CHAC: A WEATHER PATTERN CLASSIFICATION SYSTEM FOR REGIONAL CLIMATE DOWNSCALING OF DAILY PRECIPITATION

A. D'onofrio[1 *], J.-.P Boulanger[2], and E. C. Segura[3]

[1,3]Departamento de Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, UBA, Buenos Aires, Argentina
[2]Research Scientist, Institut de Recherche pour le Développement (IRD), Departamento de Ciencias de la Atmósfera y los Océanos FCEN-UBA, Buenos Aires, Argentina

1. INTRODUCTION

Throughout the last two decades the field of climate modelling has significantly progressed. In particular, the accuracy of Coupled General Circulation Models (CGCMs) have been significantly increasing, such that they can now simulate fairly well the large-scale climate features of climate. Unfortunately, their coarse spatial resolution (typically in the range from 15,000 to 160,000 km²) leaves unresolved important sub-grid scale features such as clouds, small-scale thermodynamic interactions and topography, often leading to local and regional discrepancies between models and observations. In particular, it does not allow them to properly represent local daily variability and extreme events, crucial for local impact studies. Downscaling methodologies have been developed with the purpose of filling the resolution gap between GCMs and regional and local scale prediction needed for extreme event studies and impact assessment (see for instance Menendez et al. 2008; same issue). The aim of atmospheric downscaling methodology is to use GCMs atmospheric variables as a starting point to obtain regional or local scale surface variables such as temperature and precipitation. Basically, downscaling methodologies can be divided into two main threads of investigation, one dynamical and the other statistical. Dynamical downscaling uses the information provided by GCMs as a boundary condition to generate a higher resolution simulation with a regional model (RCMs) over the spatial domain of interest. This methodology allows to add important regional information to the model such as topography, soil moisture and, more generally, surface land use and land cover. On the other hand, statistical downscaling (SDS) seeks for a relationship (called transfer function) between observed local climatic variables and large-scale circulation variables of the GCMs. Statistical techniques have proven to be as successful as dynamical techniques, being also far less computationally expensive and simpler to develop (e.g. Gutierrez et al., 2004; Hewitson, 2006). However, their major drawbacks are their tendency to under estimate extreme precipitation events, and their incapacity in predicting non-stationary behaviours in the transfer function (especially under climate and land-use/land-cover changing conditions.

Among SDS techniques, linear regression and stepwise regression are very popular, but can only be used to predict variables that are not discreet in time and space, such as temperature. However, such an approach is not suitable for daily precipitation given that a linear relationship between precipitation and weather patterns is way too far from reality due to the stochastic nature of daily precipitation. For this reason, other SDS approaches have been taken such as classification and regression trees (CART; Hughes et. al., 1993) and clustering techniques (k-means, Gutierrez et al., 2004; neural networks, Hewitson, 2006; analogues and nearest-neighbours; Mehrotra and Sharma, 2005). Data clustering is a non-linear technique that covers fields of statistical analysis, machine learning, data mining, patter recognition, neural networks, and fuzzy theory (e. g. A.K. Jain et al., 1999). Clustering techniques on SDS generate a partition of the data set into subsets called clusters, in which each cluster consists of similar weather patterns linked with its daily rainfall values. The intuitive idea behind clustering is that precipitation occurring during a specific period of time (e.g. one day) with a particular weather pattern (which dimension has to be optimized) tends to be similar as to the precipitation observed another day that has a similar weather pattern. This concept was first explored by Lorenz (1969). It must be noticed that the performance of a classifier for downscaling is then depends on three things, the amount of data available for training, the problem difficulty, and its generalization skill for prediction.

Another advantage of SDS based on classification is that they can handle very well missing data at weather station making it very suitable for Argentina analysis where some stations has gaps of information. As Hewitson and Crane (2006) pointed out, there is empirical evidence that no best statistical technique may exist and that the skill of the different approaches varies along within the location and accuracy of the weather pattern data used for training. Our objective is to develop and calibrate a statistical downscaling method valid for the region, which will be used in future studies to project climate change scenarios in Argentina and in La Plata Basin.

In this study, we analyze the strengths and weaknesses of a statistical method based on clustering techniques, using atmospheric variables from ERA-40 and observational data from Argentine stations covering a wide range of climates. While specific variables may have different weights among the large region under study, our objective is to develop a tool, which will then be used for tropical, mid- or high-latitude stations; therefore a fully operational downscaling toolbox called

---

* *Corresponding author address:* D´Onofrio Ariel E. DCAO, Ciudad Universitaria, pab II, piso 2, (1428) Ciudad Autonoma de Buenos Aires, Argentina;
e-mail: adonofri@dc.uba.ar

CHAC was made. It was written in MATLAB language and works on version 7.1 or higher. It has been tested over windows XP, Mac OS and Linux operating systems. CHAC can be downloaded from http://sourceforge.net/projects/chac.

2. DATA

ECMWF data. The observational data used is the reanalysis ERA-40 dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). The regular grid resolution of ERA-40 is of 1.125° latitude x 1.125° longitude consisting on four data atmospheric variables per day (00:00, 06:00, 12:00, 18:00 h UTC) listed in Tables 1 and 2. Given that weather station data are only available on a daily time step, only averaged atmospheric field daily values were considered. The ECMWF data used in the following cover the period starting in 1979 (the main reason of this choice is the assimilation of satellite data starting in that year, which provides more reliable data) and finalizing in year 1999, representing a total of 7670 days. As in Cavazos (2000) and Cavazos and Hewitson (2005), the selection of ERA-40 atmospheric predictors for daily precipitation is based on the rationale that they must either be dynamical, or non-dynamical components (thermal or moisture). The dynamical components are captured by the wind fields (U, V at the four levels 200, 500, 700 and 850mb), the geopotential (Z) at 500mb, the vertical velocity (w) at 500mb (for subsidence and convection) and the sea-level pressure (msl). Finally, the moisture component is represented by the relative (rh) and specific (q) humidity at low-levels (850 and 700) and by the total column water (tcw). Although long the list could also include other variables such the 500-1000hPa thickness or convection indices in future studies.

Variable	mb
U component wind (U)	200,500,700,850
V component wind (V)	200,500,700,850
Relative humidity (rh)	750,850
Specific humidity (q)	750,850
Vertical velocity (w)	500,700
Geopotential (Z)	500

Table 1: ERA-40 reanalysis list of predictor variables utilized in this study

Variable	
Total column water	(tcw)
Mean sea level pressure	(msl)

Table 2: ERA-40 two dimensional variables

Argentine weather station data. Daily precipitation time series at 39 Argentine stations covering the period 1959-2001 are used in this paper. The Argentine stations analyzed are located all over the country and have been quality-controlled before entering the CLARIS database. Figure 1 shows the stations used. Stations were subject to statistical tests to check for artificial sudden changes in the mean, outliers and trends in the monthly series (Buishand, 1982).

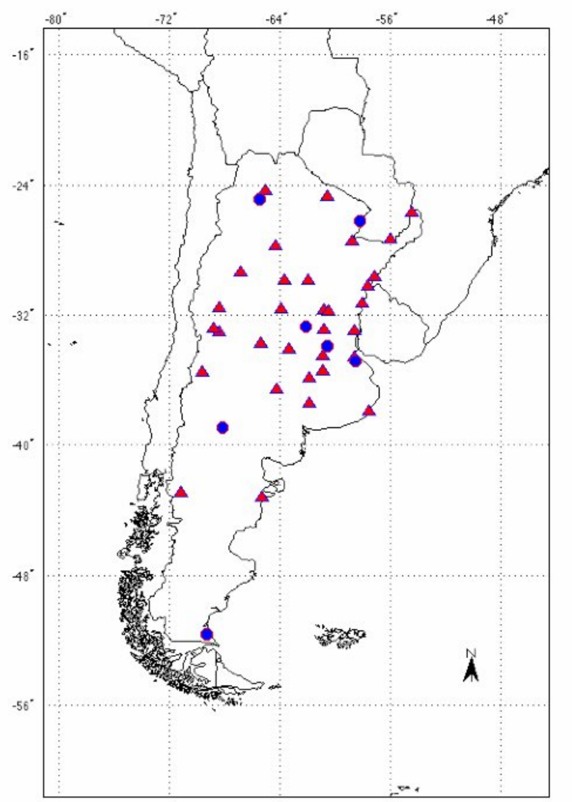


Figure 1: Map of the 39 Argentinean stations. Blue dots are stations used for the model parameter estimation.

