Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 140 (2018) 248–257

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Cyber Physical Systems and Deep Learning, CAS 2018, 5 November – 7 November 2018, Chicago, Illinois, USA

The 2015-2017 Cape Town Drought: Attribution and Prediction Using Machine Learning

Michael B. Richman[a]* and Lance M. Leslie[b]

[a] *School of Meteorology, University of Oklahoma,120 David L. Boren Blvd, Suite 5600, Norman, OK, 73072 USA*
[b] *School of Mathematical and Physical Sciences, University Technology Sydney, 15 Broadway, Ultimo, NSW 2007, Australia*

Abstract

Cape Town was declared a disaster area after the worst drought in almost a century, following its driest three consecutive wet seasons (April 1-October 31), in 2015 -2017. Cape Town's drought was extreme, with "zero day" water storage months away, causing severe water rationing to Cape Town's ~3.8 million population. The crisis extended into surrounding farmlands, as agriculture is vital for the region's economy. Possible drought causes are numerous and, aside from the decreasing wet season precipitation, the effects are exacerbated by the increasing population with associated water demand, greater agricultural acreage and land surface changes. As rainfall decreases, water management becomes critical, requiring predictions for future rainfall. Possible climate drivers associated with available Cape Town precipitation and temperature include: The Southern Annular Mode, Atlantic Meridional Mode, Indian Ocean Dipole, an Integrated Southern Hemisphere temperature index and several El Niño indices. Several variable selection techniques suggest signals in both the Atlantic and Indian Oceans contribute to Cape Town droughts. Machine learning techniques are applied to these drivers for the first time and provide encouraging predictive skill levels. Results suggest that machine learning holds promise for adapting to drought by managing water resources in Cape Town and, more generally for global locations depending solely on rainfall under a warming climate.

Keywords: Machine Learning; Variable Selection; Support Vector Machines; Drought; Global Climate Change; Attribution

* Corresponding author. Tel.: +1-405-325-1853; fax: +1-405-325-7689.
E-mail address: mrichman@ou.edu

10.1016/j.procs.2018.10.323

1. Introduction

The global warming trend reported more than a decade ago by the Intergovernmental Panel on Climate Change (IPCC) and other institutions, was a warning to Cape Town and many other major cities around the world. These cities had in common a heavy dependence on *in situ* rainfall. Combined with a warming climate, rapid population growth and projected annual rainfall decreases, additional water sources and management would be required. However, the initial and later warnings were not fully appreciated, even more so because all of Cape Town's six major dams were totally filled by the end of the 2014 wet season. However, the following three-year period from 2015-2017, all were drought years of increasing severity. Water storage declined rapidly and left the four million residents of Cape Town with dams at below 25% of capacity in January 2018 which, as the next wet season was several months away, was heading towards "Zero Day," so-named when the supply capacity falls below a pre-defined critical level of 13.5% [1]. Some pre-season rainfall in February 2018, and early-wet season rainfall in April and May, shifted Zero Day from April 2018 back to August 2018, and it might be avoided altogether in 2018 if the wet season is near normal. However, the long-term threat of future critically low water storage levels remains. As in the case of droughts in other parts of the world that similarly are so heavily reliant on rainfall, there is now a focus on analyzing Cape Town droughts in a climate-driven context, with the aim of finding ways of avoiding future Zero Days. In particular, it is important to assess whether the recent extreme drought years of 2015 were part of natural variability or were a result of global climate change exacerbated by other factors such as population growth, significant changes in land use, and management of the existing water supply.

Chapter 22 of the IPCC's Fifth Assessment Report (AR5) Working Group 2 [2] provides an assessment of the impact of climate change and other factors on Africa in general, and South Africa, in particular, prior to the commencement of the 2015-2017 Cape Town drought. Notably, [2, p.2012] stated that "A reduction in precipitation is likely over Northern Africa and the southwestern parts of South Africa by the end of the 21st century under the SRES A1B and A2 scenarios (medium to high confidence)." It also suggested that "Climate change will interact with non-climate drivers and stressors to exacerbate vulnerability of agricultural systems, particularly in semi-arid areas (high confidence)."

In this study, rainfall (in units of mm) and temperature (in Celsius) wet season annual means are used for the period 1920-2017 to determine precisely how extreme was the 2015-2017 drought (in percentiles), among the recurring droughts affecting Cape Town that are revealed by almost a century of recorded data. These data also will be used to attribute the drought to a range of climate drivers that have relevance to Cape Town, using both traditional step-wise linear regression methods and non-linear machine learning techniques.

