Manuscript prepared for J. Name
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 1 January 2016

Artificial Neural Networks and Multiple Linear Regression Model Using Principal Components to Estimate Rainfall over South America

Thalyta Soares dos Santos[a], David Mendes[a], and Roger Rodrigues Torres[b]

[a]Federal University of Rio Grande do Norte, Campus Universitário Lagoa Nova, Natal, RN, 59078-970, Brazil

[b]Federal University of Itajubá, Instituto de Recursos Naturais. , Av. BPS, 1303, Pinheirinho, Itajubá, MG, 37500-903, Brazil

Correspondence to: Thalyta Soares dos Santos (tthalysoares@gmail.com)

Abstract. Several studies have been devoted to dynamic and statistical downscaling for analysis of both climate variability and climate change. This paper introduces an application of artificial neural networks (ANN) and multiple linear regression (MLR) by principal components to estimate rainfall in South America. This method is proposed for downscaling monthly precipitation time series over South America for three regions: the Amazon, Northeastern Brazil and the La Plata Basin, which is one of the regions of the planet that will be most affected by the climate change projected for the end of the 21st century. The downscaling models were developed and validated using CMIP5 model out- put and observed monthly precipitation. We used GCMs experiments for the 20th century (RCP Historical; 1970-1999) and two scenarios (RCP 2.6 and 8.5; 2070-2100). The model test results indicate that the ANN significantly outperforms the MLR downscaling of monthly precipitation variability.

1 Introduction

The forecasting of meteorological phenomena is a complex task. The mathematical, statistical, and dynamic methods developed in recent decades help address the problem, but there is still a need to investigate new techniques to improve the results. One of these techniques is statistical down-scaling, which involves the reduction of the model's spatial scale. Downscaling techniques can be divided into two broad categories: dynamic and statistical. Dynamic techniques focus on numerical models with more detailed resolution, while statistical (or empirical) techniques use transfer functions between scales. Currently, numerical weather prediction (NWP) models can forecast various meteorological variables with acceptable accuracy (Ramírez, 2006).

Specifically, rainfall is of great interest, both for its climatic and meteorological relevance and direct effect on agricultural output, hydropower generation and other important economic factors. However, it is one of the most difficult variables to forecast, because of its inherent spatial and temporal variability (Wilson, 2002; Antolik, 2000). For this reason, the temporal and spatial scales involved are not yet solved satisfactorily by the available numerical models (Olson, 1995).

Ramos (2000), studying artificial neural networks (ANN) and multiple linear regression (MLR), found that the neural networks method performed better than linear regression, although both showed good performance for monthly and seasonal rainfall. Ramírez (2005), using observed daily rainfall in the São Paulo region, found that ANN outperformed to LMR, which showed a high bias for days without rain. Ramírez (2006) analyzed daily rainfall in Southeastern Brazil and concluded that the ANN method tended to predict moderate rainfall with greater accuracy during austral summer compared to ETA model forecasts. Mendes (2010) reported that the daily rainfall in the Amazon Basin is better represented by ANN than autocorrelation models.

In this context, the aim of this study is to conduct a statistical downscaling to estimate rainfall over South America, based on some models used in the fifth report of the IPCC (Intergovernmental Panel on Climate Change), by applying artificial neural networks and multiple linear regression using principal components.

2 Data and Methods

2.1 Data

We used monthly precipitation simulations for the austral summer (December - January - February) and winter (June - July - August) generated by ten models (Table 1) from the CMIP5 project (Coupled Model Intercomparison Project 5th Phase), obtained from the Earth System Grid Federation (ESGF) of the German Climate Computing Center (http://ipcc-ar5.dkrz.de) and the Program for Climate Model Diagnosis and Intercomparison (http://pcmdi3.llnl.gov). All model simulations for the 20th century were compared with the precipitation data of the CRU TS 3.0 (Mitchell, 2005), produced by the Climatic Research Unit (CRU) - University of East Anglia (UEA). These data cover the period from 1901 to 2005 and have spatial resolution from 0.5° x 0.5°. We used climate simulations for the 20th century (historical) in the 1970-1999 period, and projections for the 21st century (Representative Concentration Pathways - RCP 2.6 and 8.5), for the period 2070-2099, as defined by Moss (2010).

Our focus on South America is because it is one of the planet's regions that will be most affected by the climate change projected for the end of the 21st century (Marengo, 2010). According to Magrin (2014), significant trends in precipitation and temperature have been ob- served in South America (SA). In addition, changes in climate variability and in extreme events have severely affected the region. The three sub-regions evaluated in South America were defined according to the precipitation regime: the Amazon (AMZ), Northeastern Brazil (NEB), and the La Plata Basin (LPB) (Figure 1).

