1
2
3
4
5
6 **Water Resources Climate Change Projections Using Supervised**
7 **Nonlinear and Multivariate Soft Computing Techniques**
8
9

10
11 **Ali Sarhadi^{a,1}, Donald H. Burn^a, Fiona Johnson^b, Raj Mehrotra^b, Ashish Sharma^b**
12
13
14
15
16
17

18 ^a Department of Civil and Environmental Engineering, University of Waterloo, Waterloo,

19 Ontario, Canada N2L-3G1

20 ^b School of Civil and Environmental Engineering, University of New South Wales,
21 Kensington, New South Wales 2052, Australia
22
23
24
25
26
27
28
29
30
31

¹ - Corresponding Author: Sarhadi A., Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Ontario, N2L-3G1, Canada (asarhadi@uwaterloo.ca)

Abstract

Accurate projection of global warming on the probabilistic behavior of hydro-climate variables is one of the main challenges in climate change impact assessment studies. Due to the complexity of climate-associated processes, different sources of uncertainty influence the projected behavior of hydro-climate variables in regression-based statistical downscaling procedures. The current study presents a comprehensive methodology to improve the predictive power of the procedure to provide improved projections. It does this by minimizing the uncertainty sources arising from the high-dimensionality of atmospheric predictors, the complex and nonlinear relationships between hydro-climate predictands and atmospheric predictors, as well as the biases that exist in climate model simulations. To address the impact of the high dimensional feature spaces, a supervised nonlinear dimensionality reduction algorithm is presented that is able to capture the nonlinear variability among projectors through extracting a sequence of principal components that have maximal dependency with the target hydro-climate variables. Two soft-computing nonlinear machine-learning methods, Support Vector Regression (SVR) and Relevance Vector Machine (RVM), are engaged to capture the nonlinear relationships between predictand and atmospheric predictors. To correct the spatial and temporal biases over multiple time scales in the GCM predictands, the Multivariate Recursive Nesting Bias Correction (MRNBC) approach is used. The results demonstrate that this combined approach significantly improves the downscaling procedure in terms of precipitation projection.

Keywords: Statistical downscaling, dimensionality reduction, machine-learning, multivariate bias correction.

1. Introduction

It is now broadly accepted that global warming is impacting hydrological and climatological processes on regional and local scales. These changes are expected to increase extreme hydrological events and threaten water resources in different parts of the world. Therefore, the assessment of the climate change impacts on the availability of surface water resources is of particular interest to water resources managers and decision makers for mitigating the adverse impacts of global warming.

Future climate change information is derived from simulated large-scale atmospheric processes developed based on General Circulation Models (GCMs). GCMs simulate climate at coarse spatial scales, and are unable to provide information that can be directly used at the finer scales of interest to hydrologists for assessing how possible climate-change impacts on surface water availability may affect water supply (Bennett et al., 2012; Dingbao Wang, 2013). This inadequacy has been the reason for developing dynamical and statistical downscaling techniques to transfer large-scale global atmospheric variables (provided by GCMs) to regional and local hydro-climate information for use in climate change impact studies. One option for this is dynamical downscaling approaches, which are based on obtaining finer information from Regional Climate Models (RCMs) driven by boundary conditions simulated using GCMs (Najafi and Moradkhani, 2015). The limitation of these approaches is that they require expensive and complicated computations, and use biased lateral boundary inputs as the basis of their

simulations (Rocheta et al., 2014a), inputs that cannot be easily bias corrected for use. On the other hand, most commonly used and popular regression-based statistical downscaling approaches are based on empirical and quantitative relationships developed between a local hydro-climate variable and large-scale atmospheric predictors developed by reanalysis data and GCMs. The regression-based statistical downscaling is carried out in two main steps: i) deriving statistical relationships from historical climate information and hydro-climate variables of interest (developing a statistical model step), and ii) using these models to project hydro-climate variables in the future, relying on the assumption that current empirical models are applicable to GCM simulations of the future (a projection step).

Bias correction has been shown to improve the quality of GCMs for use in projecting hydro-climate variables under different climate change scenarios of the future (Mehrotra and Sharma, 2012; Ojha et al., 2013). Regarding the projection step of the statistical downscaling, the accuracy of climate change simulations is influenced by the similarity in the relationship between actual atmospheric variables and observed rainfall, as compared to simulated variables and the presumed projected rainfall. This similarity is expected to be influenced by the biases that characterise the raw GCM fields. Therefore, in the statistical downscaling processes an initial post-processing correction must be carried out on GCM outputs representing the current climate, based on the statistical characteristics of observations, to remove the difference between observed and simulated large-scale atmospheric variables. The bias correction model over a historical time period is assumed to be the same in the future, and can thus be employed on future GCM simulations (Johnson and Sharma, 2015). In addition, anomalous atmospheric circulation

patterns influence the hydrological cycle and large-scale atmospheric variables. Interannual and interdecadal variability in the large-scale climate modes are often not well represented in GCM simulations (Rocheta et al., 2014b), resulting in uncertainty and biases in projections of hydro-climate variables relating to the future. Thus, raw GCMs must also be corrected to capture the effect of low frequency variability of teleconnections on large-scale atmospheric variables (Mehrotra and Sharma, 2012).

It is therefore critical to identify the nature of these biases and develop methods to address these sources of uncertainty. Several bias correction approaches have been developed to quantify the difference between observed (or reanalysis) data and large-scale GCM-simulated variables and form the basis on which to correct biases in both current and future atmospheric GCM simulations. Commonly used bias correction procedures can be classified into two main categories. The first relies on delta change and scaling approaches, including quantile mapping, scaling, correction factor, and transfer functions (a detailed review of the various methods can be found in Johnson and Sharma (2012) and Fowler et al. (2007). All the methods in this category can be applied for post-processing either on GCM variables or outputs of downscaling models. Their main drawback is that they only take into account biases in the distribution of GCM simulations rather than biases in the representation of persistence and variability in simulations. Current climate variability is thus assumed to remain the same in the future. The second category involves approaches relying on statistical bias correction. Simple techniques in this category such as Monthly Bias Correction (MBC) (Ojha et al., 2013) correct only systematic biases in the mean and variance of GCM-simulated variables or output of downscaled processes in an independent time scale, ignoring the influence of

regional and global teleconnection signals. However, the impact of teleconnections on hydro-climate variable behaviours in large scales makes it important to properly represent the interannual and interdecadal fluctuation of climate in the raw GCM outputs. To do so, Johnson and Sharma (2012) developed a bias correction methodology by adding lag-1 correlation to the procedure to correct the representation of low frequency variability between GCM simulations and observed data. The approach corrects the distributional and persistence GCM biases from fine to progressively longer time series and is called Nested Bias Correction (NBC). An extension of NBC was proposed by Mehrotra and Sharma (2012) to enhance the representation of variability at multiple time series by reducing biases through repeating the nesting process several times (Recursive Nesting Bias Correction, RNBC method).

One of the criticisms of bias correction is that the statistical corrections do not maintain the physical relationships between different climate variables (Ehret et al., 2012; Haerter et al., 2011; Rocheta et al., 2014a). To overcome this problem Mehrotra and Sharma (2015) developed a bias correction method that can consider multiple variables and correct the cross correlations between them over a range of time scales. The Multivariate Recursive Nesting Bias Correction (MRNBC) extends the previous nesting bias correction approaches (Johnson and Sharma, 2012) and has been shown to be effective at correcting predictors for statistical downscaling leading to improved downscaled simulations. An alternative implementation could include using multiple locations rather than multiple variables to correct spatial as well as temporal dependence in the GCM simulations.

After correcting the biases that characterise raw GCM simulations, there remain a number of challenges in developing statistical downscaling models, due to the complexity of the climate system. Two main difficulties exist: i) identification of the large-scale atmospheric predictors conveying relevant climate change information, and ii) development of the right quantitative functional relationship for capturing the complex nonlinearity between target hydro-climate variables and atmospheric simulated predictors. While the first of these problems is partly due to the high dimensionality of the climate processes that lead to rainfall, the second is due to poor characterisation of the functional form. This paper attempts to address both these limitations as discussed below.