2. Data and Methods

2.1. Data

Cape Town is situated at the southwest point of South Africa (Fig. 1). As drought in Cape Town is defined as a deficit of precipitation, historical precipitation data for Cape Town were obtained for the period 1920-2017. Additionally, since Cape Town has a Mediterranean climate, with nearly all the precipitation falling in the seven month cool season (April – October), totals for the cool season were calculated. Owing to the warming climate which increases evaporation rate, the air temperature in Cape Town also was obtained. As this study is concerned with apportioning the roles of established climate drivers and global warming in the attribution of drought, a set of tropical and Southern Hemispheric climate drivers were investigated as potential predictors of drought. The relationship between drought and air temperature is not linear. At cold air temperatures, the saturated atmosphere holds less water than at warmer temperatures. However, at high air temperatures, the evaporation rate is larger, hastening the onset of drought. Once drought is initiated, the ground and vegetation become drier, and solar radiation no longer is used for photosynthesis or evaporation. Therefore, there are fewer clouds present and a higher percentage of solar radiation heats the ground in a positive feedback. For Cape Town, this relationship is indicated in the scatterplot of air temperature versus precipitation (Fig. 2).

Fig. 1. Map of South Africa showing Cape Town.

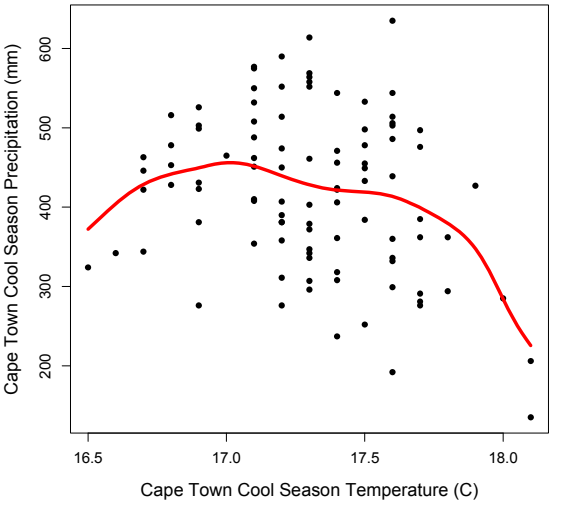

Fig. 2. Cape Town air temperature vs. precipitation for each cool season from 1920-2017. Red line is the kernel smoothed Gaussian fit.

2.2. The 2015-2017 drought in a historical perspective

With nearly a century of continuous precipitation temperature data available, the uniqueness of the recent drought can be assessed. The quantiles of the cool season precipitation time series (Fig. 3) show that the 2015-2017 drought was unique as it was the first case of three consecutive years with rainfall totals less than the 10th percentile. Previous droughts had only a single year with less than a 10 percentile precipitation (1926, 1973, 1978, 1994, 2003) or two consecutive years (2010-2011). Moreover, the incidence of severe drought can be seen to increasing with time as the percentiles of the droughts from 1994 to 2017 become progressively smaller. Over the same 1920-2017 time span, the cool-season temperature (Fig. 4) has been relatively stationary through ~1990, with only a single year (1963) reaching the 90th percentile and numerous years falling below the 15th percentile. After 1993, no year had an average cool-season temperature below the median and from 1999-2017 the average temperature was almost always above the 85th percentile. The years 2014-2017 recorded the warmest temperatures on record. This combination of increasingly persistent droughts combined with excessively warm temperatures resembles the recent California drought [3]. Moreover, both locations have a Mediterranean climate with nearly all their annual precipitation occurring in the cool season. If the cool season precipitation fails to materialize, both locations can run out of water.

