Relating Precipitation to Climatic Factors at Central Queensland, Australia: A Data Engineering Approach

Tasadduq Imam[1] and Kevin Tickle[2]

[1]CINS, CQUiniversity, Rockhampton, QLD 4702, Australia. t.imam@cqu.edu.au

[2] CINS, CQUiniversity, Rockhampton, QLD 4702, Australia. k.tickle@cqu.edu.au

Abstract

This paper contributes a data engineering framework to relate precipitation at Central Queensland in Australia to other climatic factors and ENSO. Advanced data engineering concepts including computational intelligence techniques are used to model precipitation characteristics for areas within the region. A seasonal stratification process based on standardized precipitation index, predictor selection based on mutual information, a multiple imputation technique and a computational intelligence based approach to examine the influence of ENSO have been demonstrated. An ensemble based regression approach has also been highlighted to characterize the relation between predictors and precipitation. Results indicate that a data engineering framework is effective in unraveling the inter-relationships between different factors and precipitation, and characteristics of the relation vary spatially. The outcomes are expected to aid design of regional forecasting model and relevant statistical downscaling.

Keywords—Data Engineering, ENSO, Precipitation, Regression, Regional Climate Modelling

1. Introduction

While the early works on climate change have focused on variations at global contexts, better understanding of changes at regional scale has been strongly felt in recent years (Solomon et al., 2008). Traditionally employed General Circulation Models (GCM) are not well-suited to predict climate at regional scale due to high computational complexity and limitations in conceptualizing events (Wetterhall et al., 2009; Xu, 1999). Statistical downscaling was designed to overcome these difficulties. The method comprises selecting a set of predictors from GCM simulated data and determining the value of a predictand based on these variables (Chen et al., 2010; Wetterhall et al., 2009; Wilby et al., 1999). The underlying objective of this method is to model the relation between predictand and predictors. Information about the relationship is then employed in developing a regional climate forecast model. In this article, we focus on analyzing precipitation characteristics at Central Queensland in Australia. A data engineered framework is developed to conceptualize the characteristics. However, in contrast to traditional downscaling process that works on GCM simulated data, we concentrate on deriving the relation between precipitation and other factors (i.e., predictors) from real observation data. GCM models have restrictions in projecting climate and are subject to imperfection due to high degree of uncertainty (Knutti, 2008). In recent times the GCM simulated outputs have also come under strong skepticism (Schiermeier, 2010). So we consider actual observed data in our research. More particularly, precipitation characteristics for different areas within Central Queensland are perceived and modeled using a data driven approach. Further to climatic factors, we also investigate the influence of region external factors like El Niño/Southern Oscillation (ENSO). Also, an automated method for seasonal stratification of data, a missing value imputation technique and a computational indigence based modeling process is demonstrated. The outcomes from this research are expected to aid designing a climate forecasting model for the region in a later research. In the following section (Section 2), we first provide an overview of the climatic issues at Central Queensland, followed by which, we present details on the data engineered research framework employed in our research (Section 3). Finally, Section 4 provides a summary of the findings and points to future directions.

2. Climatic Issues at Central Queensland

Central Queensland, Australia is an industrially booming and economically highly potential region, accommodating a number of industries including one of the world's largest alumina refineries at Gladstone, major Australian meat processors at Rockhampton, Queensland's largest port and largest coal fired power station at Gladstone, and sugar industries at Mackay (OCC, 2009). The region is also notable for a number of tourist attractions including parts of the Great Barrier Reef, several national parks, state forests and beaches. Further, the

