



# Application of gene expression programming for seasonal rainfall forecasting in Western Australia using potential climate indices

Farhana Islam<sup>1</sup> · Monzur Alam Imteaz<sup>1</sup>

Received: 22 September 2021 / Accepted: 21 March 2023  
© The Author(s) 2023

## Abstract

This study presents the development of rainfall forecast models using potential climate indices for the Kimberley region of Western Australia, using 100 years of rainfall and climate indices data for four rainfall stations. Three different modeling techniques: multiple linear regression (MLR), autoregressive moving average with exogenous input (ARIMAX), and gene-expression programming (GEP) were applied to develop prediction models. Preliminary analysis suggests that Western Tropical Indian Ocean (WTIO) and Southern Oscillation Index (SOI) have significant impacts on summer rainfall generation for the region. Developed models' performances were evaluated using Pearson correlation coefficient ( $r$ ), root mean square error ( $RMSE$ ), mean absolute error ( $MAE$ ), Nash–Sutcliffe efficiency ( $NSE$ ), and refined Willmot index of agreement ( $d_r$ ). It is found that the GEP model exclusively outperforms the other two alternatives. In the calibration period, the GEP model resulted in a Pearson correlation coefficient ( $r$ ) values ranging from 0.76 to 0.85, which are significantly higher than that achieved from MLR (0.32 to 0.44) and ARIMAX (0.53 to 0.83) models, while for the validation period, the correlation values for the models ranged from 0.74 to 0.87 for GEP, 0.35 to 0.51 for MLR and 0.59 to 0.77 for ARIMAX models. Considering other statistical error statistics it can be concluded that the GEP model is the best representative seasonal rainfall forecasting model for the region.

**Keywords** Prediction model · MLR · ARIMAX · GEP · WTIO · SOI · NWWA

## 1 Introduction

Reliable forecasting of rainfall variability has always been a kind of special interest in meteorology, engineering hydrology as well as in agricultural economy. Rainfall forecasting could play an important role in making investment and management decisions and risk management policies in many sectors including agriculture, water management infrastructures, coastal and disaster management, and their preparedness plans. Such forecasting has always been challenging as too many factors are involved in the generation of rainfall, therefore it is understood that the prediction outcome may not be of optimum accuracy. However, having a

forecast several months in advance can be a privilege, which may offer a certain scale of flexibility to the stakeholders to take timely decisions as well as mitigate associated risks of damage.

At present, two methods namely: dynamic method and statistical method are widely used to predict future rainfalls (Goddard et al. 2001). Dynamic models are encoded with the physics of the ocean, the interaction of the land and its atmosphere, which requires the most recent data of the present scenario and supercomputer resources to run the ensemble model and perform a computationally intensive calculation. Thus, all these intrinsic and sophisticated requirements made the dynamic method comparatively complex, expensive, and operationally time-consuming (Evans et al. 2020; Schepen et al. 2012). On the other hand, the statistical model requires long-term uninterrupted data to evaluate the relationship between the response variable and significantly contributing predictor variables. Therefore, the statistical model is relatively simple that requires less development time and supercomputer resources. While compared with dynamic models, statistical models are found to be widely

✉ Monzur Alam Imteaz  
mimteaz@swin.edu.au

Farhana Islam  
farhanaislam@swin.edu.au

<sup>1</sup> Department of Civil and Construction Engineering,  
Swinburne University of Technology, John Street, Hawthorn,  
VIC 3122, Australia

preferred over their counterparts, due to the simplicity of the formation and easy-to-use application. Furthermore, dynamic models have not shown significant prediction performance over simple statistical models despite using high-tech resources (Abbot and Marohasy 2014; Mekanik et al. 2016).

To date, several statistical techniques have widely been used to develop rainfall prediction models. These techniques include both linear and non-linear approaches. Seasonal rainfall events being a complex phenomenon, requires analyzing both linear and non-linear relationships for its prediction. Among the linear techniques, multiple linear regression (MLR) is the most popular approach that was widely used by many researchers, hydrologists, and climatologists (Hossain et al. 2018b; Islam and Imteaz 2019; Mekanik et al. 2013; Rasel et al. 2016). Among non-linear techniques, artificial intelligence (AI) based models such as an artificial neural network (ANN), adaptive neuro-fuzzy interference system (ANFIS), support vector machine (SVM), genetic programming (GP), and gene expression programming (GEP) have drawn immense attention and been successfully applied in rainfall, streamflow, and rainfall-runoff forecasting. It is understood that non-linear techniques have superior capability of explaining the underlying non-linear relationships among the variables which are found as unexplained via linear regression.

By far, ANN is the most used nonlinear statistical approach which reveals the presence of existing nonlinear relationships (either visible or hidden) among the variables. ANN has been used to model and simulate complicated time series, weather forecasting, rainfall-runoff modeling, and other hydrological and meteorological prediction models (Akhtar et al. 2009; Chiang and Chang 2009; Esha and Imteaz 2020; Hossain et al. 2020; Thirumalaiah and Deo 2000; Yilmaz et al. 2011). Despite the successful application of ANN in capturing non-linear mechanisms, researchers are often reluctant to use ANN on broad scales due to its consequential fundamental disadvantages. ANN is labeled as a black-box model as it is not capable to provide the function structure and any definite function or equation on how to calculate the output. Moreover, ANN models are deemed as complex and the outcomes are not easily interpretable (Gandomi and Alavi 2013; Hashmi et al. 2011). In parallel, GP has emerged as the most popular alternative technique to overcome the drawbacks of the ANN (Koza 1994). The main advantages of GP over ANN are its capability of capturing knowledge from the experimental data without making assumptions and finally providing a prediction equation (Alavi and Gandomi 2011). The structure of the equation is simple, which facilitates its further use for hand calculation for daily design practices (Gandomi and Alavi 2013). An extended version of GP has also received global attention in the field of structural engineering, water resources, and

hydrology, known as gene expression programming (GEP) (Ferreira 2001).

The use of GP and GEP has received great exposure in many hydrological and meteorological analyses around the world. The application of these techniques had a wide variety spanning from scouring prediction for hydraulic structures (Azamathulla 2012; Azamathulla and Ghani 2010; Azamathulla et al. 2010; Guven et al. 2008), water demand forecasting (Shabani et al. 2018), estimating evapotranspiration (Shiri et al. 2013, 2014a, 2014b), rainfall-runoff modeling (Drecourt 1999; Fernando et al. 2012a, 2012b; Khu et al. 2001; Savic et al. 1999), and spatial interpolation of data (Adhikary et al. 2016b, 2016a). A recent study showed that the GEP model offered higher efficiency in predicting specific return period events compared to the Regional Flood Estimation (RFE) method for Auckland, New Zealand (Zorn and Shamseldin 2015). In that study, the authors reported a relative error of the GEP model in flood estimation for 10 and 100 years period are 29% and 18% respectively, whereas the RFE model resulted in an error of 48% and 44% respectively. Another study used the GEP technique to model a stage-discharge relationship, where the GEP model was recommended as it outperformed traditional methods such as regression analysis and stage-discharge rating curve (Guyen and Aytak 2009). Prior to that, genetic programming was applied to forecast El Nino3.4 time series that illustrated a prediction up to 12 months in advance (De Falco et al. 2005).