3. THE STATISTICAL DOWNSCALING MODEL

3.1. The classification procedure

The starting point of the statistical downscaling methodology is mainly based on the previous work by Gutierrez et al. (2004). This methodology makes use of a clustering technique to connect atmospheric fields with time series of observed daily precipitation or temperature at local stations. The weather patterns are defined as the ensemble of the 17 ERA-40 variables listed in Tables 1 and 2 covering a specific spatial domain. The optimization of its size will be discussed later.

In a first step, the weather patterns are normalized (all variables have zero mean and unit variance) in order to prevent some atmospheric variables from dominating the following calculations.

In a second step, a Principal Component Analysis (PCA) is performed over the normalized weather patterns (Wp). PCA is a well-known statistical linear technique used to compress data by reducing the vector space of the data set in such a way that there is a minimal loss of variance. The motivation for using this technique is the assumption that Wp contains redundant predictors that will not contribute significantly to the quality of the downscaling. PCA first calculates the covariance matrix C^{nxn} of Wp^{mxn}, with C being symmetric and having $e_1,\ldots,e_n$ eigenvectors, and $\lambda_1,\ldots,\lambda_n$ real eigenvalues. The elements C_{it} represent the covariance between the variables i and t of the original space:

$$c_{it} = \left\langle \left(x_{ki} - u_i\right)\left(x_{kt} - u_t\right) \right\rangle_k$$

$$u_i = \left\langle x_{ki} \right\rangle_k \quad ,$$

$$u_t = \left\langle x_{kt} \right\rangle_k$$

A transformation matrix M^{nxn} is then constructed as follows, each row of M is an eigenvector of C, and they are ordered in decreasing order according to the magnitude of their corresponding eigenvalue. The eigenvalue corresponds to the variance of Wp projected on the eigenvector associated. The coefficients of the elements of Wp projected on the new base are called principal components (PCs^{mxn}).

The matrix calculation to generate the PCs is,

$$PCs = Wp.M^T = \begin{pmatrix} wp_{11} & \cdots & wp_{1n} \\ . & \cdots & . \\ . & \cdots & . \\ wp_{n1} & \cdots & wp_{mn} \end{pmatrix} \begin{pmatrix} e_{11} & \cdots & e_{1n} \\ . & \cdots & . \\ . & \cdots & . \\ e_{n1} & . & e_{nn} \end{pmatrix}$$

This new coordinate representation of the data allows us to select the number of PCs that retain most of the variance, generally reducing the size of Wp matrix up to 80%. This reduction has numerous advantages. Representing the information in a compressed way with a minimal loss of information is important for a classifier task due to the curse of dimensionality (*Bellman, 1961*): as dimensionality increases the data sample needed to fill the input space grows exponentially, making harder for a classifier to reach a good generalization performance if the same amount of data is available for classification. Also, an immediate reward in the reduction of the data space dimensions is the lower need in computer resources and a faster simulation speed.

In a third step, the retained PCs of weather patterns are classified. In the following, we will compare the performance of two often-used techniques: k-means (Gutierrez et al., 2004) and the SOM (Self-Organizing Maps; Hewitson and Crane, 2006). While other methods (analogue and k-nearest neighbours) have also been implemented, we only discuss these two clustering methods, which gave the best results. K-means is one of the clustering techniques most widely used, its success is based on its simplicity and fast convergence speed, making it very suitable for computer implementation, although it can be easily stuck in local minima the obtained solution is often very good. To generate the clustering, k-means starts with k vectors chosen at random, and then it tries to minimize the intra cluster variance. The clustering algorithm can be sensitive to the initial random solution, for that reason it is recommended to repeat several times the algorithm using different initial random vectors, and chose the solution with the minimum intra cluster variance. Kohonen self-organizing feature map (SOM), sometimes also called topology-preserving maps, has been created by Teuvo Kohonen (1982), and is a clustering technique combined with topological preserving and data reduction ordered map properties. It is part of the non-supervised learning neural network field and is motivated on the human nervous system. The goal of a SOM is to represent all points in the source space by points in the target space, preserving distance and proximity relationships as much as possible. The structure of a SOM consists of two layers, which are called the input layer and the output (representation) layer, with feed-forward connections from input to output and lateral connections between neurons in the output layer. The output layer has the shape of two-dimensional lattice of neurons, but also one-dimensional layer can be used, though it is not common. Each neuron has associated a weight vector the size of the input space vectors.

3.2. The simulation procedure

At this stage, the classifier algorithm divides the database of weather patterns into clusters, with the restriction that every weather pattern belongs only to one specific cluster. Given that each daily weather pattern is associated to a precipitation value observed on the same day, we can relate to each cluster a precipitation distribution.

When making a simulation with new atmospheric weather patterns as an input, each weather pattern is projected onto the clusters by computing its Euclidean distance to each of them. Only one cluster is selected (the one to which the distance is minimum). The simulated precipitation value is selected randomly in the daily precipitation distribution of the selected cluster. This random process requires performing various simulations in order to capture the range of possible precipitation amplitudes associated to a given weather pattern. For this reason, an ensemble of M simulations (a discussion on M is provided hereafter) is always generated.

3.3. The validation procedure

When validating the statistical model (considering that our input data cover the 21-year period 1979-1999), an m-fold cross-validation method is applied, meaning that the ensemble values for a specific year (e.g. 1979) make use of the clusters trained only with the other 20 years of atmospheric data. By proceeding that way, we ensure that the final validation procedure of the system is independent of the atmospheric weather patterns observed the same day and is thus truly representative of our statistical model skill.

Finally, a set of standard indices is used to validate the skill of our statistical model, one of which is the ROC (Relative Operating Characteristic) curve, (see Kharin and Zwiers, 2003). A ROC curve measures the ability of the ensemble to discriminate between two categorical precipitation events, such as days greater and lower than a certain threshold. We say that a n observed or predicted precipitation event has occurred if its precipitation value is greater than a specified threshold, otherwise, we say that the event has not occurred. The way to generate a ROC curve is as follows, let's assume we have an ensemble where each member is a deterministic simulation of size N produced over the training period, then N can be divided into two subsets, those when an event occurred and those when it does not, i.e. N=O + O´ see Table 3.

Observations	Warning	No warning	Total
Event	H	MS	O
Non Event	FA	CR	O´
Total	W	W´	N

Table 3: Definition of events used when computing the ROC curve

Let H be the number of hits, whenever an event occurred and a warning was issued, let FA be the False Alarm, whenever a warning was issued, but no event occurred, let MS be the number of missed events, which is the number of times that an event occurred but no warning was issued, and let CR be the number of correct rejections, which is the number of times that an event did not take place and no warning was made. Bearing this in mind, we can calculate the hit rate (number of occurred events well predicted) and false alarm rate (number of events predicted, which actually did not occur) as follows: $HR = \frac{H}{O}$ and $FAR = \frac{FA}{O'}$.