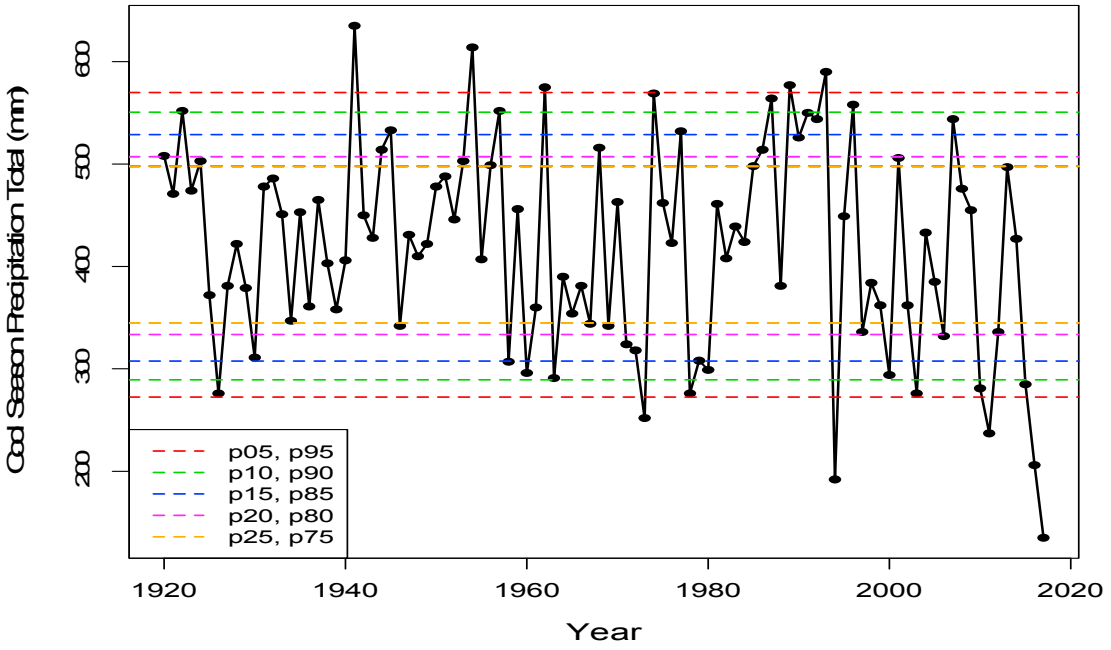

Fig. 3. Cape Town cool season precipitation. Quantiles are shown in dashed colored lines.

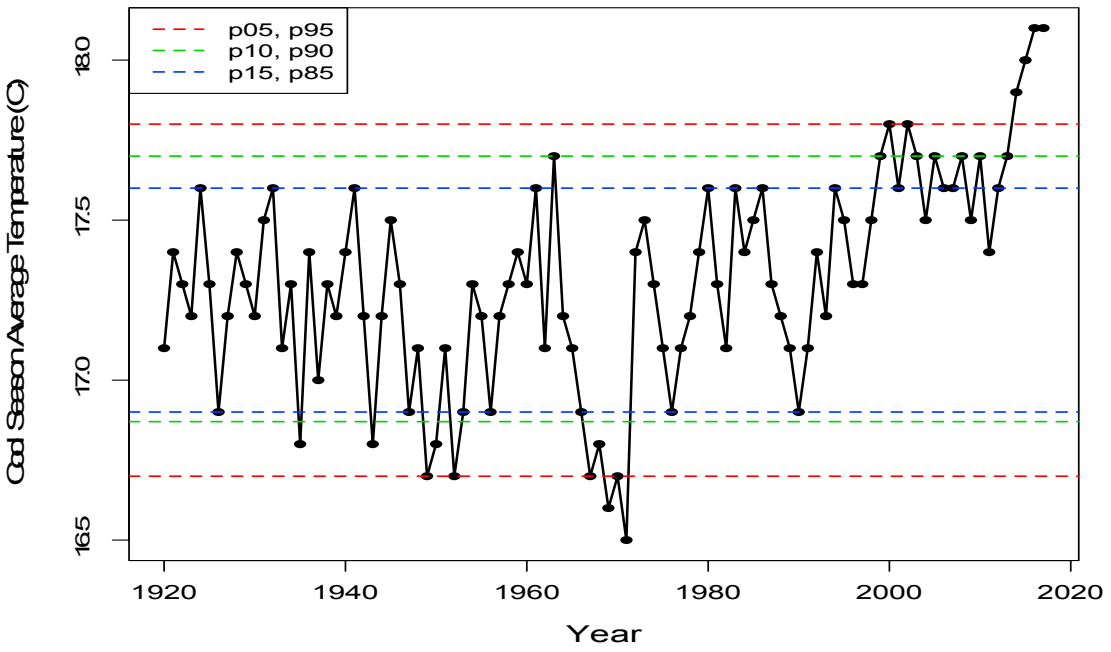

Fig. 4. Cape Town cool season average temperature. Quantiles are shown in dashed colored lines.

2.3. Potential climate drivers related to drought

Relating the large-scale climate to drought is often accomplished through the use of recurrent atmospheric modes of variability, called climate drivers. Whereas the Northern Hemisphere has a large catalogue of such drivers, fewer exist for the Southern Hemisphere. In previous studies [4], [5], [6], tropical modes, such as El Niño indices have been examined as precursors of rainfall modulation in South Africa. Additionally, the Indian Ocean Dipole and the Southern Annular Mode are sometimes evaluated. In the present study, the pool of atmospheric and oceanic climate drivers included the Southern Annular Mode (SAM), Atlantic Multidecadal Oscillation (AMO), Atlantic Meridional Mode (AMM), The Indian Ocean Dipole (DMI), Niño 3.4 sea surface temperatures (NINO), the Southern Oscillation Index (SOI) and, owing to the aforementioned relation between air temperature and precipitation, the cool season average temperature (CST), and the average temperature anomaly in the Southern Hemisphere. Additionally (SHTA), along with all two-way interactions between these climate drivers (e.g., SAMxAMO,) were considered as potential attributes.