To address the dimensionality problem in statistical downscaling processes, many studies have used conventional unsupervised dimensionality reduction methods, such as PCA, CCA, and clustering (Shashikanth et al., 2014; Tisseuil et al., 2010; Wójcik, 2015), exploring a limited sequence of subspaces from the high dimensional predictors to capture the maximum variability and the covariance structure of data without taking into account the target hydro-climate variable. These purely unsupervised techniques may throw away low variations having high predictive potential for the response variable, or keep high variance explanatory variables that are irrelevant for the task at hand. A few attempts have also used selective dimensionality-reduction methods, which cannot adequately capture the nonlinearity and interaction properties of predictors (Ahmadi et al., 2015; Hammami et al., 2012). A supervised dimensionality reduction method, called “Supervised Principal Component Analysis”, was presented as an efficient alternative by Sarhadi et al. (2015), illustrating significant improvements in the downscaled rainfall field.

169 Due to the complex nonlinear relationship existing between target hydro-climate
170 variables and large-scale atmospheric variables, standard linear methods also fail to
171 capture the nonlinear functional relationship. Therefore, to address the second challenge
172 in developing a statistical modeling step, considerable attention has been paid in the last
173 few years to nonlinear-based soft computing data-driven regression models. Machine-
174 learning methods have gained more popularity for statistical downscaling modeling.
175 Among machine learning methods, Support Vector Regression (SVR) has been widely
176 employed in hydrology for nonlinear stochastic modeling of different hydro-climatic
177 variables (Chen et al., 2012; Chen et al., 2010; Nasser et al., 2013). In recent years,
178 however, a fully probabilistic Bayesian framework of the SVR known as Relevance
179 Vector Machine (RVM) has gained more popularity in regression-based statistical
180 modeling. Ghosh and Mujumdar (2008) compared the results obtained from the SVR and
181 RVM models for projection of streamflow in a statistical downscaling process. They
182 presented the advantages of the RVM over the SVR to improve the model performance.
183 In another attempt, the authors also employed the RVM model with a fuzzy clustering
184 method to downscale GCM outputs for monsoon streamflow projections (Mujumdar and
185 Ghosh, 2008). Joshi et al. (2013) analyzed the performance of two statistical downscaling
186 frameworks to characterize the low-flow regime of three rivers in Eastern Canada. They
187 also mentioned the superiority of the RVM model to the Automated Statistical
188 Downscaling (ASD) method in terms of performance criteria. Other studies such as Bai et
189 al. (2014), Khalil et al. (2005), and Okkan and Inan (2014) have also discussed the power
190 of the RVM model compared with other learning algorithms in capturing the nonlinearity

and improving the performance accuracy of different water resources associated projections.

Overall, reducing different sources of uncertainty is crucial in statistical downscaling as it enhances the quality of the hydro-climate variable projections. Consequently, these accurate projections help hydrological and water resources studies better assess the impact of climate change on the availability and allocation of water resources for various sectors, especially drinking water supply. Therefore, the present study focuses on the existing complex and nonlinear relationship between large-scale atmospheric processes and hydro-climate variables to address the uncertainty arising from GCM simulations, high-dimensionality, and inherent nonlinear transfer functions by introducing new nonlinear and multivariate algorithms in the regression-based statistical downscaling processes. Figure 1 depicts a flowchart of the applied procedures and their relationships in the present study.

The rest of the paper is organized as follows: Section 2 explains the mathematical background of the methodologies used in the regression-based statistical downscaling, namely bias correction of multivariate atmospheric projectors, supervised dimensionality reduction, and nonlinear modeling procedure. Section 3 describes the dataset used for the statistical downscaling in the present study. The study is completed with the presentation of the results and discussion on the findings in Section 4, and the conclusions in Section 5.

2. Methodology

2.1 Bias correction

213 A brief description of the logic of the multivariate bias correction is provided with
 214 readers referred to (Mehrotra and Sharma, 2015) for the full derivation. The MRNBC is
 215 based on a multivariate autoregressive order 1 model (MAR1) (Salas, 1980) which is
 216 used to represent both the observed data and the GCM simulations. The general idea is
 217 that for all time scales of interest the GCM simulations are nested into the observed time
 218 series. In this case the observed data are the chosen atmospheric predictors from the
 219 reanalysis data set as described in Section 3. To achieve the nesting, both time series are
 220 standardized to have zero mean and standard deviation of one. Then the lag one
 221 autocorrelations and lag one and lag zero cross correlations in the GCM simulations can
 222 be corrected to match the observed correlations in time and space. Although this study
 223 uses only an MAR1 model, the possibility of correlations in higher order lags can also be
 224 checked (Molina et al., 2013).

225 Let $\mathbf{X}$ be a $p \times n$ matrix of p predictor variables with n time steps at a single location. $\mathbf{X}^h$
 226 is used to denote the observations and $\mathbf{X}^m$ the GCM variables. For the monthly
 227 correction, the parameters are allowed to vary seasonally so for month i , the data to be
 228 corrected is $\mathbf{X}_i^m$ for all years of data. The data are first standardized to form a periodic
 229 time series $\tilde{\mathbf{X}}_i^m$ and $\tilde{\mathbf{X}}_i^h$ for the observations.

230 In general the MAR1 model describes each data set as:

$$231 \quad \tilde{\mathbf{X}}_i^h = \mathbf{C}\tilde{\mathbf{X}}_{i-1}^h + \mathbf{D}\varepsilon_i \quad \text{and} \quad \tilde{\mathbf{X}}_i^m = \mathbf{E}\tilde{\mathbf{X}}_{i-1}^m + \mathbf{F}\varepsilon_i \quad (1)$$

232 where $\mathbf{C}$ and $\mathbf{D}$, are the lag zero and lag one auto and cross correlations from the
 233 observations, $\mathbf{E}$ and $\mathbf{F}$ are calculated in the same way for the standardized GCM

234 simulations and ϵ_i is a vector of p independent random variates having zero mean and the
 235 identity covariance matrix.

236 To obtain the corrected data for any particular year t the model for correction is:

$$237 \quad \mathbf{X}_{t,i}^m = \mathbf{C}_i \mathbf{X}_{t,i-1}^m + \mathbf{D}_i \mathbf{F}_i^{-1} \tilde{\mathbf{X}}_{t,i}^m - \mathbf{D}_i \mathbf{F}_i^{-1} \mathbf{E}_i \tilde{\mathbf{X}}_{t,i-1}^m \quad (2)$$

238 where $\mathbf{X}_{t,i-1}^m$ is the value in the corrected time series from the previous month in year t .

239 After correction the time series $\mathbf{X}^m$ is rescaled by the observed mean and standard
 240 deviation to give the final corrected time series $\bar{\mathbf{X}}^m$. Details on solving for the matrices
 241 $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$ and $\mathbf{F}$ are provided by Mehrotra and Sharma (2015) based on Srikanthan and
 242 Pegram (2009) and Matalas (1967).

243 Following the monthly corrections, the time series $\bar{\mathbf{X}}^m$ is aggregated to form a seasonal
 244 series and the periodic corrections described above are applied, now indexing over the 4
 245 seasons rather than 12 months to give $\bar{\mathbf{S}}^m$ where $\mathbf{S}$ refers to the seasonal matrix of
 246 simulations which is $p \times n/4$ in size. Finally this time series is aggregated to an annual
 247 time series and the correlations, standard deviation and mean are corrected to form $\bar{\mathbf{A}}^m$
 248 where $\mathbf{A}$ is the matrix of yearly data which is $p \times n/12$.