Several previous studies were conducted on explaining Australian rainfall variability that revealed a strong teleconnection between climate drivers and Australian rainfalls in different regions (Cai et al. 2011; Chowdhury and Beecham 2013; Feng et al. 2010; Fierro and Leslie 2013; Ghamariadyan and Imteaz 2020, 2021a, b; Hossain et al. 2018a; Islam and Imteaz 2019, 2020; Kirono et al. 2010; Marshall and Hendon 2014; McBride and Nicholls 1983; Mekanik et al. 2013; Risbey et al. 2009; Taschetto and England 2009; Tularam 2010). However, these teleconnections often depend on the geographical location of the site and varies with different seasons (Risbey et al. 2009). Therefore, sound knowledge of the climate drivers and their influence on localized rainfall events can facilitate predicting the trend of the seasonal rainfall. For Australia, Pacific Ocean SST anomalies have shown a high influence on rainfall generation in tropical and eastern regions, whereas, Indian Ocean SST anomalies play a key role in rainfall generation in southern and western regions. To be precise, Indian Ocean Dipole (IOD) and Southern Annular Mode (SAM) have been found as influential drivers for rainfall generation in south-eastern and western parts, Blocking highs for southern parts, and ENSO Modoki and Madden Julian Oscillation (MJO) for north-western and northern parts (Ashok et al. 2003a, 2003b; Marshall and Hendon 2014; Rasel et al.

2016; Risbey et al. 2009; Schepen et al. 2012; Taschetto and England 2009; Tibaldi et al. 1994; Ummenhofer et al. 2008). Among all these drivers, in general ENSO grouped indices were found as the major contributor to rainfall generation all over Australia (Montazerolghaem et al. 2016).

To evaluate the teleconnection between climate drivers and Australian rainfall variability, some of the studies considered the entire Australian seasonal rainfalls (Cai et al. 2011; Drosowsky and Chambers 2001; Forootan et al. 2016; Kirono et al. 2010; McBride and Nicholls 1983; Risbey et al. 2009; Schepen et al. 2012), where the rest kept their studies restricted within a zone such as Queensland (Abbot and Marohasy 2012, 2014; Tularam 2010), South Australia (Chowdhury and Beecham 2013; Kamruzzaman et al. 2017; Nicholls 2010; Rasel et al. 2016; Tozer 2014), South West Western Australia (England et al. 2006; Evans et al. 2020; Feng et al. 2010; Islam and Imteaz 2020; Ummenhofer et al. 2008), and South East and East Australia (Mekanik et al. 2013; Murphy and Timbal 2008; Verdon et al. 2004). Rainfall in different locations can be generated via the interaction among different climate drivers within the region. Under such circumstances, localized prediction can be made with maximum precision and accuracy, therefore, a localized prediction is preferred as it considers the engagement of local dominant factors, resulting in reliable model development that depicts efficient prediction performance.

Current literature suggests that most of the attempts related to seasonal rainfall forecasting in Western Australia (WA) were region-based, with a majority of them were developed for South West Western Australia (Cai and Cowan 2006; England et al. 2006; Feng et al. 2015; Smith et al. 2000; Ummenhofer et al. 2008). Apart from these, a limited number of investigations were made on Central West Western Australian (CWWA) rainfall and North West Australian (NWA) rainfall variability (Feng et al. 2013; Fierro and Leslie 2013; Lin and Li 2012; Rotstayn et al. 2012). Among the studies performed in NWA, Rotstayn et al. (2012) evaluated and confirmed the influence of aerosol and greenhouse gas for an increase in summer rainfall. This was further consolidated by Shi et al. (2008), where they investigated the dynamics of the observed trend towards increased rainfall and compared the observed trend with model forced with increasing aerosol. Their study also reported an increment in NWA rainfall due to high and low sea level pressure (SLP) anomalies. In conjunction with this, an increase in NWA summer rainfall (December to February) was found to be relative to tropical Atlantic atmospheric vertical motion and southern Indian ocean climate indices (Feng et al. 2013; Lin and Li 2012). Surprisingly to date, none of the studies considered both SST and SLP-based ENSO indices and Western Tropical Indian Ocean (WTIO) index as contributors for NWA rainfall events. This study aimed to cover that gap and investigated the influence of lagged relationships among the

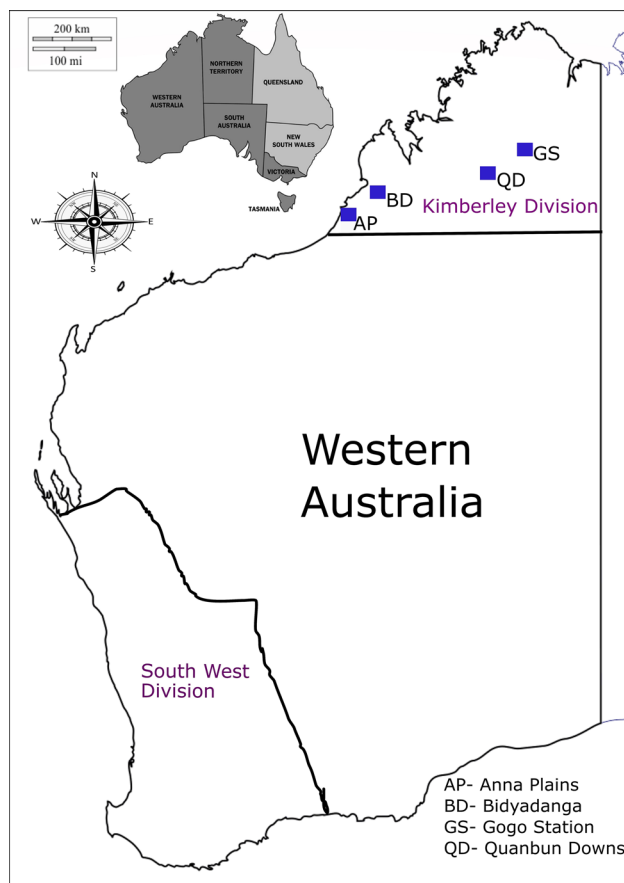
climate indices on seasonal summer rainfall (December–January–February) variability in the Kimberley region of North West Western Australia (NWWA) using three different techniques, MLR, ARIMAX, and GEP. It should be emphasized that this is the first time such a GEP technique has been used to forecast long-term seasonal rainfall in Australia. Used GEP tool will provide some equations for forecasting summer rainfall in the region several months in advance, which can be easily used by the stakeholders without having expert knowledge for different agro-economic decision-making, as well as formulating policies for the mitigation of damage due to flooding/drought.

## 2 Data and study area

The Kimberley region of Western Australia has been selected for this study due to its tropical positioning of the land in the north and diversified contribution in agriculture production, fishing and mining industry, construction, tourism, and retail trade for both the state of WA and the country. The main agricultural area in Kimberly is around 14,000 hectares around the Ord River Irrigation Area (ORIA) which makes an annual economic contribution worth 87 million Australian Dollars (AUD) in the Australian economy. Additionally, this region is popular for pastoral leases that create employment in remote areas mostly for the aboriginal community.

At present, the Kimberley region holds approximately 80% of freshwater resources in Western Australia, where most of the towns are getting their water being supplied from bore fields. Due to the pastoral nature of the inland and being dependent on limited freshwater resources, the region is vulnerable to saltwater intrusion and flooding due to sea-level rise associated with extreme weather events such as tropical cyclones during the summer season (December–January–February). This has created a demanding necessity of developing reliable rainfall prediction models for the region so that the associated adverse effects can be tackled down to save lives with minimal social and economic loss. This study considered the NWWA's Kimberley region's main rainfall season (summer rainfall events) to develop prediction models. Figure 1 and Table 1 illustrate the study area and geographical location of selected rainfall stations. Four rainfall stations from the Kimberley region were selected considering uninterrupted data availability with fewer missing values.

