8-1-2014

Statistically Downscaled North American Precipitation Using Support Vector Regression And The Big Brother Approach.

Carlos F. Gaitan

Keith W. Dixon

Venkatramani Balaji

Renee McPherson

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

STATISTICALLY DOWNSCALED NORTH AMERICAN PRECIPITATION USING SUPPORT VECTOR REGRESSION AND THE PERFECT MODEL EVALUATION FRAMEWORK

CARLOS F GAITAN (1), KEITH W. DIXON (2), RENEE MCPHERSON (3), VENKATRAMANI BALAJI (4)

(1): South Central Climate Science Center, University of Oklahoma, 201 Forrestal Road, Princeton NJ 08540 USA
(2): NOAA, Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton NJ 08540, USA
(3): South Central Climate Science Center, University of Oklahoma, 301 David L. Boren Blvd. Suite 3030, Norman, OK 73072 USA
(4): Cooperative Institute for Climate Sciences, Princeton University, 201 Forrestal Road, Princeton NJ 08540, USA

Here we introduce a novel statistical downscaling method based on machine learning algorithms which uses classification and regression trees (CART) or support vector machines for classification (SVM-C) to obtain precipitation occurrences, and then uses support vector machines for regression (SVM-R) with evolutionary strategies to obtain precipitation amounts on those days classified as rainy. The SD method was tested in terms of Peirce Skill Score (PSS) and mean absolute error skill score (MAE SS). Additionally, to test the statistical downscaling time-invariance assumption we used daily precipitation outputs from a high resolution (~25km grid spacing) global atmospheric model as predictands, and a coarsened version of the same high resolution outputs - interpolated to a ~100km grid – as predictors.

The study focuses on 16 points from different climate regions across North America. The downscaled results were evaluated in terms of historical and future to assess if the skills were time-invariant. Our results show that for 9 out 16 points the hybrid CART-SVM-R downscaling model had positive historical and future MAE SS. We also found that the SVR-C model under-predicted the total number of rainy days. The downscaled model results generally outscored the ones obtained using stepwise multiple linear regression. Future implementations will expand the predictor set aiming to improve the overall MAE SS.

INTRODUCTION

Statistical downscaling techniques, like the ones implemented in the present work, are often used to refine the coarse resolution outputs from global climate models (GCMs) or reanalysis products as the spatial resolution of these models is insufficient to resolve many local scale phenomena occurring at a much higher scale, like convective and orographic precipitation [1], and accurate local estimates of meteorological variables are often needed by ecologists, biologists, engineers, and hydrologists, among others. However, the statistical downscaling techniques rely in several assumptions in order to generate future values of the climatological variable of interest; one of them is the time-invariance relationship between predictors and the predicted local variables required by the climate change impact studies.

Support vector regression [2] has shown to be an effective downscaling technique when used to get finer scale local precipitation over India [3], but its model output is highly dependent on the values of multiple hyper-parameters, often optimized via an extensive grid search. Alternatively, one could use evolutionary strategies to obtain these parameters, decreasing the computing time. Similarly, classification and regression trees (CART) ensembles have been extensively used in numerous applications [4, 5] showing to be a fast and efficient classifier, often outscoring popular methods like discriminant classification, naïve-Bayes classification and k-nearest neighbors [6]. Similarly, support vector machines for classification (SVM-C) have been shown to perform well in a variety of settings [7], and according to James, Witten [8] are often considered one of the best classifiers.

Here we tested the classification capabilities of CART and SVM-C, and the regression skills of SVM-R to obtain downscaled precipitation. The experimental setup involves using precipitation outputs from a coarsened version of the C360 HIRAM as predictors, and precipitation outputs from the C360 HIRAM as predictands. This methodology allows us to validate the downscaled results versus historical and future values of the predictand. The study includes 16 points from different climate regions across North America

Please use your paper number as the filename for identification. The organizing committee reserves the right to return/reject papers that do not fully comply strictly with the format instructions given below.

METHODS & DATA

Statistical downscaling is a post-processing technique often used to refine the coarse resolution GCM output so the statistical characteristics of these time-series are more alike to local/finer scale observation-based datasets. Our methodology is certainly not the only way to downscale precipitation occurrences and amounts, and multiple downscaling methods including weather generators, classification techniques, constructed analogues methods, and other linear and nonlinear regression approaches can be used to tackle the problem.

We used the following downscaling procedure: First, we trained the model using the historical (H2 or H3) coarse resolution data as predictors and the high-resolution GCM pseudo-observations as predictands; then we computed the model's validation error on independent

historical data using 2-fold cross-validation [9]; and finally, we used the coarsened outputs of two different model simulations as inputs of the downscaling model in order to obtain downscaled future projections.