2.2 Methods

2.2.1 Artificial Neural Networks (ANN)

An ANN is a system inspired by the operation of biological neurons with the purpose of learning a certain system. The construction of an ANN is achieved by providing a stimulus to the neuronal model, calculating the output and adjusting the weights until the desired output is achieved. An entry is submitted to the ANN along with a desired target, a defined response for the output (when this is the case, the training is regarded as supervised). An error field is built based on the difference between the desired response and the output of the system. The error information is used as feedback for the system, which adjusts its parameters in a systematic way, in other words, the backpropagation error algorythm is used to train the network. According to Alsmadi et al. (2009) the backpropagation architecture is the most popular, effective, and easy to learn model for complex, multilayered networks. This network is used more than all others combined. This algorithm has a first phase with a functional propagation signal (feedforward) and a second phase with the backpropagation of the error (backpropagation).

In the first phase, the functional signal based on the inputs propagates through the network until generating an output, with the weights of synapses remaining fixed. In the second phase, the output is compared with a target, producing an error signal. The error signal propagates from the output to the input and the weights are adjusted in such a way as to minimize the error. The process is repeated until the performance is acceptable. As such, the performance of the ANN is strongly dependent on the data source.

A first part of the data is used for training, the second is used for cross- validation, and the third part is used for testing. The architecture of the ANN used in the present study can be found in Figure 2. It consists of an input, a hidden layer and an output layer. The number of intermediate units was obtained through trial and error. During the training, the performance of the ANN is also assessed within the validation set.

The structure of the ANN used here involves training of 11 predictors (10 outputs of the models plus the observation data) as input to the network, and the best network performance is selected. We therefore expect that the ANN will be able to provide more reliable values (through the error analysis between the simulated values) than when using only climate models.

2.2.2 Multiple Linear Regression using Principal Components

Multiple linear regression (MLR) is a statistical technique that consists of finding a linear relationship between a dependent (observed) variable (and more than one independent variable (outputs of the GCMs). A multiple regression model can be represented by the following equation:

$$Y_i = a + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m + C, \tag{1}$$

Where Y_i is the dependent variable, X_1, X_2, $\cdots$, X_m the independent variables, a is the intercept, b_1,b_2 and b_m are the multiple regression coefficients, to be estimated by the least squares method (Wilks, 1995), and C is the error term.

In spite of their obvious success in many applications, MLRs present multicollinearity when employed with climatic variables. In this regard, the parameter estimation errors can be incorrectly interpreted Leahy (2000). To resolve this problem, we used principal components (PCs). This method seeks to reduce the number of variables through orthogonal transformations, and to remove the multicollinearity of the independent variables. The PCs of the explanatory variables are therefore a new set of variables with the same information as the original variables, but uncorrelated.

The Multiple Linear Regression is commonly used in various research areas, and is widely accepted by the scientific community. The Artificial Neural Networks are still being inserted in science, especially when it comes to climate studies. Our intention is to show advantages of using Artificial Neural Networks to the weather. The advantages of the Artificial Neural Networks stands out: The nonlinearity inherent Networks that allows this technique can perform functions that a linear program (such as Multiple Linear Regression) can not. In addition, a neural network can be designed to provide information not only about which particular pattern select, but also on the confidence in the decision.

3 Results and Discussion

3.1 Validation of the ANNs

After using the precipitation simulations for the period 1970-1999 with the ANNs, we obtained a final error after a number of interactions, which ranged from 1 to 600 (Figure 3). One of the difficulties of using ANNs consists of identifying the best stopping point for training Haykin (2001), because the training error starts out with a maximum value, decreases rapidly and then levels off, indicating there is no more error to correct. In the summer, the network became stable more rapidly, indicating that the GCMs employed converge to the same pattern of precipitation.

With respect to winter, the networks remained unstable for a longer time before finding the minimum error. The NEB region should be highlighted, which required the largest number of iterations, around 600. This is possibly related to the greater variability of rainfall in this season (Figure 3).

According to Villanueva (2011), it is assumed that the three sets (training, validation and testing) contain independent samples, and that they are well capable of representing the problem being addressed. One should therefore expect that good performance on the validation set will imply good performance of the testing set. In this study, the validation values were closest to the test values in summer.

3.2 Validation of the MLR by PC

To validate the MLR, the following assumptions need to be met: i) the residuals must have random distribution around mean zero (homoscedasticity); ii) the residuals should have a normal distribution; and iii) variance must be homogeneous (Silva, 2014).

