Water Resour Manage (2018) 32:1777–1794
https://doi.org/10.1007/s11269-018-1903-y

A Comparative Assessment of Models to Predict Monthly Rainfall in Australia

Adil M. Bagirov[1] · **Arshad Mahmood**[1]

Received: 7 February 2017 / Accepted: 3 January 2018 /
Published online: 17 January 2018

Abstract Accurate rainfall prediction is a challenging task. It is especially challenging in Australia where the climate is highly variable. Australia's climatic zones range from high rainfall tropical regions in the north to the driest desert region in the interior. The performance of prediction models may vary depending on climatic conditions. It is, therefore, important to assess and compare the performance of these models in different climatic zones. This paper examines the performance of data driven models such as the support vector machines for regression, the multiple linear regression, the k-nearest neighbors and the artificial neural networks for monthly rainfall prediction in Australia depending on climatic conditions. Rainfall data with five meteorological variables over the period of 1970–2014 from 24 geographically diverse weather stations are used for this purpose. The prediction performance of each model was evaluated by comparing observed and predicted rainfall using various measures for prediction accuracy.

Keywords Rainfall prediction · Prediction models · Regression analysis · Prediction performance

1 Introduction

Rainfall is the most important hydro-climate variable because of its significance for sustainable water management (Chowdhury and Beecham 2013). The accurate prediction of water

✉ Adil M. Bagirov
a.bagirov@federation.edu.au

Arshad Mahmood
a.mahmood@federation.edu.au

[1] Faculty of Science and Technology, Federation University Australia, Ballarat, Australia

availability is immensely beneficial for making management decisions and can assist with the sustainable operation of water resource systems.

Rainfall prediction is a challenging task because of the dynamic nature of climate phenomena and random fluctuations involved in the physical process. Such a prediction is particularly challenging in Australia where in a long-term analysis the rate of change in the frequency and intensity of rainfall extremes can often be greater than the rate of change for average rainfall (Garnaut 2008).

Rainfall prediction models can be classified into two categories: physical and data driven models. Physical models use the physical laws to model the relevant processes that contribute to rainfall process. Data driven models use historical data to make predictions. The most frequently applied models are Multiple Linear Regression (MLR) (Chattopadhyay et al. 2010; Mekanik et al. 2013), Artificial Neural Networks (ANNs) (Abbot and Marohasy 2012, 2014; Chattopadhyay et al. 2010; Mekanik et al. 2013), and k-Nearest-Neighbours (k-NN). Studies have shown that data driven models, in general, give better results than the physical models (Abbot and Marohasy 2012, 2014).

Rainfall predictions are made for some time periods which include weekly, monthly and seasonal predictions. In rainfall prediction, the month is used to define the start, duration and end of the rainy season. Moreover, monthly rainfall data provide more accurate an intra-year rainfall distribution than seasonal rainfall data (Omotosho et al. 2000). Such information may help to significantly improve decisions with regard to irrigation needs and their timings and also decisions on water conservation strategies for dams and on operation of water infrastructure (Omotosho et al. 2000; Sharma et al. 2013). Accurate prediction of streamflow a month ahead is essential information to help water resource managers for efficient planning (Wang et al. 2011).

Australia's climate is highly variable having high rainfall tropical regions in the north and the driest desert region in the interior. The performance of prediction models alter depending on climatic zones. Therefore, comparative assessment of models depending on meteorological variables and climatic conditions is important. Such a comparison may help a decision maker to choose appropriate models depending on input variables and climatic zones. There are several papers on the comparison of rainfall prediction models (see, for example, Abbot and Marohasy 2012; Aksoy and Dahamsheh 2009; Mekanik et al. 2013). To the best of our knowledge the comparison of models depending on meteorological variables and climatic zones has not been studied.

The aim of this paper is a comparative assessment of the performance of various data-driven prediction models depending on meteorological variables and climatic conditions. These models include linear SVMReg, SVMReg with RBF kernel, MLR, k-NN and ANNs with one hidden and without hidden layer. Rainfall data with five meteorological variables (maximum and minimum temperatures, evaporation, vapour pressure and solar radiation) over the period of 1970 – 2014 from 24 geographically diverse weather stations across Australia are used for evaluation of models. Prediction performance of models was evaluated by comparing observed and predicted rainfall using performance measures Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Efficiency (CE).

There are several aspects that distinguish this paper from other papers on the comparison of rainfall prediction models. First, in this paper we consider most well-known prediction models whereas other papers consider only very few of them. Second, we use most important meteorological variables as input variables which have not been considered together as input variables to compare rainfall prediction models. Third, the performance of prediction models is compared using data from weather stations distributed over all Australia and located in different climatic zones.

2 Models and Methods

2.1 Support Vector Machines for Regression

Consider the training data $X = \{(x^1, y_1), (x^2, y_2), \ldots, (x^k, y_k)\} \subset R^n \times R$, where x^i is an input vector and $y_i \in R$ is a corresponding output, $i = 1, \ldots, k$, k is the number of observations in the training set. Given an $\varepsilon > 0$, the aim of SVMReg is to find a function $f(x)$ that has at most ε deviation from the targets y_i for all the training data. In the linear SVMReg, the regression function f is written as: $f(x) = w^T x + b$, where $w \in R^n$ is the weight vector, $b \in R$ and T stands for transpose of a vector. w and b are estimated by solving the following minimization problem (Collobert and Bengio 2001; Müller et al. 1997; Smola and Schölkopf 2004):

$$\begin{cases} \text{minimize} & R = \frac{1}{2} w^T w + C \sum_{i=1}^{k} (\xi_i + \xi_i^*), \\ \text{subject to} & y_i - [w^T x^i + b] \leq \varepsilon + \xi_i, \\ & [w^T x^i + b] - y_i \leq \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0. \end{cases} \tag{1}$$

Here $C > 0$ is a penalty parameter which determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. ξ_i and ξ_i^* are slack variables introduced to deal with infeasibility. The linear SVMReg model is extended to the non-linear SVMReg model using kernel functions, for example, the radial basis function (RBF) (see Collobert and Bengio 2001; Müller et al. 1997; Smola and Schölkopf 2004, for details). There have been several applications of the SVMReg model for rainfall prediction, see, for example, Feng et al. (2015), Kisi and Cimen (2012), Lin et al. (2009), Mercer et al. (2013), and Nayak and Ghosh (2013).

2.2 Multiple Linear Regression

MLR is an extension of a simple linear regression method, where two or more independent variables are used to predict one dependent variable through the least squares method. A general form of the MLR model can be presented as: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$, where y is the dependent variable, $x_1, \ldots, x_p$ are independent variables, β_0 is y-intercept, $\beta_1, \ldots, \beta_p$ are regression coefficients of the corresponding independent variables and ε is the noise in the data. MLR models have been used for rainfall prediction in Aksoy and Dahamsheh (2009), Ramirez et al. (2005), and Shukla et al. (2011).

2.3 *k*-Nearest Neighbours Method

The k-NN method is a non parametric statistical pattern recognition procedure, extended to time series prediction in Yakowitz and Karlsson (1987) (see, also Al-Qahtani and Crone 2013).