Statistical downscaling methods for precipitation occurrence

Tree-based models

Tree-based models belong to the sequential decision making family of algorithms. In general, these models divide the input space into rectangular regions according to whether $xi \leq B$, or $xi > B$, where B is a parameter of the model [9], and each region has a separate model to predict or to classify the target variables. Tree-based models are divided in two categories: a) classification trees and b) regression trees; although the term classification and regression tree (CART) usually refers to both types of trees.

We used classification trees to model the precipitation occurrence process, and to obtain smaller classification errors we used bootstrap aggregation of 500 trees; where every tree in the ensemble is grown on an independently drawn bootstrap replica of the input data [10]. In general, bootstrap allows us to obtain maximum likelihood estimates of standard errors and other quantities where no formulas are available.

Statistical downscaling methods for precipitation amounts

Support vector machines for regression (SVM-R)

As mentioned earlier, support vector machines (SVM) were originally designed for nonlinear classification problems [11], and only few years later Vapnik [2] introduced an SVM extension aimed to solve regression problems. This extension is known as support vector regression (SVM-R). Recent applications of support vector machines to environmental sciences problems include forecasting sulphur-dioxide (SO2), precipitation and surface temperature [12], seasonal winter extreme precipitation forecasts [13] and statistical downscaling of precipitation [3, 14], among many others. To obtain the best hyper-parameters the user often needs to implement a 2-D or 3-D grid search.

To statistically downscale the precipitation amounts we used the e1071 package for R which uses the LIBSVM implementation [7]. We selected the optimal hyper-parameters using evolutionary strategies (ES), following Eiben and Smith [15], and the grid search procedure proposed by Fan, Chen [16]. For a more detailed explanation of SVM-R using evolutionary strategies we recommend to read the work of Lima, Cannon [12].

To prevent fitting the model to noise we divided the dataset in two parts, one for training the model and the other one for evaluation, this second dataset is commonly referred as validation dataset in the machine learning community, even thou technically speaking the validation of natural systems is impossible as the Earth is an open system and the model results are non-unique [17].

Evolutionary strategies

Evolutionary strategies had been used recently in conjunction with other machine learning/artificial intelligence methods for parameter optimization and to increase the models' performance. In general the parameters used by the evolutionary strategies have the ability to

co-evolve with the solutions so the algorithm can self-adapt. Specifically, our implementation used uncorrelated mutation with P step sizes following Eiben and Smith [15], with 3 nearest neighbors and knn regression to estimate the noise level, 150 offspring and 30 generations; also we initialized the gamma parameter using the Cherkassky and Ma [18] estimate following Lima, Cannon [12].

Study area and data

Daily atmospheric model outputs were obtained from a set of GFDL C360 HIRAM model ("C360") experiments (see http://www.gfdl.noaa.gov/hiram). We used output from two 30-year long historical C360 model runs from 1979-2008 to train and validate the downscaling method using 2-fold cross validation. The transform functions derived from the historical period were applied to C360 climate change projection simulations – more specifically, to a pair of three member ensembles covering the period 2086-2095. For both the historical and future periods, the raw high resolution C360 output served as pseudo observations (predictands) and the smoothed by- interpolation data sets served as predictors.

The downscaling model used CART or SVM-C to obtain the precipitation occurrences and SVM-R to obtain the precipitation amounts. Comparisons between the predictands and the downscaled results were made in terms of mean absolute error skill score (MAE SS) using the nearest coarse-resolution predictor as benchmark. In particular, the CART model used an ensemble of 500 classification trees to obtain the precipitation occurrences, and the SVM-R model used a Gaussian kernel and evolutionary strategies to determine the values of the three hyper-parameters needed to perform the nonlinear regression.

Table 1. Selected grid-points

Number	**ID**	**Location**
1	Seattle	Washington
2	San Francisco	California
3	Yosemite	California
4	Bakersfield	California
5	Aspen	Colorado
6	Las Cruces	New Mexico
7	Winnipeg	Manitoba
8	Norman	Oklahoma
9	Corpus Christi	Texas
10	Iowa City	Iowa
11	Baton Rouge	Louisiana
12	Sudbury	Ontario
13	Charlotte	North Carolina
14	Miami	Florida
15	Saranac Lake	New York
16	South Brunswick	New Jersey

Results

The statistically downscaled precipitation occurrences were compared against a binary predictand. The predictand is zero when the daily precipitation is less than 0.127 mm and 1 otherwise, while the statistically downscaled precipitation amounts from the hybrid model were compared against the coarsened predictors, used as benchmark.