Figures 4 and 5 show that the residuals versus adjusted values meet the assumption of homoscedasticity. With respect to the QQ-plot, the quantiles of the residuals versus the normal distribution indicate that all regions present normality in the residuals. Given that the closer the residuals are to the line, the closer they are to having normal distribution. The employed data therefore fit the MLR by the PC model. Based on the PC analysis (Table 2), one can see that in summer for the AMZ region, the accumulated proportion explains around 77% in NEB and 80% in PC6, while in winter, the PC1 of the AMZ explained 71% and PC3 explained 72% in NEB, thus representing the greatest variability of precipitation in these regions. In general, one can observe that a smaller number of climate models were required in winter to capture the variance of precipitation in these regions. Similar behavior of PCs in both seasons stands out in the LPB region, which may be due to the failure of GCMs to capture the variance of precipitation in this region.

Tables 3 and 4 show the Pearson's correlation coefficients at significance level of 5% between the ANNs and the observed data, and between the MLR by PC and observed data, respectively. One can see that in both downscaling methods used, the highest correlations occur in winter in all regions under study, indicating that the models are better able to represent the variability of precipitation during this season.

Ramírez (2006) performed statistical downscaling for the precipitation forecast for the Southeast of Brazil, using ANNs and MLR with the ETA model. The results suggested that the precipitation forecasts using ANNs performed better in winter than in summer, since the synoptic forcing is more pronounced and the deep convective activity is less common. One can also observe that in the regions NEB (ANN x Obs) and LPB (MLR x Obs), the correlations of 38 and 20%, respectively, were not statistically significant. The lowest correlation occurred in the LPB region. Seth (2010) stated that the mean of the set of models reveals weaker moisture transport east of the Andes, which may be one of the factors that induce underestimation of precipitation in this region.

3.3 Downscaling scenarios

Table 5 presents the results of the monthly precipitation simulation for the end of this century (2071-2100) based on the ten GCMs described previously in the RCP scenarios 8.5 and 2.6, in relation to the reference period 1971-1999 (observation) for the two downscaling methods.

In both scenarios, and employing both ANNs and MLR, an increase of precipitation in the summer and a decrease in the winter can be observed. These results corroborate the findings of Mendes (2010), who used ANNs and autocorrelation to study changes in monthly precipitation for the Amazon Basin in scenarios A2, A1B and B1, derived from five models of the CMIP3, used in the IPCC AR4. The authors found an increase in precipitation in the summer months and a reduction in winter.

In the NEB region (Table 5), an increase of precipitation in summer of around 30% was observed. With respect to winter, one can see a reduction of 40% in the higher forcing scenario (RCP 8.5), and of 10% (RCP2.6) in the lower climate forcing. The IPCC AR4 revealed CMIP5 precipitation projections for end of century (2081?2100) of increase precipitation from October to March over the southern part of Southeast Brazil and the La Plata Basin. From April to September, the CMIP5 ensemble projects precipitation increases over the La Plata Basin and northwestern SA near the coast (Stocker, 2013). According to Magrin (2014) seasonal scales, rainfall reductions during winter and spring in southern Amazonia may indicate a late onset of the rainy season in those regions and a longer dry season. The changes are more intense for the late 21-st century and for the RCP8.5 when compared to scenario RCP 2.6, as can be seen in Table 5.

4 Conclusions

This paper investigated the applicability of artificial neural networks and multiple linear regression analysis by principal components, as temporal downscaling methods for the generation of monthly precipitation over South America (for current years and future scenarios). Both the ANN and MLR methods provided good fit with the observed data. This indicates that ANNs are a viable alternative for the modeling of precipitation in time series. ANNs can be compared with the statistical model, and this indicates that the networks are a potentially competitive tool.

The future scenarios used (RCP 2.6, lower climate forcing, and RCP 8.5, higher climate forcing) indicate an increase in precipitation in summer and a reduction in precipitation during winter according to both the methods used.

In general, the results showed that the use of ANNs produced more accurate results than MLR by PC, which can be attributed to the fact that ANNs perform tasks that a linear program is unable to do. In addition, one of the advantages of ANNs is their capacity for temporal processing, and thus their ability to incorporate not only concurrent, but also several predictive values, as inputs without any additional effort.

Acknowledgements. We are grateful to CAPES and PPGCC/UFRN for financial support.

References

Alsmadi, M. K. S., Omar, K. B. and Noah, S. A: Back propagation algorithm: the best algorithm among the multi-layer perceptron algorithm. International Journal of Computer Science and Network Security, 9(4), 378-383, 2009.

Antolik, M. S.: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. Journal of Hydrology, 239(1), 306-337, DOI = 10.1016/S0022-1694(00)00361-9, 2000.

Haykin, Simon S. Redes neurais. Bookman, Porto Alegre, 2001.