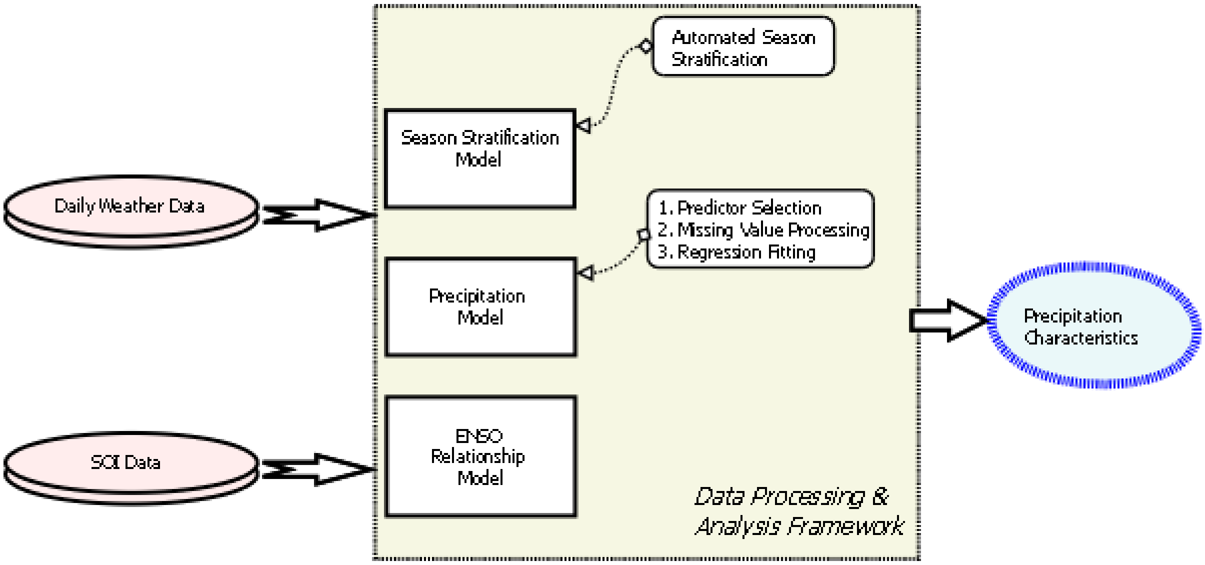
Automated Season Stratification
Season Stratification Model
Daily Weather Data
1. Predictor Selection
2. Missing Value Processing
3. Regression Fitting
Precipitation Model
Precipitation Characteristics
ENSO Relationship Model
SOI Data
Data Processing & Analysis Framework

Figure 1. Research Framework for Precipitation Analysis

region is covered by a number of major rivers including the Fitzroy, Pioneer, Plane and Waterpark rivers. With increasing industrialization, potential impact on its climate is, hence, a growing area of concern for relevant authorities and communities (DERM, 2006; DERM, 2008). The latest comprehensive works in this regard (CSIRO, 2007; OCC, 2009) have projected the region's climate through an ensemble of a set of circulation models. But a comprehensive investigation of the area's climatic features, in particular precipitation properties, is yet lacking. Daily weather records for the area are maintained by the Bureau of Meteorology (BoM, 2009), but insightful analysis employing advanced data engineering technique like computational intelligence are still missing. This study will address these issues in the context of precipitation characterization for Central Queensland.

3. Data Engineering Research Framework

Data engineering (Wolkenhauer, 2001) constitutes conceptualizing and modeling of a system through processing the relevant data (including analysis by statistical methods and computational intelligence algorithms). In our research, we employ a data engineering based research framework as outlined in Figure 1. The data used in our research include the daily weather data from climate stations (i.e., active weather monitoring stations with long term record) and the monthly Southern Oscillation Index (SOI). As per the Bureau of Meteorology (BoM), there are 3 climate stations in Central Queensland, respectively located at Rockhampton (latitude: -23.3753, longitude: 150.4775), Gladstone (latitude -23.8553, longitude: 151.2628) and Mackay (latitude: -21.1172, longitude: 149.2169). So, we focus on these 3 regions due to the availability of long period data. The time span covered by these stations, however, vary. To ensure a consistent time scale across the regions, weather records spanning the period Jan 1960 – Oct 2009 are used in our analysis. Three data driven models are developed. A novel algorithm for seasonal stratification is designed to categorize the months, using precipitation characteristics, across the different seasons. Regression based modeling is then performed, along with relevant data analysis including predictor selection and missing value processing, to model precipitation. Lastly, a computational intelligence approach is employed to relate ENSO to precipitation. The outcome of the research framework is a knowledgebase characterizing precipitation at Central Queensland.

