AGU FALL MEETING
San Francisco | 15–19 December 2014

INRS Université d'avant-garde

Downscaling using Probabilistic Gaussian Copula Regression model

Mohamed Ali Ben Alaya*[1], Fateh Chebana[1] & Taha Ouarda[2]
[1] *INRS-ETE, University of Quebec,* [2] *Masdar Institute of science and technology*

1. Introduction

Context: Atmosphere–ocean general circulation models (AOGCMs) are useful to simulate large-scale climate evolutions. However, AOGCM data resolution is too coarse for regional and local climate studies. Downscaling techniques have been developed to refine AOGCM data and provide information at more relevant scales. Among a wide range of available approaches, regression-based methods are commonly used for downscaling AOGCM data.

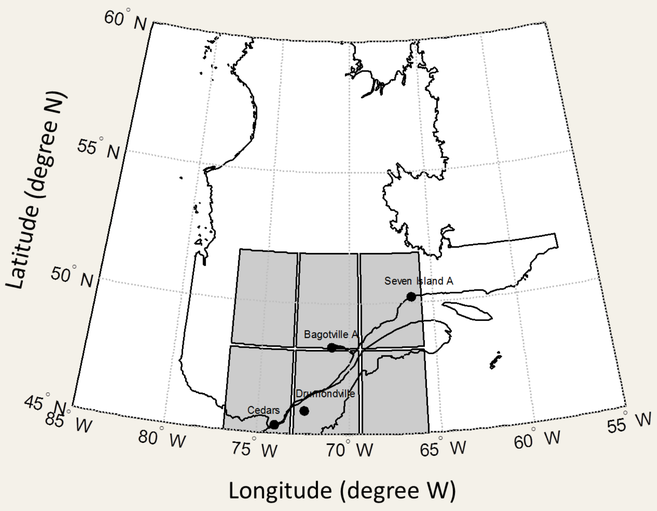
Motivation: When several variables are considered at multiple sites, regression models are employed to reproduce the observed climate characteristics at small scale, such as the temporal variability and the relationship between sites and variables.

- Limitations of traditional regression-based approaches:
  - The underestimation of the temporal variability and the poor representation of extreme events.
  - The assumption of normality of data.
  - The inconsistency between downscaled and observed relationships between sites and variables.

Objective: Introducing a Probabilistic Gaussian Copula Regression (PGCR) model to address the limitations of traditional regression-based approaches in a downscaling perspective.

2. Data series and study area

The study area is located in Quebec (Canada), in latitudes between 45° N and 60° N and longitudes between 65° W and 75° W.



- Predictands series Y: daily Tmax , Tmin and precipitation at 4 stations.
- Predictors X: 6 CGCM3 grid, for each grid point, 25 NCEP predictors are provided (150 predictors). A PCA is performed and the first 40 components are retained as predictors.
- All data cover the period between Jan 1st 1961 and 31 Dec 31 2000.

3. Methodology

Probabilistic Regression

The distribution of each predictand at the observed sites is represented by an appropriate probability density function (PDF), and then a regression model with outputs are parameters of the PDF is employed.

Example: a conditional normally distributed response would have two outputs, one for the conditional mean and one for the conditional variance.

Gaussian Copula

A copula is a multivariate distribution whose marginals are uniformly distributed on the interval [0,1]. A Gaussian copula C is defined as :

$$C(w;C) = \Phi_q\left\{\Phi^{-1}(w_1),\ldots,\Phi^{-1}(w_q);C\right\}$$

where Φ is the standard normal cumulative distribution function and Φ_q is the cumulative distribution function for a multivariate normal vector $w = (w_1,\ldots,w_q)$ having zero mean and covariance matrix C.

Probabilistic Gaussian Copula Regression (PGCR)

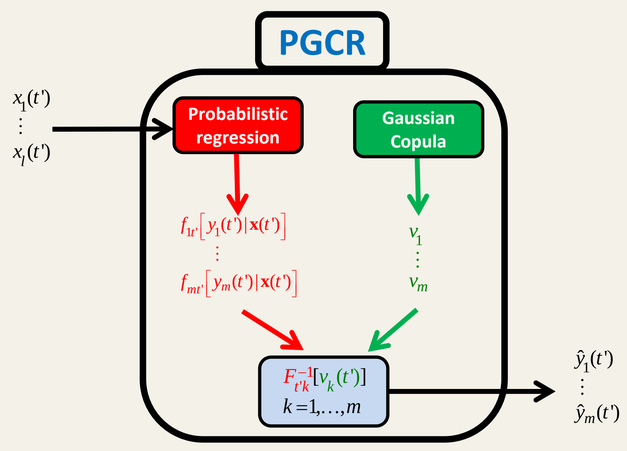
PGCR

Probabilistic Regression + Gaussian Copula

Conditional mean & variance — Dependency

- Multisite
- Miltivariable
- Non constant variance
- Probability distribution (discrete & continuous)