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V. and Midgley, P. M.: Climate change 2013: The physical science basis. Intergovernmental Panel on Climate Change, Working Group I Contribution to the IPCC Fifth Assessment Report (AR5), Cambridge University Press, New York, 2013.

Leahy, K.: Multicollinearity: When the solution is the problem. In: Rud OP, editor. Data mining cookbook: Modelling data for marketing, risk and costumer relationship management. New York: John Wiley & Sons, 106–108, 2001.

Magrin, G. O., Marengo, J. A., Boulanger, J. P., Buckeridge, M. S., Castellanos, E., Poveda, G., Scarano, F. R. and Vicuna, S.: Central and South America. In: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 1499–1566, 2014.

Marengo, J. A., Ambrizzi, T., Da Rocha, R. P., Alves, L. M., Cuadra, S. V., Valverde, M. C., Torres, R. R., Santos, D. C. and Ferraz, S. E.: Future change of climate in South America in the late twenty-first century: intercomparison of scenarios from three regional climate models. Climate Dynamics, 35(6), 1073-1097, DOI = 10.1007/s00382-009-0721-6, 2010.

Mendes, D. and Marengo, J. A.: Temporal downscaling: a comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. Theoretical and Applied Climatology, 100(3-4), 413–421, DOI = 10.1007/s00704-009-0193-y, 2010.

Mitchell, T. D. and Jones, P. D. An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International journal of climatology, 25(6), 693–712, DOI = 10.1002/joc.1181, 2005.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., and Wilbanks, T. J. The next generation of scenarios for climate change research and assessment. Nature, 463(7282), 747–756, DOI = 10.1038/nature08823, 2010.

Olson, D. A., Junker, N. W., and Korty, B. Evaluation of 33 years of quantitative precipitation forecasting at the NMC. Weather and Forecasting, 10(3), 498–511, DOI = 10.1175/1520-0434(1995)010<0498:EOYOQP>2.0.CO;2, 1995.

Ramos, A. M.: Desagregação espacial da precipitação simulada por modelos atmosféricos no Nordeste do Brasil. Master's thesis, Federal University of Paraíba. 96p., 2000.

Seth, A., Rojas, M., and Rauscher, S. A.: CMIP3 projected changes in the annual cycle of the South American Monsoon. Climatic Change, 98(3-4), 331–357, DOI = 10.1007/s10584-009-9736-6, 2010.

da Silva, A. G. and Silva, C. M. S.: Improving Regional Dynamic Downscaling with Multiple Linear Regression Model Using Components Principal Analysis: Precipitation over Amazon and Northeast Brazil. Advances in Meteorology, v. 2014, Article ID 928729, 9 pages, DOI = 10.1155/2014/928729, 2014.

Ramírez, M. C., Ferreira, N. J., and Velho, H. F. C.: Linear and nonlinear statistical downscaling for rainfall forecasting over southeastern Brazil. Weather and forecasting, 21(6), 969–989, DOI =10.1175/WAF981.1, 2006.

Ramírez, M. C. V., de Campos Velho, H. F., and Ferreira, N. J.: Artificial neural network technique for rainfall forecasting applied to the São Paulo region. Journal of Hydrology, 301(1), 146-162, DOI = 10.1016/j.jhydrol.2004.06.028, 2005.

Villanueva, W. J. P.: Síntese automática de redes neurais artificiais com conexões à frente arbitrárias. Master's thesis, Federal University of Campinas. 220p., 2011.

Wilks, D. S.: Statistical methods in the atmospheric sciences: An Introduction. Academic press, San Diego, 1995.

Wilson, L. J., and Vallée, M.: The Canadian updateable model output statistics (UMOS) system: Design and development tests. Weather and forecasting, 17(2), 206–222, DOI = 10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2, 2002.