For preliminary analysis and model development purposes, 100 years of monthly rainfall data were collected from the Australian Bureau of Meteorology website (<http://www.bom.gov.au/climate/data/>). Also, 100 years (1916 to 2015) of climate indices data was collected from the climate explorer website ([!\[\]\(3e2231b1ad3ca8da8658228c00dd08e0\_img.jpg\) Springer](http://climexp.knmi.</a></p>
</div>
<div data-bbox=)



**Fig. 1** The geographical location of the study area and selected rainfall stations

nl/). Climate drivers namely Southern Oscillation Index (SOI) (SLP based), ENSO indices Nino3.4, Nino4, Nino3 (SST based), El Nino Modoki index (EMI), Dipole Mode Index (DMI), and Western Tropical Indian Ocean (WTIO) have been selected, extracted, and utilized to determine the significance of the correlations between rainfall and individual climate indices. A brief description of SOI and WTIO has been presented in Table 2 as these two climate indices showed a significant correlation with NWWA summer rainfall. A detailed description of the rest of the climate indices can also be found in the previous study of Islam and Imteaz (2019). For model development, the entire dataset was partitioned into two sets: calibration or training set (1916–1985) and validation or testing set (1986–2015), as using a partitioning ratio of 70:30 for calibration and validation data set is recommended for such model development (Ferranti 2012; Vaze et al. 2012).

Australian Bureau of Meteorology being a federal government authority, collects and maintains rainfall and other weather data with very high integrity and accuracy. Several weather data including the rainfall data used in this study have been used by numerous researchers for many climate related researches. Out of more than 8000 active rainfall stations, some stations experienced some missing data. However, the selections selected in this study was having no missing data as well as longer coverage of data period (> 100 years). Climate indices data is managed by the World Meteorological Organization (WMO) with highest level of accuracy.

The selection of variables for model development was followed by the selection of analytical methodology that involved a stepwise selection of analytical approach

**Table 1** Overview of the selected rainfall stations in the Kimberley region

Station no	Station name	Latitude	Longitude	Elevation (m)	Annual mean rainfall (mm)	Summer rainfall (mm)	Autumn rainfall (mm)	Winter rainfall (mm)	Spring rainfall (mm)
3028	Anna plains	19.25° S	121.49° E	10	417.92	266.03	122.38	24.98	7.79
3030	Bidyadanga	18.68° S	121.78° E	11	521.56	329.29	149.29	25.85	8.10
3014	Gogo station	18.29° S	125.59° E	150	487.30	349.17	103.3	14.35	31.76
3022	Quanbun downs	18.38° S	125.23° E	100	500.80	356.5	110.1	13.55	34.81

**Table 2** Brief description of influential climate indices in this study

Climate indices name	Abbreviation	Description	Ocean
Southern oscillation index	SOI	SOI is the measure of sea level pressure (SLP) difference between Darwin and Tahiti	Pacific
Western tropical indian ocean	WTIO	The WTIO SST anomaly index is an indicator of the surface temperatures in a cross-equatorial region spanning the western tropical Indian Ocean. It is one-half of the Dipole Mode Index, an indicator of the east–west gradient in sea surface temperatures in the Indian Ocean. It is calculated with SSTs in the box 50°E–70°E, 10°S–10°N	Indian

beginning with linear technique, followed by time series analysis and then, gene-expression programming.

### 3 Methodology

#### 3.1 Multiple linear regression

Multiple linear regression (MLR) is a statistical measure that evaluates the strength of the linear relationship between the dependent and two or more independent variables. The dependent variable is known as the response variable while the independent variable is known as the predictor. The mathematical expression of MLR is presented below in Eq. (1):

$$Y = c + \beta_1 X_1 + \beta_2 X_2 + e \quad (1)$$

where,  $Y$  is the dependent variable (i.e. rainfall),  $X_1$  and  $X_2$  are the independent variables (i.e., lagged WTIO and lagged SOI);  $\beta_1$  and  $\beta_2$  are the regression model coefficients;  $c$  is constant, and  $e$  is an error.

The prediction efficiency of an MLR model is often tested with the goodness of fit and multicollinearity checks. A goodness of fit value is usually tested with statistical parameters namely Pearson correlation coefficient ( $r$ ), root mean square error (RMSE), mean absolute error (MAE), and refined Willmott index of agreement ( $d_r$ ). For all these parameters, a value close to 1 represents a good fit. Multicollinearity is the association of the residuals (autocorrelation) derived from the predictors during the regression process, thus, can undermine the prediction efficiency. Tolerance ( $T$ ) and variance inflation factor ( $VIF$ ) values are good indicators of multicollinearity among the predictors and can be utilized to avoid exaggerated predictive performance of a model. In addition to that, Durbin-Watson (DW) test can also be used to detect the multicollinearity present among the predictors. A detailed description of the MLR methodology and associated testing techniques for the model performance can be found in Islam and Imteaz (2019).

#### 3.2 Univariate autoregressive integrated moving average with exogenous input (ARIMAX)

ARIMA model is used to predict future value considering the influence of their past, current value, and past errors. It is the combination of 'AR', 'I', and 'MA', where 'AR' stands for Auto-Regressive, 'I' stands for Integrated, a time series which needs to be differenced to make a non-stationary series to stationery, 'MA' stands for Moving Average. This model can be used to analyze and forecast univariate time series data. The expression of the model

function consists of two segments, one is seasonal and the other one is non-seasonal. Generally, it is expressed as  $(p,d,q)*(P, D, Q)$ . Where 'p' represents the non-seasonal auto-regressive, 'd' represents non-seasonal differencing, and 'q' represents the non-seasonal moving average. P, D, and Q represent the same for the seasonal segment (Corporation 2013; Adamowski et al. 2012). This study used the non-seasonal segment only as no seasonality was found in the time series.

An ARIMA model is termed as ARIMAX, whenever any exogenous input or predictors are included in a conventional ARIMA model (Kamruzzaman et al. 2013). In the ARIMAX model development for this study, two kinds of input orders were necessary: ARIMA order (dependent variable: summer rainfall) and Transfer function order (predictors or exogenous input: lagged climate indices). A detailed description of these two orders can be found in the previous study of Islam and Imteaz (2020). The mathematical expression for the ARIMAX model is presented below in Eq. (2):

$$\Delta Y_t = \varepsilon_t + \sum_{i=1}^p \varphi_i \Delta Y_{t-i} - \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \sum_{m=1}^M \beta_m X_{t-m} \quad (2)$$

where,  $\varphi_1 \dots, \varphi_p$  and  $\theta_1 \dots, \theta_q$  are the parameters;  $\varepsilon_1, \varepsilon_{t-1}$  are white noise errors and  $\beta_1 \dots, \beta_m$  are the parameters of independent variables input  $X_1$  and  $t$  is the time.

ARIMAX model development follows three steps (Box and Jenkins 1976; Cryer and Chan 2008):

Step1: Identification: In this step, the raw data is checked to verify whether the data is stationary or not. If the data set is found as non-stationary, differencing is performed to make it stationary.

Step 2: Parameter Estimation and Selection: In this step correlograms of the autocorrelation function (ACF) and partial autocorrelation function (PACF) are explored to choose the accurate 'AR' and 'MA' order. The 'AR' order relay on the lag of PACF cut and the 'MA' order relay on the lag of ACF cut. However, decision-making on their order is not that simple as several trials and errors are required to select the appropriate order. Some general guidelines can be followed in the selection of AR and MA orders as discussed in the previous study of Islam and Imteaz (2020).

Step 3: Diagnostic Check: Model adequacy is validated using diagnostic checks where the residual of the ARIMAX model should satisfy the requirement of being white noise. This requirement can be verified in two ways, one is drawing a residual ACF and PACF plot and checking on the spikes. If the spikes stay between the boundary lines (by at least 95%), it indicates the residual is white noise. Another way of such a check is to carry out a Ljung-Box test, in which if the p-value is more than 0.05, the null hypothesis gets

verified as being white noise (Ljung and Box 1978). A successful deployment of all these three steps and subsequent verification can provide sufficient evidence of the model's forecasting capability.