249 According to Srikanthan and Pegram (2009), the corrections at each time aggregation can
 250 be applied in a single correction step as follows:

$$251 \quad \bar{\mathbf{X}}_{i,s,t}^m = \left(\frac{\bar{\mathbf{X}}_{i,s,t}^m}{\mathbf{X}_{i,s,t}^m} \right) \left(\frac{\bar{\mathbf{S}}_{s,t}^m}{\mathbf{S}_{s,t}^m} \right) \left(\frac{\bar{\mathbf{A}}_t^m}{\mathbf{A}_t^m} \right) \mathbf{X}_{i,s,t}^m \quad (3)$$

252 There are some further details in the bias correction that ensure optimum results. A three
 253 step correction procedure is used to correct biases firstly in the mean, then the standard

deviation and finally the correlations. This ensures that the future climate change signal is not affected by the bias correction. The bias corrections are applied three times in the recursive nature suggested by Mehrotra and Sharma (2012) to achieve even better performance in the bias corrected simulations. To correct the future GCM projections, the statistics from the historical period GCM simulations and the reanalysis data are applied for the corrections (Johnson and Sharma, 2012). This allows the GCM projections to evolve according to the impacts of climate assuming that the biases are stationary over time.

2.2 Dimensionality reduction procedure

Most dimensionality reduction approaches used in statistical downscaling are unsupervised, projecting a sequence of subspaces capturing the maximum variation of the covariance of atmospheric predictors. These methods ignore the dependency existing between the response variable (predictand) and the predictors, leading to reduced accuracy in the projection of future behaviour of hydro-climate variables. However, in downscaling processes, seeking directions along which the dependency between the response variable and explanatory variables (predictors) is maximized is preferable. For this purpose, in the present study, a supervised method called “Supervised PCA” is used to extract principal components having maximal dependency on the response hydro-climate variable.

Supervised PCA is based on the Hilbert-Schmidt Independence Criterion (HSIC), used to measure the dependence between a target variable and atmospheric predictors. Two random variables are considered statistically independent if and only if any continuous

276 function of the two variables is uncorrelated. Let us assume $\mathcal{Z} := \{x_1, y_1), \dots, (x_n, y_n)\} \subseteq$
 277 $\mathcal{X} \times \mathcal{Y}$ to be a series of n independent observations drawn from the joint probability
 278 distribution $P_{\mathcal{X}, \mathcal{Y}}$. $\mathcal{F}$ and $\mathcal{G}$ are also considered as the separable Reproducing Kernel
 279 Hilbert Spaces (RKHS) of real-valued functions from data x and y to $\mathbb{R}$, with universal
 280 kernel $K(.,.)$ and $L(.,.)$, respectively (Barshan et al., 2011). In this case, an empirical
 281 estimate of HSIC is defined as

$$HSIC(\mathcal{Z}, \mathcal{F}, \mathcal{G}) := (n - 1)^2 \text{tr}(KHLH) \quad (4)$$

282 where $H, K, L \in \mathbb{R}^{n \times n}$, $k_{ij} := k(x_i, x_j)$, $L_{ij} := l(y_i, y_j)$, and $H_{ij} := I - n^{-1}ee^T$ (the
 283 centering matrix, ‘ e ’ is a vector for all ones).

284 The objective in Supervised PCA is to maximize the dependence between two random
 285 variables or enhance the estimate value of HSIC, i.e, $\text{tr}(KHLH)$. Suppose for
 286 atmospheric predictors there is a set of n data points $\{X_i\}_{i=1}^n$ with p features, stacked in
 287 the $p \times n$ matrix X . The hydro-climate target variable is also Y in the $l \times n$ matrix. The
 288 goal is finding the manifolds in which the dependency between the projected explanatory
 289 data $U^T X$ and the target variable is maximized. According to the empirical HSIC, the
 290 objective function can be formulated as

$$\text{tr}(HKHL) = \text{tr}(U^T XHLHX^T U) \quad (5)$$

291 The above formula can be written as an optimization problem:

$$\begin{aligned} & \underset{U}{\operatorname{argmax}} \quad \text{tr}(U^T XHLHX^T U) \\ & \text{subject to} \quad U^T U = I \end{aligned} \quad (6)$$

292 The optimization problem has a closed form solution. If the symmetric and real matrix
 293 denotes $Q = XHLHX^T$, the optimal solution will be $U = [u_1, u_i, \dots, u_d]$, where u_i is the
 294 corresponding eigenvector to the i -th largest eigenvalue.

295 However, the relationship between high dimensional large-scale atmospheric predictors
 296 and hydro-climate variables is not linearly separable. Thus, nonlinear transformations are
 297 required in the learning dimensionality reduction algorithm to capture nonlinear
 298 variability. One efficient method that can address the nonlinear variability is to use
 299 kernels to capture the similarity measure between two data points. By using a nonlinear
 300 mapping function, the predictors can be mapped from $x \rightarrow \Phi(x)$. Thus, the nonlinear
 301 kernel for a feature matrix will be $K = \Phi(x)^T \Phi(x)$. Accordingly, Supervised PCA is
 302 formulated as

$$\begin{aligned} \underset{U}{\operatorname{argmax}} \quad & \operatorname{tr}(U^T \Phi(x) H L H \Phi(X)^T U) \\ \text{subject to} \quad & U^T U = I \end{aligned} \quad (7)$$

303 The solution can be found by using Singular Value Decomposition (SVD)
 304 on $\Phi(x) H L H \Phi(X)^T$. It should be noted that for a kernel function, $\Phi(X)^T \Phi(X)$ can be
 305 efficiently computed, without computing $\Phi(X)$ explicitly. The transformation matrix U
 306 can be further represented as a linear combination of the projected data points, $U =$
 307 $\Phi(X)\beta$ (Barshan et al., 2011). Thus, the objective function can be rewritten as

$$\begin{aligned} \underset{U}{\operatorname{argmax}} \quad & \operatorname{tr}(\beta^T K H L H K \beta) \\ \text{subject to} \quad & \beta^T K \beta = I \end{aligned} \quad (8)$$

308 Here β can be computed as the top d generalized eigenvectors of $(K H L H K, K)$. This
 309 method is a nonlinear kernelized version of Supervised PCA and is called Supervised

Kernel PCA. More details about the application of Supervised linear and nonlinear PCA in statistical downscaling processes can be found in Sarhadi et al. (2015).

After reducing the dimensionality of the atmospheric predictors, there are still complex and nonlinear relationships between the target hydro-climate variable and new projected explanatory atmospheric predictors extracted using Supervised PCA methods. Therefore, linear based models fail to capture the nonlinear functional relationships. Because of this, in developing a statistical modeling step, nonlinear-based machine-learning methods are used to capture the nonlinear transfer function.

2.3 Nonlinear modeling procedure

To capture the nonlinear relationship between the target hydro-climate variable and the transformed dimension-reduced atmospheric predictors, two nonlinear machine-learning methods, Support Vector Regression (SVR) and Relevance Vector Machine (RVM), are applied in the present study. The machine-learning methods are developed based on the theory of Support Vector Machines (SVMs). According to this theory, statistical learning models are formed based on implementation of structural risk minimization principles. The principles help minimize both empirical risk and the confidence interval of the learning machine to achieve a generalization and an optimum network structure (Vapnik, 1998). The notion of machine-learning is based on mapping the original datasets into higher dimensional, or infinite dimensional, feature space (Hilbert space) using nonlinear transformation functions (kernel functions) to produce simpler relationship from complex unknown relationships between a set of input variables and the target variable. Using this

method minimizes both complexity and prediction error simultaneously (Raghavendra and Deka, 2014).

2.3.1 Support Vector Regression (SVR)

Given an input dataset as predictors and a target variable as a predictand, a function $f(X)$ should be developed to describe the inherent nonlinear relationship between the dataset and predictand. This function can be used later to project the target variable from generated new input data (here principal components extracted by Supervised PCA techniques). The standard nonlinear function in the SVR is formulated as

$$\hat{y} = f(X) = \sum_{i=1}^n w_i K(X_k, X) + b \quad (9)$$

where $K(X_k, X)$ is a kernel function, based on which functions are defined in the training dataset.

However, despite its robustness and efficiency, SVR suffers from several drawbacks, which are addressed in the Sparse Bayesian Learning (SBL) algorithm known as Relevance Vector Machine (RVM) (Tipping, 2001).