The proposed PGCR model relies on a probabilistic regression framework to specify the marginal distribution for each downscaled variable at a given day through AOGCM predictors, and handles multivariate dependence between sites and variables using a Gaussian copula.



- $x(t') = (x_1(t'),\ldots,x_I(t'))$: predictors values on a day t' from the validation period.
- m: number of predictand variables.
- $v = [v_1,\ldots,v_m]$: generated uniform random variables using the gaussian copula.
- $f_{tk}[y_k(t)\,|\,x(t)]$: conditional PDF of the k^{th} predictands on a day t'.
- $F_{tk}(y_k(t))$: Conditional cumulative distribution function of the k^{th} predictands on a day t'.

4. Model Calibration & Testing

❖ Calibration

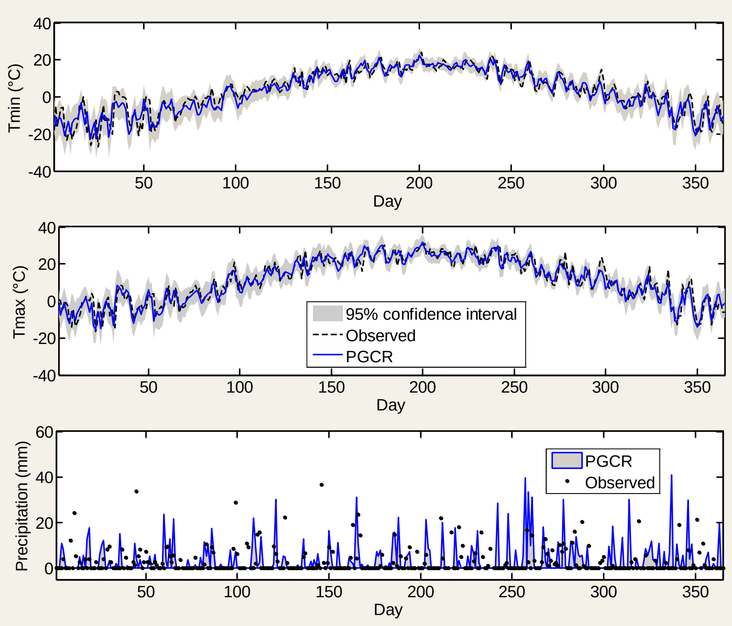
- Data from 01-01-1961 to 31-12-1990.
- Conditional distribution choices for the probabilistic regression model:

- Tmax and Tmin: Normal distributions.
- Precipitation occurrences (Poc): Bernoulli distribution.
- Precipitation amounts (Pam): Gamma distribution.

❖ Validation

- Data from 01-01-1991 to 31-12-2000.
- Statistical criteria: RMSE, ME and the difference between observed and modeled variances D.
- Scatter plot of cross-correlations.
- Comparaison with Multivariate Multiple Linear Regression (MMLR), and Multivariate Multisite Statistical Downscaling Model (MMSDM).