### 3.3 Gene expression programming

GEP is a combination of the principles of genetic algorithms (GA) and genetic programming (GP). The basic disagreement between these three algorithms (GA, GP, and GEP) is their nature or the way of representing chromosomes. GA utilizes a linear string of fixed length of chromosomes and GP utilizes non-linear entities of tree-based chromosomes with different sizes and shapes (parse tree) and GEP is encoded as a simple linear string of fixed length chromosomes and expressed as nonlinear entities of different sizes and shapes (Ferreira 2001).

In GEP, genotype and phenotype are the two vital types of entities that are structurally and functionally different from each other. In genotype, chromosomes are simple small linear entities composed of one or more genes, where replication, mutation, recombination, and transposition can be performed easily. In phenotype, Expression Trees (ETs) are the algebraic or mathematical expression of the genetic information encoded in respective chromosomes. Genetic code is incorporated with the symbol of the chromosome and the terminal functions. This genetic code decides the structural organization of the function and terminals in expression trees. Moreover, GEP genes are combined with two elements one is the head and another one is the tail. The head encoded the functions for expression. It represents both the function set (F) and the terminal set (T). On the other hand, the tail represents the only terminal set (T). This terminal set from the tail acts as a reservoir for an argument required by the function that can be used in the head while there is a shortage of terminals. Therefore, the head contains functions, variables, and constants but the tail contains only variables and constants. For any problem, head (h) length can be selected manually, and tail length (t) needs to be calculated using the following Eq. (3):

$$t = h(n - 1) \quad (3)$$

where, n is the number of variables/arguments required by the functions, h is the head length, t is the tail length. For example, any gene consists of function [Q, \*, /, -, +, a, b], head length is selected as 10 and the number of arguments is 2, in that scenario, the tail length is  $t = 10 * (2 - 1) + 1 = 11$ . Therefore, the length of the genes is  $10 + 11 = 21$ . GEP represented two types of languages: one is related to genes and the other one is the language of ETs. The system that decides the structure of ETs and their interactions and provides the sequence of genes is called the Karva language.

#### 3.3.1 An overview of gene expression programming algorithms

The process of GEP begins with the random origination of chromosomes of the primary population. These chromosomes are decoded into computer programs and their fitness is evaluated by the appropriate fitness function. From the outcome of the fitness test, individuals are selected to reproduce with a further modification that leaves progeny with new offspring. In the reproduction phase, genomes or chromosomes are improved by several genetic operators namely replication, mutation, inversion sequence (IS) transposition, root insertion sequence (RIS) transposition, gene transposition, one-point and two-point recombination, and gene recombination. Elaborative details about the functionality of these genetic operators can be found in Ferreira (2001). The reproduced individuals then go through the same development procedure such as genomes expression, engagement of the chosen environment, and reproduction with further modification. The procedure is repeated until an adequate result has been obtained. An overview of the gene expression programming (GEP) algorithm has been presented in Fig. 2. Further to this, an explanation of the GEP genes formation framework or structure has been presented in the following subsections.

**3.3.1.1 Open reading frames (ORFs) and genes** Open reading frames (ORFs) are a coding order of the gene, in terms of biology: begins with a start codon, continues with amino acid codons, and finishes with termination codons. ORFs are the sequences upstream from the start codon and sequences downstream from the stop codon, thus presents the framework/structure of GEP genes. In GEP, the beginning site is the first spot of a gene however, the termination/stop may not necessarily be the last spot. Therefore, for some instances, the GEP genes may have a noncoding downstream region from the termination or stop point. This noncoding downstream region permits the modification of genomes through genetic operators for producing accurate programs. For example, if an equation as given below in Eq. (4) is considered:

$$\sqrt[3]{(a + b) * \left[ c * \left( \frac{d}{e} \right) \right]} \quad (4)$$

The following ORF diagram should be the structure of the formation and chromosome of the equation in terms of ETs as presented in Fig. 3.

In Fig. 3 (a), "Q" is the cube root function, a, b, c, d, and e are the terminals, and +, - and \* are the addition, subtraction, and multiplication function. The expression

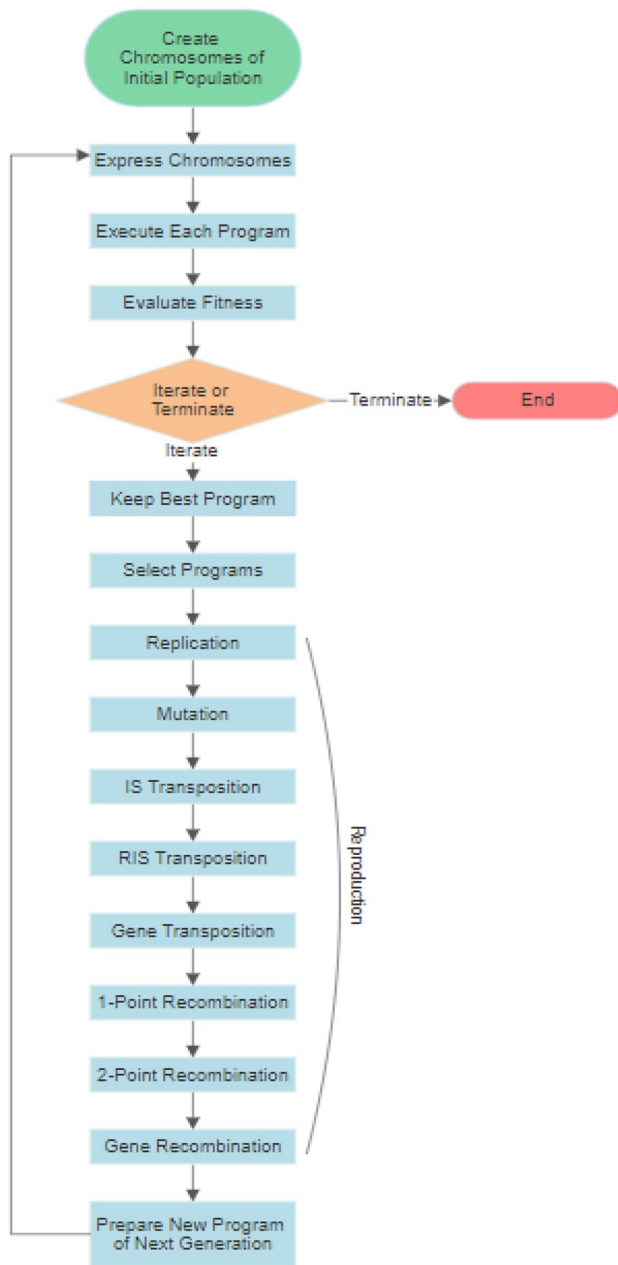


Fig. 2 An overview of Genetic Expression Programming (GEP)

presented in Fig. 3 (b) is called ORFs Karma expression, which started at “Q” (position 0) and terminating at “e” (position 9). The way of reading the expression tree is: left to right, then from top to bottom. The starting point of the ORFs is related to the root of the ETs that creates the first line of the ET. Considering the number of arguments of each element, the next line continues by generating the requisite number of nodes. For example, the root of the ET presented in Fig. 3 will start with “Q” at position 0. As the cube root function has only one argument, the next