Next, we briefly describe k-NN for univariate time series. Consider a finite time series $u(t), t = 1, 2, \ldots, m$ without input variables. In the first step, series is transformed into equal length d-dimensional feature vectors: $u^d(t) = (u(t), u(t-1), \ldots, u(t-(d-1)))$. Here $d < m$ is a predetermined integer called embedding dimension and $t \geq d$. In next step either a set of $m - d$ overlapping vectors with $t = (d, d+1, \ldots, m-d)$ or a set of m/d non-overlapping vectors with $t = (d, 2d, \ldots, m-d)$ is defined. These vectors are called d-histories.

In the third step either the distance or correlation between the last observed vector $u^d(m) = (u(m), u(m-1), \ldots, u(m-(d-1)))$ and all d-histories is computed. Here any

distance function can be used, however in k-NN predominantly the Euclidean distance is used. In the fourth step the calculated distances are ranked and k vectors having lowest distance from the target feature vector are selected. Then these k feature vectors are used for a prediction, often using a simple arithmetic mean with equal weights. In case of correlations, k vectors having highest correlation with the feature vector are used to form a prediction. The k-NN method for univariate time series is extended to multivariate time series by extending construction of vectors for each input variable.

2.4 Artificial Neural Networks

ANNs consist of simple neurons, and links that process information in order to find relationship between inputs and outputs. ANNs take input, apply the activation function to combine the input into a single value and to produce an output. The activation function generally consists of the combination and the transfer functions. The combination function assigns weights to inputs and combines the weighted inputs in a single value. The transfer function produces an output. The sigmoid, hyperbolic tangent and step functions are widely used as the transfer functions. There exist many algorithms to train neural networks, but the back propagation algorithm and its variations are the most computationally efficient (see, for example, Haykin (2001) for more details). The application of ANNs for rainfall prediction can be found in Abbot and Marohasy (2012), Abbot and Marohasy (2014), Aksoy and Dahamsheh (2009), Awan and Bae (2014), Karamouz et al. (2008), Lorrai and Sechi (1995), Mekanik et al. (2013), Ramirez et al. (2005), and Shukla et al. (2011).

3 Data

Historical monthly rainfall data was taken from the Scientific Information for Land Owners (SILO) available at www.longpaddock.qld.gov.au/silo/. SILO is an enhanced climate database hosted by the Queensland Government Department of Science, Information Technology and Innovation. The data is reliable and quality checked.

There are six major climatic zones in Australia: temperate, grassland, desert, tropical, subtropical and equatorial (Australian weather and seasons 2013). We selected two weather stations from the tropical zone, two from subtropical, five from desert, seven from temperate and eight from grassland zones. The number of stations depends on areas of zones. The equatorial zone is not considered as its area is small.

We used data of six meteorological variables from 24 weather stations for the period January 1970 - December 2014 to develop prediction models. Meteorological variables used in this study are: Monthly rainfall, Maximum temperature (TMax), Minimum temperature (TMin), Evaporation (Evap), Vapour pressure (VP), and Solar radiation (Rad). These variables were selected because they are interdependent and influence precipitation. There are 540 records for each weather station. The geographic details as well as climatic zones of these stations and descriptive statistics of the monthly rainfall are given in Table 1 and a location map is given in Fig. 1. The average monthly rainfall varies across these sites from 15.07 mm to 125.87 mm.

Correlations between rainfall and input variables for each weather station are given in Table 2. In this table for each station we also present the number of high (H) (between -1 and -0.5 and between 0.5 and 1), medium (M) (between -0.5 and -0.3 and between 0.3 and 0.5), low (L)(between -0.3 and -0.1 and between 0.1 and 0.3) correlations and the number of no correlations (N) (between -0.1 and 0.1). In all locations there is at least low

Table 1 Geographic details, climatic zones, elevation (m.), minimum, maximum and average of monthly rainfall values for weather stations

Station name	Classification	Latitude	Longitude	Elev.	Min.	Max.	Mean
Victoria							
Annuello	Grassland	−34.85	142.78	52	0.00	261.40	27.20
Dookie	Temperate	−36.37	145.70	185	0.00	227.60	47.23
Orbost	Temperate	−37.69	148.46	41	1.90	425.00	72.06
Cape Otway	Temperate	−38.86	143.51	82	1.60	218.20	77.80
New South Wales							
Warren	Grassland	−31.50	147.69	192	0.00	258.80	41.49
Yamba	Subtropical	−29.43	153.36	27	0.10	629.40	125.87
Moss Vale	Temperate	−34.54	150.38	675	0.40	527.00	75.60
Wilcannia	Desert	−31.56	143.37	75	0.00	252.30	24.31
Queensland							
Palmerville	Tropical	−16.00	144.08	203	0.00	813.20	89.69
Richmond	Grassland	−20.73	143.14	211	0.00	664.20	42.46
Boulia	Desert	−22.91	139.90	161	0.00	464.90	22.74
Fairymead	Subtropical	−24.79	152.36	5	0.00	1143.40	88.22
Northern Territories							
Katherine	Tropical	−14.46	132.26	106	0.00	937.70	91.43
Newery	Grassland	−16.05	129.26	101	0.00	810.00	69.17
Henbury	Desert	−24.55	133.25	432	0.00	376.10	22.68
Alexandria	Grassland	−19.06	136.71	274	0.00	565.40	37.76
South Australia							
Marree	Desert	−29.65	138.06	50	0.00	203.30	15.07
Blinman	Grassland	−31.09	138.68	615	0.00	294.20	28.86
Koppio	Temperate	−34.41	135.82	173	0.00	204.00	42.92
Port Elliot	Temperate	−35.53	138.69	10	0.00	152.60	41.14
Western Australia							
Ningaloo	Grassland	−22.70	113.67	10	0.00	293.60	20.85
Wiluna	Desert	−26.59	120.23	521	0.00	271.60	24.29
Dowerin	Grassland	−31.19	117.03	273	0.00	171.20	28.52
Peppermint Grove	Temperate	−34.44	119.36	60	0.00	308.60	58.75

correlation between rainfall and some input variables. In most locations, more specifically, in 20 out of 24 there is at least one medium correlation. Finally, in 14 out 24 locations there are high correlations between rainfall and some input variables. These observations justify the use of meteorological variables for rainfall prediction.

4 Implementation and Evaluation of Models

Statistical package R-Version 3.2.2 is used to implement all models. R is an environment for statistical computing and graphics including time series analysis, clustering, classification, modeling and statistical tests (R Core Team 2013).