3.1 Seasonal Stratification

Climate analysis at a geographical location is generally performed in terms of seasons identified from historical experiences. Researchers have recognized four seasons for Australia - summer (Dec-Feb), autumn (Mar-May), winter (Jun-Aug) and spring (Sep-Oct) (Hennessy et al., 1999). For precipitation analysis, two seasons are often considered- dry and wet season (Wetterhall et al., 2009). However, season structure varies from places to places. Also, it has been noted that seasons understood from historical experiences may not match seasons perceived from observed data (Tripathi et al., 2006; Winkler et al., 1997). In this article, we employ an algorithmic approach to determine the seasons at Central Queensland automatically from data. The pseudo-code for the algorithm is illustrated in Algorithm 1. An outstanding feature of the algorithm is the use of Standardized Precipitation Index (SPI) to characterize the months. SPI is a relatively modern and well popular measure of drought conditions and is calculated using the probabilistic distribution of precipitation (Guttman, 2007; McKee et al., 1993; McKee et al., 1995). The measure is well suited to compare precipitation across spatial and temporal boundaries, and has been applied to a number of scenarios including flood risk monitoring (Seiler et al., 2002) and drought probability detection (Türke et al., 2009). Use of SPI in characterizing months for automatic season detection, to the best of our knowledge, is novel. The Algorithm 1 first calculates SPI for each of the months of every year from the total precipitation during the month. The SPI values are then discretized into 3 intervals: [-∞,-0.99), [-0.99,0.99), [0.99, ∞) and the corresponding climatic conditions are regarded as *dry*, *normal* and *wet* conditions respectively. Then, based on the number of years a particular month has sustained *dry*, *normal* and *wet* conditions, the algorithm proceeds to characterize the month as part of *dry* or normal or *wet* season. If the aggregate count for *dry* and *normal* conditions, and the aggregate count for *wet* and *normal* conditions for a particular month both exceed 80% of the total number of years, then *normal* condition is prevalent during that month and it's considered part of *normal* season (i.e., the corresponding season is neither dry nor wet). Otherwise, if the aggregate count for *dry* and *normal* conditions exceed that for *wet* and *normal* conditions, the corresponding month is categorized as part of *dry* season (i.e., occurrence of rainfall is rare). If none of the previous conditions is fulfilled, the month is considered part of *wet* season (i.e., rainfall is frequent). In Figure 2, we present the outcome of seasonal stratification for the three regions in Central Queensland It's notable that for all the three regions, December-March comprises the wet season. While July-August comprises dry season for all three areas, the length and time-span for dry seasons vary across the regions. Also, between *dry* and *wet* seasons, *normal* seasons are observed for each of the areas. April is normal season for all areas, while precipitation characteristics during May, June, October and November vary. For two of the regions, May, October and November are part of the normal season and June is part of the dry season. Overall, we conclude that at Central Queensland, in terms of precipitation characteristics, December-March and June-September are *wet* and *dry* seasons respectively, while between these two seasons, two *normal* seasons appear during April-May and October-November.

Algorithm 1. Pseudo-code for seasonal stratification

```
Let, $D$ be the climate dataset containing daily records including Precipitation.
    1. Calculate monthly total precipitation for each year and each month from $D$
    2. Calculate Standardized Precipitation Index (SPI) for each of the months, considering the monthly total
       precipitation as input to SPI calculation algorithm (Wheatley, 2010)
    3. Partition the SPI values based on 3 intervals: [-∞,-0.99), [-0.99,0.99), [0.99, ∞) and consider these 3 intervals
       indicating dry, normal and wet conditions respectively. Based on this partitioning, assign to each month a
       variable indicting whether it has been dry, normal or wet.  Let these be denoted by $M_{i,y,}$ where $i=1..12$ for
       the 12 months and $y$ are the different years under consideration.
    4. For $i=1..12$:
                Let $c_d$ = count of years for which $M_{i,y,}$ = dry
                Let $c_n$ = count of years for which $M_{i,y,}$ = normal
                Let $c_w$ = count of years for which $M_{i,y,}$ = wet

                Let $S_i$ denotes a variable that indicate the final decision of the algorithm regarding whether
                month $i$ shall be considered as dry, normal or wet

                Let, $sm_1 = c_d + c_n$   and   $sm_2 = c_w + c_n$   and  $sm = c_d + c_w + c_n$

                if ($sm_1/sm > 0.8$)  and ($sm_2/sm > 0.8$)  then $S_i$ = normal
                else:
                    if ($sm_1 > sm_2$)  then $S_i$ = dry
                    else $S_i$ = wet
    5. Return $S_i$ for $i=1..12$
```

▤ dry   ▦ wet   □ normal

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Rockhampton	▦	▦	▦	□	▤	▤	▤	▤	▤	□	▦	▦
Gladstone	▦	▦	▦	□	□	▤	▤	▤	▤	□	□	▦
Mackay	▦	▦	▦	□	□	□	▤	▤	▤	▤	□	▦

Figure 2. Seasonal stratification for the regions in Central Queensland.