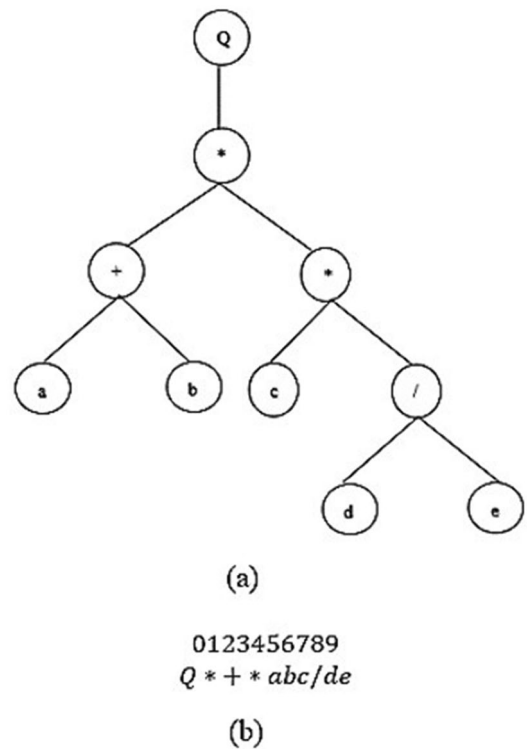


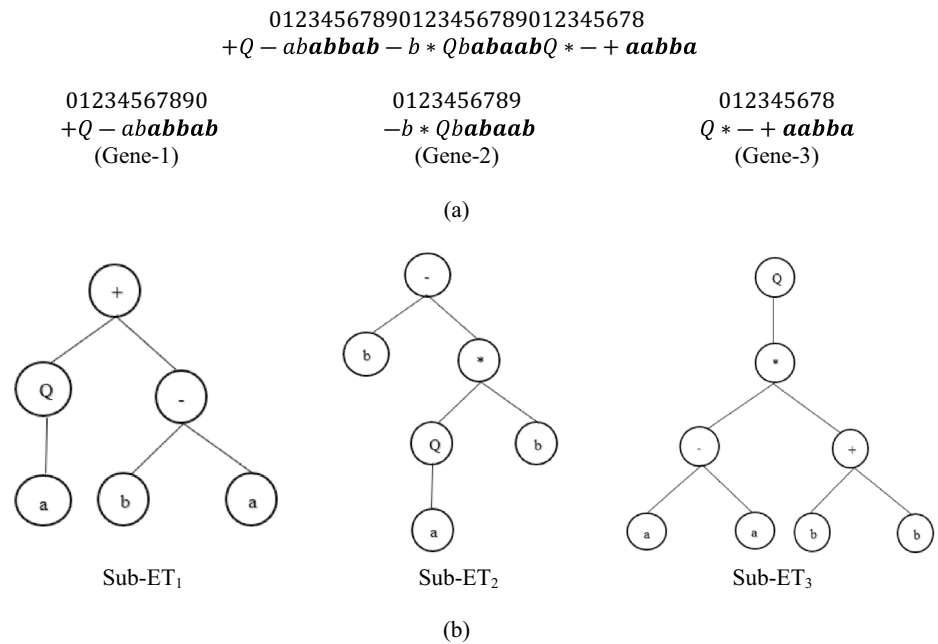
Fig. 3 Example of an ORFs: a expression tree; b chromosome (genotype)

line continues with one node at position 1, which is “\*”. This multiplication factor required two arguments thus the next line continued with “+” and “\*” at node positions 2 and 3, respectively. For node position 2, the next line got filled up with terminal/ leaf nodes “a” and “b” at positions 4 and 5. On the other hand, for node position 3, the next line got one terminal/ leaf node “c” at position 6 and one argument node “/” at position 7. At this stage, the argument node at position 7 required two more branches which are terminal/ leaf nodes “d” and “e” at positions 8 and 9, respectively. This terminates the growth of the ET as no further offspring are grown.

**3.3.1.2 Multigenic Chromosomes** Chromosomes in GEP usually contain more than one gene of equal length. As the entity of a complex individual requires complex genes, it is necessary to have multigenic genomes to develop complex entities. For any problem in GEP, the gene number and the head number are selected by the user. Each gene encoded the sub-expression tree (sub-ET), and these sub-ETs are interconnected to create a more complex entity.

Figure 4 presents a scenario where the total gene length is 29, comprised of three genes, each terminating at different position (with tails denoted as bold) based on the fitness to derive the output equation of a complex problem. In

**Fig. 4** Expression of multigenic chromosomes **a** three genes and **b** sub-ETs encoded by three genes



this scenario, a multigenic chromosome consists of three genes, forming three open reading frames for sub-ETs as each ORF constructed a particular sub-ET. While the first element of each gene is in position 0, the position also indicates the end of each ORF. In Fig. 4, the first ORF terminates at position 5 in sub-ET<sub>1</sub>, the second ORF ends at position 5 in sub-ET<sub>2</sub>, and the final one terminates at position 7 in ET<sub>3</sub>.

### 3.4 GEP model development for rainfall forecasting

GEP methodology has been applied to develop models to represent relationships between climate indices and rainfall. The GEP form of the prediction model can be presented as in Eq. (5):

$$Y = f(X_1, X_2, \dots, X_n) \tag{5}$$

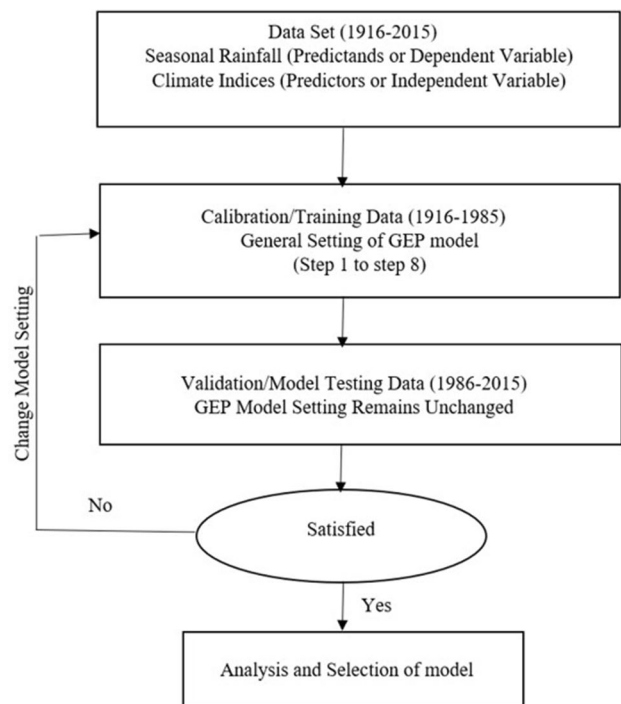
where,  $Y$  is the dependent or response variable (seasonal summer rainfall),  $X_1, X_2, \dots, X_n$  are the predictors or independent variables (large-scale climate indices).

The methodology for generating the GEP model can be presented in the following flowchart presented in Fig. 5.

The major steps that followed to predict seasonal summer rainfall using GEP are given below:

- a. Random generation of the initial population of the chromosome (genotype) where the length of the chromosome is fixed.
- b. Individual chromosome in the initial population is translated into phenotypes that are expressed by an Expression Tree (ET).

- c. Selection of a best-suited fitness function namely correlation coefficient, mean absolute error, and relative error to evaluate the performance of the developed program.
- d. Selection of the terminal (T) set, and function (F) set to generate chromosome. The selection of these two functions entirely depends on the nature of the problem, users' understanding, and the trial-and-error process.