2.3.2 Relevance Vector Machine (RVM)

RVM is a fully probabilistic Bayesian learning framework. Unlike the SVR model, the most compelling feature of the RVM algorithm is its ability to utilize dramatically fewer basis functions, which leads to deriving accurate prediction models and avoiding over-fitting problems (Khalil et al., 2006; Tipping, 2001).

A description of the mathematical formulations and the theoretical background of the RVM model are found in Tipping (2000; 2001).

Once the two machine-learning methods have been applied in the modeling step, a cross-validation method is used to evaluate and compare each model's performance. A two-fold cross-validation procedure is used for training and testing. In this procedure, the datasets (including inputs and the target datasets) are randomly split into two non-overlapping subsets (a shuffling is implemented on the data array and then the dataset is split into two training and testing subsets). Each of the nonlinear regression models is then trained based on training subset. The validation of the model is then examined based on testing subset assumed as unobserved data points (validation). A sampling 10-fold cross validation method is used for tuning of the machine-learning parameters (gamma, cost, and error in SVR and kernel width in RVM). The training and testing process is carried out based on a chosen kernel. Gaussian (radial basis), Polynomial, and Laplacian kernel functions are used as nonlinear functions in both SVR and RVM models in this study. To enhance the performance of the models in predicting extreme events (low and high magnitude values) a clustering method is also applied for both regression models (Sarhadi et al., 2015).

To evaluate the goodness of fit of the machine-learning models, the most important performance criteria, including the correlation coefficient (R^2), Nash-Sutcliffe model efficiency (NSE), root mean squared error (RMSE), and mean absolute error (MAE), are used to assess the correspondence of the predicted and observed monthly-based precipitation sequences for the testing dataset.

2.4 Projection for a future climate

After developing an empirical model and adjusting its function parameters from the bias corrected, dimension-reduced large scale atmospheric predictors, the next phase in the statistical downscaling processes is to project these for future climate simulations to the response of interest.

Therefore, to project the impact of climate change on the behavior of the target variable for the future, the simulated GCMs for different climate change scenarios are first corrected using the MRNBC method in the preprocessing step of downscaling. Subsequently, the bias corrected atmospheric predictors are imported into the supervised dimensionality reduction model (Supervised PCA) developed in the second step of the downscaling process to generate principal components. Consequently, the principal components are substituted in the best-selected machine-learning method to downscale the target variable for long term projection. The complete downscaling procedure is illustrated in Figure 1.

3. Case study and Dataset

3.1 Study area

The capital of Iran, Tehran, has been selected as the domain for the present research (Figure 2). This megacity, with its population of more than 14 million, is the largest and the most-populated city in Iran, and in western Asia. The consistent rapid expansion of the city and population has led to several environmental issues, particularly surface water resources shortages. Additionally, the occurrence of severe dry spells and mismanagement of surface water resources in recent years have made this shortage much more complicated, so that water authorities are obliged to ration water. Therefore, it is

essential to have a plan for better management of water resources in relation to the future likelihood of extreme hydrological events, which could lead to dry spells and increase the water shortage in this city. Thus, the present study investigates the impact of future climate change on the availability of surface water resources for different purposes, especially the water supply in this megacity.

3.2 Observations

A monthly precipitation time series recorded at a synoptic meteorological station (Mehrabad) in Tehran city is used as the predictand in the present study. These data, collected by the Iran Meteorological Organization (IRIMO), span 1951 to 2011 (60 years). The annual mean precipitation in this site is 235 mm, which highlights the semi-arid climate. Tehran's climate is largely influenced by its geographical location, surrounded by Alborz Mountains in the north and the central flat plains in the south. Seasonal precipitation is varied in this city and influenced by the monsoon phenomenon, so that summer and winter are characterized as dry seasons, while spring and fall are almost lush, with the main precipitation occurring in these seasons.

National Center for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data are employed as a proxy of observed large-scale atmospheric predictors. Reanalysis NCEP/NCAR datasets created by assimilating climate observations are available from 1948 to present on a regular grid format with a spatial resolution of $2.5^{\circ} \times 2.5^{\circ}$. In the pre-processing phase of the statistical downscaling procedure, NCEP/NCAR data are considered as observations and are employed as a benchmark for correcting systematic biases in the different GCMs. In the next phase, variables from the NCEP/NCAR dataset act as atmospheric predictors for developing the

empirical model, which forms a basis for projecting the hydro-climate predictand of interest (precipitation here) for the following decades. Therefore, the selection of atmospheric predictors from the reanalysis data is important, as they should be able to not only represent climate change signals and have a significant association with the predictand, but also be realistically simulated by GCMs for future climate change scenarios (Hewitson and Crane, 2012). Accordingly, the main identified atmospheric predictors in this study are the precipitation (PRECFLUX), mean (AIR), maximum (TMAX), and minimum air temperature (TMIN) variables, mean sea level pressure (SLP), surface specific (SHUM) and relative humidity (RHUM), and geopotential height (HGT) at three pressure levels (1000, 500, 250 hPa). All these predictors are extracted from nine grid cells surrounding the study site (illustrated in Figure 2), with the time period overlapping the predictand time series (1951-2011) in the monthly-based temporal resolution.

3.3 CMIP5 Climate Models

For the climate change scenarios, the CMIP5 multi-model ensemble, conducted in support of the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5) is used for future simulation of the atmospheric projectors. Two time periods of the multi-model ensemble are applied: i) 1951-2005 corresponding to the recoded observation data (labeled ‘Present’), and ii) 2005-2100 for the future simulations of climate change (labeled ‘Future’). A list of the CMIP5 GCMs used in the present study for simulation of the twentieth and twenty-first century climates is given in Table 1. The horizontal resolution of the models ranges from 0.9° to 3.75° , with a median of 2.1° . Each model is thus re-gridded by bilinear interpolation method to match the resolution of the

NCEP/NCAR reanalysis data ($2.5^{\circ} \times 2.5^{\circ}$) for all of the corresponding present and future components.

The simulations of the CMIP5 models for the present time period are forced with time-evolving observed concentrations of greenhouse gases, aerosols, solar irradiance, and ozone all initiating from an arbitrary point of a quasi-equilibrium control run (Scoccimarro et al., 2013). The future climate change simulations are employed for the twenty-first century through radiative forcing scenarios. The most-recent climate change scenarios, known as ‘Representative Concentration Pathways’ (RCPs), are designed to provide a consistent combination of future population growth, social and economic developments with specific radiative forcing paths (Taylor et al., 2012). Three considered radiative forcing scenarios in the present study are RCP2.6, in which the radiative forcing is estimated to increase to about 3 W/m^2 by year 2100 and decline afterwards, and also 4.5 W/m^2 and 8.5 W/m^2 in the other two scenarios, RCP4.5 and RCP8.5, respectively (Scoccimarro et al., 2013; Taylor et al., 2012).

4. Results and discussion

4.1 Bias correction

To correct the biases in GCMs of the multi-model ensemble, NCEP/NCAR-based atmospheric projectors are used as the bench mark. The bias corrected results are presented in Table 2 where it is clear that the MRNBC has led to substantial improvements in the representation of the statistics of each of the predictors compared to the raw GCM simulations for the current climate. Due to space limitations, results are

provided for a single GCM (bcc-csm1-1) but the performance of the bias correction is similar for all GCMs.

Figures 3, 4 and 5 show the performance of the bias correction across all GCMs, grid cells and, in the case of the monthly level statistics, all 12 months. Again it is clear that the bias correction leads to substantial improvements in all three statistics for the current climate. The observed monthly means (and therefore seasonal and annual means) are almost exactly reproduced. The variance and persistence (standard deviation at the monthly and annual level) also show good improvements although there are some remaining biases compared to the mean corrections, for example the positive bias in all models at all locations for Sea Level Pressure variance. This is because the corrections are applied in the multivariate setting so it is not possible to correct each variable perfectly as well as correcting the inter-variable relationship. Larger reductions in bias could be expected if the corrections were applied separately for each variable but this would not preserve the physical relationship between variables which is important as they form the inputs to the downscaling models.

