two indices through Fuzzy Logic.
4. Vulnerability is defined by categorical variables: Low Vulnerability, Moderate Vulnerability, High Vulnerability. These Categorical Variables are transformed into Fuzzy Sets.

WAVELET ANALYSIS FOR DATA QUALITY .

The wavelets are suitable to analyse certain nonstationary time series and classes of autocorrelated processes. They are especially useful for the examination of the characteristics of time series on different scales and are already used in various fields of application.

Detecting inhomogeneities and correct them! Making a comparison between candidate series and series that have been constructed by a weighted average of some selected station series.

We apply the Discrete Wavelet Transform (DWT) for trend estimation and shift detection. This estimate is used to establish a test, which is suitable for correlated data non homogeneity detection. Furthermore wavelets are applied to examine the behaviour of the data on different scales (frequency components):

H0: No Trend (No Shift)
x
H1: Significant Trend (Shift Exists)

The test assume that the observed time series can be modelled additively by a deterministic trend (shift: change intervention or break point) component plus a realisation of a stochastic process. Using the DWT, the data vector is decomposed into a component, representing the variability on large scales, and another for small scales.

HOMOGENISATION.

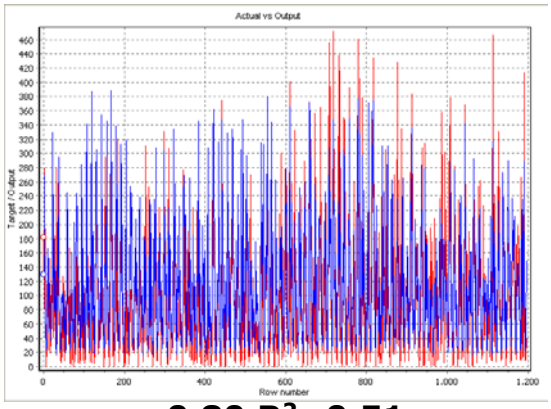
To validate this work-algorithm, the diagnostic of homogenised rainfall was accomplished. The spatial distribution of rainfall is summarised by the subjective descriptive four-moment measures: Mean, Variance, Skewness and Kurtosis, giving support to spatial pattern recognition. A number of homogeneity tests with kinds to detect non homogeneities are employed (methods currently used – Chow & Pettitt, SNHT Alexandersson, Range Buishand test, Von Neumann ratio and Craddock tests) and the effect of natural variability is established taking into account ensembles of consecutive years. The significance of the adjustments was tested using the "Robust Modified Wilcoxon Rank Sum Test". The "Customized Kendall τ Test" is used as a non-parametric method to test the significance of trends. As expected, this robust reconstruction method has good performance, since more information is introduced in the decision-making system.

SUMMARY.

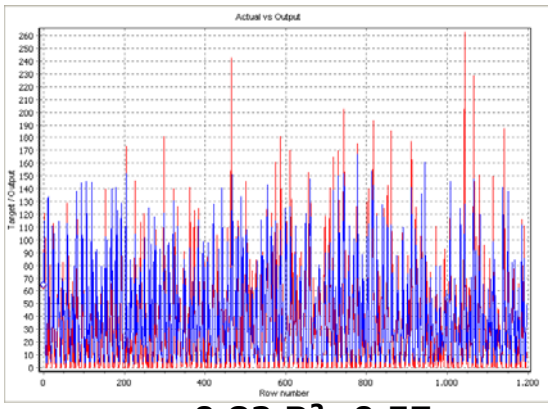
This work consists on the reconstruction of weather time series for Brazilian localities, substantiating the spatial consistency, apparent cycles and respective trends. This approach recognises very well the mutual dependence between spatiotemporal rainfall variability.