**Fig. 5** Flow chart of the methodology used in this study



**Table 3** Initial setting of the GEP model

Initial setting	Symbol or value	Arity
<b>Function set</b>		
Addition	+	2
Subtraction	−	2
Multiplication	*	2
Division	/	2
Square root	Sqrt	1
Exponential	Exp	1
Natural logarithm	Ln	1
Logarithm of base 10	Log	1
Inverse	Inv	1
x to the power of 2,3	$x^2, x^3$	1
Cubic root	3Rt	1
Minimum of 2 inputs	Min 2	2
Maximum of 2 inputs	Max 2	2
Average of 2 inputs	Avg 2	2
Sine	Sin	1
Cosine	Cos	1
<b>General setting</b>		
Chromosome	30	
Genes number	3–6	
Head size	7–10	
Linking function	Addition	
Fitness function error type	RMSE with parsimony Pressure	
<b>Genetic operator</b>		
Optimal evaluation		
Mutation rate	0.00138	
Inversion rate	0.00546	
IS transposition	0.00546	
RIS transposition	0.00546	
One-point recombination rate	0.00277	
Two-point recombination rate	0.00277	
Gene recombination rate	0.00277	
Gene transposition rate	0.00277	
Numerical constants	± 10	

e. Selection of the structural organization of the chromosome, which is the combination of head length (h), gene number, and genetic operators. The reproduction of the chromosome is performed by utilizing best-performing individuals' programs through a genetic operator such as replication, mutation, transposition, and recombination.

- f. Selection of linking functions such as addition, multiplication, subtraction, or division, respectively. It must be selected before the program runs to obtain the equation by connecting all the subtrees.
- g. The development of a new generation program through reproduction is the last step.
- h. Re-application of steps 2 and 7 until the selected termination benchmark is reached.

This study considered root mean square error (RMSE) with parsimony pressure as a fitness function to evaluate the model fitness. Parsimony pressure in the model ensures that the developed model is not overfitted and is in the best-fit conditions. The predictability in the parsimonious model is more accurate compared to the generally developed model. Besides, the selection of the terminal and function set is also of great importance for better prediction model development. The terminal set contains independent variables that get selected from the correlation analysis. The selection of a functional set is usually performed considering the nature of the problem, simplicity to use, and past evidence of the function as an efficient and effective tool. For this study, the climate indices that showed the highest significant correlation with seasonal rainfall were selected as a terminal set, where the selected functional set has been presented in Table 3. Table 3 also presents the genetic operators used to create genetic variation in the chromosome population. As required for the problem encountered, random models were generated with a combination of the function set and terminal set until they reached a valid solution. As suggested in Ferreira (2001) and Guven and Aytok (2009), the chromosome gene number was set as a minimum of 3 to a maximum of 6, the head length was set from 6 to 10, and “addition” was considered as the linking function.

### 3.5 Performance metrics

Development of prediction models require evaluating the model performances; thus, several statistical metrics were used such as root mean square error (RMSE), mean absolute error (MAE), Nash–Sutcliffe efficiency (NSE), and refined Willmot index of agreement ( $d_r$ ). Among them, *RMSE* and *MAE* are the most prominent method of error measurement in hydroinformatics, where a lower value of *RMSE*, and *MAE* indicates a better predictability performance of the model (Saigal and Mehrotra 2012; Singh et al. 2005; Shabani et al. 2018).

Nash–Sutcliffe Efficiency (NSE) was measured to evaluate the predictability skill of a developed hydrological model as it assesses its goodness of fit (Nash and Sutcliffe 1970). NSE is calculated using the following Eq. (6):

$$NSE = 1 - \frac{\sum_{i=1}^n (P_i - O_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (6)$$

where,  $P_i$  is the predicted value of the  $i$ th observation,  $O_i$  is the observed value of the  $i$ th observation;  $\bar{O}$  observed mean value and  $n$  is the number of observations. The  $NSE$  value ranges from 0 to 1, where a value of 1 means the developed model is a perfect fit and it has a perfect predictive skill. On the other hand,  $NSE$  value equal to 0 indicates the modeled values are as accurate as of the observed mean value; where a value of  $NSE < 0$  indicates there is a severe error in the data and the observed mean is a better predictor compared to the developed model. Thus,  $NSE$  value close to 1 ensures the predictive skill of a model (Gupta and Kling 2011; McCuen et al. 2006).

Refined Willmot Index of Agreement ( $d_r$ ) is another new statistical parameter introduced by Willmott et al. (2012) to evaluate the skillfulness of the developed model. It specifies the sum of the magnitudes of the differences between the predicted and observed deviations from the observed mean relative to the sum of the magnitudes of the perfect model ( $P_i = O_i$ , for all  $i$ ) and observed deviations from the observed mean. The refined index of agreement ( $d_r$ ) can be calculated using the following Eq. (7):

$$d_r = \left\{ \begin{array}{l} 1 - \frac{\sum_{i=1}^n (P_i - O_i)}{c \sum_{i=1}^n (O_i - \bar{O})}, \text{ when} \\ \sum_{i=1}^n (P_i - O_i) \leq c \sum_{i=1}^n (O_i - \bar{O}) \\ \frac{c \sum_{i=1}^n (O_i - \bar{O})}{\sum_{i=1}^n (P_i - \bar{O})} - 1, \text{ when} \\ \sum_{i=1}^n (P_i - O_i) \geq c \sum_{i=1}^n (O_i - \bar{O}) \end{array} \right\} \quad (7)$$

where,  $P_i$  is the predicted value of the  $i$ th observation,  $O_i$  is the observed value of the  $i$ th observation,

$\bar{O}$  is the observed mean value and  $n$  is the sample size. A “ $c$ ” value equal to 2 is suggested in the equation. The “ $d_r$ ” value ranges from  $-1$  to  $+1$ , where a positive value indicates a good fit, while a negative value indicates the opposite.

## 4 Results and discussion

### 4.1 Preliminary analysis

In this study, a single correlation or Pearson correlation ( $r$ ) was used to evaluate the lagged relationship between climate indices and seasonal rainfalls of Western Australia. Four rainfall stations from the Kimberley region of NWWA were selected to conduct this study. At first, single correlation

analyses were performed between climate indices and seasonal rainfall to identify potential predictors. It was observed that for Kimberley, maximum rainfall occurred in the summer season. This study evaluates the influences of selected climate indices on summer rainfall for the Kimberley region. Climate indices with statistical significance (at 1% and 5% levels) were considered for further analysis. All these analyses were performed using the *IBM SPSS Statistics 26* software package.

#### 4.1.1 Single correlation analysis

Once the rainfall data and climate data were extracted from the database, bivariate correlation or single correlation analysis was performed to evaluate the lagged relationship between the climate indices and the rainfall events. Seasonal summer rainfall and monthly values of climate indices namely SOI, WTIO, DMI, Nino3.4, Nino3, Nino4, and EMI were considered for the analysis. For the climate indices, lagged monthly values (March<sub>(n-1)</sub> to November<sub>n</sub>) was used, where ‘ $n$ ’ is the year for which the seasonal summer rainfall is to be predicted, and  $(n-1)$  is the immediately previous year. The outcome of the single correlation analysis is presented in Table 4.

From the correlation analysis, it was observed that SLP based ENSO index (i.e., SOI) showed great influence on NWWA summer rainfall for the selected rainfall stations. This outcome has been found consistent with the findings of Fierro and Leslie (2013), as they mentioned that SOI has the most robust relationship with November to April rainfall. On the other hand, SST-based climate indices (i.e., Nino3.4, Nino3, Nino4, and EMI) showed very little influence on summer rainfall. Moreover, DMI which is the indicator of IOD did not show any influence at all (except for the station-Quanbun Downs). This finding is aligned with the available literature, where the researchers demonstrated that SLP based climate index (SOI) has influence on NWWA summer rainfall and SST based ENSO, ENSO Modoki (EMI) index, and IOD has no significant impact on it. However, tropical Indian ocean indices may have a positive impact (Lin and Li 2012; Shi et al. 2008). Furthermore, the data presented in Table 4 also confirmed that WTIO has a significant correlation with summer rainfall for all the selected stations in NWWA.