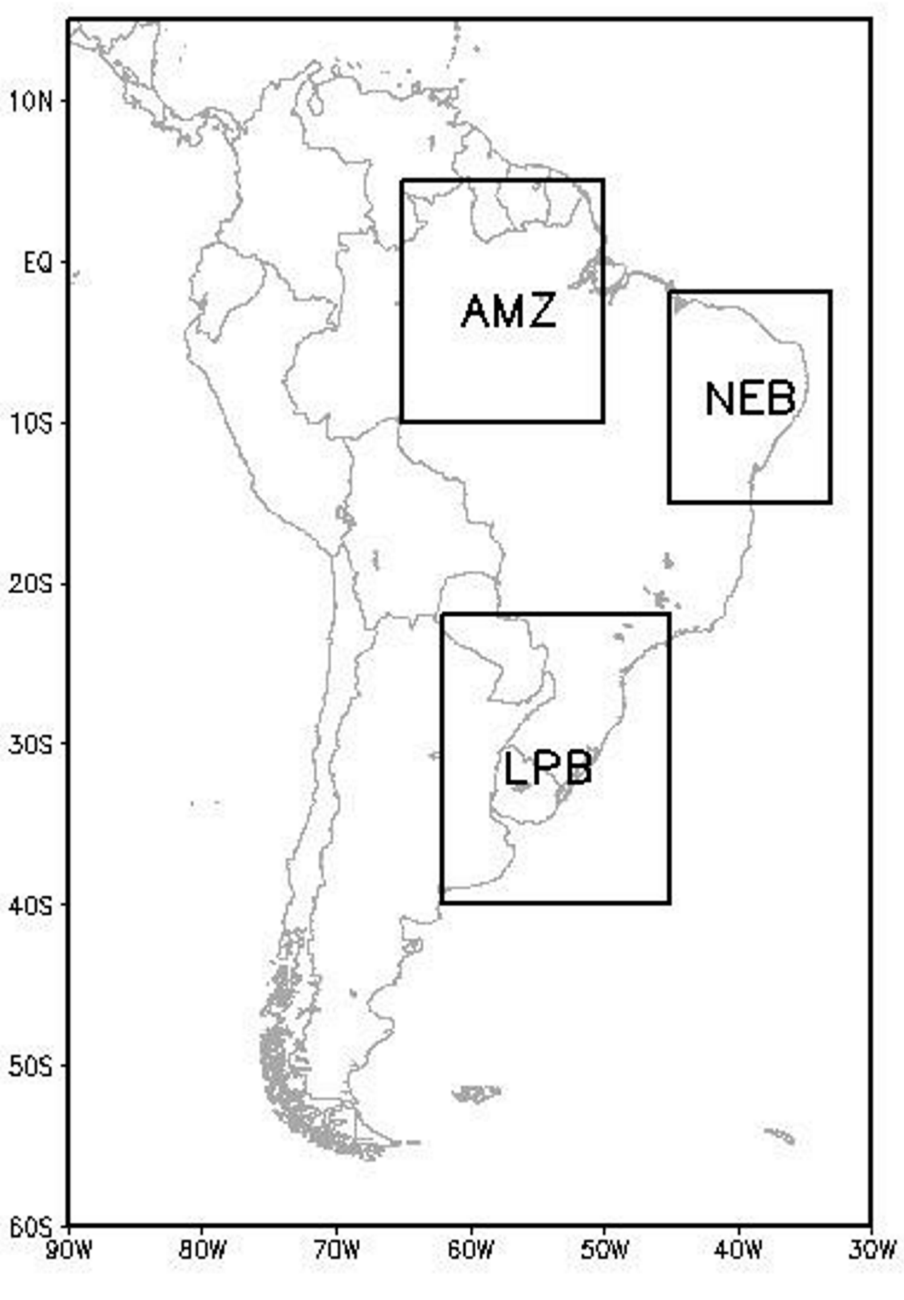

Figure 1. Illustration of the study areas of the defined regions.

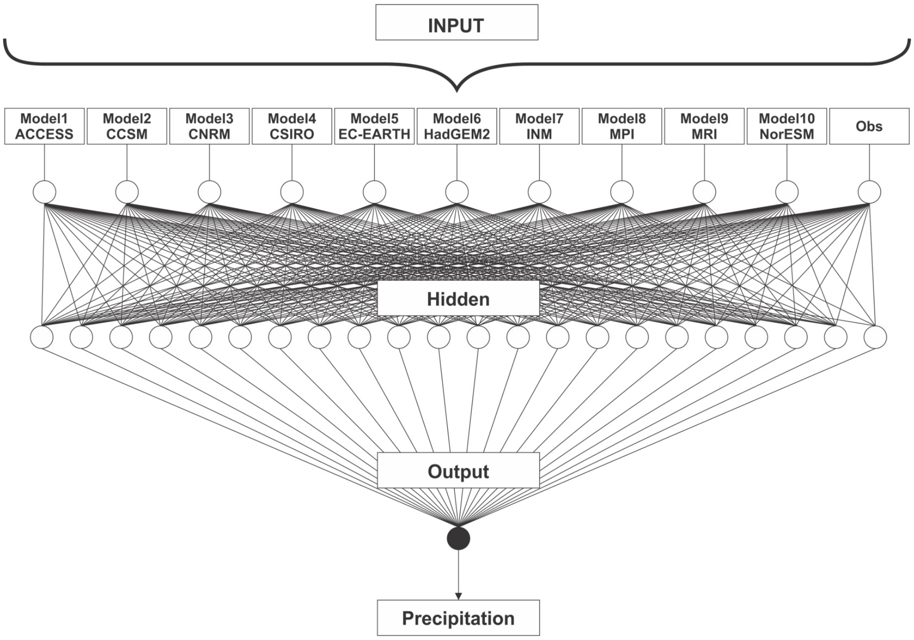

Figure 2. Structure of the artificial neural network.