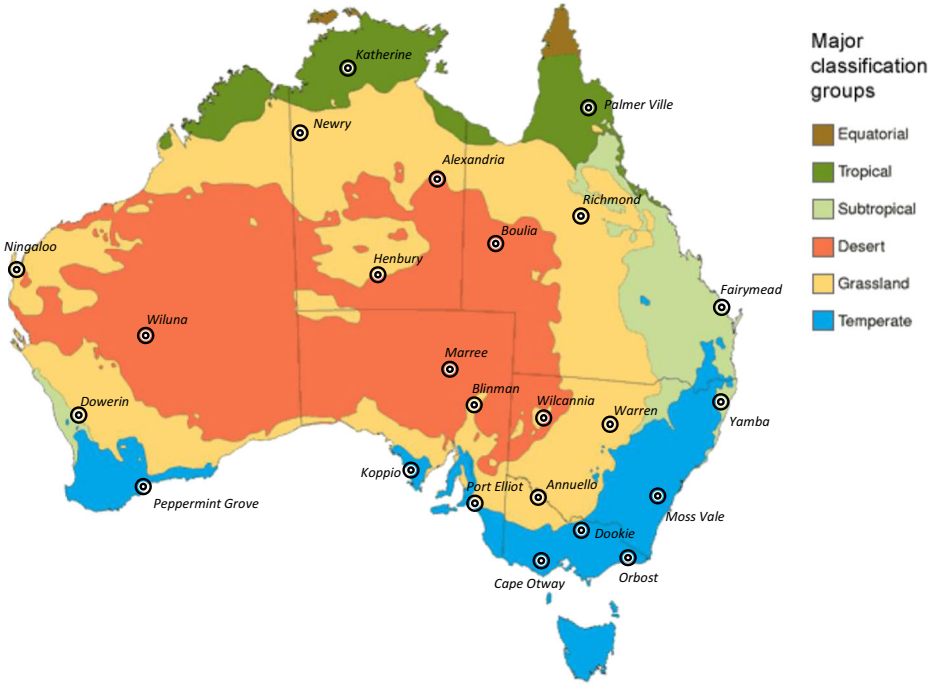

Fig. 1 Location map

We use the R package *nnet* for ANNs (Venables and Ripley 2002), *kknn* for k-NN (Schliep and Hechenbichler 2016) and *e1071* for SVMReg (Meyer et al. 2015). In implementing k-NN, the most important step is the selection of the number of neighbours. Different values were evaluated ranging from 1 to 12 and the model with the minimum RMSE value was selected. We implement ANNs both without hidden layer and with one hidden layer, linear SVMReg and SVMReg with the RBF kernel function.

Prediction models were developed using all combinations of input variables without repetition. There are total of fifteen such combinations. Then the best combination for each model is selected according to prediction performance measures described in next subsection. All models were developed for each weather station using training data sets consisting of 360 records and evaluated by using test data sets consisting of 180 records.

Prediction performance of models was evaluated by comparing observed and predicted rainfall using three measures of prediction accuracy calculated from the test sets: the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the Coefficient of Efficiency (CE). It is well-known that MAE is less sensitive to outliers than RMSE. The small values of RMSE and MAE indicate small deviations of the predictions from actual observations.

CE, proposed in Nash and Sutcliffe (1970), is a normalized statistic that determine the relative magnitude of the residual variance and data variance. CE ranges from $-\infty$ to 1. An efficiency $CE = 1$ means a perfect prediction. An efficiency of 0 indicates that the model predictions are as accurate as the mean of the observed data and an efficiency $-\infty < CE < 0$ occurs when the observed mean is a better predictor than the model.

Table 2 Correlations between monthly rainfall and input meteorological variables

Station name	TMax	TMin	Evap.	VP	Rad.	H	M	L	N
Temperate zone									
Dookie	−0.23	−0.04	−0.26	0.12	−0.25	0	0	4	1
Orbost	−0.17	0.01	−0.11	0.02	−0.12	0	0	2	3
Cape Otway	−0.17	0.01	−0.11	0.02	−0.12	0	0	3	2
Moss Vale	−0.02	0.16	−0.09	0.22	−0.15	0	0	3	2
Koppio	−0.67	−0.54	−0.63	−0.40	−0.63	4	1	0	0
Port Elliot	−0.64	−0.54	−0.61	−0.42	−0.58	4	1	0	0
Peppermint Grove	−0.62	−0.48	−0.60	−0.39	−0.57	3	2	0	0
Grassland zone									
Annuello	−0.09	0.07	−0.12	0.32	−0.11	0	1	2	2
Warren	0.04	0.22	−0.03	0.45	−0.03	0	1	1	3
Richmond	0.20	0.47	−0.05	0.66	−0.12	1	1	2	1
Newry	0.20	0.56	−0.17	0.72	−0.38	2	1	2	0
Alexandria	0.20	0.45	−0.08	0.67	−0.22	1	1	2	1
Blinman	−0.15	−0.02	−0.18	0.38	−0.19	0	1	3	1
Ningaloo	−0.23	0.00	−0.35	0.22	−0.43	0	2	2	1
Dowerin	−0.52	−0.37	−0.51	−0.10	−0.55	3	1	1	0
Desert zone									
Wilcannia	0.00	0.16	−0.05	0.52	−0.05	1	0	1	3
Boulia	0.18	0.33	0.03	0.61	−0.06	1	1	1	2
Henbury	0.09	0.25	0.01	0.59	−0.06	1	0	1	3
Marree	0.07	0.21	0.03	0.49	0.00	0	1	1	3
Wiluna	0.08	0.23	−0.01	0.51	−0.12	1	0	2	2
Tropical and Subtropical zones									
Yamba	0.17	0.29	−0.14	0.35	−0.21	0	1	4	0
Fairymead	0.34	0.44	0.14	0.50	0.05	1	2	1	1
Palmerville	0.05	0.63	−0.26	0.75	−0.27	2	0	2	1
Katherine	0.13	0.55	−0.39	0.70	−0.57	3	1	1	0

5 Results and Discussion

All models are trained using data from Jan 1970 to Dec 1999 and tested using data from Jan 2000 to Dec 2014 with each combination of input variables in all 24 locations. Negative predicted values were adjusted to zero rainfall before the calculation of performance measures. The best combination of input variables for each model was determined using test data and RMSE and MAE as primary performance measures.

Tables 3 and 4 summarize the prediction performance of models with best combinations of input variables. In tables best results among all models are highlighted in bold.

Results for the temperate zone are presented in Table 3 and illustrated in Fig. 2. These results show that SVMReg(RBF) and ANN(1) models outperform other models. According to all performance measures SVMReg(RBF) provides best predictions at four out of seven stations and all of them are coastal stations. ANN(1) gives best results at Koppio and Dookie. At Moss Vale, these two models demonstrate the similar performance.