4.2 Dimensionality reduction and modeling based on reanalysis data

To establish a statistical relationship between atmospheric predictors and the target monthly precipitation, the eight large-scale atmospheric variables (mentioned in the previous section) are extracted from nine grid cells of the NCEP/NCAR reanalysis data surrounding the study site. Thus for explanatory atmospheric variables, a high dimensional input matrix with dimension 72×732 (where 732 represents monthly recorded precipitation observations) is available. The results of the Pearson's and the

Spearman's rank correlation analyses show that all the atmospheric projectors and the target precipitation observations are significantly correlated at 1% significance level.

Due to the approximation of the statistical properties of the neighbor grid cells inducing redundancy and collinearity among explanatory variables, the modeling procedure will give rise to inadequate results in terms of performance accuracy. Therefore, to reduce the dimension of the predictors and improve prediction accuracy in the statistical downscaling, two supervised dimensionality reduction methods (Supervised-PCA and Kernel Supervised-PCA) are employed. Since the process of the dimensionality reduction is a supervised and learning based algorithm, a random two-fold split is performed on the dataset (including the target and projectors), taking 75% for training and 25 % for testing. For both Supervised-PCA and Kernel Supervised-PCA an RBF kernel is used for projection of dataset on the target variable. After estimating the transformation matrix, 10 new features or principal components from 72 explanatory variables are extracted. Using the training subset, two nonlinear machine-learning methods (SVR and RVM) are then fitted to the target variable Y and the dimension-reduced data $Z=10$ to develop statistical downscaling regression models. To ensure consistency between the dimensionality reduction and the downscaling modeling processes and to evaluate the effectiveness of each supervised dimensionality reduction, the two-fold cross validation procedure in the modeling is also simultaneously carried out on the same split data set from the dimensionality reduction step.

In this way, two statistical downscaling models are constructed using the target precipitation Y , which depends on a specific subset of the reduced dimension features, projected by Supervised-PCA and Kernel Supervised PCA separately. Using the same

515 tuned parameters in the training part of the two supervised dimensionality reduction, low
516 dimensional projections of the testing dataset are subsequently retained. The encoded
517 reduced dimension testing data $Z=10$ from the two supervised PCA methods (as the
518 validation projectors) are then imported to the already developed machine-learning
519 models.

520 The results of the performance criteria for the two developed machine-learning models
521 using the encoded reduced dimension testing data (after tuning parameters— C , σ , ε in
522 SVR, and kernel width in RVM—using 10 fold-cross validation) for the best selected
523 kernel functions are presented in Table 3. In the SVR models, RBF kernel-based models
524 act as the best selected regression models for both dimension-reduced input datasets.
525 However, the SVR model developed through the Kernel Supervised PCA method,
526 showing better results in terms of performance criteria, outperforms the Supervised PCA-
527 based SVR model. In addition, reducing the number of support vectors in Kernel based
528 dimensionality reduction model indicates better performance of the kernel-based SVR
529 model in terms of overfitting. Nevertheless, both SVR models suffer from overfitting,
530 using many support vectors relative to the number of observations, leading to the poor
531 predictive performance of the models.

532 Alternatively, using a fully probabilistic Bayesian learning algorithm, the RVM model,
533 which explicitly penalizes excessively complex models, results in a significant reduction
534 in the number of the relevant vectors (specified for RVM) and addressed overfitting
535 problem using both dimensionality reduction results as input (Table 3). It is worth noting
536 that the probabilistic reasoning of the RVM model leads to this superiority of this model
537 in minimizing the possibility of overfitting and computational time as well. As given in

Table 3, the RVM model with input predictors based on Supervised PCA and RBF kernel function reduces the overfitting, which is not achieved by the SVR models. However, this model still cannot mimic the performance criteria obtained by the SVR models. Using a set of kernel based dimensionality reduction projectors as input, the RVM model not only employs the least relevant vectors, but also outperforms all the other models in terms of performance accuracy. Therefore, from the results of the both machine-learning models, it can be concluded that nonlinear transformation of the original atmospheric projectors into a higher dimensional space (using Kernel Supervised PCA) will give rise to capturing complex nonlinear dependence between the target precipitation and the atmospheric projectors, and of course, leading to reducing complexity in machine-learning modeling and minimizing the possibility of overfitting in learning procedure. The observed and predicted precipitation obtained from the best dimensionality reduction method (Kernel Supervised PCA) and the best machine-learning model (RVM) for testing dataset is illustrated in Figure 6b and the scatter plot between them in Figure 6a. The results indicate a good agreement between observed and predicted precipitation time series and also demonstrate the good performance of the combination of the RVM model and the Kernel Supervised PCA method for the modeling section in statistical downscaling. Although the presented model is able to capture extreme minimum and maximum precipitation events, in very rare cases, it cannot completely mimic extreme recorded monthly precipitation events. The reason might be related to the essence of regression based statistical downscaling models that cannot explain entire variance of the downscaled response variable (Ghosh and Mujumdar, 2008; Tripathi et al., 2006).

4.3 Historical precipitation representation in GCMs

In this section, to assess the impact of the bias corrections on each GCM-based input atmospheric projector of the downscaling model, the representations of precipitation downscaled with different ensemble GCM models are extracted for the historical period and compared with observed precipitation during the same time period (1951-2005).

Doing so, the same raw atmospheric projectors used in the modeling step, and the bias-corrected ones derived from the previous section for each GCM model, are used as input to the best dimensionality reduction method from the modeling step (Kernel Supervised PCA). Using the same tuned parameters derived in the modeling step for the Kernel Supervised PCA, low dimensional projectors of the both raw and corrected historical atmospheric projectors are retained. The derived transformed projectors (in the same reduced-dimension as the modeling section, $Z=10$) for CMIP5 GCM models are imported to the best selected machine-learning model (RVM) for probabilistic precipitation projection of the multiple-GCM-models for the historical time period. The results of reproduced precipitations from both bias-corrected and raw GCM models are compared against observed precipitation for the same time period, in the form of the empirical cumulative distributions in Figure 7. As illustrated in Figure 7 (part a), a significant bias exists between observed and projected precipitation for the raw GCM models across the different exceedance probabilities, especially rare high-magnitude observations. This error, related to variability and persistence biases in atmospheric projectors, is significantly reduced for the corrected multiple-GCM-derived precipitation through using the MRNBC bias correction model. As current precipitation variability in the GCMs should be the same as that in the observed data, Figure 7 (Part b) depicts that all the

GCM-derived projected precipitations follow similar distributional behavior and show a good fit with observed precipitation. Thus, the MRNBC technique in the bias correction procedure is able to remove the difference between observed and the raw simulated multiple-GCMs through the addressing variability and biases in serial and cross dependence large-scale atmospheric projectors, thereby enhancing the quality of GCMs. Accordingly, relying on the MRNBC model is expected to improve the performance of precipitation projection under different climate change scenarios for the upcoming decades.

4.4 Future precipitation projections

To project the impact of climate change on the future precipitation behavior, 15 different GCMs of the CMIP5 multi-model ensemble are used under three different forcing emission scenarios. Table 1 shows the availability in the CMIP5 archive of each GCM for each emission scenario. To project future precipitation for each GCM model under climate change scenarios, the same three operational procedures explained in the previous section for the historical dataset (i.e., bias correction, dimensionality reduction, and nonlinear machine-learning based modeling) are implemented on the future simulated atmospheric projectors in two temporal time scales, near-term (2015-2040) and long-term (2015-2100).

To better explain the projection changes and understand how the bias-correction procedure can influence the behavior of projected precipitation, a linear trend is calculated from the annual mean of projected precipitation derived from raw and bias-corrected-multi GCM models under three forcing scenarios for the two temporal scales

(Table 4). According to the results, the precipitation trend is quite variable across the different GCMs, but the majority of the models indicate a steadily decreasing trend in the twenty-first century under each climate change scenario. This is consistent for both the raw and bias-corrected models showing that the bias correction has not affected the average direction of the climate change signal, even though the absolute magnitude of the future projections is different.