Considering that an ensemble is actually a set of deterministic simulations, a critical threshold P_{cr}, is introduced and a warning is fired on a specific day every time the percentage of the ensemble's precipitation values that are defined as events exceeds this threshold. In this case, the hit and false alarm rates can be expressed as a function of P_{cr} in the following way,

$$HR(P_{cr}) = \Pr\{\Omega_P / E = 1\} = \int_{\Omega_P} f(P/E=1)dP$$

$$FAR(P_{cr}) = \Pr\{\Omega_P / E = 0\} = \int_{\Omega_P} f(P/E=0)dP$$

where Ω_P denotes forecast probabilities $P>P_{cr}$ and $f(P/E)$ is the conditional probability density function of probability forecasts

Finally, the ROC curve is made by plotting the resulting hit rates versus false alarm rates as P_{cr} varies from 0 to 1. In every ROC curve, a forecast with no skill corresponds to the diagonal connecting the points (0,0) and (1,1), and a perfect forecast generates a ROC curve going through the points (0,0),(1,0) and (1,1), furthermore, a forecast with no skill has equal hit rate and false alarms, while a perfect forecast is characterized by a hit rate equal to 1 and a false alarm equal to 0.

The index mostly used in the present study is the ROC skill score (RSS), which is based on the area under the ROC curve as follows: RSS=2*(area-0.5). Then a no-skill forecast has a ROC area of 0.5 and a RSS equal to 0, and a perfect forecast has a ROC area of 1, and a RSS of 1. Moreover a forecast A is better than a forecast B, if the area under the ROC curve of A is greater than B i.e. if the RSS A is larger than the one of B. As a final comment, to compare the skill of simulated distributions of dry/wet spells and daily amount, we will display PP-plot and QQ-plots.

3.4 Model parameter choice

In order to test the statistical model parameter choices (clustering technique, number of clusters, domain size, ensemble size), we evaluate the model skill in a subset of weather stations (7) representative of very different climate conditions over Argentina (see blue circles in Figure 1). Formosa (Formosa Province) is located at the frontier between Argentina and Paraguay and is characterized by a near-tropical climate with very strong convection systems and a well-marked rainy season during the South American Monsoon (austral summer). Salta (Salta Province) is located in the northwest region of Argentina and is characterized by a dry climate. Pergamino (Buenos Aires Province) is in the core region of agriculture production in Argentina. Its annual mean precipitation exceeds 1000mm. Marcos Juarez (Cordoba Province) is a semi-arid region and an agricultural marginal region. However, the increase of precipitation during the last 30 years considerably favoured the growth of cereal crops in the area (G. O. Magrin et. al. (2005), B. Smit and M. W. Skinner, (2004)). Ezeiza (Buenos Aires Province) is an airport station very close to the Capital of Argentina (Buenos Aires). The climate is influenced by convection and frontal systems during the summer. Monthly precipitation amounts are quite equally distributed during the year (Boulanger et al., 2005). Neuquén (Neuquén province) is very close to the Andes and located in a mid-latitude region strongly influenced by frontal systems. Rio Gallegos (Santa Cruz Province) is one of the stations located in the south of Argentina (near 50°S) and is representative of a mid-latitude climate.

This subset of stations is analyzed in detail to evaluate the sensitivity of the method to the location of the station and thus develop a method relatively skilful all over the country. For the sake of clarity (and number of figures), results mostly present averaged RSS among the 7 stations.

Clustering technique and number of clusters. Figure 2 displays the RSS score averaged over the 7 selected stations as a function of the number of clusters (actually expressed as the mean number of observations per cluster i.e. number of daily observations divided by the number of clusters) using two different clustering techniques (k-means in red and SOM in blue). Following Hewitson and Crane (2006), we selected a domain size on the order of 1000kmx1000km around each station to extract the atmospheric variables. Ensembles of 100 simulations are performed to compute the ROC scores. It appears that the k-means mean curve is always higher than the SOM curve suggesting the k-means clustering technique is more effective in covering the data space than SOM. Indeed, the topological order imposed by the SOM method strongly limits the SOM clusters to explore low-density regions of the data space (Fig. 3). This behaviour then affects the representativity of the SOM clusters and their generalization skill. Although a thorough work on the SOM parameters could reduce such a limitation, the good results of the k-mean technique, which is fast and easy to implement, lead us to select this method as the clustering method. Moreover, regardless the clustering technique, it is found that the ROC curve for precipitation occurrence or for different thresholds of precipitation (25%, 50% or 75% of the precipitation CDF) displays a similar curve mostly independent of the station under study (not shown).

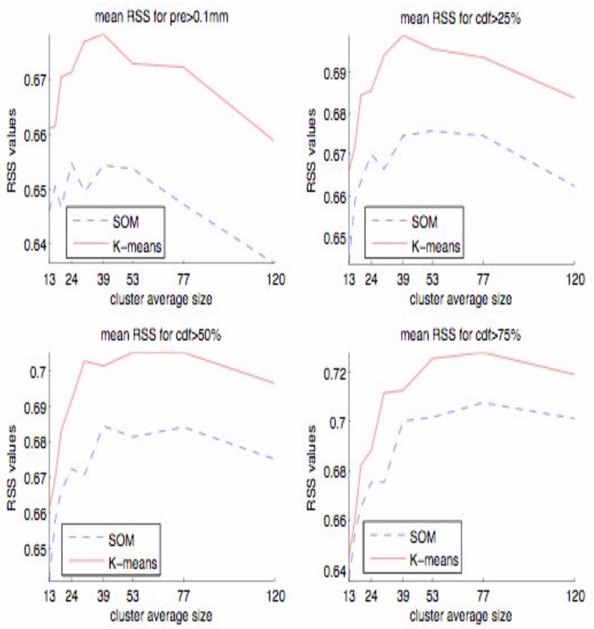


Figure 2: ROC values averaged over the seven selected stations as a function of the mean number of observations per cluster, for domain size of 4x4 and ensemble size of 100 members. The blue-dashed curve represents results from the SOM classification, while the red-solid curve represents results from k-means classification. The RSS is computed for four different criteria: (upper-left) wet/dry day; (upper-right) daily amount larger than the 25th percentile of the rainfall amount cumulative distribution function (CDF); (lower-left) daily amount larger than the 50th percentile; (lower-right) daily amount larger than the 75th percentile.

The optimal number of observations per cluster can be estimated from Figure 2 to be between 39 and 53 observations, although the decrease in the RSS value is very slow as the number of observations increases, and the differences are not highly significant in terms of daily precipitation simulations. The selected number of observations per cluster is 39 (equivalent to 196 clusters for our dataset). This value is much smaller than the one (90) selected by Hewitson and Crane (2006), but as said above the RSS values are not very different for 39 or 90 observations per cluster.

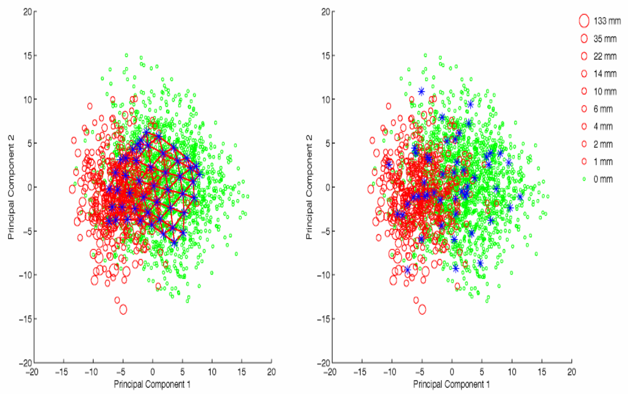


Figure 3: Representation of the cluster nodes (blue stars) computed by the SOM (left) and k-means (right) technique) in the two-dimensional space of the first two principal components of the data space for the station Marcos Juarez in the period 1990-1996, for domain size of 4x4. The green dots represent dry days, red circles represent wet days, and the size of the red circles is a function of the intensity of the daily rainfall.