Table 3 Prediction performance of models in the temperate and grassland zones

Stations	Measures	SVMReg (linear)	SVMReg (RBF)	ANN (0)	ANN (1)	MLR	$K-$NN
Temperate zone							
Moss Vale	RMSE	46.48	45.68	46.56	**44.69**	46.51	47.48
	MAE	31.03	**29.92**	32.82	31.52	34.72	33.47
	CE	0.32	0.34	0.32	**0.37**	0.32	0.29
Koppio	RMSE	23.01	22.89	22.75	**22.24**	22.75	22.39
	MAE	17.08	16.93	17.37	**16.61**	17.37	17.05
	CE	0.51	0.51	0.52	**0.54**	0.52	**0.54**
Port	RMSE	19.82	**18.71**	19.60	19.62	19.60	19.08
Elliot	MAE	14.99	**14.25**	14.96	15.01	14.96	14.45
	CE	0.51	**0.56**	0.52	0.52	0.52	0.54
Dookie	RMSE	26.57	26.48	26.30	**25.34**	25.91	28.32
	MAE	18.37	18.13	18.64	**17.36**	18.08	19.63
	CE	0.35	0.35	0.36	**0.41**	0.38	0.26
Orbost	RMSE	37.57	**36.71**	**36.70**	41.54	36.68	37.53
	MAE	26.97	**26.83**	28.05	32.59	28.03	29.56
	CE	0.30	**0.33**	**0.33**	0.15	**0.33**	0.30
Cape	RMSE	33.58	**31.27**	32.10	33.32	33.41	32.84
Otway	MAE	25.33	**23.81**	24.80	25.72	25.59	24.95
	CE	0.35	**0.44**	0.41	0.36	0.36	0.38
Peppermint	RMSE	35.80	**35.22**	35.55	36.62	36.41	37.07
Grove	MAE	23.52	**22.78**	24.25	25.62	25.20	24.19
	CE	0.38	**0.40**	0.39	0.35	0.36	0.34
Grassland zone							
Warren	RMSE	28.57	25.13	25.13	**24.88**	27.59	28.13
	MAE	19.77	**18.05**	18.58	18.56	19.85	21.19
	CE	0.50	0.61	0.61	**0.62**	0.53	0.51
Newry	RMSE	84.98	68.68	75.09	69.18	75.10	**67.75**
	MAE	43.20	36.84	40.39	**36.10**	40.31	37.17
	CE	0.50	**0.68**	0.61	0.67	0.61	**0.68**
Alexandria	RMSE	55.62	43.09	40.97	**40.53**	46.38	44.40
	MAE	24.52	**19.82**	20.84	20.28	23.21	22.72
	CE	0.45	0.67	0.70	**0.71**	0.61	0.65
Richmond	RMSE	41.88	34.94	35.81	**34.50**	39.21	35.67
	MAE	25.30	**21.47**	25.66	24.91	26.12	22.27
	CE	0.56	0.69	0.68	**0.70**	0.61	0.68
Blinman	RMSE	21.26	21.59	**21.08**	23.90	**21.08**	23.73
	MAE	**14.52**	14.73	15.02	17.56	15.02	17.56
	CE	0.37	0.35	**0.38**	0.20	**0.38**	0.21
Annuello	RMSE	24.16	24.05	23.27	**23.10**	23.46	23.84
	MAE	13.94	13.93	13.46	**13.32**	13.72	14.87
	CE	0.19	0.20	0.25	**0.26**	0.23	0.21

Table 3 (continued)

Stations	Measures	SVMReg (linear)	SVMReg (RBF)	ANN (0)	ANN (1)	MLR	$K-$NN
Ningaloo	RMSE	32.11	27.43	**25.70**	34.55	29.02	30.29
	MAE	14.09	**12.79**	13.53	18.67	14.54	15.19
	CE	0.32	0.51	**0.57**	0.22	0.45	0.40
Dowerin	RMSE	21.23	20.66	**20.02**	20.67	20.07	21.24
	MAE	13.18	**12.87**	12.91	13.40	12.88	14.13
	CE	0.33	0.37	**0.40**	0.37	**0.40**	0.3

Table 4 Prediction performance of models for monthly rainfall prediction in desert, tropical and subtropical zones

Stations	Measures	SVMReg (linear)	SVMReg (RBF)	ANN (0)	ANN (1)	MLR	*K*NN
Desert zone							
Wilcannia	RMSE	23.43	22.32	21.47	24.73	**21.29**	22.96
	MAE	13.02	12.89	12.77	16.80	**12.31**	13.99
	CE	0.43	0.48	0.52	0.36	**0.53**	0.45
Henbury	RMSE	32.80	27.64	**24.95**	27.92	30.39	29.80
	MAE	16.77	**16.26**	17.48	18.65	18.21	17.67
	CE	0.35	0.54	**0.62**	0.53	0.44	0.46
Boulia	RMSE	29.27	**23.76**	24.34	24.59	28.78	26.20
	MAE	16.86	**14.46**	15.85	15.99	17.11	16.50
	CE	0.42	**0.62**	0.60	0.59	0.44	0.54
Marree	RMSE	**13.26**	13.84	15.23	17.14	14.98	15.12
	MAE	8.73	**8.63**	11.09	10.97	10.86	10.77
	CE	**0.44**	0.39	0.26	0.06	0.28	0.27
Wiluna	RMSE	28.02	29.34	**27.43**	27.69	28.33	27.61
	MAE	20.53	17.37	**17.10**	17.36	17.88	17.64
	CE	0.26	0.19	**0.29**	0.28	0.25	0.29
Tropical and subtropical zones							
Katherine	RMSE	77.19	**60.15**	62.51	64.85	68.26	60.51
(Tropical)	MAE	44.88	**34.26**	38.15	38.28	42.32	36.09
	CE	0.67	**0.80**	0.78	0.77	0.74	**0.80**
Palmerville	RMSE	90.77	71.67	74.43	72.34	83.00	**69.80**
(Tropical)	MAE	54.15	41.03	43.38	42.64	51.02	**40.91**
	CE	0.51	0.69	0.67	0.69	0.59	**0.71**
Yamba	RMSE	77.48	76.48	76.14	**75.11**	76.06	78.15
(Subtropical)	MAE	**54.57**	54.65	56.20	55.25	56.64	60.53
	CE	0.34	0.36	0.37	**0.38**	0.37	0.33
Fairymead	RMSE	105.46	99.81	98.59	98.00	101.59	**91.28**
(Subtropical)	MAE	48.85	47.64	45.26	51.61	49.71	**47.04**
	CE	0.19	0.27	0.29	0.30	0.25	**0.39**