Bias correction does however lead to changes in the magnitude of precipitation reduction. For example, before bias correction, the average precipitation change ranges across the different scenarios from $-2.92 \text{ mm}(10\text{yr})^{-1}$ to $-3.59 \text{ mm}(10\text{yr})^{-1}$ and from -0.66 to $-5.08 \text{ mm}(10\text{yr})^{-1}$ for the near-and long-term periods respectively. After employing bias correction on all the GCMs, the precipitation reduction becomes much greater for the near-term, so that average precipitation changes range from $-1.20 \text{ mm}(10\text{yr})^{-1}$ to $-11.30 \text{ mm}(10\text{yr})^{-1}$ and for the long-term it varies from -0.87 to $-3.43 \text{ mm}(10\text{yr})^{-1}$. Interestingly, the RCP2.6 case has the largest reduction in the near-term period, which may be attributed to the availability of more GCMs for this scenario. For the two high emission scenarios the expected trends are still expected to be negative and of the order of around 10 mm reduction per decade to the current trend for recorded precipitation. Slightly smaller reductions in precipitation are expected over the long-term period. Generally the largest reductions are projected in the near-term under all three emission scenarios.

Overall, the results indicate a more negative trend after employing the bias-correction procedure in the statistical downscaling for both near-term (2015-2040) and long-term (2015-2100) periods. More models agree on the direction of climate change signal after applying bias correction. Similar results were found by (Johnson and Sharma, 2015) in

the context of drought modeling where agreement across the GCMs was improved following bias correction with more consistency in the direction of the change in drought frequency. Therefore, it is worth noting that the applied bias-correction model (MRNBC) has potential to improve the results of future projections through addressing biases in persistence, which may represent interannual variability arising from large-scale climate modes, and also biases across the atmospheric projectors. Thus, the bias-correction procedure constitutes an important step in the post-processing analysis of large-scale atmospheric projectors in regression-based statistical downscaling and can play a vital role in reducing the uncertainty of hydro-climate variable projections.

Due to the importance of the seasonal precipitation variation in water resources applications, seasonal behavior changes of projected precipitation are also considered under different climate change scenarios. Figure 8 illustrates the seasonal changes of derived average projected precipitation based on post-processed (bias-corrected) CMIP5 models under three different emission scenarios for near and long-term periods. The mean of observed precipitation is also shown in Figure 8 (as blue marks) for each season to compare the future projected precipitation changes derived from different scenarios to historical recorded precipitation for the study site. The results indicate a decrease of precipitation for winter, compared with the historical time period, under all of the scenarios in both near and long-term periods, as less reduction is seen in the near period. The mean reduction of projected precipitation in winter ranges from -6.10% under RCP2.6 to -17.71% under the same scenario. However, over the long-term period the same reduction pattern is also projected for spring, with variation from -2.63% to -7.84% under different forcing scenarios for both time periods. The changes for summer are

generally small as this is the dry season but the majority of the scenarios still show reductions in projected precipitation. In fall, no significant changes are projected based on the ensemble averages over the short-term period, even precipitation under RCP2.6 tends to increase. Unlike the short-term period, more reduction is seen in the long-term period for all the scenarios, as the fluctuation varies from -0.12% to -3.20%. Although the most precipitation reduction is projected to occur in winter and spring, other characteristics of precipitation (especially rainfall), i.e., distribution and intensity, may potentially be significantly influenced by climate change under different scenarios. Thus, in addition to studying only the amount of precipitation, it is also important that these two factors are considered and taken into account when projecting the availability of surface water resources at the study site.

Overall, according to the results, the precipitation for Tehran city is clearly projected to be influenced by climate change in future decades, and the availability of surface water resources will potentially be severely decreased in the near and long-term. Therefore, water resources authorities and managers need to provide near and long term plans for mitigating adverse consequences arising from water shortages and surface water resources reduction under the impact of climate change in future decades.

5. Conclusions

Projecting the impact of climate change on the probabilistic behavior of hydro-climate variables in fine local scales is highly complicated due to the existing complex and nonlinear relationship between climate-associated processes and the target hydro-climate variables of interest. This complexity results in different sources of uncertainty, which

676 influence the projection accuracy of the global warming impacts on the response variable.
677 To improve the accurate projection of hydro-climate variables, this study has presented a
678 comprehensive methodology for enhancing the predictive power of regression-based
679 statistical downscaling through addressing different sources of uncertainty arising from
680 biases in raw GCMs, high dimensional space atmospheric projectors, and nonlinearity
681 between hydro-climate predictands and atmospheric projectors.

682 The results in the dimensionality-reduction section demonstrate that a kernelized form of
683 supervised dimensionality reduction technique is able to efficiently reduce the impact of
684 the high dimensional space of atmospheric projectors in terms of performance accuracy
685 in statistical downscaling. This improvement is achieved through modeling the nonlinear
686 variability and interdependency among bias-corrected atmospheric projectors by taking
687 into account the dependency between the target precipitation variable and explanatory
688 projectors.

689 Subsequently, the application of nonlinear data-driven machine-learning methods proves
690 the high efficiency of the Bayesian learning algorithm (RVM) in capturing the
691 nonlinearity between dimension-reduced atmospheric projectors and the target variable in
692 the modeling section. The superiority of this model in addressing the complex nonlinear
693 relationships gives rise to minimizing the possibility of overfitting and reducing
694 computational burden in the downscaling modeling as well.

695 Unlike traditional univariate bias correction approaches, the current study demonstrated
696 the usefulness of the Multivariate Recursive Nesting Bias Correction (MRNBC) approach
697 on simultaneously correcting biases and variability in multiple-GCM-derived variables
698 over multiple time scales for regression-based statistical downscaling. Since multiple

projectors in statistical downscaling represent a number of grid cells surrounding a specific study site, it is important to extend the scale-dependent climate model biases to account for the biases in the spatial cross dependence attributes among multi projectors as well. Thus, employing this procedure leads to reducing the uncertainty and improving the future projection accuracy of hydro-climate variables of interest.

It should be noted that the proposed methodology is not restricted to precipitation and can be used for other hydro-climate variables as well. The application of the proposed methodology in the regression-based statistical downscaling in the study site reduces the impact of different sources of uncertainty and results in more accurate climate change impact assessments on hydro-climate variables in future. Using more-accurate projections, decision makers of the capital of Iran will be able to better define long-term and effective proactive strategies to be adopted in response to potential future changes. They will thus be better able to mitigate adverse consequences arising from global warming that might threaten the availability of surface water resources in this megacity. However, whether the method is able to globally outperform other methodologies is a challenge that needs be addressed in future works.

Acknowledgments

We thank two anonymous reviewers whose suggestions helped improve the paper. The authors gratefully acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC). We also acknowledge the CMIP5 climate coupled modelling groups, for producing and making their model output (listed in Table 1 of this paper) available, the U.S. Department of Energy's Program for Climate Model

Diagnosis and Intercomparison (PCMDI), which provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. The CMIP5 model outputs used in the present study are available from http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html. We also thank the Iran Meteorological Organization (IRIMO) for providing rainfall data recorded at the Tehran synoptic station.