3.2 Precipitation Model

This section focuses on the design of a precipitation model (a model that associates precipitation amount to other climatic factors through statistical and computational intelligence based techniques, and thereby conceptualizes the underlying relationship). In statistical downscaling, this is an important step and different techniques like multiple-linear regression, support vector machine and artificial neural network have been explored in this context (Hessami et al., 2008; Tripathi et al., 2006; Wetterhall et al., 2009; Wilby et al., 2002). As mentioned earlier, we use real observational data (instead of a GCM simulated data) in our aim to characterize precipitation at Central Queensland. Development of a precipitation model from this data poses a number of challenges – selecting predictors for analysis, missing value processing and fitting a regression model. We present further details on these under the following headings:

Predictor Selection

The collected daily weather data for the three regions in Central Queensland contained a number of redundant or erroneous attributes. Information for some of the variables (for instance, relative humidity) was available at three-hour time resolution only and a daily statistic for these attributes was missing. For some of the attributes, very limited information was available. In our analysis, we ignore the erroneous and very limited information attributes, and compute missing daily statistics from the available three-hourly measures using an aggregation function (for instance, relative humidity for a day is computed by averaging available three hourly observations). Further, we consider that human feeling of climate is an important issue when characterizing climate. So, we compute and include daily average values for apparent-temperature (Steadman, 1979), a stress index indicating the effect of humid and hot conditions on human body. Thus preprocessed dataset contains 28 attributes. In Table 1, we detail the attributes and outline the symbolic notations used to represent these variables for subsequent analysis. From the preprocessed dataset, we analyze the relationship between precipitation and other factors to identify predictors for precipitation modeling. Correlation coefficient is often used in this context (Tripathi et al., 2006; Yin et al., 2009). But the measure is not well-suited to capture non-linear dependence between predictands and predictors. In our research, we use, in addition to correlation coefficient, mutual information to select predictors. Mutual information is a measure motivated from information theory and is capable of representing non-linear dependence (Veyrat-Charvillon and Standaert, 2009). For predictor selection, we employ a combination of Spearman correlation coefficient and mutual information. Attributes having correlation coefficient of at least |0.4| and the five top attributes in terms of mutual information with respect to precipitation are considered as the features having moderate to strong statistical relationship with precipitation. The chosen predictors are union of these two sets of attributes. In Table 2, we show the predictors identified for the three areas across the three seasons. It's notable that, the number and characteristics of the predictors for the three areas vary. However, daily average for total cloud amount, low cloud amount and relative humidity are positively correlated to precipitation for all three areas and seasons, while solar exposure has statistical association for the wet season. Attributes like wind speed and direction and various air pressure statistics, that are often identified as potential drivers of precipitation at varied geographical location and in statistical downscaling of GCM (Barry and Chorley, 2003; Chang, 2006; Wetterhall et al., 2009), appear to have limited impact on precipitation at Central Queensland.

Missing Value Processing

A challenging issue in our research is the occurrence of missing values for varied attributes and observations. Missing values pose significant issue in data analysis and different methods for imputing missing values have been proposed in literature (Harel and Zhou, 2007). But it's unclear which missing value imputation method is particularly suitable for multivariate temporal data like the climate dataset used in our research. To handle missing values in our experiments, we employ a multiple imputation strategy. The strategy is outlined in Algorithm 2. Considering values for each of the predictors as a time-series, missing values are imputed using two techniques-Amelia-II (Honaker et al., 2010), a multiple imputation process, is executed to generate 5 imputed time-series and two cubic spline interpolation techniques are used to generate two other time series. Amelia-II utilizes a robust method for multiple imputation, while cubic splines allow single imputation by fitting a well understood mathematical structure. The noteworthy feature of the Algorithm 2 is combination of the seven missing value imputed time series into a single time series. For each of the imputed time series, mutual information between the predictor and precipitation is calculated and the one with maximum mutual information replaces original measurement values of the predictor. The rationale behind the algorithm is: multiple imputation results in multiple possible values for missing data and thereby decreases uncertainty, while combination based on maximum mutual information reduces inter-independence between predictors and predictand and is expected to lead to better fitting for the regression model detailed under the next heading.

Table 1. List of attributes for the pre-processed dataset.

Attribute Symbol	Physical Interpretation
SE	Solar exposure in MJ/m^2
Prec	Precipitation in mm.
MaxT, MinT, AvgT	Maximum, minimum and average temperature in ^{o}C
MaxDwT, MinDwT, AvgDwT	Maximum, minimum and average dew point temperature in ^{o}C
MaxWbT, MinWbT, AvgWbT	Maximum, minimum and average wet bulb temperature in ^{o}C
HrBrSun	Span of bright sunshine in hours
MaxWS, AvgWS	Maximum and average of daily wind speed in km/h
DirMaxWs	Direction of maximum wind flow in degrees
AvgHpSeaL, AvgHpStnL	Average daily mean sea level and station level pressure in hPa
MaxHpVp, MinHpVp, AvgHpVp	Maximum, minimum and average vapor pressure in hPa
MaxHpSat, MinHpSat, AvgHpSat	Maximum, minimum and average saturated vapor pressure in hPa
AvgClA, AvgLoClA	Average total cloud amount and low cloud amount in eighths
AvgVis	Average visibility in km.
AvgR	Average relative humidity in percentage
AvgAT	Average apparent temperature in ^{o}C

Table 2. Predictors for the three areas. Predictors are sorted on correlation coefficient and mutual information in descending order respectively. * marked predictors have negative correlation.