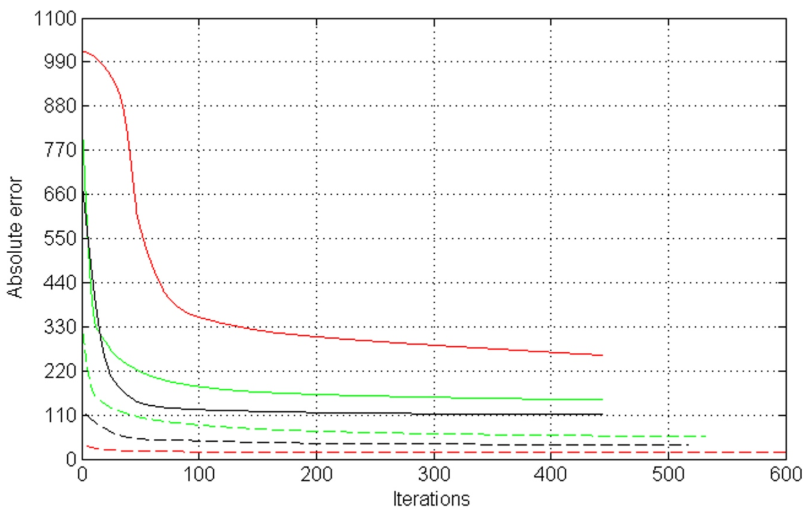

Figure 3. Absolute error as a function of the number of iterations, AMZ (green), NEB (red) and LPB (black). Continuous lines represent the summer period for each region and the dashed lines represent winter.