Domain size. Given the previous result, at each of the seven stations, daily weather patterns are defined for different spatial domain sizes and then clustered. Ensembles of 100 simulations using k-means classifier are performed to compute the RSS. In the following, we will define 1x1 as the vertical at the grid point closest to the weather station, 2x2 as the four nearest grid points to the station, 4x4 as the 16-grid point square including the 2x2 square, 6x6 as the 36-grid point square including the 4x4 area, etc… In terms of spatial resolution, 1x1 is a profile, 2x2 is a 1.125° size square, 4x4 is a 3.375° size square, 6x6 is a 5.625° size square etc… Thus, we computed the RSS-averaged values from the RSS values at all seven stations for four different ROC criteria (0.1mm/day, 25%, 50%, and 75% of the station precipitation CDF distribution). The results are displayed in Figure 4. While the RSS value is of similar amplitude for domains smaller than 6x6, the large domains (8x8 and larger) present decreasing RSS values. The RSS value of the vertical profile (1x1) is always found to be smaller than the 2x2, 4x4 or 6x6 domain sizes, meaning that local- to synoptic-scale information improves the relationship between atmospheric conditions and precipitation simulation at the weather station. We finally decided to select a 4x4 domain size (3.375° x 3.375°). Therefore, our result highlights a very small-radius of atmospheric variability around the station suggesting a strong role of moisture components in the weather pattern clustering. Indeed, if the atmospheric variables are only dynamical variables (blue dashed curve in Figure 4), the optimum domain size is larger and extends up to 10x10 or 12x12 domain sizes highlighting the importance of larger scale patterns to define dynamical weather patterns. However, it is worth pointing out that the RSS values of the method, when based only on dynamical variables, are significantly smaller than the RSS values of the method when based on the dynamical and non-dynamical atmospheric variables. This suggests that even if the large-scale circulation is clearly a strong constraint on local precipitation, what defines ultimately the occurrence and amplitude is strongly related to non-dynamical local- and synoptic-scale processes.

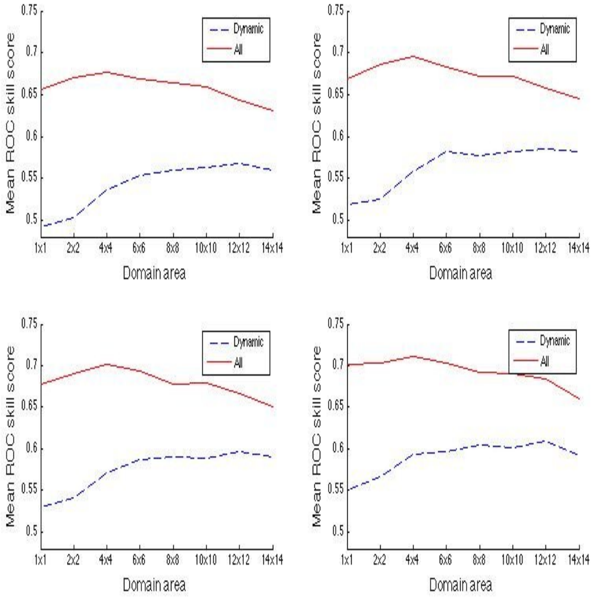


Figure 4: ROC skill score (RSS) averaged over the seven selected stations as a function of the domain area around the station, using k-means classifier and ensemble size of 100 members. The red-solid curve represents results when using all the dynamical and non-dynamical variables. The blue-dashed curve represents results when using only dynamical variables. RSS is computed for four different criteria: (upper-left) wet/dry day; (upper-right) daily amount larger than the 25th percentile of the rainfall amount cumulative distribution function (CDF); (lower-left) daily amount larger than the 50th percentile; (lower-right) daily amount larger than the 75th percentile.

Ensemble size. Previous results were obtained computing ensembles of 100 members. Here we are interested in analyzing the sensitivity of the ROC calculation to the number of members of the ensemble, in order to reduce the computational time without losing predictive skill. In Figure 5, the RSS score averaged between the 7 selected stations is displayed for different types of events (Pre>0.1mm/day, > 25%, 50% or 75% of the distribution). Briefly, it appears that the computation of the RSS score converges for 100 simulations or more per ensemble. In fact, we found (for different experiments not shown here) that the convergence happens for a number of simulations 2 to 3 times the average number of observations per cluster. This can easily be explained by the fact that the weather patterns have been selected at least twice in the ensemble. Therefore, in the following, all ensembles will have 100 members, even if very small improvements (~0.01) in the ROC score could be reached when doubling the number of simulations.

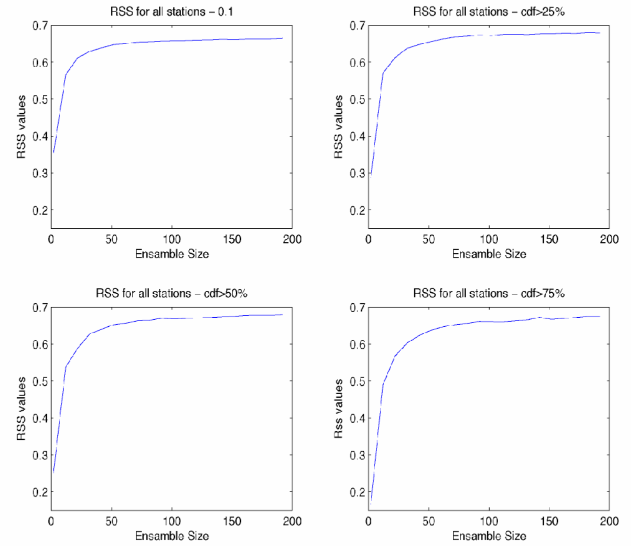


Figure 5: ROC skill score (RSS) averaged over the 39 stations as a function of the ensemble size, from 2 to 200 simulations, using k-means classifier and domain size of 4x4. RSS is computed for four different criteria: (upper-left) wet/dry day; (upper-right) daily amount larger than 25th percentile of the rainfall amount cumulative distribution function (CDF); (lower-left) daily amount larger than the 50th percentile; (lower-right) daily amount larger than 75th percentile.

4. MODEL SKILL IN ARGENTINA

Once we founded the optimal parameters for the set of 7 stations, we evaluate the general skills of the method over the entire database, from now on all results are generated by the SDS with the optimal parameters, meaning that the clustering algorithm will be k-means, the domain size around each station is of 4x4, and the ensemble size is of 100 members. To plot the results, instead of defining precipitation events in terms of a threshold in mm/day identical for all stations, we defined four types of daily events at each station. These events are wet days (amount larger than 0.1mm/day) and precipitation amounts larger than 25%, 50% or 75% of the amount distribution. These thresholds are proper to each station in order to compute the number of events predicted. Then the hit rates and false alarm rates are computed over all the stations in order to derive the ROC curve and generate the RSS value. As displayed in Figure 6, whatever the threshold considered, the mean RSS is around 0.72 highlighting the good results of the statistical method either for low or high rainfall amounts. Moreover, as a difference to computing the RSS score for events larger than, for example 10mm/day, which for a country like Argentina would reduce the evaluation of the model skills to wet regions, our method of computing events relative to, a percentile of the rainfall amount distribution, ensures that the RSS is indeed representing the skill of the method over all stations to simulate local rainfall amounts.