The precipitation occurrence results show the CART method outscoring the SVM-C in terms of number of days with precipitation during the historical period (H2 & H3). When comparing both classification methods against the pseudo-observed number of days with precipitation – determined by the number of days with precipitation > 1 mm in the HIRAM (Fig. 1), we see that even though both methods under-predicted the number of rainy days, CART provided a closer agreement with the pseudo-observations (specially in Aspen and Saranac Lake). The results also show that SVM-C was especially unsuccessful when downscaling precipitation occurrences in Las Cruces and Baton Rouge, partly because the number of rainy days was unbalanced between classes (rain/no-rain). These results suggest that the CART method should be preferred versus the SVM-C to downscale precipitation amounts under our experimental setup.

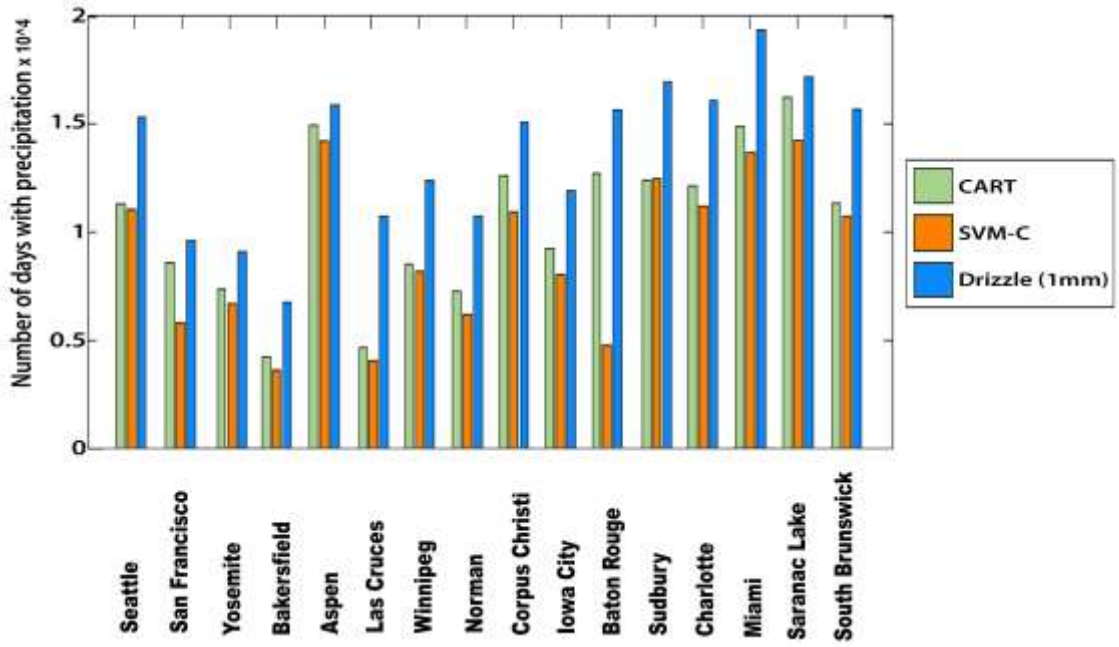

Figure 1. Number of days with precipitation (historical period)

To determine how effective the SVM-R downscaling technique was, we compared the downscaled results versus the coarsened GCM precipitation in terms of MAE Skill Score. A positive skill score implies better agreement than the coarsened GCM, and negative values indicate that the coarsened GCM output agreed more than the downscaled output when compared to the pseudo-observations. In other words, there is value added by the post-processing technique where the MAE SS are positive, and where the skill scores are negative the coarsened GCM represents a better agreement to the pseudo-observations than the downscaled values.

Conclusions and Recommendations

We evaluated the downscaled results in terms of historical and future MAE SS to assess if their skills were time-invariant. The results show that for 9 out 16 points the hybrid CART-SVM-R downscaling model had positive historical and future MAE SS. We also found that the CART

model under predicted the total number of rainy days, thus affecting the MAE SS, the length of the wet/dry spells and the yearly precipitation amounts.