This research work summarizes a procedure used to create serially complete daily temperature and precipitation datasets (1951–2000) for Brazil. Determining target and estimator stations by scanning the quality of individual station records, reconciling metadata (including observation times and station locations), and categorizing observation

times proved to be time consuming but necessary. Estimating the missing data values and cross validating the results proved to be relatively straightforward once preparatory work was accomplished. Our results show that the efficacy of the estimation procedure and thus the reliability of the estimated missing values are dependent on a number of factors. For all three meteorological parameters the selection and quantity of surrounding stations are critically important to the results of the interpolations. We feel that the pre-selection of surrounding stations, based on their relationship with the station to be estimated, is an integral first step.



ρ=0.80 R^2=0.51



ρ=0.82 R^2=0.57

Fig.2: Daily Rainfall Reconstruction of an arbitrary Brazilian meteorological station.

The conclusion highlight the use of climate proxies response as potential weather predictor. The use of ANN has been recognised recently as a promising way of making predictions on time series, detecting irregular behaviour.

ACKNOWLEDGMENTS. Grateful thanks to Alex Grechanowski for kindly provide us the software NeuroIntelligence - neural network software for professionals.

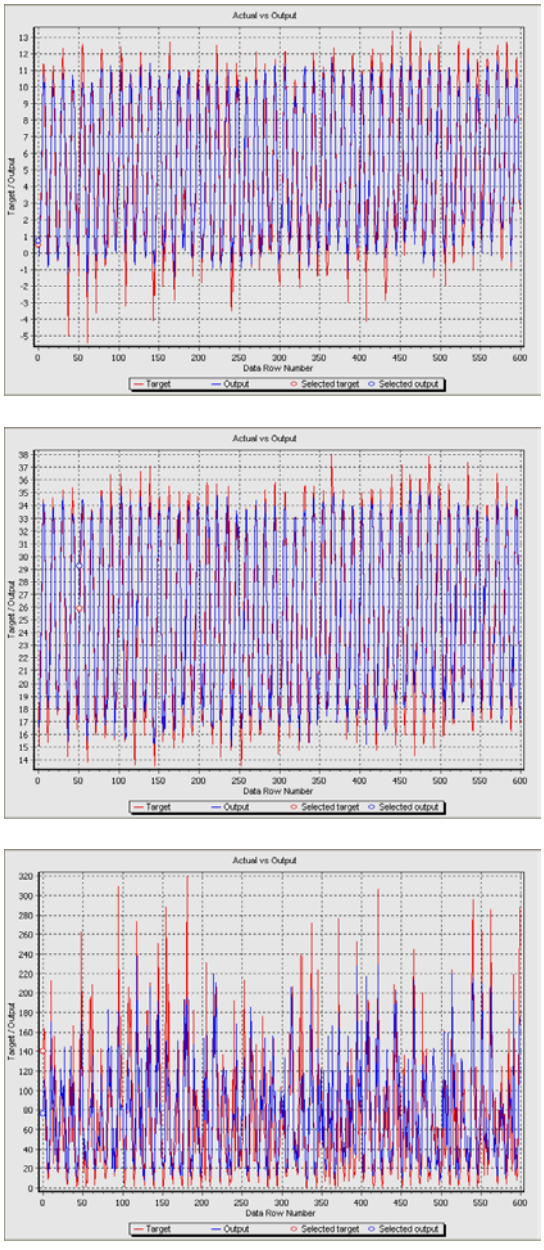


Fig.3: Monthly TMIN, TMAX and Rainfall Reconstruction of an arbitrary Brazilian meteorological station.

REFERENCES.

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Oxford: Oxford University Press.

Briggs, W. M., Wilks, D. S., 1996. Estimating monthly and seasonal

distributions of temperature and precipitation using the new CPC long-range forecasts. J. Climate, 9: 818–839.

DeGaetano, A. T., K. L. Eggleston, andW.W. Knapp, 1993: A method to produce serially complete daily maximum and minimum temperature data for the Northeast. NRCC Research Publication RR 93-2, 9 pp.

Huth, R., and I. Nemesova, 1995: Estimation of missing daily temperatures: Can a weather categorization improve its accuracy? J. Climate, 8, 1901–1916.