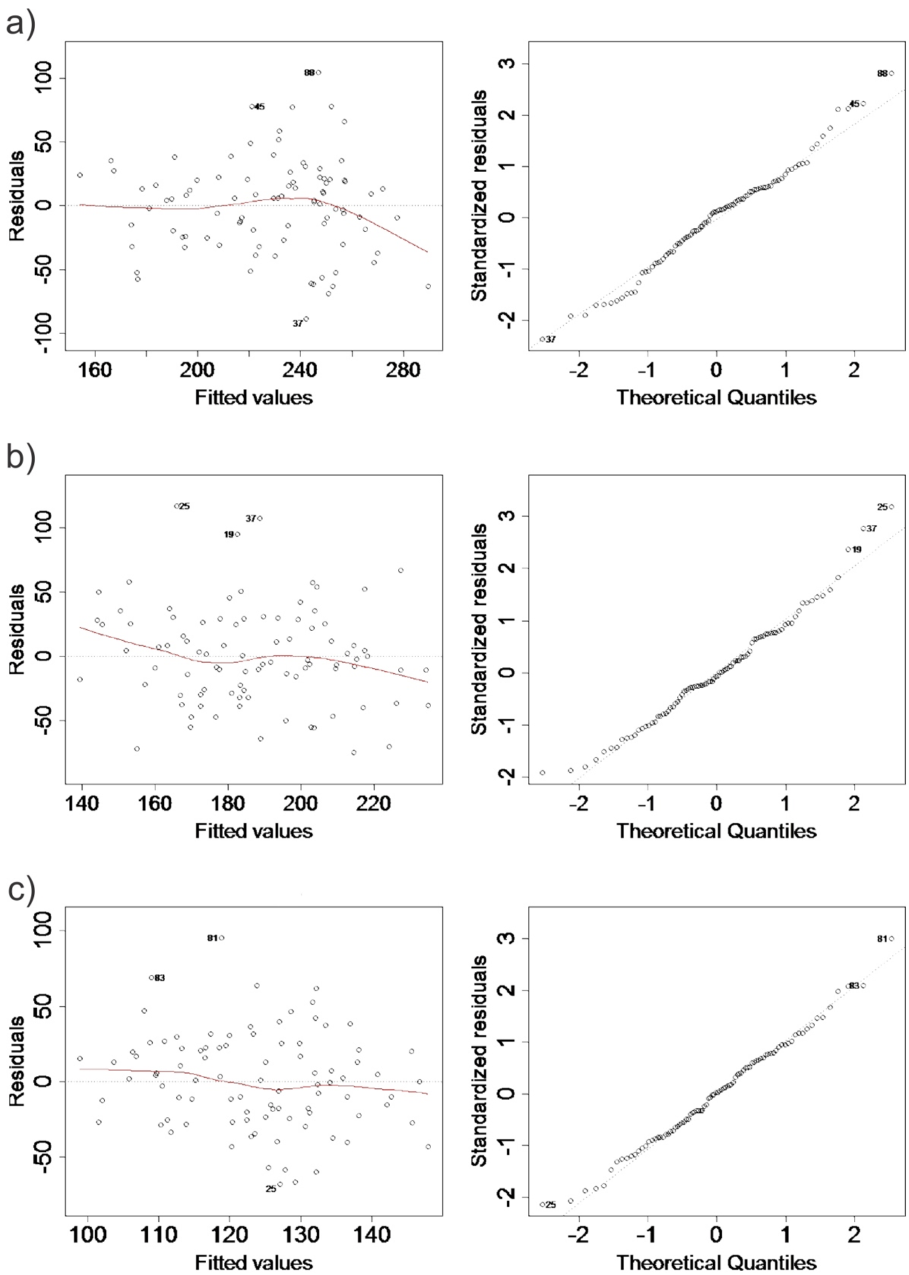

Figure 4. Residuals x Fitted Values and Theoretical Quantiles, for the summer. (a) AMZ, (b) NEB, and (c) LPB..

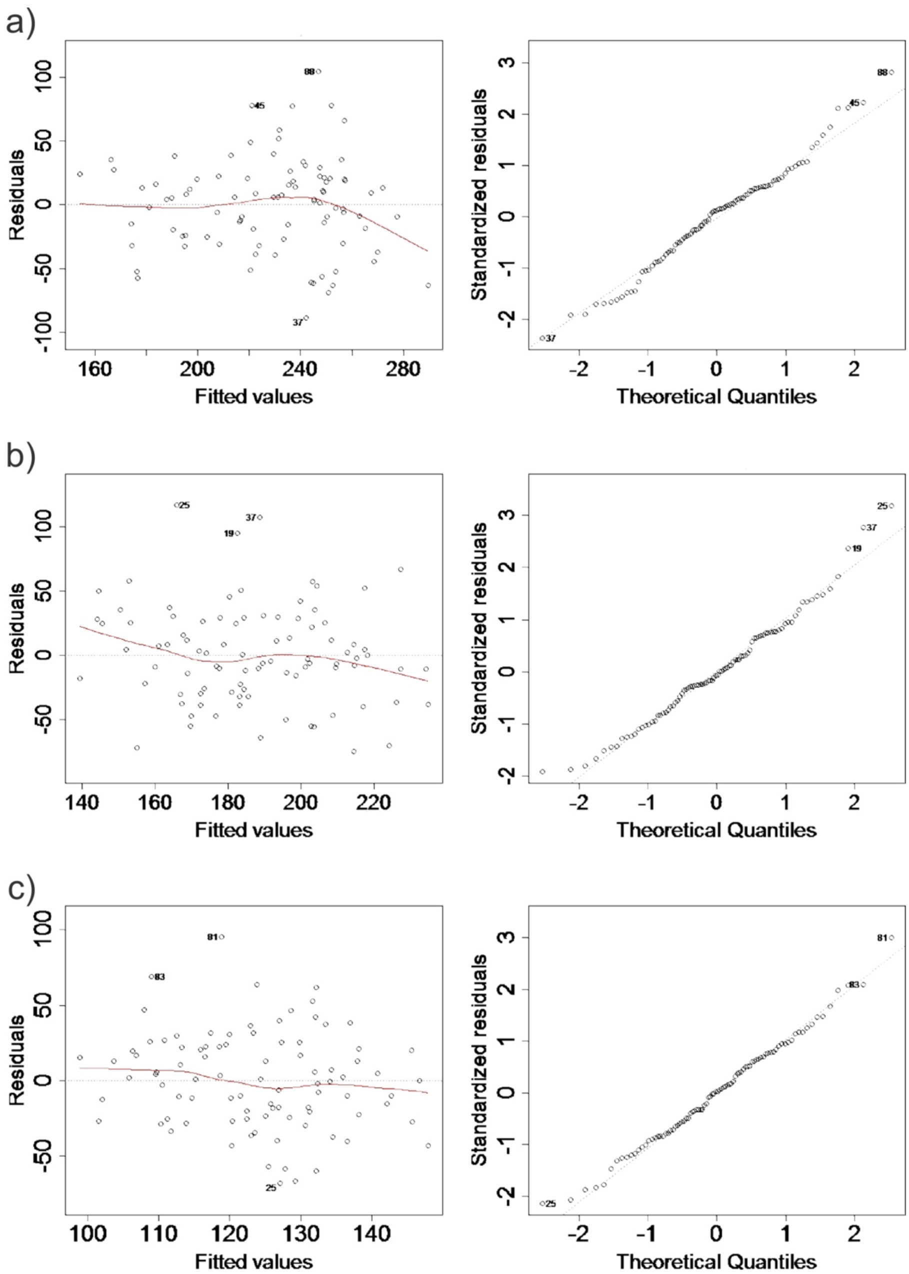

Figure 5. The same as in Fig. 4, but for winter.