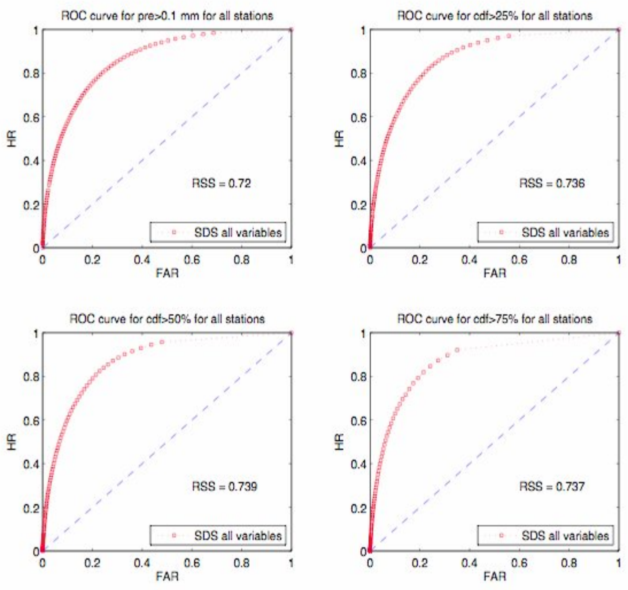


Figure 6: ROC curve computed from all the stations in the database, using k-means classifier, domain size of 4x4 and ensemble size of 100 members. The ROC curve is computed for four different criteria: (upper-left) wet/dry day; (upper-right) daily amount larger than the 25th percentile of the rainfall amount cumulative distribution function (CDF); (lower-left) daily amount larger than the 50th percentile; (lower-right) daily amount larger than the 75th percentile.

In order to study climate impacts (especially on hydrology and agriculture), it is important that a weather generator or a statistical downscaling method reproduce different distributions such as the rainfall amount, the dry spell and the wet spell distributions. Figure 7 displays such comparisons using pp-plots and qq-plots. It is worth pointing out that, while the classification method is based on the entire dataset (i.e. the seasonal cycle is not explicitly separated before the classification), we made all plots by comparing observed and simulated monthly distributions at all stations allowing to evaluating the representation of the seasonal cycle by the method. Finally, considering that our simulations are ensembles of 100 deterministic simulations, we computed the simulated distributions as follows. First, for each month of each simulation of each station, we computed the dry and wet sequence, and the rainfall amount CDF. Second, for each month and at each station, we averaged the CDFs between all the members of the ensemble before comparing to the observed monthly CDFs.

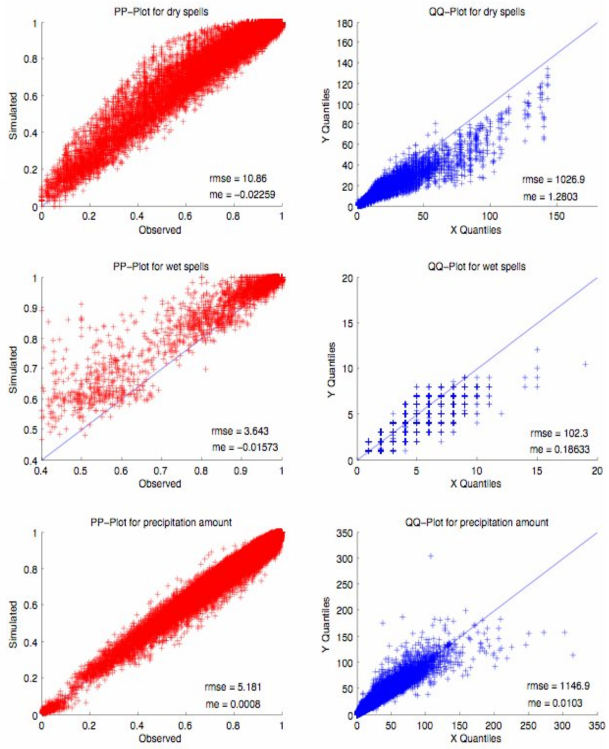


Figure 7: (a: upper-left) PP-Plot of the dry sequences; (b: upper-right) QQ-Plot of the dry sequences; (c: middle-left) PP-plot of the wet sequences; (d: middle-right) QQ-plot of the wet sequences; (e: lower-left) PP-plot of the rainfall amounts; (f: lower-right) QQ-plot of the rainfall amounts generated by using k-means classifier, domain size of 4x4 and ensemble size of 100 members. All plots are computed comparing the distributions of all twelve months at all stations, the root-mean square error (RMSE) and the model data error misfit (ME) are computed also.

First, the method is found to represent fairly well the dry spell distributions for all probabilities (Fig. 7a) although it clearly underestimates the dry spell lengths longer than 40 days by 30% to 50% (Fig. 7b). This bias is quite systematic in most of the weather generators (Wilks, 1999; Boulanger et al., 2007). A potential improvement of the method would be to take into account the weather pattern and precipitation history in the computation of the precipitation occurrence probability.

Second, coherently with the previous point, the statistical model significantly overestimates (Fig. 7c) the short rainfall spells, while it underestimates very long wet spells (longer than 10 days; Fig. 7d)

Third, rainfall amount distributions as displayed by Figures 7e-f seem to be fairly well simulated by the k-means method confirming its skill in simulating different amplitudes of rainfall events. Only in Figure 7f, it seems that the method does not reproduce very large specific daily rainfall events (larger than 200mm/day). This can actually be explained by the fact that all simulations over a specific year make use of a classification of weather patterns observed during all the period except this specific year. Therefore, very large individual events (not observed during the rest of the period) are unlikely to be simulated by the statistical method.

When analyzing more closely the model sensitivity over different regions of Argentina, the RSS values results at each station (Figure 8) confirm that the RSS values are very similar whatever the threshold suggesting that the statistical model skills are good either for low and large rainfall amounts. However, Figure 8 also displays regional differences. In particular, lower RSS values (0.55-0.6) are mainly observed near the Andes. We believe that such low scores can be explained by two reasons. First, these regions are quite dry and thus small errors in precipitation occurrence or amounts (in terms of % of the distribution) can lead to lower scores. Second, the quality of the reanalysis can also be lower near the Andes, where abrupt topography change is a clear challenge for atmospheric models. We believe that both reasons can lead to a transfer function not as skilful in relating weather patterns to local daily precipitation.

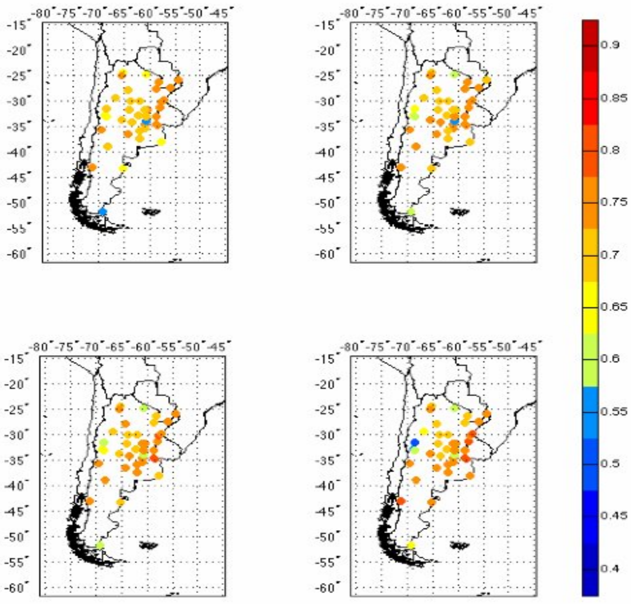


Figure 8: ROC skill score (RSS) at each station, using k-means classifier, domain size of 4x4 and ensemble size of 100 members. RSS is computed for four different criteria: (upper-left) wet/dry day; (upper-right) daily amount larger than the 25th percentile of the rainfall amount cumulative distribution function (CDF); (lower-left) daily amount larger than the 50th percentile; (lower-right) daily amount larger than the 75th percentile.