3. Attribute selection and the potential for predicting drought

3.1. Attribute selection

Given the large number of potential attributes, some selection techniques were required to determine those that had the best performance in predicting the cool season precipitation in a cross-validation framework. Both linear (correlation-based feature selection - Cfs) and non-linear (radial basis function support vector regression-SVR-feature selection) methods were assessed using ten-fold cross-validation. The first type of attribute evaluator considered the individual predictive ability of each feature along with the degree of redundancy between by searching the space of attribute subsets by greedy hill climbing method [7]. Both forward (GreF) and backward (GreB) searches were used. The second type of attribute selection used the subset evaluation with a genetic search algorithm (Gen) as described by [8]. Additionally, a scatter-search algorithm (Scatt) was used to examine the space of attribute subsets [9]. Finally, a set of wrapper subset methods [10] were used with SVR employing a radial basis function kernel as a classifier for best forward, backward, genetic and scatter search evaluations (SVRGreF, SVRGreB, SVRGen, SVRScatt). The number of times each attribute appeared in the ten folds was for a given method is shown in Table 1. The goal was to find those variables that emerged across several types of evaluation. The results are summarized for each column of Table 1 by the first four central moments. Owing to the instance of more folds with zero or small percentages for each of the attributes, the four Cfs selection methods tends to have a smaller means than the support vector counterparts. The standard deviations are comparable. The skewness statistic, for all but the genetic algorithm, for the Cfs were larger than those for SVR methods as there were more solutions with zero or low percentage folds in those methods; hence the outliers were those features with a high percentage of folds. The kurtosis of each Cfs selection method, other than the genetic algorithm is larger for the Cfs methods. In this context, kurtosis is a good proxy for column simplicity (i.e., more very small and very high percentage folds), so a positive kurtosis is associated with those techniques that select only a few attributes (e.g., greedy forward selection and greedy scatter search). The SVR-based algorithms all are platykurtotic, suggesting less simplicity and selecting more attributes than their correlation-based counterparts.

Analysis of Table 1 by row reveals those attributes that emerge across most or all methods (i.e., CST, SAM, AMO, DMI, DMIxCST, CSTxAMO). Such attributes are characterized by large means, small standard deviations and near-zero or negative skewness. Those attributes that tend to emerge for either Cfs or SVR techniques alone (e.g., DMIxAMM, SHTAxCST, SHTAxAMO, SAMxAMO, AMOxSOI, AMMxSOI) have moderate means and larger standard deviations and negative kurtosis. In both cases, these attributes should be retained for further analysis. Additionally, certain attributes never/rarely appear under any selection method (e.g., SOI; NINO, DMIxSAM, DMIxSOI, SHTAxSAM, SHTAxNINO, CSTxSOI, SAMxNINO, and SAMxSOI) and have small means. Such attributes can be discarded, a priori.

.

Table 1. Attribute selection for eight different methods. The percent of folds selecting the attribute is shown for each of the eight methods (columns 2-9). Statistics (mean, standard deviation, skewness and kurtosis) for each row and column are shown.