Area	Season	Predictors Correlated with coefficient of at least |0.4|	Predictors with top statistical dependence indicated by mutual information
Rockhampton	dry	*AvgLoClA*	*AvgR, AvgLoClA, AvgClA, MaxHpVp, MaxDwT*
	normal	*AvgLoClA, AvgR, AvgClA*	*AvgLoClA, AvgR, AvgClA, AvgDwT, AvgHpVp*
	wet	*AvgR, AvgClA, AvgLoClA, SE*[*]	*AvgR, AvgClA, AvgLoClA, SE, MinHpVp*
Gladstone	dry	*AvgLoClA*	*AvgLoClA, AvgClA, AvgR, MaxHpVp, MaxDwT*
	normal	*AvgClA, AvgLoClA*	*AvgLoClA, AvgClA, AvgR, SE, AvgVis*
	wet	*AvgR, AvgClA, AvgLoClA*	*AvgR, AvgClA, AvgLoClA, SE, MaxSatVp*
Mackay	dry	*AvgR, AvgClA, AvgLoClA, HrBrSun*[*]	*AvgLoClA, AvgR, HrBrSun ,AvgClA, MinWbT*
	normal	*AvgClA, AvgLoClA, AvgR, SE*[*], *HrBrSun*[*]	*HrBrSun, AvgLoClA, AvgClA , SE, AvgR,*
	wet	*AvgLoClA, AvgClA, AvgR, MaxWS, MaxT, MaxSatVp, AvgVis*[*], *SE*[*], *HrBrSun*[*]	*AvgLoClA, AvgClA, HrBrSun , SE, AvgR*

Algorithm 2. Pseudo-code for missing value estimation

```
Let, $D$ be the climate dataset
    1. Partition $D$ into $D_{dry}$, $D_{normal}$ and $D_{wet}$ subsets based on dry, normal and wet season respectively. Let these partitions contain precipitation records (denoted by Prec), and predictors (denoted by $V_k$) having moderate to strong correlation or high mutual information, as indicated in Table 2, for the respective seasons
    2. Let, $i \in \{dry, normal, wet\}$
    3. For each $D_i$ :
            For each predictor $V_k$:
                Set, $M_{1..5}$ = 5 missing value imputed time series for $V_k$ generated by applying Amelia-II

                Set, $M_6$ and $M_7$ be two time series with missing values generated by cubic interpolation using (Forsythe et al., 1977) and natural splines on $V_k$

                Set $V_k' = arg_j$ (maximizing mutual information between $M_j$ and Prec)

            Set $D_i'$ = the dataset comprising $V_k'$ as predictors and Prec
    4. Return $D_i'$ for $i \in \{dry, normal, wet\}$
```

Fitting Regression Model

With predictors chosen and missing values imputed in the previous steps, we, here, reflect on fitting a regression model to daily rainfall quantity and thus establish a relationship between precipitation and other climatic factors. Different researchers have used different regression strategies and methods in this context (Tripathi et al., 2006; Wetterhall et al., 2009; Wilby et al., 2002). It's, however, unclear which method is particularly suitable for data from a geographical location. In this research, we consider two popular techniques - multiple linear regression (Weisberg, 2005) and Support Vector Regression (Fan et al., 2005; Vapnik, 2000). Multiple linear regression (MLR) comprises fitting a linear model to the predictand and the objective is to estimate coefficients for the predictors that best fits the observed values. Support Vector Regression (SVR), on the other hand, is a theoretically solid technique that takes as input a regularization parameter and a kernel function, and produces model with high generalization capability. To fit a regression model to the precipitation information at Central Queensland, we use both of the methods and compare the performance. Experiments are conducted for each season at each of the areas (thus, we work with 9 datasets, representing the 3 seasons at the 3 areas). Predictors are chosen based on the predictor selection process outlined earlier. We experiment with two different versions of the same data. In one version all incomplete records are removed, while the other version of data comprises missing information imputed using Algorithm 2. For SVR regression we use RBF kernel, which is well matched for representing non-linear relationship. For each of the 9 datasets, we randomly select 90% of the records as training samples and report performance in terms of root mean squared error (RMSE). The training parameters for SVR are optimized using 10-fold cross-validation. In addition to MLR and SVR, we test the suitability for an equal-weighted ensemble of the two techniques. The experiment outcomes are detailed in Table 3. A particular characteristic notable from the outcome is the suitability of the ensemble approach. For datasets with missing values imputed, the ensemble method results in the least RMSE (i.e., better generalization) for six datasets. However, for datasets with incomplete information removed, SVR wins in five cases, while the ensemble method is better for four datasets. The ensemble method is also the second-best for datasets in which it is not the best-performing strategy. MLR, noticeably, has performed relatively poor compared to the other methods. Overall, we conclude that for the areas in Central Queensland, an ensemble model comprising equal weights for a MLR based model and a SVR model is the well-suited regression model in relating precipitation to other climatic factors. Symbolically the model is represented as follows:

$$Y_E = 0.5 \bullet Y_L + 0.5 \bullet Y_S^*. \tag{1}$$

Y_L is the prediction from MLR model and Y_S^* is the prediction from SVR model, where the SVR model is trained using RBF kernel and optimal learning parameters derived by cross-validation.

3.3 Influence of ENSO

El Niño/Southern Oscillation (ENSO) is a global climatic event that influences precipitation at several localities including Australia (Holbrook et al., 2009). In this subsection, we explore the impact of this event at the areas in Central Queensland. Values of monthly SOI, an index representing the impact of ENSO (Ropelewski and Jones, 1987; Trenberth, 1984), is associated to the monthly total precipitation. Apriori (Agrawal et al., 1996), a well known rule mining algorithm, is employed to unravel the association. The results are highlighted in Table 4. Only the three rules with top confidence values (i.e., the most likely rules) are included. We note the differing influence of SOI for the different regions. In Rockhampton and Gladstone, influence of SOI is similar for both the dry and normal season. During wet season Rockhampton has sustained rainfall when SOI values are in the range [10,30). Gladstone, on the other hand, has experienced notable rainfall during wet seasons for SOI values in the range [5,30). This implies that precipitation at Gladstone, during wet season, is more sensitive to positive values of SOI than Rockhampton. A similar notable characteristic is observed for Mackay. We observe that precipitation in this area is impacted by SOI values in the range [5,30) for all the seasons, implying that precipitation characteristics at Mackay is very sensitive to SOI. Another noteworthy observation is SOI in the range [-30,-5) has led to the least precipitation for all the areas, implying drought condition in the region is effected by highly negative values for SOI.

4. Summary and Conclusion

In this article, we have characterized precipitation at Central Queensland from different perspectives. A research framework outlining a season stratification approach using standardized precipitation index, a predictor selection strategy, a missing value imputation technique and a regression model relating precipitation to other climatic factors have been contributed. Further, seasonal influences of ENSO have been examined. Overall, we observe that Central Queensland sustains there seasons in terms of precipitation characteristics with Dec-Mar being the wet season. Predictors related to cloud amount, relative humidity and solar exposure principally influence precipitation in the region. An equal weighted ensemble of linear and SVR regression models explain well the underlying relation between precipitation and predictors. Also, areas within the region are influenced by ENSO differently, with precipitation at Mackay being more sensitive to positive SOI. In a future research, we shall focus on designing a regional forecast model using the framework developed and the characteristics discovered.

Acknowledgements

The authors would like to thank the Bureau of Meteorology, Australia for providing the climate datasets.

References

Agrawal R., Mannila H., Srikant R., Toivonen H., Verkamo A. (1996) Fast Discovery of Association Rules. Advances in Knowledge Discovery and Data Mining 12:307-328.

Barry R., Chorley R. (2003) Atmosphere, Weather, and Climate. Routledge.

BoM. (2009) Australia's Reference Climate Station Network, Bureau of Meteorology. URL: http://www.bom.gov.au/climate/change/reference.shtml.

Chang M. (2006) Forest Hydrology: An Introduction to Water and Forests. CRC Press.

Chen H., Guo J., Xiong W., Guo S., Xu C. (2010) Downscaling GCMs Using the Smooth Support Vector Machine Method to Predict Daily Precipitation in the Hanjiang Basin. Advances in Atmospheric Sciences 27:274-284.

CSIRO. (2007) Climate Change in Australia, CSIRO Tech Report. URL: http://climatechangeinaustralia.gov.au/technical_report.php.