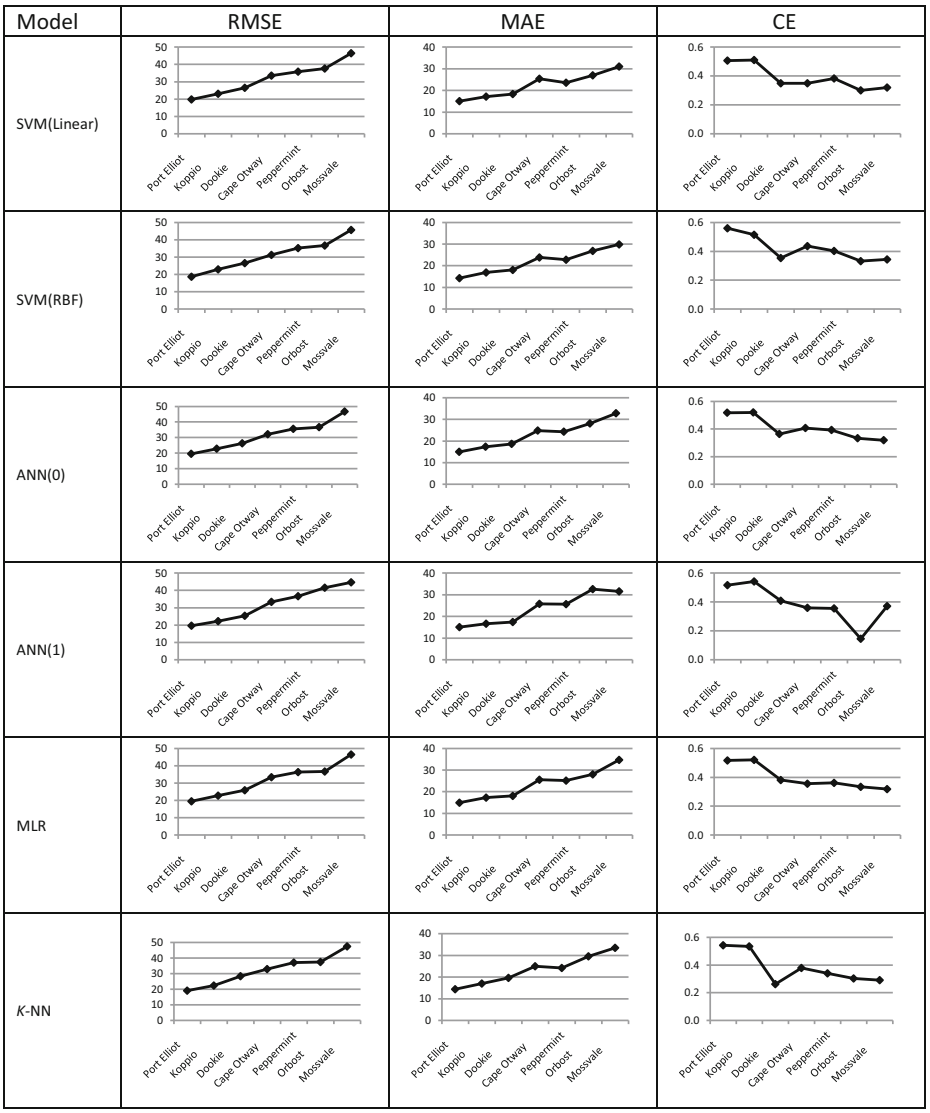

Fig. 2 Graphical display of the performance of models in the temperate zone

The best predictions with SVMReg(RBF) for Port Elliot and Cape Otway had inputs TMax, TMin, Evap and Rad; for Peppermint Grove TMax, TMin and VP; for Moss Vale TMax, TMin, Evap and VP; while for Orbost the best combination was the full set of five variables. The most accurate predictions with ANN(1) model for Koppio had TMax and TMin as inputs; for Dookie Rad was the only input variable; while for Moss Vale the best combination was the full set of five variables.

Results presented in Fig. 2 show that according to RMSE and MAE all models provide best predictions at Port Elliot and worst predictions at Moss Vale. CE indicates that all

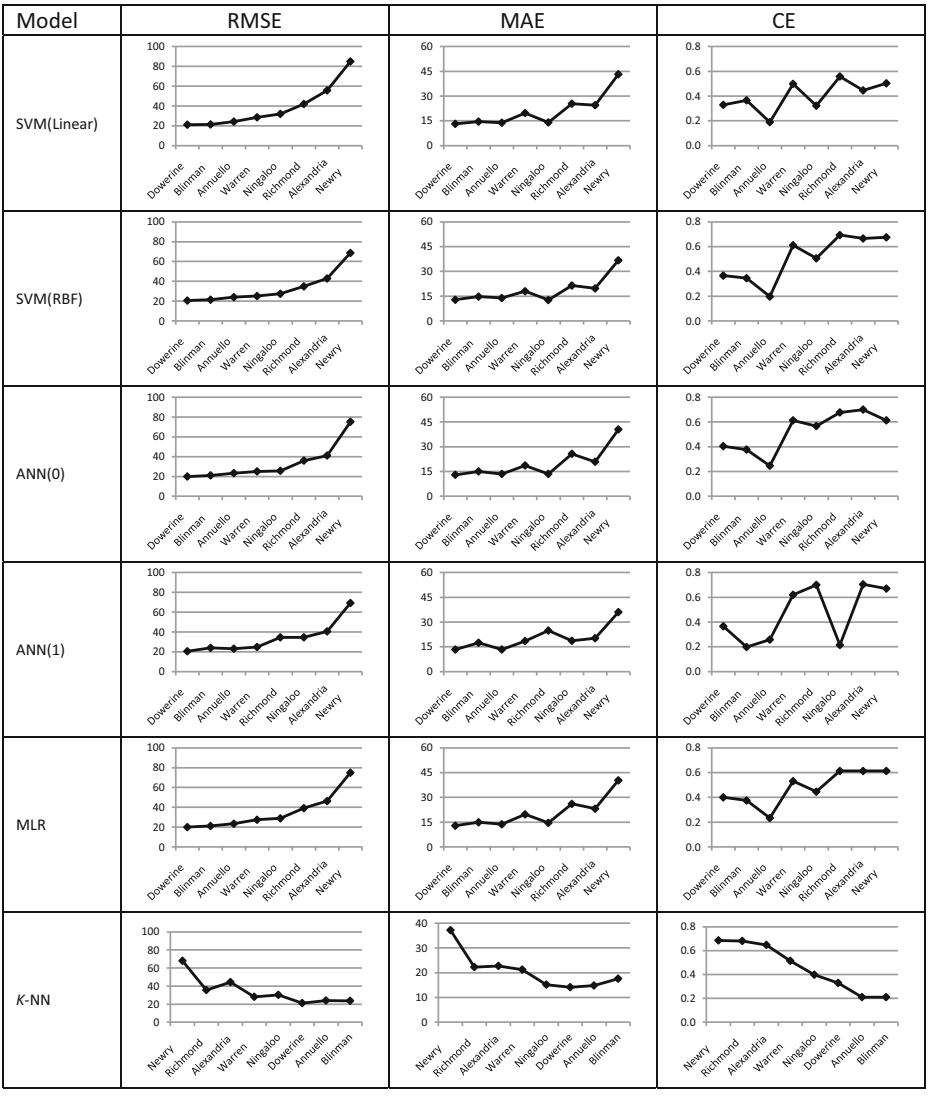

Fig. 3 Graphical display of the performance of models in grassland zone

models performed well at Port Elliot and Koppio, while worse at Orbost and Moss Vale. Models failed to predict extreme rainfall values at all locations.

Table 3 presents monthly rainfall prediction results and Fig. 3 illustrates the performance of models in the grassland zone. For this zone we also include a visual comparison of observed and predicted rainfall over the test period which is given in Fig. 6.

These results show that the SVMReg(RBF) and ANN(1) models outperform other models at most locations. However, the performance of other models are not always significantly different from that of SVMReg(RBF) and ANN(1). At least one performance measure indicates that ANN(0) is the best at three locations; MLR at two; SVMReg(linear) and $k-$NN at one location.

Results presented in Fig. 3 show that the performance measures RMSE and MAE give different results than CE. RMSE and MAE indicate that with respect to some tolerance all models have the lowest prediction error at Dowerin and the highest prediction error at Newry. According to CE all models, except ANN(1), provide predictions with the lowest error at Richmond and the highest error at Annuello. The ANN(1) model predictions have the lowest error at Ricmond and Alexandria and the highest error at Blinman. Figure 6 demonstrates that all models follow the series patterns at Newry and Alexandria, however, this is not true for Warren. Models failed to predict extreme rainfall values at all three locations.