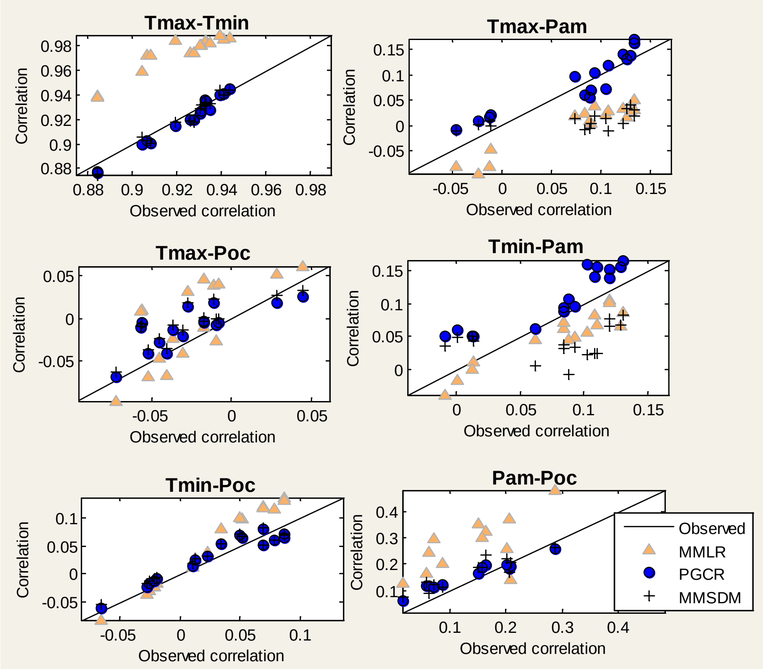
5. Results



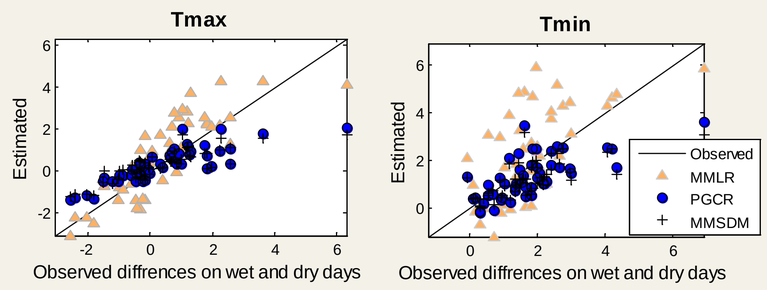
Example of PGCR results at cedars station during 1991.

Quality assessment of the estimated series for GCQR, MMLR and MMSDM during the validation period (1991–2000) for the four weather stations. Criteria are ME, RMSE, and differences between observed and modeled variance D. For PGCR and MMSDM models criteria were calculated from the conditional mean.

		ME			RMSE			*D*		
Station	Model	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)
Cedars	PGCR	0.55	0.22	**−0.14**	3.28	3.70	**6.35**	**−1.34**	**−0.85**	**−1.16**
	MMLR	0.55	0.22	2.59	3.28	3.70	7.02	8.81	9.21	45.49
	MMSDM	0.55	0.22	1.97	3.30	3.71	6.48	−1.84	−2.83	17.92
Drummondville	PGCR	**0.46**	0.49	**0.48**	3.31	4.08	**5.42**	**−3.83**	6.31	**10.69**
	MMLR	0.47	0.49	2.44	3.31	4.08	6.14	7.22	18.15	34.53
	MMSDM	0.47	0.49	0.97	3.32	4.10	5.57	−4.10	**3.47**	11.27
Seven Islands	PGCR	0.20	**−0.52**	**−0.67**	3.18	3.59	**5.57**	**5.51**	**2.12**	**−10.02**
	MMLR	0.19	−0.53	2.11	3.17	**3.58**	5.99	16.36	13.42	33.31
	MMSDM	0.19	−0.53	0.72	3.17	3.60	5.59	6.5	2.34	13.43
Bagotville	PGCR	−0.05	0.14	**0.87**	3.53	3.85	**6.13**	**1.12**	**−0.28**	**24.40**
	MMLR	−0.05	0.14	2.37	3.53	3.84	6.72	15.42	12.76	41.51
	MMSDM	−0.05	0.14	1.11	3.53	3.85	6.24	3.48	−1.77	28.16



Scatter plot of observed and modeled cross-predictand correlations. Correlation values of PGCR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.



Observed versus modeled differences of daily temperatures on wet days and dry days.

❖Results indicate the superiority of the proposed PGCR over both the multivariate multiple linear regression model and the multivariate multisite statistical downscaling model in term of the three statistical criteria.

❖In terms of reproducing spatial and inter-variable properties, both PGCR and MMSDM models provide interesting results.

6. Conclusion

❖A PGCR model is proposed for the downscaling of AOGCM predictors to multiple predictands at multi-sites simultaneously and to preserve relationships between sites and variables.

❖PGCR offers a simple way to reproduce relationships of multiple variables as well as to introduce exogenous forcing covariates.

❖The modular structure of the proposed PGCR is mathematically rich, making it a valuable tool in hydrometeorology and climate research analyses where often non-normally distributed random variables, like precipitation, wind speed, cloud cover, humidity, are involved.