Attribute/Method	GreF	GreB	Gen	Scatt	SVRGreF	SVRGreB	SVRGen	SVRScatt	$\bar{x}$	*s*	*skew*	*kurt*
CST	100	100	80	100	90	100	100	100	96.2	7.4	-1.5	0.4
SHTA	0	0	0	0	10	10	40	0	7.5	13.9	1.7	1.3
SAM	60	90	100	60	80	70	80	50	73.8	16.9	0.1	-1.3
AMO	100	80	80	90	70	90	70	80	82.5	10.4	0.3	-1.1
AMM	0	20	60	10	40	50	50	30	32.5	21.2	-0.2	-1.5
DMI	100	100	90	100	80	90	90	90	92.5	7.1	-0.3	-1.1
NINO	0	0	10	0	0	10	10	10	5.0	5.3	0	-2.1
SOI	0	0	0	0	0	0	0	0	0.0	0.0	NaN	NaN
DMIxSHTA	0	10	30	0	40	30	30	20	20.0	15.1	-0.2	-1.6
DMIxCST	80	90	90	90	70	70	80	40	76.2	16.9	-1.2	0.3
DMIxSAM	0	0	0	0	0	10	20	10	5.0	7.6	1.0	-0.6
DMIxAMO	10	10	10	10	20	30	50	10	18.8	14.6	1.3	0.2
DMIxAMM	20	30	60	20	80	90	80	90	58.8	30.9	-0.3	-1.8
DMIxNINO	0	0	60	0	30	50	30	50	27.5	24.9	0.0	-1.8
DMIxSOI	0	0	10	0	0	10	20	0	5.0	7.6	1.0	-0.6
SHTAxCST	30	80	100	40	70	70	80	60	66.2	22.6	-0.3	-1.1
SHTAxSAM	0	0	30	0	10	20	20	0	10.0	12.0	0.5	-1.5
SHTAxAMO	10	20	60	20	70	90	80	80	53.8	32.0	-0.3	-1.8
SHTAxAMM	0	0	10	0	20	40	30	20	15.0	15.1	0.4	-1.4
SHTAxNINO	20	10	0	10	0	20	10	0	8.8	8.3	0.2	-1.5
SHTAxSOI	0	0	50	0	10	20	40	10	16.2	19.2	0.8	-1.1
CSTxSAM	0	0	70	0	30	30	40	0	21.2	25.9	0.7	-0.9
CSTxAMO	60	60	90	60	80	100	100	100	81.2	18.9	-0.2	-1.9
CSTxAMM	0	0	0	0	10	20	40	20	11.2	14.6	0.9	-0.6
CSTxNINO	0	0	0	0	50	50	60	30	23.8	26.7	0.2	-1.9
CSTxSOI	0	0	0	0	0	0	10	0	1.2	3.5	2.1	2.4
SAMxAMO	10	50	100	20	80	60	90	100	63.8	35.0	-0.4	-1.5
SAMxAMM	10	10	50	10	60	80	60	70	43.8	29.2	-0.2	-1.8
SAMxNINO	0	0	0	0	0	10	30	10	6.2	10.6	1.5	0.7
SAMxSOI	0	0	0	0	0	10	30	10	6.2	10.6	1.5	0.7
AMOxAMM	0	0	10	0	20	30	40	20	15.0	15.1	0.4	-1.4
AMOxNINO	0	0	0	0	10	20	50	10	11.2	17.3	1.5	0.7
AMOxSOI	0	0	0	10	70	90	60	80	38.8	39.8	0.1	-2.0
AMMxNINO	0	0	50	0	30	30	20	30	20.0	18.5	0.1	-1.4
AMMxSOI	20	50	40	10	80	90	80	80	56.2	30.7	-0.4	-1.6
NINOxSOI	0	0	10	0	40	60	40	40	23.8	23.9	0.2	-1.7
$\bar{x}$	17.5	22.5	37.5	18.3	37.5	45.8	48.9	37.5				
s	31.7	34.4	36.8	31.5	31.9	32.7	28.2	35.0				
skew	1.8	1.3	0.4	1.7	0.2	0.3	0.2	0.6				
kurt	1.6	-0.1	-1.4	1.4	-1.6	-1.4	-1.1	-1.2				

To examine the similarity of the variable selection methods, an average linkage cluster analysis of the Euclidean distance was performed on the ensemble of methods. The dendrogram (Fig. 5) shows all but the genetic algorithm search correlation-based feature subset selection methods are clustered. The SVM-based methods are in a separate cluster. This reinforces the analysis of the aforementioned subjective analysis of the percentage of folds in Table 1 by each selection method.

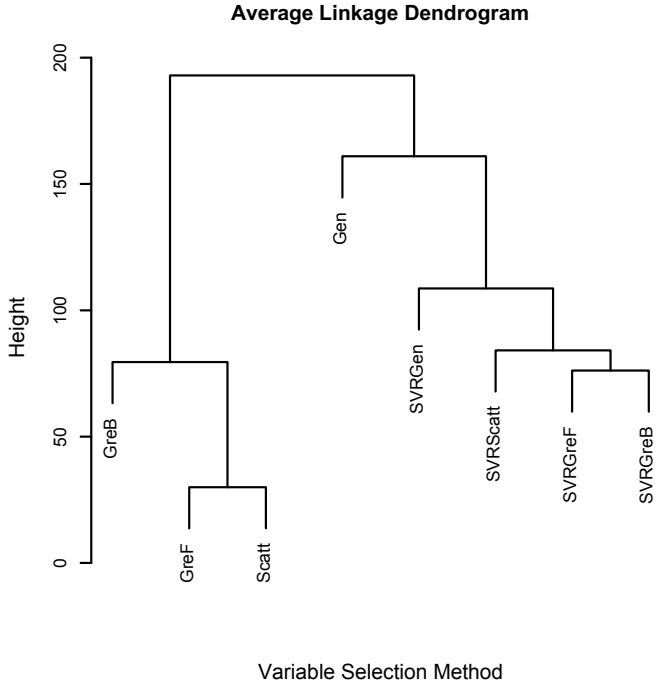

Fig. 5. Average Linkage cluster dendrogram.