DERM. (2006) Central Queensland Regional Water Supply Strategy Report, Department of Environment and Resource Management. URL: http://www.derm.qld.gov.au/water/regionalsupply/central_queensland/report.html.

DERM. (2008) Clean and Healthy Air for Gladstone, Department of Environment and Resource Management. URL: http://www.derm.qld.gov.au/environmental_management/air/clean_and_healthy_air_for_gladstone/index.html.

Fan R., Chen P., Lin C. (2005) Working Set Selection Using Second Order Information for Training Support Vector Machines. The Journal of Machine Learning Research 6:1918.

Forsythe G., Malcolm M., Moler C. (1977) Computer Methods for Mathematical Computations. Prentice-Hall Englewood Cliffs, NJ.

Guttman N. (2007) Accepting the Standardized Precipitation Index: A Calculation Algorithm. JAWRA Journal of the American Water Resources Association 35:311-322.

Harel O., Zhou X. (2007) Multiple Imputation: Review of Theory, Implementation and Software. Statistics in Medicine 26:3057.

Hennessy K., Suppiah R., Page C. (1999) Australian Rainfall Changes, 1910-1995. Australian Meteorological Magazine 48:1-13.

Hessami M., Gachon P., Ouarda T., St-Hilaire A. (2008) Automated Regression-Based Statistical Downscaling Tool. Environmental Modelling & Software 23:813-834.

Holbrook N., Davidson J., Feng M., Hobday A., Lough J., McGregor S., Risbey J. (2009) El Niño–Southern Oscillation. A Marine Climate Change Impacts and Adaptation Report Card for Australia 2009:978-1.

Honaker J., King G., Blackwell M. (2010) Amelia: Amelia Ii: A Program for Missing Data, R package version 1.2-15. URL: http://CRAN.R-project.org/package=Amelia.

Knutti R. (2008) Should We Believe Model Predictions of Future Climate Change? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 366:4647.

McKee T., Doesken N., Kleist J. (1993) The Relationship of Drought Frequency and Duration to Time Scales. pp. 184.

McKee T., Doesken N., Kleist J. (1995) Drought Monitoring with Multiple Time Scales. pp. 15-20.

OCC. (2009) Climate Change in the Central Queensland Region, Office of Climate Change. URL: http://www.climatechange.qld.gov.au/__data/assets/pdf_file/0004/23998/CQ_RegSumm_20090713_web.pdf.

Ropelewski C., Jones P. (1987) An Extension of the Tahiti-Darwin Southern Oscillation Index. Monthly Weather Review 115:2161-2165.

Schiermeier Q. (2010) IPCC Flooded by Criticism. Nature 463:596.

Seiler R., Hayes M., Bressan L. (2002) Using the Standardized Precipitation Index for Flood Risk Monitoring. International Journal of Climatology 22:1365-1376.

Solomon S., Qin D., Manning M., Chen Z., Marquis M., Averyt K., Tignor M., Miller H. (2008) Climate Change 2007: The Physical Science Basis. Cambridge University Press Cambridge and New York.

Steadman R. (1979) The Assessment of Sultriness. Part I: A Temperature-Humidity Index Based on Human Physiology and Clothing Science. J Appl Meteorol 18:861-873.

Trenberth K. (1984) Signal Versus Noise in the Southern Oscillation. Monthly Weather Review 112:326–332.

Tripathi S., Srinivas V., Nanjundiah R. (2006) Downscaling of Precipitation for Climate Change Scenarios: A Support Vector Machine Approach. Journal of Hydrology 330:621-640.

Türke, scedil M., Tatl, inodot H. (2009) Use of the Standardized Precipitation Index (Spi) and a Modified Spi for Shaping the Drought Probabilities over Turkey. International Journal of Climatology 29:2270-2282.

Vapnik V. (2000) The Nature of Statistical Learning Theory. Springer Verlag.

Veyrat-Charvillon N., Standaert F. (2009) Mutual Information Analysis: How, When and Why? Cryptographic Hardware and Embedded Systems-CHES 2009:429-443.

Weisberg S. (2005) Applied Linear Regression. Wiley-Blackwell.

Wetterhall F., Bárdossy A., Chen D., Halldin S., Xu C. (2009) Statistical Downscaling of Daily Precipitation over Sweden Using GCM Output. Theoretical and Applied Climatology 96:95-103.

Wheatley J. (2010) Visualizing Drought, Biospherica. URL: http://joewheatley.net/visualizing-drought/.