Table 1. List of models from the CMIP5 dataset used in this study.

Acronym	Model	Resolutions
ACCESS	ACCESS1.0	1.3° x 1.9°
CCSM	CCSM4	0.9° x 1.3°
CNRM	CNRM-CM5	1.4° x 1.4°
CSIRO	CSIRO-Mk3-6-0	1.9° x 1.9°
EC-EARTH	EC-EARTH	1.1° x 1.1°
HadGEM-ES	HadGEM2-ES	1.3° x 1.9°
INM	INMCM4	1.5° x 2.0°
MPI	MPI-ESM-LR	1.9° x 1.9°
MRI	MRI-CGCM3	1.1° x 1.1°
NorESM	NorESM1-M	1.9° x 2.5°

Table 2. Proportion and cumulative proportion of variance for the indicated regions. Left column for summer and right column for winter.

	SUMMER										WINTER									
	AMZ										AMZ									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Proportion of variance	0.24	0.13	0.12	0.11	0.09	0.08	0.07	0.06	0.06	0.04	0.71	0.07	0.05	0.04	0.03	0.03	0.02	0.02	0.02	0.01
Cumulative proportion	0.24	0.37	0.49	0.60	0.69	0.77	0.84	0.90	0.96	1.00	0.71	0.78	0.83	0.87	0.90	0.93	0.95	0.97	0.99	1.00
	NEB										NEB									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Proportion of variance	0.32	0.13	0.10	0.09	0.08	0.08	0.06	0.07	0.04	0.03	0.54	0.10	0.08	0.07	0.06	0.04	0.04	0.03	0.02	0.02
Cumulative proportion	0.32	0.45	0.55	0.64	0.72	0.80	0.86	0.93	0.97	1.00	0.54	0.64	0.72	0.79	0.85	0.89	0.93	0.96	0.98	1.00
	LPB										LPB									
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Proportion of variance	0.16	0.15	0.14	0.10	0.11	0.09	0.07	0.06	0.06	0.06	0.16	0.15	0.12	0.11	0.10	0.09	0.08	0.07	0.06	0.06
Cumulative proportion	0.16	0.31	0.45	0.55	0.66	0.75	0.82	0.88	0.94	1.00	0.16	0.31	0.43	0.54	0.64	0.73	0.81	0.88	0.94	1.00

Table 3. p-value and Pearson's correlation coefficient at the level of significance of 5% between the ANNs and observed data from the CRU in all regions under study.

	p-value			Correlation coefficient		
	AMZ	NEB	LPB	AMZ	NEB	LPB
Summer	$1.60e^{-07*}$	0.08	0.01^{*}	0.61	0.38	0.18
Winter	$5.28e^{-10*}$	$1.02e^{-10*}$	$1.20e^{-06*}$	0.77	0.69	0.49

∗Significance 5%.

Table 4. p-value and Pearson's correlation coefficient at the level of significance of 5% between the MLR by PCs and observed data from the CRU in all regions under study.

	p-value			Correlation coefficient		
	AMZ	NEB	LPB	AMZ	NEB	LPB
Summer	$1.35e^{-10*}$	$2.69e^{-04*}$	0.06	0.52	0.27	0.20
Winter	$1.44e^{-18*}$	$9.56e^{-14*}$	0.00^{*}	0.62	0.60	0.33

∗Significance 5%.

Table 5. Change in monthly precipitation in terms of an increase or decrease by the end of this century (2071-2100) in the scenarios RCP 8.5 and 2.6, in relation to the reference period 1971-1999 (observation), in mm.month^{-1} and percentage.

		RCP 8.5		RCP 2.6	
		ANN(mm/%)	MLR (mm/%)	ANN(mm/%)	MLR (mm/%)
AMZ	Summer	20.0 / 14.1	23.1 /16.5	18.8 / 13.3	22.4 / 15.8
	Winter	-9.3 / -12.2	-9.9 / -13.9	-0.5 / -0.7	-3.1 / -4.9
NEB	Summer	55.2 / 36.2	47.1 / 30.9	48.0 / 33.1	40.0 / 27.5
	Winter	-6.6 / -42.7	-6.9 / -44.5	-1.81 / -9.41	-2.06 / -10.7
LPB	Summer	7.26 / 5.63	5.7 / 4.42	5.56 / 4.4	4.01 / 3.15
	Winter	-2.79 / -4.17	-3.67 / -5.48	-3.09 / -4.63	-3.09 / -4.56