### 4.2 MLR model development

#### 4.2.1 Multiple linear regression analysis

From this outcome of single correlation analyses, various MLR model sets with a different combination of lagged indices (WTIO-SOI, WTIO-Nino4, WTIO-Nino3, and

DMI-SOI) were developed. The description of the model sets has been presented in Table 5.

The outcome of the MLR model sets has been presented in Table 6. The multiple linear regression model output showed that Pearson correlation ( $r$ ) has increased compared to the single correlation analyses. From Table 6, it is observed that the WTIO-SOI model showed the highest correlation compare to the other combination model. Therefore, the lagged WTIO-SOI model has been considered as the best model for the selected rainfall stations. The best model for each of the rainfall stations with associated regression coefficient, Pearson correlation ( $r$ ), Durbin-Watson (D-W), Tolerance (T), and VIF values are presented in Table 7. It can be observed that all these models have satisfied the requirements of having no autocorrelations among the residuals and the predictors.

A validation test was carried out to evaluate the appropriateness of the model selection in the calibration period. Statistical parameters such as Pearson correlation ( $r$ ),

$RMSE$ ,  $MAE$ , and refined Willmot index of agreement ( $d_r$ ) were calculated. A Comparative demonstration of the statistical parameters for the MLR model in both calibration and validation period has been presented in Table 8. From Table 8, it is noticeable that comparatively high Pearson correlation ( $r$ ) and refined Willmot index of agreement ( $d_r$ ) values are evident in the validation period, where  $RMSE$  and  $MAE$  values were relatively low if compared to the calibration period.

### 4.3 ARIMAX model development

#### 4.3.1 Exogenous input/ predictors, ARIMA order, and transfer function input selection

In ARIMAX model development, climate indices that exhibit significant correlation in single correlation analysis were selected as exogenous input. Several ARIMAX model sets were developed with different lagged climate indices

**Table 4** Pearson correlation ( $r$ ) of lagged climate indices with summer rainfall

Climate indices	Kimberley rainfall stations			
	Pearson correlation ( $r$ )			
	Anna plains	Bidyadanga	Gogo station	Quanbun downs
WTIO <sub>Mar</sub>	0.30*	–	–	–
WTIO <sub>April</sub>	0.34**	0.25*	–	–
WTIO <sub>May</sub>	0.29*	–	0.27*	–
WTIO <sub>June</sub>	–	–	0.27*	–
WTIO <sub>July</sub>	–	0.24*	0.32**	–
WTIO <sub>Aug</sub>	0.25*	0.24*	0.33**	–
WTIO <sub>Sep</sub>	–	–	0.32**	–
WTIO <sub>Oct</sub>	0.24*	–	–	0.27*
WTIO <sub>Nov</sub>	–	–	–	–
SOI <sub>Mar</sub>	–0.27*	–	–0.27*	–
SOI <sub>April</sub>	–	–	–	–
SOI <sub>May</sub>	–	0.24*	–	–
SOI <sub>June</sub>	–	–	–	–
SOI <sub>July</sub>	–	–	–	–
SOI <sub>Aug</sub>	–	–	–	–
SOI <sub>Sep</sub>	–	0.31**	–	–
SOI <sub>Oct</sub>	–	–	–	–
SOI <sub>Nov</sub>	–	0.32**	–	–0.27*
Nino3.4 <sub>(from Mar to Nov)</sub>	–	–	–	–
Nino4 <sub>Mar</sub>	0.24*	–	–	–
Nino3 <sub>June</sub>	–	–	–	0.24*
EMI <sub>(from Mar to Nov)</sub>	–	–	–	–
DMI <sub>Sep</sub>	–	–	–	0.24*

\* Correlation is significant at the 0.05 level (2-tailed)  
 \*\* Correlation is significant at the 0.01 level (2-tailed)

combinations (WTIO-SOI, WTIO-Nino4, WTIO-Nino3, and DMI-SOI) to evaluate their predictability performance. The IBM SPSS Statistics 26 software package was used for all of these analyses.

In the identification stage, summer rainfall and climate indices data were analyzed. Rainfall and climate indices were found as non-stationery and rainfall patterns as non-seasonal. However, in the ARIMAX model, data sets are needed to be stationary, therefore differencing (d) of the data was performed. Figure 6 depicts the data condition before and after differencing was performed for the station—Anna Plains. A similar approach has also been applied for the rest of the rainfall stations and selected climate indices.

To select the AR and MA order in the ARIMAX model, ACF and PACF plots were drawn for the selected rainfall stations. AR order is selected from the PACF plot and MA order is selected from the ACF plot, considering the spike outside of the boundary lines and some other guidelines to select the appropriate order. Figure 7 presents the ACF and PACF plots with respective lag numbers for rainfall station—Anna Plains. ARIMAX (0,1,1) order was found as appropriate for Anna Plains and a similar approach has been applied for the rest of the rainfall stations. Table 9 presents the selected ARIMA orders and transfer function orders for all the rainfall stations considered in this study.

**Table 5** Multiple regression model sets for selected rainfall stations in the Kimberley region

Model combination	Different lagged months for each rainfall station			
	Anna plains	Bidyadanga	Gogo station	Quanbun downs
WTIO-SOI	Mar-Mar, April-Mar, May-Mar, Aug-Mar, Oct-Mar	April-May, April-Sep, April-Nov, July-May, July-Sep, July-Nov, Aug-May, Aug-Sep, Aug-Nov	May-Mar, June-Mar, July-Mar, Aug-Mar, Sep-Mar	Oct-Nov
WTIO- Nino4	Mar-Mar, April-Mar, May-Mar, Aug-Mar, Oct-Mar	–	–	–
WTIO-Nino3	–	–	–	Oct-June
DMI-Nino3	–	–	–	Sep-June
DMI-SOI	–	–	–	Sep-Nov

**Table 6** Pearson Correlation (r) results with the different model sets in MLR

Region	Station name	Pearson correlation (r) for different model sets				
		DMI-Nino3	DMI-SOI	WTIO-SOI	WTIO-Nino3	WTIO-Nino4
Kimberley	Anna plains	–	–	0.36	–	0.34
	Bidyadanga	–	–	0.44	–	–
	Gogo station	–	–	0.36	–	–
	Quanbun downs	0.28	0.30	0.32	0.29	–

**Table 7** Summary of the regression model

Station name	Predictors	Constant	Coefficient	Pearson correlation (r)	R <sup>2</sup>	Standard error of estimate (σ)	Durbin-Watson (D-W)	Multicollinearity	
								Tolerance (T)	VIF
Anna plains	WTIO <sub>April</sub>	82.81	71.69	0.36	12.96%	60.22	1.68	0.79	1.26
	SOI <sub>Mar</sub>		–8.51						
Bidyadanga	WTIO <sub>Aug</sub>	109.37	91.28	0.44	19.36%	51.92	1.78	0.96	1.04
	SOI <sub>Nov</sub>		27.01						
Gogo station	WTIO <sub>Aug</sub>	113.07	50.27	0.36	12.96%	43.95	1.80	0.83	1.20
	SOI <sub>Mar</sub>		–7.12						
Quanbun downs	WTIO <sub>Oct</sub>	110.48	38.98	0.32	10.24%	51.89	2.03	0.82	1.21
	SOI <sub>Nov</sub>		–10.55						