Determining the underlying dimensionality and structure of the feature selection ensemble was accomplished by application of principal component analysis (PCA). The correlation structure of the fold information in Table 1 served as the basis for the analysis. Three eigenvalues statistically separated from zero were present, suggesting three unique principal components (PCs). The eigenvectors were post-multiplied by the square root of their corresponding eigenvalue to create PC loadings. The three PC loadings were rotated orthogonally and their pairwise plots (Fig. 6) showed the PC loadings were not aligned with the PC axes, occupying the space between the axes. Such a configuration is indicative of an overly orthogonal frame, so an oblique rotation (Promax) was used. The Promax PC loadings were plotted (Fig. 6), showing near-perfect alignment of the PC axes with the clusters of attributes. The interpretation of Fig 6 is that the three dimensions are all SVR selection methods on PC1, all but the genetic Cfs algorithm on PC2 and the Cfs genetic algorithm on PC3. Such an interpretation is consistent with the cluster analysis, but the PCA adds value as a specific dimensionality emerges and the Promax rotation finds an excellent simple structure [11] with specific Promax PC loadings weighting for each technique (Table 2).

3.2. Prediction of drought

Owing to previous success in prediction of California drought [3], SVR was applied to assess the potential to predict Cape Town precipitation. The cool-season precipitation served as the target and the aforementioned climate drivers passing the variable selection from section *3.1* serve as attributes. Both polynomial and radial basis function SVR were employed. A ten-fold cross validation was employed to gather statistics on the goodness of the model. Attributes with near-zero weights were discarded to construct a more parsimonious model, removing a single attribute and re-

testing the model. The SVR polynomial model maximized the 10-fold cross validation with the following attributes: CST, AMM, DMI, DMIxAMM, CSTxAMO, CSTxAMM and SAMxAMM. The cross-validated statistics are shown in Table 3 as well as the model errors in recent drought years. The same attributes were used for the RBF SVR with nearly equal accuracy in the cross-validation (Table 3). The predicted and observed values for the various folds in leave one of cross-validation are shown in Fig. 7.

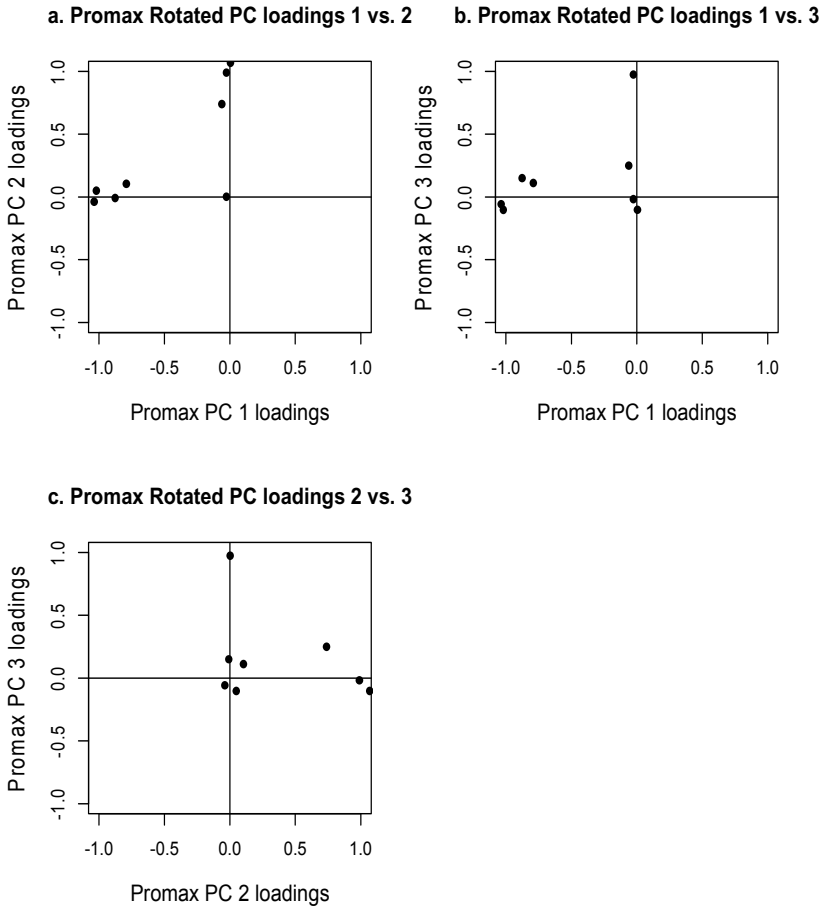

Fig. 6. Biplots for the 3 PC Promax solution.