In order to better describe such differences, we displayed in Figure 9 different daily precipitation time series with low RSS (Mendoza~0.6 and Rio Gallegos~0.5) or large RSS scores (Gualeguaychú on the Uruguay River ~0.75; and Marcos Juarez~0.70). At the first two stations, the 60-day filtered daily precipitation time series display very low precipitation amplitudes (lower than 4mm/day in Mendoza and 2mm/day in Rio Gallegos). In Mendoza, the seasonal cycle is quite strong and fairly well simulated by the model. However, the model fails to capture some periods such as 1988, 1991, 1996 and 1998. In Rio Gallegos, the seasonal cycle is not well defined, and the interannual variability is large. The statistical model clearly fails in representing the observed variability and, instead, displays a relatively smooth annual and interannual variability.

On the contrary, at Gualeguaychú and Marcos Juarez stations with larger rainfall amplitudes the statistical method captures fairly well their strong annual and interannual variability. It is worth pointing out that even if the method fails in simulating some specific large peaks (related to individual large events), the green and pink curves (10th and 90th percentiles) follow pretty well the observed blue curve demonstrating that the ensemble does not present too large a spread in its simulations.

As a final validation test, we compared the results of the ensemble generated by our method with the ensemble generated by selecting any observed precipitation value of the same month in the training period as the predicted precipitation value (Fig. 10). This method allows to representing the seasonal cycle but has no skill in simulating the interannual variability. The series displayed in Figure 10 present a very weak year-to-year variability due to the stochastic nature of the simulation. Indeed, we only perform 100 simulations (to compare with our method) although each month has around 600 days of observations. The comparison between Figures 9 and 10 shows that the series generated as in Fig. 10 has no skill in representing the interannual variability of precipitation, while our method captures such differences between years of low or large amplitude in precipitation. As a consequence, the skill of our method is not related to a skill in simulating the seasonal cycle but really in capturing the interannual variability in precipitation associated to the interannual variability in the weather patterns. This result is confirmed by comparing the RSS values of the ensembles displayed in Figures 9 and 10 (Table 4), where it clearly appears that the skill of the ensembles generated by the seasonal cycle method (Fig. 10) ranges between 0.09 and 0.30 while the skill of our method ranges between 0.52 and 0.76.

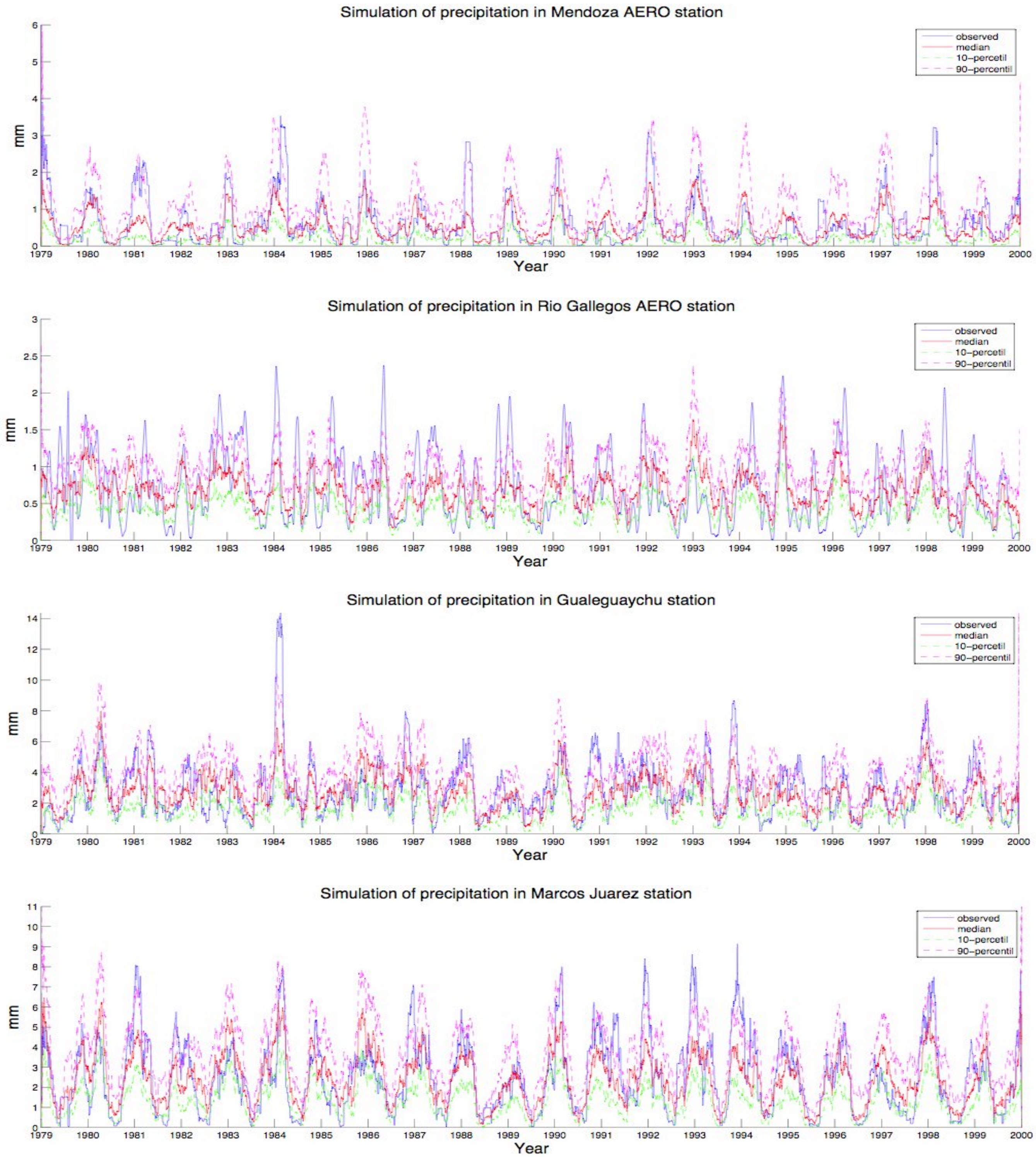
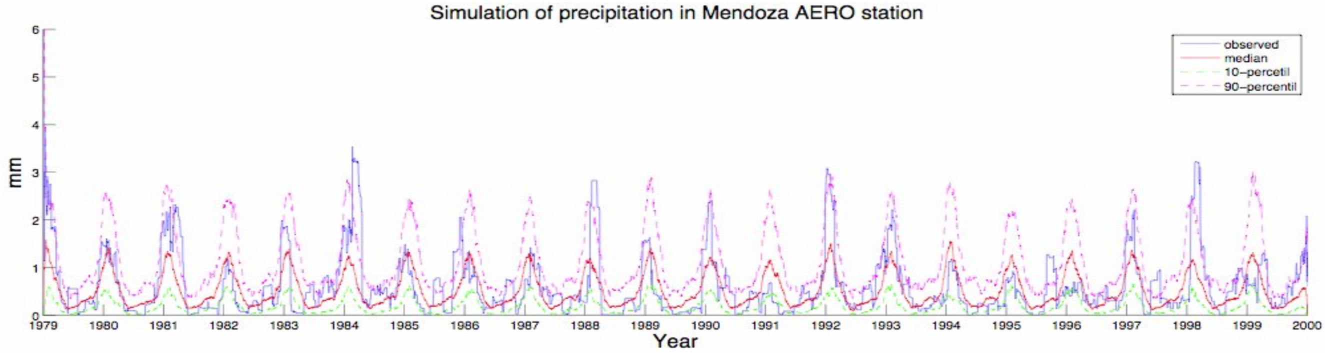
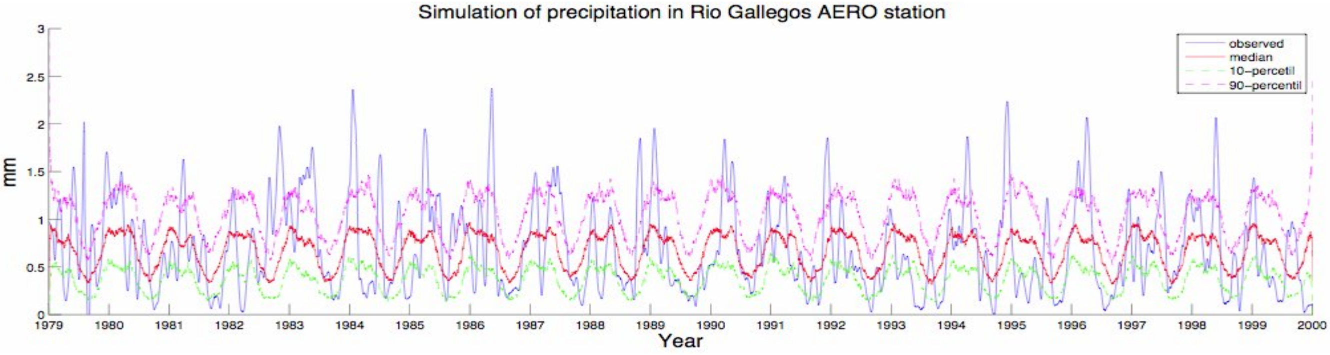
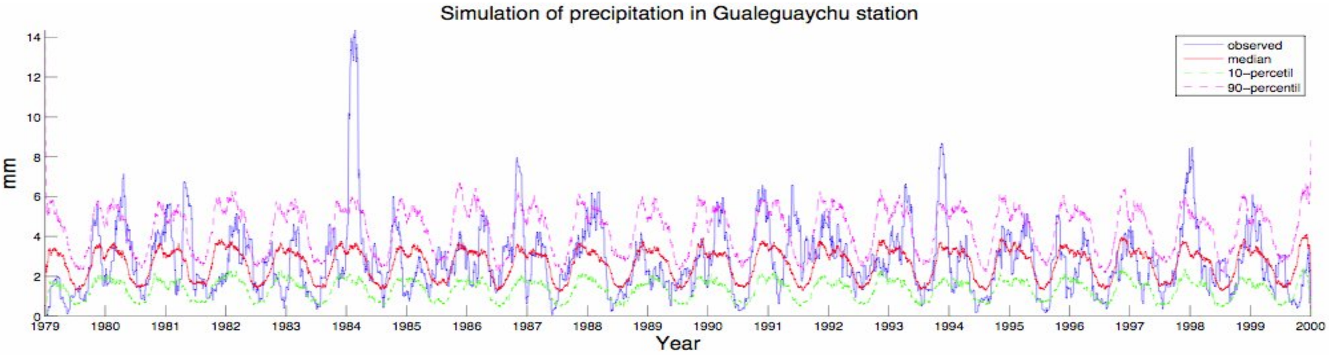