Contact Information

Mohamed Ali Ben Alaya, PhD student
490 rue de la couronne
G1K 9A9, Québec, Canada
Tel: 418 654 2430#4468
Email: mohammed_ali.ben_alaya@ete.inrs.ca

Downscaling using Probabilistic Gaussian Copula Regression model

Mohamed Ali Ben Alaya*[1], Fateh Chebana[1] & Taha Ouarda[2]

[1] *INRS-ETE, University of Quebec,* [2] *Masdar Institute of science and technology*

December 2014

1. Introduction

Context: Atmosphere–ocean general circulation models (AOGCMs) are useful to simulate large-scale climate evolutions. However, AOGCM data resolution is too coarse for regional and local climate studies. Downscaling techniques have been developed to refine AOGCM data and provide information at more relevant scales. Among a wide range of available approaches, regression-based methods are commonly used for downscaling AOGCM data.

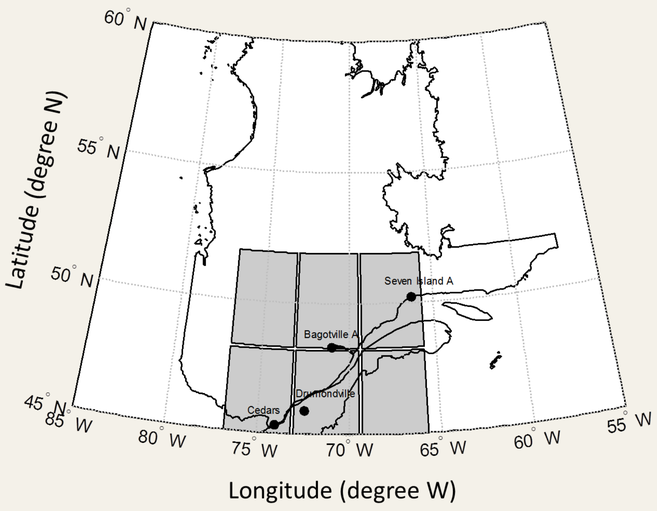
Motivation: When several variables are considered at multiple sites, regression models are employed to reproduce the observed climate characteristics at small scale, such as the temporal variability and the relationship between sites and variables.

- Limitations of traditional regression-based approaches:
  - The underestimation of the temporal variability and the poor representation of extreme events.
  - The assumption of normality of data.
  - The inconsistency between downscaled and observed relationships between sites and variables.

Objective: Introducing a Probabilistic Gaussian Copula Regression (PGCR) model to address the limitations of traditional regression-based approaches in a downscaling perspective.

2. Data series and study area

The study area is located in Quebec (Canada), in latitudes between 45° N and 60° N and longitudes between 65° W and 75° W.



- Predictands series Y: daily Tmax , Tmin and precipitation at 4 stations.
- Predictors X: 6 CGCM3 grid, for each grid point, 25 NCEP predictors are provided (150 predictors). A PCA is performed and the first 40 components are retained as predictors.
- All data cover the period between Jan 1st 1961 and 31 Dec 31 2000.

3. Methodology

Probabilistic Regression

The distribution of each predictand at the observed sites is represented by an appropriate probability density function (PDF), and then a regression model with outputs are parameters of the PDF is employed.

Example: a conditional normally distributed response would have two outputs, one for the conditional mean and one for the conditional variance.

Gaussian Copula

A copula is a multivariate distribution whose marginals are uniformly distributed on the interval [0,1]. A Gaussian copula C is defined as :

$$C(w;C) = \Phi_q\left\{\Phi^{-1}(w_1),\ldots,\Phi^{-1}(w_q);C\right\}$$

where Φ is the standard normal cumulative distribution function and Φ_q is the cumulative distribution function for a multivariate normal vector $w = (w_1,\ldots,w_q)$ having zero mean and covariance matrix C.

Probabilistic Gaussian Copula Regression (PGCR)

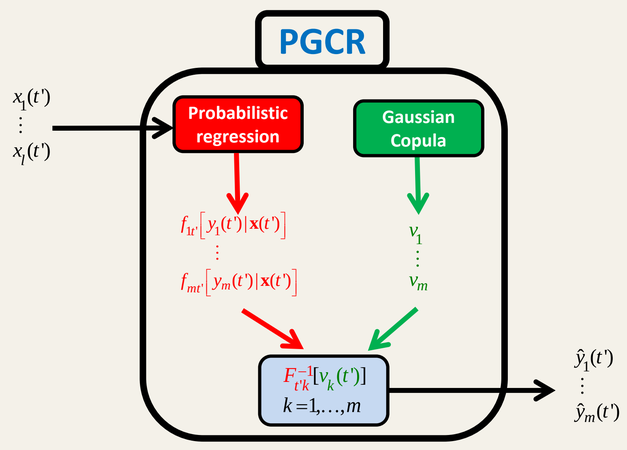
PGCR

Probabilistic Regression + **Gaussian Copula**

Conditional mean & variance — Dependency

- Multisite
- Miltivariable
- Non constant variance
- Probability distribution (discrete & continuous)

The proposed PGCR model relies on a probabilistic regression framework to specify the marginal distribution for each downscaled variable at a given day through AOGCM predictors, and handles multivariate dependence between sites and variables using a Gaussian copula.