Table 2. Promax rotated PC (RPC) loadings for 3 PC solution.

Attribute Selection Method	RPC 1	RPC 2	RPC 3
GreF	0.0	1.07	-0.10
GreB	-0.06	0.74	0.25
Gen	-0.03	0.0	0.97
Scat	-0.03	0.99	-0.02
SVRGreF	-0.88	-0.01	0.15
SVRGreB	-1.02	0.05	-0.10
SVRGen	-0.79	0.10	0.11
SVRScat	-1.04	-0.04	-0.06

Table 3. SVR model accuracy in 10 fold cross-validation (columns 2 and 3) and in predicting the 2015-2017 cool-season rainfall (columns 4-6)

Model/Statistic	Correlation	MAE (mm)	2015 error (mm)	2016 error (mm)	2017 error (mm)
Polynomial SVR (Poly, E=1, C=1)	0.561	73.04	-20.0	+52.3	+116.0
Radial Basis Function SVR (G=. 02, C= 50)	0.562	72.61	-27.0	+32.9	+115.9

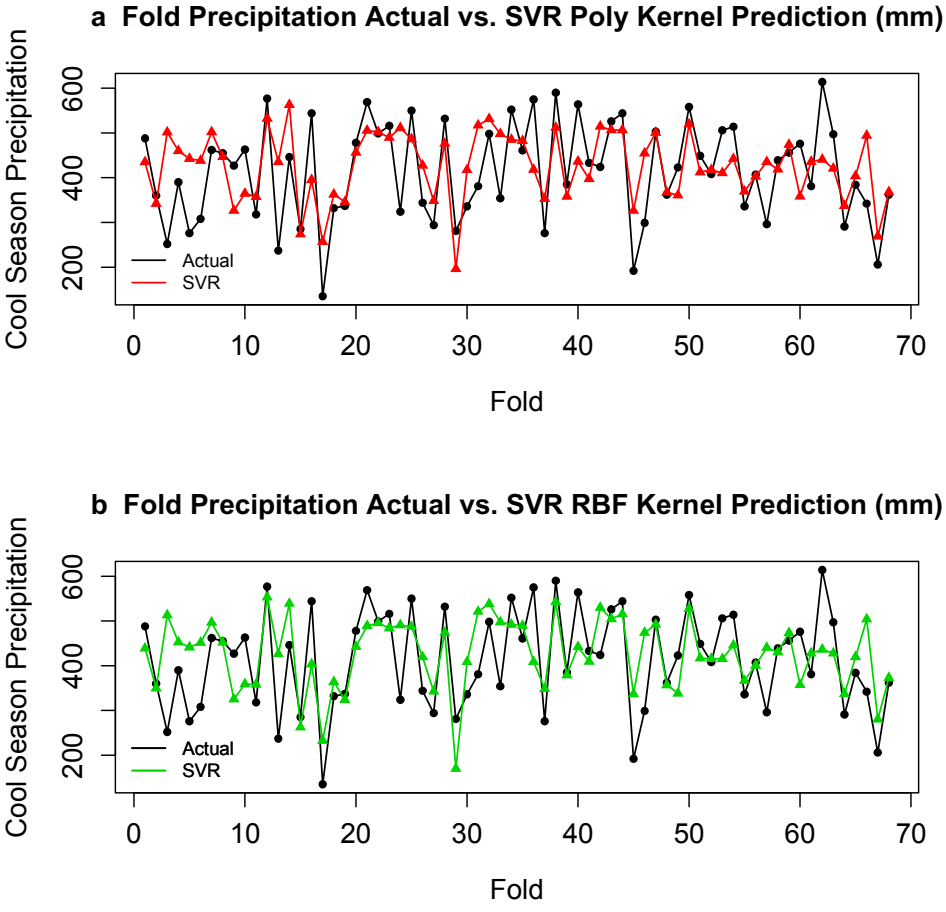

Fig. 7. Precipitation for each fold using leave one out cross-validations for (a) SVR polynomial and (b) SVR RBF models.