References

- Ahmadi, A., Han, D., Kakaei Lafdani, E., Moridi, A., 2015. Input selection for long-lead precipitation prediction using large-scale climate variables: a case study. *Journal of Hydroinformatics*, 17(1): 114. DOI:10.2166/hydro.2014.138
- Bai, Y., Wang, P., Li, C., Xie, J., Wang, Y., 2014. A multi-scale relevance vector regression approach for daily urban water demand forecasting. *Journal of Hydrology*, 517: 236-245.
- Barshan, E., Ghodsi, A., Azimifar, Z., Zolghadri Jahromi, M., 2011. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7): 1357-1371. DOI:10.1016/j.patcog.2010.12.015
- Bennett, K.E., Werner, A.T., Schnorbus, M., 2012. Uncertainties in Hydrologic and Climate Change Impact Analyses in Headwater Basins of British Columbia. *Journal of Climate*, 25(17): 5711-5730. DOI:10.1175/jcli-d-11-00417.1
- Chen, H., Xu, C.-Y., Guo, S., 2012. Comparison and evaluation of multiple GCMs, statistical downscaling and hydrological models in the study of climate change impacts on runoff. *Journal of Hydrology*, 434: 36-45.
- Chen, S.-T., Yu, P.-S., Tang, Y.-H., 2010. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. *Journal of Hydrology*, 385(1-4): 13-22. DOI:10.1016/j.jhydrol.2010.01.021
- Dingbao Wang, X.C., 2013. Constructing Comprehensive Datasets for Understanding Human and Climate Change Impacts on Hydrologic Cycle. *Irrigation & Drainage Systems Engineering*, 02(01). DOI:10.4172/2168-9768.1000106
- Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., Liebert, J., 2012. HESS Opinions" Should we apply bias correction to global and regional climate model data?". *Hydrology and Earth System Sciences Discussions*, 9(4): 5355-5387.
- Fowler, H.J., Blenkinsop, S., Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12): 1547-1578. DOI:10.1002/joc.1556
- Ghosh, S., Mujumdar, P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources*, 31(1): 132-146.
- Haerter, J., Hagemann, S., Moseley, C., Piani, C., 2011. Climate model bias correction and the role of timescales. *Hydrology and Earth System Sciences*, 15(3): 1065-1079.
- Hammami, D., Lee, T.S., Ouarda, T.B.M.J., Lee, J., 2012. Predictor selection for downscaling GCM data with LASSO. *Journal of Geophysical Research: Atmospheres*, 117(D17): n/a-n/a. DOI:10.1029/2012jd017864
- Hewitson, B.C., Crane, R.G., 2012. Large-scale atmospheric controls on local precipitation in tropical mexico *Geophysical Research Letters*, 19(18): 1835-1838.

- Johnson, F., Sharma, A., 2012. A nesting model for bias correction of variability at multiple time scales in general circulation model precipitation simulations. *Water Resources Research*, 48(1). DOI:10.1029/2011wr010464
- Johnson, F., Sharma, A., 2015. What are the impacts of bias correction on future drought projections? *Journal of Hydrology*, 525: 472-485.
- Joshi, D., St-Hilaire, A., Daigle, A., Ouarda, T.B.M.J., 2013. Databased comparison of Sparse Bayesian Learning and Multiple Linear Regression for statistical downscaling of low flow indices. *Journal of Hydrology*, 488: 136-149. DOI:10.1016/j.jhydrol.2013.02.040
- Khalil, A., Almasri, M.N., McKee, M., Kaluarachchi, J.J., 2005. Applicability of statistical learning algorithms in groundwater quality modeling. *Water Resources Research*, 41(5).
- Khalil, A.F., McKee, M., Kemblowski, M., Asefa, T., Bastidas, L., 2006. Multiobjective analysis of chaotic dynamic systems with sparse learning machines. *Advances in Water Resources*, 29(1): 72-88. DOI:10.1016/j.advwatres.2005.05.011
- Matalas, N.C., 1967. Mathematical assessment of synthetic hydrology. *Water Resources Research*, 3(4): 937-945. DOI:10.1029/WR003i004p00937
- Mehrotra, R., Sharma, A., 2012. An improved standardization procedure to remove systematic low frequency variability biases in GCM simulations. *Water Resources Research*, 48(12): n/a-n/a. DOI:10.1029/2012wr012446
- Mehrotra, R., Sharma, A., 2015. Correcting for systematic biases in multiple raw GCM variables across a range of timescales. *Journal of Hydrology*, 520: 214-223. DOI:10.1016/j.jhydrol.2014.11.037
- Molina, J.-L., Pulido-Velázquez, D., García-Aróstegui, J.L., Pulido-Velázquez, M., 2013. Dynamic Bayesian networks as a decision support tool for assessing climate change impacts on highly stressed groundwater systems. *Journal of Hydrology*, 479: 113-129.
- Mujumdar, P., Ghosh, S., 2008. Modeling GCM and scenario uncertainty using a possibilistic approach: Application to the Mahanadi River, India. *Water Resources Research*, 44(6).
- Najafi, M.R., Moradkhani, H., 2015. Multi-Model Ensemble Analysis of Runoff Extremes for Climate Change Impact Assessments. *Journal of Hydrology*(In Press). DOI:<http://dx.doi.org/10.1016/j.jhydrol.2015.03.045>
- Nasser, M., Tavakol-Davani, H., Zahraie, B., 2013. Performance assessment of different data mining methods in statistical downscaling of daily precipitation. *Journal of Hydrology*, 492: 1-14. DOI:10.1016/j.jhydrol.2013.04.017
- Ojha, R., Nagesh Kumar, D., Sharma, A., Mehrotra, R., 2013. Assessing Severe Drought and Wet Events over India in a Future Climate Using a Nested Bias-Correction Approach. *Journal of Hydrologic Engineering*, 18(7): 760-772. DOI:10.1061/(asce)he.1943-5584.0000585
- Okkan, U., Inan, G., 2014. Statistical downscaling of monthly reservoir inflows for Kemer watershed in Turkey: use of machine learning methods, multiple GCMs and emission scenarios. *International Journal of Climatology*.

- Raghavendra, N., Sujay, Deka, P.C., 2014. Support vector machine applications in the field of hydrology: A review. *Applied Soft Computing*, 19: 372-386. DOI:10.1016/j.asoc.2014.02.002
- Rocheta, E., Evans, J., Sharma, A., 2014a. Assessing atmospheric bias correction for dynamical consistency using potential vorticity. *Environmental Research Letters*, 9(12). DOI:10.1088/1748-9326/9/12/124010
- Rocheta, E., Sugiyanto, M., Johnson, F., Evans, J., Sharma, A., 2014b. How well do general circulation models represent lowfrequency rainfall variability? *Water Resources Research*, 50(3): 2108-2123. DOI:10.1002/2012WR013085
- Salas, J.D., 1980. *Applied modeling of hydrologic time series*. Water Resources Publication.
- Sarhadi, A., Burn, H.D., Yang, G., Ghodsi, A., 2015. *Advances in Projection of Climate Change Impacts Using Supervised Nonlinear Dimensionality Reduction Techniques*. *Climate Dynamics* Under review.
- Scoccimarro, E., Gualdi, S., Bellucci, A., Zampieri, M., Navarra, A., 2013. Heavy Precipitation Events in a Warmer Climate: Results from CMIP5 Models. *Journal of Climate*, 26(20): 7902-7911. DOI:10.1175/jcli-d-12-00850.1
- Shashikanth, K. et al., 2014. Comparing statistically downscaled simulations of Indian monsoon at different spatial resolutions. *Journal of Hydrology*, 519: 3163-3177. DOI:10.1016/j.jhydrol.2014.10.042
- Srikanthan, R., Pegram, G.G.S., 2009. A Nested Multisite Daily Rainfall Stochastic Generation Model. *Journal of Hydrology*, 371(1): 142-153. DOI:10.1016/j.jhydrol.2009.03.025
- Taylor, K.E., Stouffer, R.J., Meehl, G.A., 2012. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4): 485-498.
- Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector. *Journal of Machine Learning Research*, 1: 211- 244.
- Tisseuil, C., Vrac, M., Lek, S., Wade, A.J., 2010. Statistical downscaling of river flows. *Journal of Hydrology*, 385(1-4): 279-291. DOI:10.1016/j.jhydrol.2010.02.030
- Tripathi, S., Srinivas, V., Nanjundiah, R.S., 2006. Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, 330(3): 621-640.
- Vapnik, V.N., 1998. *Statistical learning theory*. new yourk. Wiley.
- Wójcik, R., 2015. Reliability of CMIP5 GCM simulations in reproducing atmospheric circulation over Europe and the North Atlantic: a statistical downscaling perspective. *International Journal of Climatology*, 35(5): 714-732. DOI:10.1002/joc.4015

Figure 1. Schematic flowchart of the methodology

Figure 2. Study area and the location of surrounding 9 grid cells

Figure 3 Monthly mean values for each predictor variable for all GCMs and all grid locations for raw and bias corrected GCM simulations. Each plot contains 1620 points representing 15 GCMs, 9 grid locations and 12 months. R^2 values are provided for both cases below the title for each panel.