Kaplan, A.; Kushnir, Y., and Cane, M. A. 2000: Reduced Space Optimal Interpolation of Historical Marine Sea Level Pressure: 1854 – 1992. Journal of Climate, 13, 2987-3002.

Knippertz, P., Christoph, M., Speth, P., 2003. Long-term precipitation variability in Morocco and the link to the large-scale circulation in recent and future climates. Meteorology and Atmospheric Physics, 83: 67–88.

Kyriakidis, P. C. and Journel, A. G., 1999: Geostatistical space-time models: a review. Mathematical Geology, 31 (6), 651-684.

Michelangeli, P., Vautard, R., Legras, B., 1995. Weather regimes: recurrence and quasi-stationarity. Journal of the Atmospheric Sciences, 52: 1237–1256.

Murphy, J., 1999. An evaluation of statistical and dynamical techniques for downscaling local climate. Int. J. Climatology, 12: 2256–2284.

Shumway, R. H. and Stoffer, D. S., 2000: Time series analysis and its applications. New York: Springer-Verlag.

Wilks, D. S., 1996. Statistical significance of long-range optimal climate normal temperature and precipitation forecasts. J. Climate, 9: 827–839.

Zorita, E., Hughes, J. P., Lettemaier, D. P., von Storch, H., 1995. Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. J. Climate, 8: 1023–1042.

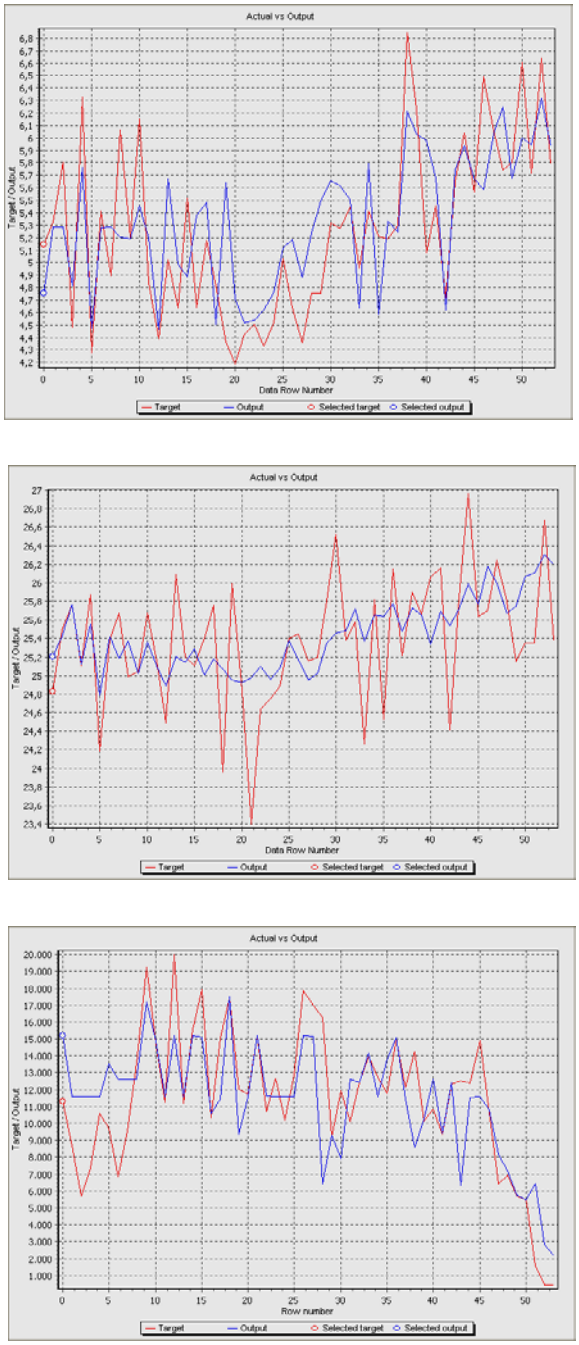


Fig.4: Annual TMIN, TMAX and Rainfall Reconstruction of an arbitrary Brazilian meteorological station.