4. Conclusions and Future Work

By detailed application of an ensemble of attribute selection techniques and support vector regression, cross-validated prediction of precipitation was made. The correlation between the actual and modeled precipitation was ~0.56 for the two SVR methods tested. The key result is that the traditional use of the El Niño sea surface temperatures (in a region known as Niño 3.4), or the El Niño atmospheric response (known as the Southern Oscillation Index), for precipitation prediction was found to be ineffective for this Cape Town study. Linear regression of each of those two attributes (not shown herein) gave R^2 values of 0.004 and 0.001, respectively. The decreased attribution and predictive capacity of ENSO phases, which for a long period of time have formed the basis for seasonal and climate outlooks, is alarming. To mention just one example of many, the Australian Bureau of Meteorology's seasonal tropical cyclone,

temperature and precipitation outlooks all rely heavily on ENSO phase projections [12]. Adding additional attributes and applying machine learning techniques to winnow the number of predictors can be at least moderately effective in increasing the predictive capability [13]. The detection of other climate drivers now is of paramount importance to increase the accuracy of climate projections. Although arduous, this can be accomplished through a grid search of attributes in the global ocean and atmosphere, applying rank correlation between the attribute at each grid location and Cape Town precipitation as an importance metric.

Based on our findings in the present study, future work will be carried out on temperature predictions for Cape Town and, more generally, for the Southwest Cape region of South Africa. As mentioned in the Introduction, this region has been identified as one on which has been identified by the IPCC as being vulnerable to the combined impacts of lower rainfall, increasing temperature, and indirect or non-climate factors, such as rapid population growth and massive land use changes. All these factors should be the subject of future research in leading to extended droughts. The long term, limited definition of drought in terms of precipitation deficits has now been recognized and studies addressing drought that include other factors now are necessary. To do so requires, among other factors, an increased emphasis on Southern Hemisphere climate driver modes, and on current and projected population growth and land use change estimates.

Acknowledgements

The authors wish to thank the School of Mathematical and Physical Sciences, University of Technology Sydney, New South Wales, Australia for financially supporting this study.

References

[1] Maxmen, A. (2018) “As Cape Town Water Crisis Deepens, Scientists prepare for ‘Day Zero’,” Nature, 24 January 2018, (www.nature.com/articles/d41586-018-01134-x)

[2] IPCC (2014). Climate Change 2014, Working Group 2: “Africa”. [Core Writing Team Niang, I. and Ruppel, O.C]. Chapter 22. Geneva, Switzerland, 167pp.

[3] Richman, M.B. and Leslie, L.M. (2015). “Uniqueness and causes of the California drought.” Complex Adaptive Systems. *Procedia Computer Sciences*, Elsevier, **61**, 428-435.

[4] Philippon, N., Rouault, M., Richard, Y. and Favre, A. (2012). “The influence of ENSO on winter rainfall in South Africa”. International J. of Climatology, 32, 2333–2347.

[5] Tadross, M and Johnston, P. (2012). “Climate Systems Regional Report: Southern Africa”. In Sub-Saharan African Cities: A Five-City Network to Pioneer Climate Adaptation through participatory Research and Local Action, ICLEI – Local Governments for Sustainability – Africa Climate Systems Regional Report: Southern Africa, 38pp.

[6] Kane R. (2009). “Periodicities, ENSO effects and trends of some South African rainfall series: an update” *S Afr J Sci.,* 105,199–207.

[7] Hall, M. A. (1998) “Correlation-based Feature Subset Selection for Machine Learning”. PhD dissertation Hamilton, New Zealand.

[8] Goldberg, D. E. (1989). “Genetic algorithms in search, optimization and machine learning”. Reading, MA. Addison-Wesley.

[9] Lopez, F. G., Torres, M.G. Batista, B.M., Moreno-Perez, J.A. and Moreno-Vega, J.M. (2006). “Solving feature subset selection problem by a Parallel Scatter Search”. European J. of Operations Research, 169(2),477-489.

[10] Ron Kohavi, R. and John, G. H. (1997). “Wrappers for feature subset selection”. Artificial Intelligence. 97(1-2), 273-324.

[11] Richman, M. B. (1986). “Rotation of principal components”. International J. of Climatology, 6, 293-335.

[12] Bureau of Meteorology, http://www.bom.gov.au/climate/ahead/

[13] Ramsay, H.A., Richman, M.B. and Leslie, LM. (2014). “ Seasonal Tropical Cyclone Predictions Using Optimized Combinations of ENSO Regions: Application to the Coral Sea Basin”, J. Climate, 27, 8527-8542.