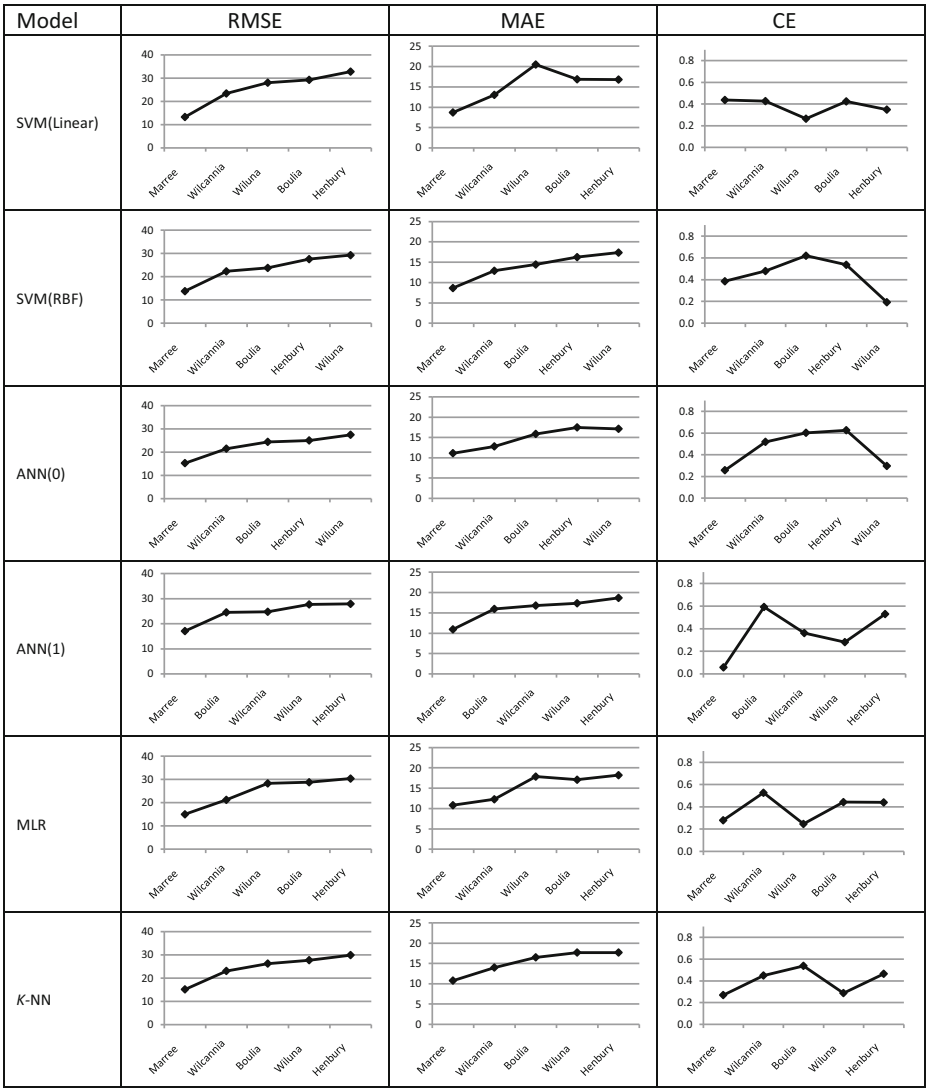

Fig. 4 Graphical display of the performance of models in desert zone

The SVMReg(RBF) model produced best predictions with inputs TMax, TMin and VP at Newry and Ningaloo, with inputs TMax, TMin, VP, Rad at Warren, Richmond, Annuello and Dowerin and with the full set of five inputs at Alexandria and Blinman.

Table 4 presents results for monthly rainfall predictions in desert zone and Fig. 4 illustrates the performance of models. One can see that overall, in the desert zone the ANN(0) and SVMReg(RBF) models produce better predictions than other models. SVMReg(linear) provides the best prediction for Marree and MLR gives the best prediction for Wilcannia weather station. The subset of best input variables strongly depends on location. For example, the SVMReg(RBF) model gave best predictions for Henbury with inputs Evap, VP and

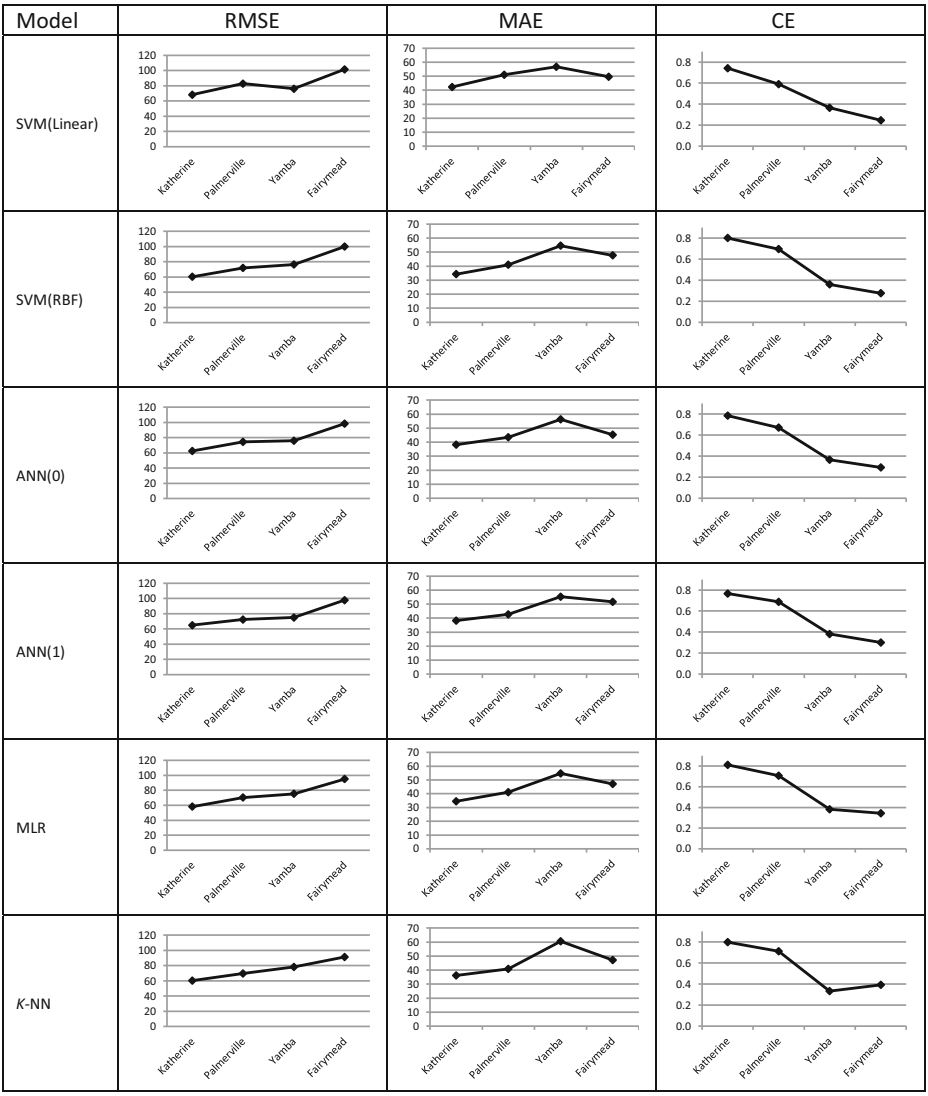

Fig. 5 Graphical display of the performance of models in tropical and subtropical zones