Figure 4 Monthly standard deviation values for each predictor variable for all GCMs and all grid locations for raw and bias corrected GCM simulations. Each plot contains 1620 points representing 15 GCMs, 9 grid locations and 12 months. R^2 values are provided for both cases below the title for each panel.

Figure 5 Annual standard deviation values for each predictor variable for all GCMs and all grid locations for raw and bias corrected GCM simulations. Each plot contains 135 points representing 15 GCMs and 9 grid locations. R^2 values are provided for both cases below the title for each panel.

Figure 6. Observed and projected monthly precipitation from combination of RVM and Kernel Supervised PCA for testing subset

Figure 7. Empirical cumulative distribution functions of reproduced precipitation from, a) raw-multiple-GCM models, and b) bias-corrected multiple-GCM models for the historical time period (1951-2005) against observed precipitation in the same time period.

Figure 8. Seasonal changes of projected precipitation after bias correction under different emission scenarios relative to the historical (1951-2011) mean for that season. The near-term time periods are illustrated in green color, as the long-term ones are in blue.

Table1. List of CMIP5 multi-models applied in the present study

Number	Model	Modeling center	Spatial Resolution	Data Duration			
				Historical	RCP2.6	RCP4.5	RCP8.5
1	BCC-CSM1.1	BCC (China)	$1^{\circ} \times 1.33^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
2	CanESM2	CCCMA (Canada)	$2.8^{\circ} \times 2.8^{\circ}$	1948-2014	-	2006-2100	-
3	CCSM4	NCAR (USA)	$0.9^{\circ} \times 1.25^{\circ}$	1948-2014	2006-2100	2006-2100	-
4	CNRM-CM5	CNRM-CERFACS (France)	$1.5^{\circ} \times 1.5^{\circ}$	1948-2014	2006-2100	-	-
5	CSIRO-Mk3.6.0	CSIRO-QCCCE (Australia)	$1.875^{\circ} \times 1.875^{\circ}$	1948-2014	2006-2100	-	-
6	GFDL-ESM2M	NOAA GFDL (USA)	$2^{\circ} \times 2.5^{\circ}$	1948-2014	-	2006-2100	2006-2100
7	GISS-E2-R	NASA GISS (USA)	$2^{\circ} \times 2.5^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
8	HadGEM2-ES	MOHC (UK)	$1.25^{\circ} \times 1.875^{\circ}$	1948-2014	2006-2100	-	-
9	INM-CM4	INM (Russia)	$1.5^{\circ} \times 2.0^{\circ}$	1948-2014	-	2006-2100	2006-2100
10	IPSL-CM5A-MR	IPSL (France)	$1.25^{\circ} \times 2.5^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
11	MIROC5	MIROC (Japan)	$2.8^{\circ} \times 2.8^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
12	MIROC-ESM	MIROC (Japan)	$2.8^{\circ} \times 2.8^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
13	MIROC-ESM-CHEM	MIROC (Japan)	$2.8^{\circ} \times 2.8^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
14	MRI-CGCM3	MRI (Japan)	$1.125^{\circ} \times 1.125^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100
15	NorESM1-M	NCC (Norway)	$1.875^{\circ} \times 2.5^{\circ}$	1948-2014	2006-2100	2006-2100	2006-2100

Table 2. Root mean square error (RMSE) in estimates of annual statistics from GCM simulations compared to NCEP/NCAR estimates. RMSE calculated across 9 grid cells for BCC-CSM1-1 GCM. Results for other GCMs available on request.

Statistic	GCM case	AIR	RHUM	SHUM	SLP	TMAX	TMIN	HGT3
Annual Mean	Raw	2.83	12.05	4.3	3.69	2.69	4	17.68
	MRNBC	0.01	0.17	0.01	0.03	0.02	0.03	0.22
Annual Standard Deviation	Raw	0.11	1.23	0.38	0.7	0.21	0.3	1.76
	MRNBC	0.02	0.49	0.06	0.24	0.08	0.1	0.77
2 year Standard Deviation	Raw	0.11	1.42	0.38	0.86	0.23	0.31	2.75
	MRNBC	0.02	0.44	0.06	0.2	0.06	0.08	0.43

Table 3. Results of the performance criteria for combination of the machine-learning methods and the dimensionality reduction methods in modeling step based on testing dataset

Machine-learning method	Dimensionality reduction method	Kernel	No of support or relevant vectors	R^2	RMSE	NSE	MAE
SVR	S-PCA	<i>RBF</i>	188	0.65	2.95	0.61	2.37
	Kernel S-PCA	<i>RBF</i>	135	0.76	1.45	0.75	1.32
RVM	S-PCA	<i>Laplacian</i>	37	0.68	2.66	0.63	2.33
	Kernel S-PCA	<i>RBF</i>	12	0.79	1.39	0.78	1.27

- Bold signifies the best selected combination model

Table 4. Linear trends for projected precipitation (mm (10yr)⁻¹) in near-term (2015-2040) and long-term (2015-2100) periods, before and after bias-correction under different climate change scenarios

Model	RCP2.6				RCP4.5				RCP8.5			
	Near-Term		Long-Term		Near-Term		Long-Term		Near-Term		Long-Term	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
BCC-CSM1.1	-8.69	-16.68	-1.07	-2.25	-9.61	-30.10	-4.80	-1.37	-25.63	-20.53	-6.41	-2.14
CanESM2	-1.11	-5.17	-0.07	-5.89	-9.24	-26.74	-3.58	-1.46	-	-	-	-
CCSM4	0.12	-27.40	-1.85	-1.23	0.58	-27.08	-0.17	0.59	-	-	-	-
CNRM-CM5	6.76	17.98	1.38	11.05	12.31	24.90	0.92	3.62	-	-	-	-
CSIRO-Mk3.6.0	-19.40	-32.08	-0.25	0.78	-	-	-	-	-2.34	-31.74	-2.80	-3.52
GFDL-ESM2M	-	-	-	-	-19.98	-5.68	-6.25	-2.03	4.46	-8.97	-3.90	-1.69
GISS-E2-R	-4.09	-9.89	-0.82	-0.99	19.49	40.60	-1.69	-0.34	3.46	13.85	-1.39	1.07
HadGEM2-ES	10.11	-9.89	2.57	-0.99	-	-	-	-	-	-	-	-
INM-CM4	-	-	-	-	-6.03	15.00	-4.63	-2.63	1.25	-21.91	-3.80	-4.82
IPSL-CM5A-MR	4.72	8.49	2.26	2.46	-7.29	23.37	-0.14	1.87	7.90	-0.61	-6.60	-5.17
MIROC5	-3.43	-17.27	-3.15	-3.63	-13.72	-12.21	-1.92	-3.13	6.49	3.54	-5.76	-3.60
MIROC-ESM	-22.05	-51.31	-3.72	-5.93	-10.34	-30.75	-8.56	-5.36	3.84	36.30	-9.23	-5.16
MIROC-ESM-CHEM	-8.51	-39.31	-3.34	-2.33	10.38	20.22	-4.78	-3.96	-26.45	-33.73	-5.52	-4.42
MRI-CGCM3	4.88	18.81	1.80	1.24	-1.05	23.66	-0.74	0.01	-2.84	6.87	-1.48	-2.38
NorESM1-M	-2.09	16.81	-2.25	-4.44	-12.16	-30.75	-4.95	-5.57	-2.20	-5.17	-8.94	-5.89
Average	-3.29	-11.30	-0.66	-0.87	-3.59	-1.20	-3.18	-1.52	-2.92	-5.65	-5.08	-3.43
No of negative models	8/15	9/15	9/15	9/15	9/15	7/15	12/15	9/15	5/15	7/15	11/15	10/15

* It should be noted that the precipitation linear trend analysis for the historical records (1951-2011) shows an increasing trend of 7.29 mm (10yr)⁻¹.