Figure 9: Two-month filtered time series of observed daily precipitation (blue-solid curve), of simulated ensemble median (red-solid curve), of simulated ensemble 10th percentile (green-dashed curve) and of simulated ensemble 90th percentile (pink-dashed curve). From up to bottom, the time series correspond to Mendoza, Rio Gallegos, Gualeguaychú and Marcos Juarez. Ensembles were generated using k-means classifier, domain size of 4x4 and ensemble size of 100 members.


Simulation of precipitation in Mendoza AERO station


Simulation of precipitation in Rio Gallegos AERO station


Simulation of precipitation in Gualeguaychu station

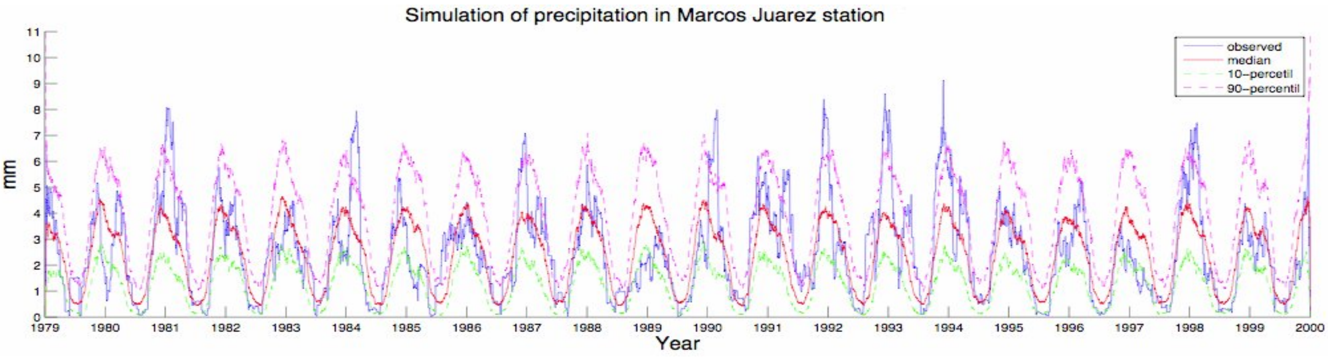

Simulation of precipitation in Marcos Juarez station

Figure 10: Two-month filtered time series of observed daily precipitation (blue-solid curve), of simulated ensemble median (red-solid curve), of simulated ensemble 10th percentile (green-dashed curve) and of simulated ensemble 90th percentile (pink-dashed curve). Generated byFrom up to bottom, the time series correspond to Mendoza, Rio Gallegos, Gualeguaychú and Marcos Juarez. Ensembles were generated using k-means classifier, domain size of 4x4 and ensemble size of 100 members.

	RSS values for Method 1				RSS values for Method 2			
Station	Pre>0.1mm	cdf>25%	cdf>50%	cdf>75%	Pre>0.1mm	cdf>50%	cdf>50%	cdf>75%
Mendoza	0.62	0.64	0.62	0.57	0.19	0.22	0.23	0.30
Rio Gallegos	0.52	0.60	0.61	0.64	0.13	0.16	0.16	0.15
Galeguaychú	0.75	0.76	0.76	0.76	0.09	0.11	0.12	0.11
Marcos Juarez	0.71	0.70	0.70	0.71	0.22	0.30	0.30	0.29

Table 4: Comparison of RSS values between: Method 1: our SDS method using k-means, 4x4 domain size and 100 members ensemble. Method 2: Using any observation of the same month in the training period as the predicted precipitation value. The RSS values are computed for four different criteria: wet/dry day; daily amount larger than the 25th percentile of the rainfall amount cumulative distribution function (CDF); daily amount larger than the 50th percentile; daily amount larger than the 75th percentile.

5. CONCLUSION AND PERSPECTIVES

Very few studies (e.g. Solman and Nuñez, 1999; Gutierrez et al., 2004) have developed and used statistical downscaling methods to predict daily precipitation or temperature at local scale (weather station) in South America. The present study is a contribution to fill this gap. We developed and calibrated a weather pattern clustering technique using the 1979-1999 ERA-40 atmospheric reanalysis and a set of 39 quality-controlled weather stations in Argentina. The variety of climates observed in Argentina gives us the opportunity to develop a method valid either for the north of the country (tropical climate), for the south (mid- to high-latitude climate), for the west (near the Andes), for the east (Atlantic ocean coast) and for the interior of the country.

First, the method parameters were optimized based on the RSS:

(i) *The clustering technique*: A comparison between the k-means and the SOM (Self-Organizing Map) clustering methods led to the conclusion that k-means has two main advantages over SOM. First, k-means has a better generalization skill than SOM, this is due to the topological order property of SOM that reduces the possibility for the clusters to represent well all the data space (especially regions of low density) and therefore the generalization skill of the projection method. Secondly, k-means algorithm has less parameters than SOM and therefore its application is straightforward, on the other hand SOM has more parameters making harder to reach a good generalization without finding the optimum values for this specific problem.