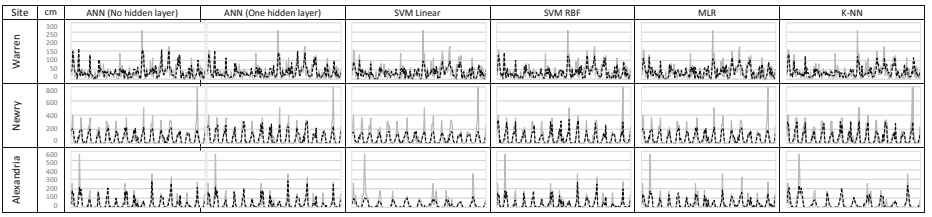

Fig. 6 Observed rainfall (grey line) vs model predictions (dotted line) for grassland zone

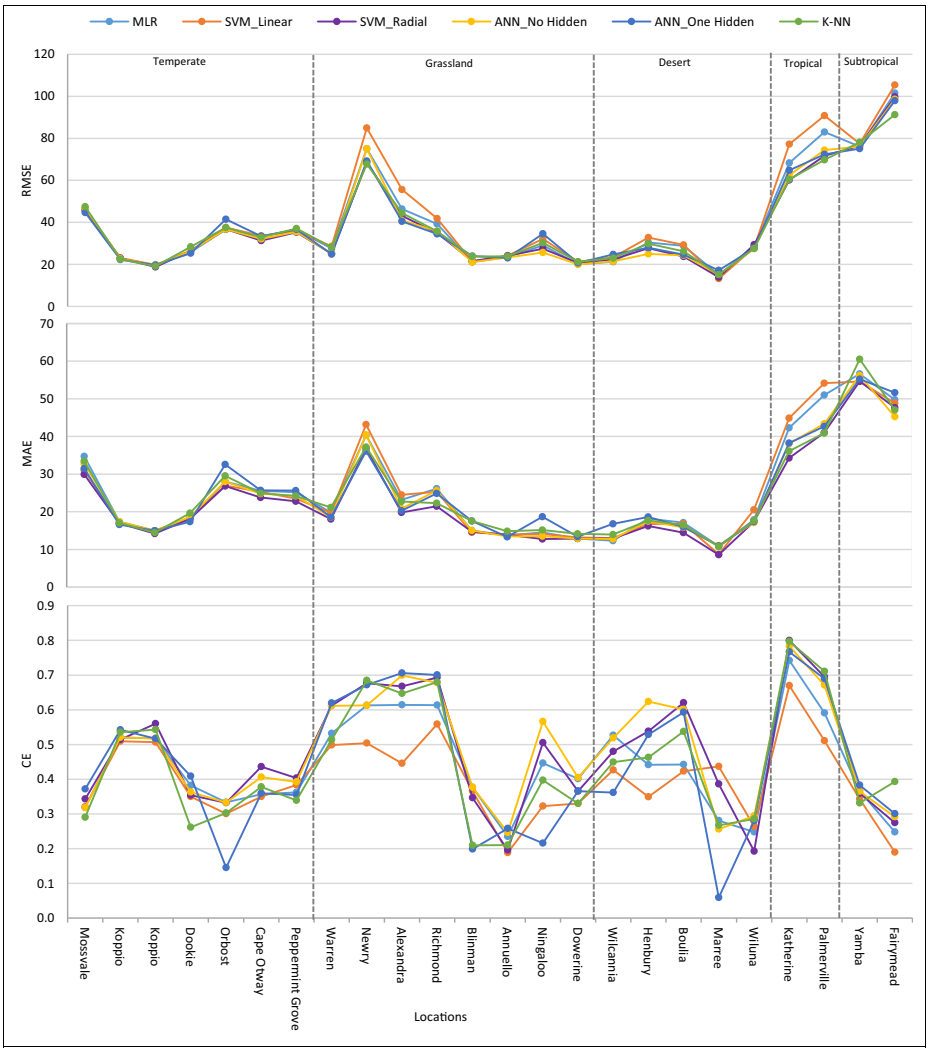

Fig. 7 Scatter plot of the performance measures for all 24 locations

Rad; for Boulia with TMax, TMin, VP, and Rad; and for Marree with the full set of input variables.

The performance measures RMSE and MAE imply that all models give predictions with the lowest error at Marree and with the highest error at Henbury and Wiluna. The performance measure CE is not in full agreement with RMSE and MAE. According to it the best performance of models is at Boulia and the worst performance is at Marree and Wiluna.

Monthly rainfall predictions and illustration of the performance of models for tropical and subtropical classification zones are given in Table 4 and Fig. 5, respectively.

Results show that the k-NN model gives the best predictions at Palmerville and Fairymead. At Katherine k-NN model's results are similar to the best results obtained by SVMReg(RBF). According to the performance measures RMSE and CE, ANN(1) gives best predictions at Yamba. Again the subset of input variables providing the best performance of models depends on a location. For example, the SVMReg(RBF) model gives best predictions at Katherine with inputs TMax, TMin, Evap and VP; at Palmerville with inputs TMax, TMin and Rad and at Yamba with the full set of input variables.

Figure 5 shows that there is an agreement between all three performance measures in determining a location with the best prediction results. All of them indicate Katherine. However, there is some inconsistency in determining a location with the worst prediction results. RMSE determines Fairymead, MAE Yamba and CE Fairymead (except the $k-$NN model) as the location with the worst prediction results. $k-$NN gives the worst prediction at Yamba (Fig. 6).

The scatter plot of three performance measures for all 24 locations and models is given in Fig. 7. This figure shows that RMSE and MAE give similar results on the quality of prediction in all climatic zones while CE not always follows their patterns. There is some disagreement between RMSE and MAE on one side and CE on the other side on the quality of predictions.

6 Conclusions

This paper reports results on a comparison of monthly rainfall prediction models using meteorological input variables. Data from 24 weather stations distributed over five climatic zones in Australia are used for this purpose. This data set consists of 540 records (from January 1970 to December 2014) and six meteorological variables, one output variable: rainfall and five input variables: maximum and minimum temperatures, vapour pressure, evaporation and solar radiation. The use of different climatic zones allowed to study the performance of the prediction models depending on different climate and hydrological regimes.

Six prediction models: SVMReg(linear), SVM with the RBF kernel function (SVMReg(RBF)), ANN without hidden layer (ANN(0)), ANN with one hidden layer (ANN(1)), k-NN and Multiple Linear Regression (MLR) were selected for comparison. All the selected models were developed for each weather station using training sets and evaluated using test sets. The prediction performance of models was evaluated by comparing observed and predicted rainfall using performance measures RMSE, MAE and CE.

Based on obtained results we can draw the following conclusions:

1. Among all six models, SVMReg(RBF) and ANN(1) are most accurate for rainfall prediction. Although k-NN and ANN(0) models give the best predictions for some locations, they are not as accurate as SVMReg(RBF) and ANN(1) for many other locations.