Wilby R., Hay L., Leavesley G. (1999) A Comparison of Downscaled and Raw GCM Output: Implications for Climate Change Scenarios in the San Juan River Basin, Colorado. Journal of Hydrology 225:67-91.

Wilby R., Dawson C., Barrow E. (2002) Sdsm--a Decision Support Tool for the Assessment of Regional Climate Change Impacts. Environmental Modelling & Software 17:145-157.

Winkler J., Palutikof J., Andresen J., Goodess C. (1997) The Simulation of Daily Temperature Time Series from GCM Output. Part II: Sensitivity Analysis of an Empirical Transfer Function Methodology. Journal of Climate 10:2514-2532.

Wolkenhauer O. (2001) Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis. Wiley-Interscience.

Xu C. (1999) Climate Change and Hydrologic Models: A Review of Existing Gaps and Recent Research Developments. Water Resources Management 13:369-382.

Yin Z., Cai Y., Zhao X., Chen X. (2009) An Analysis of the Spatial Pattern of Summer Persistent Moderate-to-Heavy Rainfall Regime in Guizhou Province of Southwest China and the Control Factors. Theoretical and Applied Climatology 97:205-218.

Table 3. RMSE for three types of model fit on data. Results for both missing data deletion and imputation are shown and the best performing method's statistic is underlined.

		Missing Data Deleted			**Missing Data Imputed**		
		MLR	**SVR**	**Ensemble**	**MLR**	**SVR**	**Ensemble**
Rockhampton	dry	3.17	2.98	2.98	3.53	3.45	3.41
	normal	9.85	9.83	9.77	9.40	9.33	9.30
	wet	7.79	7.18	7.10	8.95	8.65	8.52
Gladstone	dry	6.38	6.52	6.40	7.21	7.27	7.21
	normal	4.30	3.30	3.68	9.49	9.82	9.59
	wet	7.67	3.36	4.87	13.82	13.98	13.67
Mackay	dry	3.41	3.08	3.08	2.75	2.40	2.45
	normal	22.03	21.29	21.58	14.61	14.78	14.60
	wet	18.02	14.83	15.95	20.78	18.65	19.25

Table 4. Influence of ENSO. Rules with the three highest confidence values are shown. The number beside the rule indicates confidence. Symbol *I* and *Prec* denote monthly SOI and monthly precipitation respectively.

Area	**Season**	**Rules**		

Rockhampton	dry	I=[10,30) =>
		Prec=[17.4,228.4] *0.75*
			I=[-30,-5) =>
			Prec=[0.0, 17.4) *0.62*
				I=[0,5) =>
				Prec=[17.4,228.4] *0.60*

	normal	I=[5,10) =>
		Prec=[40.2,303.8] *0.71*
			I=[-30,-5) =>
			Prec=[0.0, 40.2) *0.70*
				I=[10,30) =>
				Prec=[40.2,303.8] *0.64*

	wet	I=[10,30) =>
		Prec=[76.4,660.2] *0.62*
			I=[-30,-5) =>
			Prec=[1.6, 76.4) *0.57*
				I=[-5,0) =>
				Prec=[1.6, 76.4) *0.57*

Gladstone	dry	I=[10,30) =>
		Prec=[23.4,220.3] *0.79*
			I=[-30,-5) =>
			Prec=[0.0, 23.4) *0.70*
				I=[0,5) =>
				Prec=[23.4,220.3] *0.62*

	normal	I=[5,10) =>
		Prec=[44.2,316.4] *0.75*
			I=[-30,-5) =>
			Prec=[0.0, 44.2) *0.70*
				I=[10,30) =>
				Prec=[44.2,316.4] *0.61*

	wet	I=[-30,-5) =>
		Prec=[0.0, 90.3) *0.65*
			I=[10,30) =>
			Prec=[90.3,709.8] *0.62*
				I=[5,10) =>
				Prec=[90.3,709.8] *0.62*

Mackay	dry	I=[5,10) =>
		Prec=[24.3,392.1] *0.59*
			I=[10,30) =>
			Prec=[24.3,392.1] *0.58*
				I=[0,5) =>
				Prec=[0.0, 24.3] *0.56*

	normal	I=[10,30) =>
		Prec=[76.6,545.6] *0.60*
			I=[5,10) =>
			Prec=[0.0, 76.6] *0.58*
				I=[-30,-5) =>
				Prec=[0.0, 76.6] *0.52*

	wet	I=[10,30) =>
		Prec=[191,1159] *0.72*
			I=[-30,-5) =>
			Prec=[0, 191] *0.65*
				I=[5,10) =>
				Prec=[191,1159] *0.62*