(ii) *The number of clusters or average number of observations per cluster*: The RSS was found to be optimum for an average of 39 or more observations per cluster. For lower values, the generalization skill of the clustering analysis reduces significantly the RSS. For very large values, the same applies as too many weather states are mixed in a same cluster leading to a loss in predictive information. We decided to use 196 clusters equivalent to an average of 39 observations per cluster.

(iii) *The domain size*: First (not shown), no significant sensitivity to the location was found in the computation of an optimum domain size to extract the atmospheric variables and thus to define the weather pattern. It was found that the optimum domain size was relatively small (~4° to 6° square around the station). Moreover, the same analysis reducing the atmospheric variables to dynamical only variables led to the conclusion that the optimum domain size depends on the atmospheric variables. Indeed, dynamical variables define an optimum transfer function if the domain size is larger (~10°) suggesting that our results with all variables together is strongly influenced by local- to synoptic-scale non-dynamical variables (moisture component). It also suggests that dynamical and non-dynamical variables should be defined onto different domains.

(iv) *Ensemble size*: As any stochastic method, it is necessary to compute an ensemble of simulations. The RSS is found to converge rapidly to a maximum value when the number of simulations is larger than twice or thrice the average number of observations per cluster. This result is coherent with the fact that, in such a case, for each day, the ensemble of simulated values does explore all the distribution of daily values in each cluster.

Although Argentina is a country with a variety of climates (from tropical in the north to mid- to high-latitude in the south), our parameter optimization is found to be homogeneous all over the country despite some lower results near the Andes and south, probably due to a topography effect, which may affect the transfer function between ERA-40 reanalysis near-Andes weather patterns and daily precipitation. The RSS for different criteria (wet/dry days, days with rainfall larger than the 25th, 50th or 75th percentiles of the rainfall amount CDF) is found to be relatively stable (around 0.72 for the entire database and for each criterion). More interestingly, while the method does not explicitly separate the seasonal cycle characteristics, the simulated monthly CDFs of wet spells, dry spells and rainfall amounts do compare fairly well to the observed CDFs. The major weakness of the method is its underestimation of either long dry or wet spells. Such a weakness is similar to other methods or to weather generators. Finally, the analysis of the observed and simulated time series allows concluding that the method has less skill in arid regions than in semi-arid or wet regions. Indeed, in arid regions, small errors in the simulation of the amplitude of rainfall events (defined as percentiles of the distribution) do strongly affect the statistical scores. In the opposite, in wetter regions, the regimes of precipitation (weak vs. large amplitude) and their variability (seasonal, interannual) are much better captured and simulated by the statistical method.

To conclude, the present method could be improved in the following ways. First, as we found that dynamical and non-dynamical variables relate to the local precipitation through weather patterns of different spatial sizes, an improvement would thus be to combine different spatial domains associated to the different types of variables (dynamical and non-dynamical). Second, in this first version, during a simulation, a daily weather pattern is projected onto the clusters and only one “winning” cluster is considered (the one with the lowest distance). Then to simulate the precipitation, all observations belonging to the winning cluster are equally probable. The introduction of the weather pattern and/or precipitation history in conditioning the probability of each precipitation observation may improve our simulations, at least the simulations of long dry and wet spells. Last, it could be possible to relax the “winning” cluster condition by defining a probability function based on the distance to all the clusters. This “relaxation” complementary to the second possible improvement, may allow to better represent the probability of specific rainfall events and, in particular, extreme events. Such improvements will be tested in the near future to the La Plata Basin whether stations in order to provide climate change scenarios. This work will be part of the CLARIS LPB Project (2008-2012) funded by the 7th Framework programme of the European Commision.

Acknowledgements

We wish to thank the European Commission 6th Framework programme for funding the CLARIS Project (Project 001454) during the 3-year duration of the project. Jean-Philippe Boulanger wants to thank the Centre National de la Recherche Scientifique (CNRS) for the administrative coordination of the project, the Institut de Recherche pour le Développement (IRD) for its constant support, and the University of Buenos Aires and its “Department of Atmosphere and Ocean Sciences” for welcoming him during the entire duration of the project. We are grateful to the European Center for Medium Range Weather Forecast (ECMWF) for providing the ERA-40 dataset to the CLARIS Project, and to the IPSL for archiving the ERA-40 dataset and helping us in the data extraction.

REFERENCES

A. K Jain, M. N. Murty, P. J. Flynn. Data Clustering: A Review. ACM Computing Surveys, Vol. 31, Number. 3, 265-322, September 1999.

Barry Smit , Mark Skinner, Adaptation options in agriculture to climate change: a typology. Migration and Adaptation Strategies for Global Change, Vol. 7:Number 1:85-114 , 2004.

Boulanger, J.-P., Fernando Martinez, Olga Penalba and Enrique Carlos Segura, Neural network based daily precipitation generator (NNGEN-P), Clim. Dyn., DOI 10.1007/s00382-006-0184-y, 2007.

Belmman, R. Adaptive Control Process : A Guided Tour. Princenton University Press, New Jersey, (1961).

Cavazos, T., Using Self-Organizing Maps to Investigate Extreme Climate Events: An Application to Wintertime Precipitation in the Balkans, *J. Clim*. 13, 1718-1732, 2000

Cavazos T, and B. C. Hewitson, Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. *Climate Research* Vol. 28:95-107, 2005.

Gutiérrez J. M, Cano R., Cofiño A. S. and C. Sordo, Analysis and Downscaling Multi-Model Seasonal Forecast using Self-Organizing Maps, Tellus, 2004.

Hewitson B.C., Crane R.G, Consensus between GCM climate change projections with empirical downscaling: Precipitation downscaling over South Africa, Int. J. Climatology, 2006.

Hughes, J., Lettemaier, D. P., and Guttorp, P. A stochastic approach for assessing the effect of changes in regional circulation patterns on local precipitation. Water Resources Research, 29:3303–3315, 1993.

Kharin, V. V., and F. W. Zwiers, On the ROC score of probability forecasts, J. Clim., 16(24) :4145, 2003.

Kohonen, T, Self-Organizing formation of topologically correct feature maps. Biological Cybernetics, 43(1):59-69, 1982

Lorenz, E. N. Atmospheric predictability as revealed by naturally occurring analogues. Journal of Atmospheric Sciences, 26:636-646, 1969.

Magrin, G. O. , Travasso M. I. , Rodriguez G. R. , Changes in Climate and Crop Production During the 20th Century in Argentina. Climatic Change 72:229-249, 2005.

Mehrotra, R., and A. Sharma, A nonparametric nonhomogeneous hidden Markov model for downscaling of multisite daily rainfall occurrences, *J. Geophys. Res*., 110, D16108, doi: 10.1029/2004JD005677, 2005.

Solman S. A., Nuñez Mario N., Local estimates of global climate change: A statistical downscaling approach. Int. J. Climatology 19:835-861, 1999.

Vrac, M., M. L. Stein, K. Hayhoe, and X.-Z. Liang, A general method for validating statistical downscaling methods under future climate change, *Geophysicial Research Letters*, Vol. 34, L18701, doi:10.1029/2007GL030295, 2007

Wilby, R. L., T. M. L. Wigley, D. Conway, P. D. Jones, B. C. Hewitson, J. Main, and D. S. Wilks , Statistical Downscaling of General Circulation Model Output: A Comparison of Methods, WATER RESOURCES RESEARCH, Vol. 34, No. 11, 2995–3008, 1998.

Wilks, D. S., Interannual variability and extreme-value characteristics of several stochastic daily precipitation models, *Agric and Forest Meteor*., 93, 153-169, 1999.