- $x(t') = (x_1(t'),\ldots,x_I(t'))$: predictors values on a day t' from the validation period.
- m: number of predictand variables.
- $v = [v_1,\ldots,v_m]$: generated uniform random variables using the gaussian copula.
- $f_{tk}[y_k(t)\,|\,x(t)]$: conditional PDF of the k^{th} predictands on a day t'.
- $F_{tk}(y_k(t))$: Conditional cumulative distribution function of the k^{th} predictands on a day t'.

4. Model Calibration & Testing

❖ Calibration

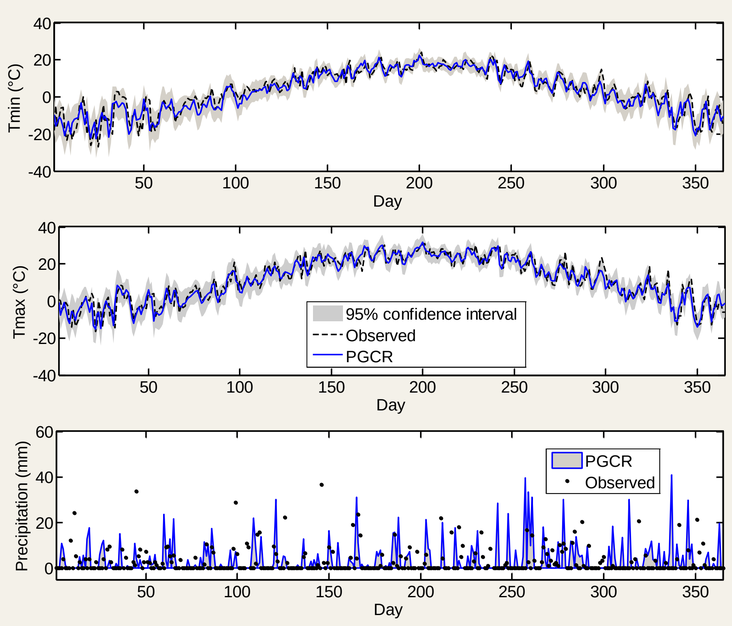
- Data from 01-01-1961 to 31-12-1990.
- Conditional distribution choices for the probabilistic regression model:

- Tmax and Tmin: Normal distributions.
- Precipitation occurrences (Poc): Bernoulli distribution.
- Precipitation amounts (Pam): Gamma distribution.

❖ Validation

- Data from 01-01-1991 to 31-12-2000.
- Statistical criteria: RMSE, ME and the difference between observed and modeled variances D.
- Scatter plot of cross-correlations.
- Comparaison with Multivariate Multiple Linear Regression (MMLR), and Multivariate Multisite Statistical Downscaling Model (MMSDM).