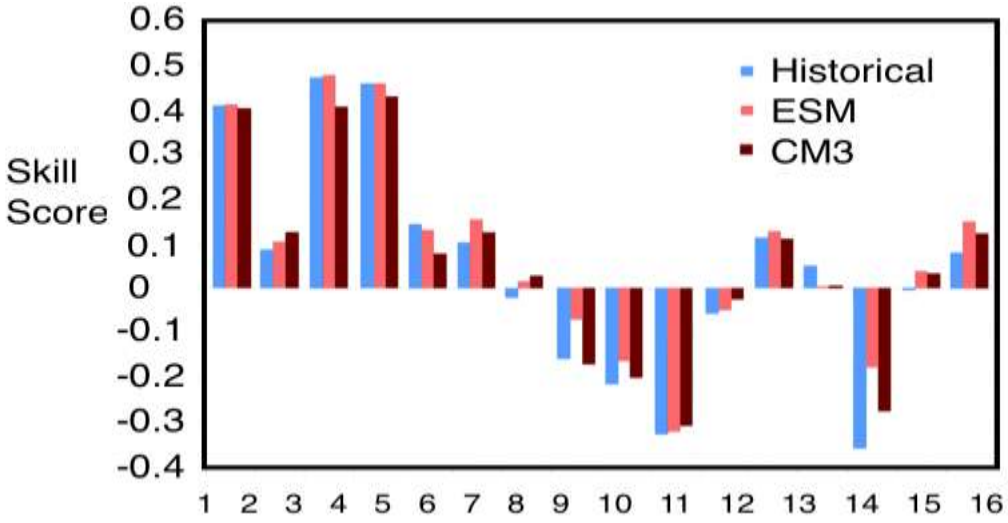

Figure 2. Historical and future mean absolute error skill score. Red bars represent the future simulations

Future implementations will test other classification methods (e.g. support vector classification, drizzle threshold, k-nearest neighbor) and will expand the predictor variables aiming to improve the overall MAE SS. Ongoing work includes the development of statistical downscaling models using Bayesian neural networks, quantile regression and different types of linear and nonlinear regression. We aim to learn more about the methods' strengths and limitations, and specially learn about the future behavior of different extrapolation techniques, as some of the future GCM outputs may be outside of the historical period range used during training.

Acknowledgements
We wish to thank the National Oceanographic and Atmospheric Administration for sharing the GFDL C360 HIRAM model outputs.

REFERENCES

1. Wilby, R.L. and T.M.L. Wigley, *Downscaling general circulation model output: a review of methods and limitations.* Progress in Physical Geography, 1997. **21**(4): p. 530-548.
2. Vapnik, V., *Statistical Learning Theory*. 1998, NY: John Wiley. 732.
3. Tripathi, S., V. Srinivas, and R. Nanjundiah, *Downscaling of precipitation for climate change scenarios: A support vector machine approach.* Journal of Hydrology, 2006. **330**(3-4): p. 621-640.
4. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning*. Second ed. 2009: Springer.

5. Hsieh, W.W., *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. 2009, Cambridge, UK: Cambridge University Press.
6. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**: p. 123-140.
7. Chang, C.-C. and C.-J. Lin, *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**: p. 27.
8. James, G., et al., *An Introduction To Statistical Learning. With Applications in R*. Springer Texts in Statistics, ed. G. Casella, S. Fienberg, and I. Olkin. 2011, New York: Springer.
9. Bishop, C.M., *Pattern Recognition and Machine Learning*. 2006, Cambridge, U.K.: Springer.
10. Preisendorfer, R.W., *Principal Component Analysis in Meteorology and Oceanography*. Developments in Atmospheric Science, ed. C.D. Mobley. Vol. 17. 1988, Amsterdam: Elsevier. 425.
11. Vapnik, V., *The Nature of Statistical Learning Theory*. 1995, Berlin: Springer.
12. Lima, A.R., A.J. Cannon, and W.W. Hsieh, *Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy.* Computers & Geosciences, 2013. **50**: p. 136-144.
13. Zeng, Z., et al., *Seasonal prediction of winter extreme precipitation over Canada by support vector regression.* Hydrology and Earth System Sciences, 2011. **15**(1): p. 65-74.
14. Chen, S.T., P.S. Yu, and Y.H. Tang, *Statistical downscaling of daily precipitation using support vector machines and multivariate analysis.* Journal of Hydrology, 2010. **385**(1-4): p. 13-22.
15. Eiben, A.E. and J.E. Smith, *Introduction to Evolutionary Computing*. Natural computing series. 2003, Berlin: Springer. 300.
16. Fan, R.E., P.H. Chen, and C.J. Lin, *Working set selection using second order information for training support vector machines.* Journal of Machine Learning Research, 2005. **6**: p. 1889-1918.
17. Oreskes, N., K. Shrader-Frechette, and K. Bellitz, *Verification, validation, and confirmation of numerical models in the Earth sciences* Science, 1994. **263**(5147): p. 641-646.
18. Cherkassky, V. and Y. Ma, *Practical selection of SVM parameters and noise estimation for SVM regression.* Neural Networks, 2004. **17**(1): p. 113-126.