Two linear models, SVMReg(linear) and MLR, in general, are not accurate models for rainfall prediction. The SVMReg(RBF) and ANN(1) models are especially accurate in temperate, grassland and desert zones.

2. In tropical and subtropical zones the $k-$NN model is the most suitable model for monthly rainfall predictions where this model obtained best predictions or close to the best predictions. In these zones prediction errors by all models are higher than those for other climatic zones because of higher rainfall variability and extreme values.
3. All six models at all locations, with a very few exceptions, fail to predict extreme rainfalls.
4. Prediction performance of all six models varies considerably both within and across climatic zones. In tropical and subtropical zones, predictions have a large deviation from the actual rainfall observations.
5. The performance measures RMSE and MAE give approximately similar results, while in some locations CE provides opposite results to that of by RMSE and MAE. This is very clear from the scatter plot of the performance measures for all 24 locations given in Fig. 7. One reason for such a behavior of performance measures is extreme rainfall values. In the case of large number of extreme rainfall values the RMSE measure is better than the MAE measure as in this case the former measure takes into account the effect of these values.
6. Results show that both RMSE and MAE should be considered as primary measures to identify a subset of best input variables, that is the subset of input variables which provides the best prediction. This is due to the fact that these measures allows to determine almost the same subset of input variables across all weather stations for a given climatic zone, whereas for the CE measure this subset varies significantly even within a climatic zone.
7. We use results from papers (Abbot and Marohasy 2012, 2014) to compare the performance of the ANN model with that of presented in this paper. These two papers and the current paper use the data from the same weather stations in Queensland, Australia. However, the sets of input variables are not the same. The comparison is based on the RMSE measure and it shows that there is no any significant difference in the performance of ANN presented in these papers. However, this comparison cannot be considered conclusive as data used are not the same. There are no results with other models on similar data sets and therefore, it is not possible to compare their performance.

Rainfall is a very complex climate variable. It is controlled by physical processes involving random fluctuations. Relationship between rainfall and climatic or meteorological variables is highly nonlinear. Results confirm that data-driven modelling presents a powerful approach for rainfall prediction. Models which are able to capture nonlinearities are most suitable for such predictions. Our results on the SVMReg(RBF) and ANN(1) models confirm this conclusion. However, results from this paper also show that mainstream models are not always successful for rainfall predictions and there is a need for better models. Such models should be able, in particular, to predict extreme rainfall events which are real challenge for existing models.

Acknowledgements This research by Dr. A. Bagirov was supported by the Australian Research Counsil's Discovery Projects funding scheme (Project No. DP140103213). The authors would like to thank the Editor and two anonymous referees for their comments that helped to significantly improve the quality of the paper.

References

Australian weather and seasons (2013) http://www.australia.gov.au/about-australia/australian-story/austn-weather-and-the-seasons

Abbot J, Marohasy J (2012) Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv Atmos Sci 29(4):717–730

Abbot J, Marohasy J (2014) Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos Res 138:166–178

Aksoy H, Dahamsheh A (2009) Artificial neural network models for forecasting monthly precipitation in Jordan. Stochastic Environ Res Risk Assess 23:917–931

Al-Qahtani F, Crone S (2013) Multivariate *k*-nearest neighbour regression for time series data - A novel algorithm for forecasting UK electricity demand. In: The 2013 international joint conference on neural networks (IJCNN), pp 1–8

Awan J, Bae D (2014) Improving ANFIS based model for long-term dam inflow prediction by incorporating monthly rainfall forecasts. Water Resour Manag 28(5):1185–1199

Chattopadhyay G, Chattopadhyay S, Jain R (2010) Multivariate forecast of winter monsoon rainfall in India using SST anomaly as a predictor: neurocomputing and statistical approaches. Compt Rendus Geosci 342(10):755–765

Chowdhury R, Beecham S (2013) Influence of SOI, DMI and Niño 3.4 on South Australian rainfall. Stoch Environ Res Risk Assess 27:1909–1920

Collobert R, Bengio S (2001) SVMTorch: Support vector machines for large-scale regression problems. J Mach Learn Res 1:143–160

Feng Q, Wen X, Li J (2015) Wavelet analysis-support vector machine coupled models for monthly rainfall forecasting in arid regions. Water Resour Manag 29(4):1049–1065

Garnaut R (2008) The Garnaut climate change review: final report. Cambridge University Press, Cambridge

Haykin S (2001) Neural networks: a comprehensive foundation. Upper Saddle River, N.J.: Prentice Hall

Karamouz M, Razavi S, Araghinejad S (2008) Long-lead seasonal rainfall forecasting using time-delay recurrent neural networks: a case study. Hydrol Process 22(2):229–241

Kisi O, Cimen M (2012) Precipitation forecasting by using wavelet-support vector machine conjunction model. Eng Appl Artif Intell 25(4):783–792

Lin GF, Chen GR, Wu MC, Chou YC (2009) Effective forecasting of hourly typhoon rainfall using support vector machines. Water Resour Res 45(8):W08440. https://doi.org/10.1029/2009WR007911

Lorrai M, Sechi G (1995) Neural nets for modelling rainfall-runoff transformations. Water Resour Manag 9(4):299–313

Mekanik F, Imteaz M, Gato-Trinidad S, Elmahdi A (2013) Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes. J Hydrol 503:11–21

Mercer A, Dyer J, Zhang S (2013) Warm-season thermodynamically-driven rainfall prediction with support vector machines. Procedia Comput Sci 20:128–133

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F (2015) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. https://CRAN.R-project.org/package=e1071. R package version 1.6–7

Müller KR, Smola A, Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V (1997) Predicting time series with support vector machines. In: Gerstner W, Germond A, Hasler M, Nicoud JD (eds) Artificial neural networks – ICANN'97. ICANN 1997. Lecture notes in computer science, vol 1327. Springer, Berlin, Heidelberg, pp 999–1004

Nash J, Sutcliffe J (1970) River flow forecasting through conceptual models part I-A discussion of principles. J Hydrol 10(3):282–290

Nayak M, Ghosh S (2013) Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. Theor Appl Climatol 114(3-4):583–603

Omotosho J, Balogun A, Ogunjobi K (2000) Predicting monthly and seasonal rainfall, onset and cessation of the rainy season in West Africa using only surface data. Int J Climatol 20:865–880

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Ramirez M, de Campos V, Haroldo F, Ferreira N (2005) Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region. J Hydrol 301(1):146–162

Schliep K, Hechenbichler K (2016) kknn: Weighted *k*-Nearest Neighbors. https://CRAN.R-project.org/package=kknn

Sharma V, van de Graaff S, Loechel B, Franks D (2013) Extractive resource development in a changing climate: learning the lessons from extreme weather events in Queensland. National Climate Change Adaptation Research Facility, Gold Coast, p 110

Shukla R, Tripathi K, Pandey A, Das I (2011) Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach. Atmos Res 102(1-2):99–109

Smola A, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222

Venables W, Ripley B (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4

Wang E, Zhang Y, Luo J, Francis HS, Chiew F, Wang Q (2011) Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data. Water Resour Res 47:1–13

Yakowitz S, Karlsson M (1987) Nearest neighbor methods for time series, with application to rainfall/runoff prediction. In: Advances in the statistical sciences: stochastic hydrology. Springer, Berlin, pp 149–160