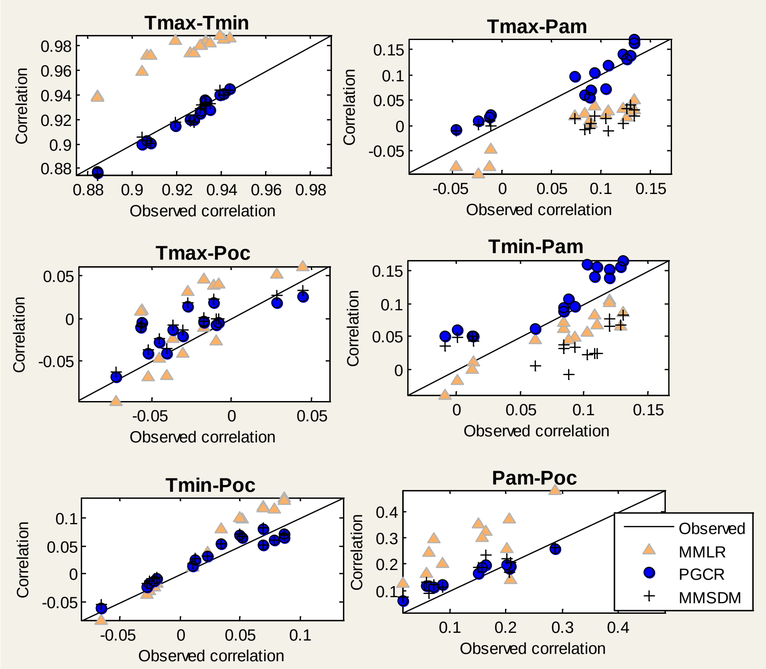
5. Results



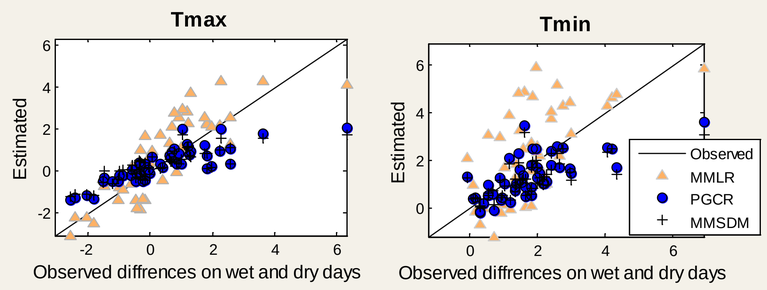
Example of PGCR results at cedars station during 1991.

Quality assessment of the estimated series for GCQR, MMLR and MMSDM during the validation period (1991–2000) for the four weather stations. Criteria are ME, RMSE, and differences between observed and modeled variance D. For PGCR and MMSDM models criteria were calculated from the conditional mean.

		ME			RMSE			*D*		
Station	Model	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)
Cedars	PGCR	0.55	0.22	**−0.14**	3.28	3.70	**6.35**	**−1.34**	**−0.85**	**−1.16**
	MMLR	0.55	0.22	2.59	3.28	3.70	7.02	8.81	9.21	45.49
	MMSDM	0.55	0.22	1.97	3.30	3.71	6.48	−1.84	−2.83	17.92
Drummondville	PGCR	**0.46**	0.49	**0.48**	3.31	4.08	**5.42**	**−3.83**	6.31	**10.69**
	MMLR	0.47	0.49	2.44	3.31	4.08	6.14	7.22	18.15	34.53
	MMSDM	0.47	0.49	0.97	3.32	4.10	5.57	−4.10	**3.47**	11.27
Seven Islands	PGCR	0.20	**−0.52**	**−0.67**	3.18	3.59	**5.57**	**5.51**	**2.12**	**−10.02**
	MMLR	0.19	−0.53	2.11	3.17	**3.58**	5.99	16.36	13.42	33.31
	MMSDM	0.19	−0.53	0.72	3.17	3.60	5.59	6.5	2.34	13.43
Bagotville	PGCR	−0.05	0.14	**0.87**	3.53	3.85	**6.13**	**1.12**	**−0.28**	**24.40**
	MMLR	−0.05	0.14	2.37	3.53	3.84	6.72	15.42	12.76	41.51
	MMSDM	−0.05	0.14	1.11	3.53	3.85	6.24	3.48	−1.77	28.16



Scatter plot of observed and modeled cross-predictand correlations. Correlation values of PGCR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.



Observed versus modeled differences of daily temperatures on wet days and dry days.

❖Results indicate the superiority of the proposed PGCR over both the multivariate multiple linear regression model and the multivariate multisite statistical downscaling model in term of the three statistical criteria.

❖In terms of reproducing spatial and inter-variable properties, both PGCR and MMSDM models provide interesting results.

6. Conclusion

❖A PGCR model is proposed for the downscaling of AOGCM predictors to multiple predictands at multi-sites simultaneously and to preserve relationships between sites and variables.

❖PGCR offers a simple way to reproduce relationships of multiple variables as well as to introduce exogenous forcing covariates.

❖The modular structure of the proposed PGCR is mathematically rich, making it a valuable tool in hydrometeorology and climate research analyses where often non-normally distributed random variables, like precipitation, wind speed, cloud cover, humidity, are involved.

Contact Information

Mohamed Ali Ben Alaya, PhD student
490 rue de la couronne
G1K 9A9, Québec, Canada
Tel: 418 654 2430#4468
Email: mohammed_ali.ben_alaya@ete.inrs.ca