**Table 8** Model description of the selected MLR model both in calibration and validation period

Station name	Model set	Calibration period				Validation period			
		Pearson correlation ( $r$ )	RMSE	MAE	Refined Willmot index of agreement ( $d_r$ )	Pearson correlation ( $r$ )	RMSE	MAE	Refined Willmot index of agreement ( $d_r$ )
Anna plains	WTIO <sub>April</sub> -SOI <sub>Mar</sub>	0.36	59.07	43.06	0.53	0.42	62.68	50.50	0.56
Bidyadanga	WTIO <sub>Aug</sub> -SOI <sub>Nov</sub>	0.44	61.93	47.31	0.54	0.51	59.72	49.10	0.56
Gogo station	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	0.36	43.95	35.88	0.54	0.35	54.29	46.47	0.52
Quanbun downs	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	0.32	51.89	41.80	0.52	0.36	57.33	48.81	0.54

#### 4.3.2 ARIMAX model development and selection of best forecast model

Once all the requirements of the ARIMAX model set up were satisfied, several ARIMAX models with the combination of influential indices namely WTIO-SOI, WTIO-Nino4, WTIO-Nino3, and DMI-SOI were developed. Table 10 presents a different combination of ARIMAX model sets with respective Pearson correlation ( $r$ ) values.

From Table 10, it is eminent that the WTIO-SOI model combination depicted the highest correlation statistics compare to the other combination sets except for Anna Plains. For Anna Plains, both WTIO-SOI and WTIO-Nino4 showed a good correlation, however, to keep the model set simple and consistent with other rainfall stations, WTIO-SOI model set was selected for further model development.

The statistical performance of the best models during the calibration period has been presented in Table 11. For Anna Plains, Bidyadanga, Gogo Station, and Quanbun Downs, the model sets with the highest correlation values are WTIO<sub>Aug</sub>-SOI<sub>Mar</sub>, WTIO<sub>Aug</sub>-SOI<sub>May</sub>, WTIO<sub>Aug</sub>-SOI<sub>Mar</sub>, and WTIO<sub>Oct</sub>-SOI<sub>Nov</sub> with correlation ( $r$ ) values of 0.83, 0.68, 0.65, and 0.53. Except for Quanbun Downs, all the remaining rainfall stations showed good predictability for at least four months lead time. This confirms the selected models' prediction capability at least four months in advance. Many other models were also found with longer lead times but having lower correlation values, or higher errors for these selected stations, hence they were not chosen as the best model.

Once the best model got selected, a diagnostic check was carried out to verify the accuracy of the developed model. To check the autocorrelation of the residuals, a Ljung-Box test was performed. From this test, it was found that the residuals are being white noise for all the rainfall stations as the  $p$ -values for all the selected models were found as greater than 0.05 (Ljung and Box 1978). Another approach for such a check was to draw a residual ACF and PACF plot and check for the spikes. If the spikes are found to stay between the boundary lines (at least by 95%), it indicates the residual is white noise. Figure 8 presents the

evidence that all the spikes are within the boundary lines, thus, no autocorrelation is present among the residuals.

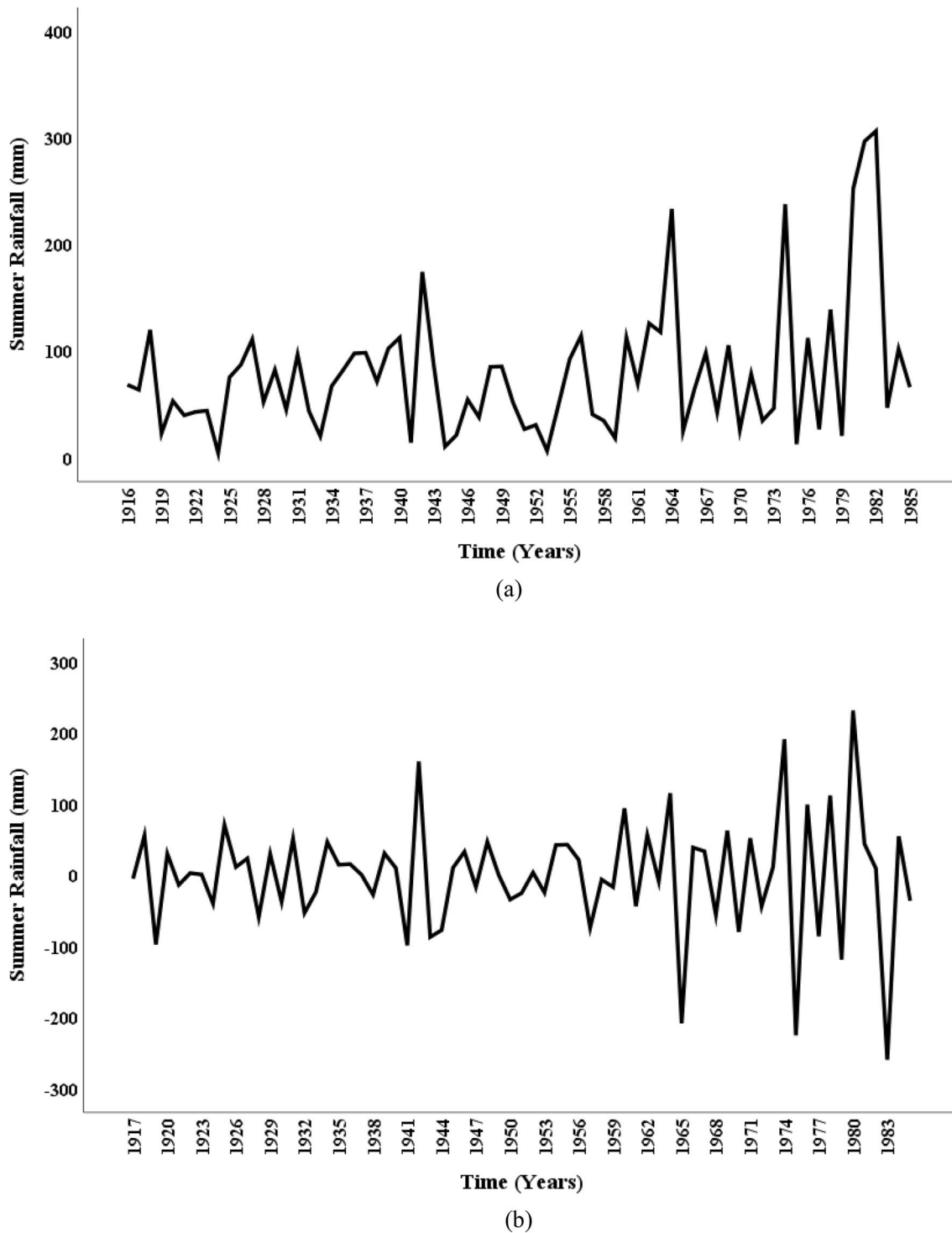
Once the ARIMAX model got developed, a validation test was performed for the selected model set. Table 12 presents the model description for the developed ARIMAX models in both calibration and validation periods. In the validation tests, the Pearson correlation ( $r$ ) increased significantly for all the rainfall stations except Anna Plains. An increase in refined Willmott index of agreement ( $d_r$ ) was also observed for the same. Similarly, a reduction in error values is also an indicator of the models' prediction performance as observed, in particular for Bidyadanga.

#### 4.4 GEP model

For GEP model development, the most influential predictors were selected from the correlation analyses between summer rainfall and climate indices. Among four rainfall stations, Bidyadanga and Gogo Station's summer rainfall exhibited a significant correlation with WTIO and SOI. For Anna Plains, WTIO, SOI, and Nino4, and for Quanbun Downs, WTIO, DMI, SOI, and Nino3 showed a significant correlation. Considering the facts, several model sets having different combinations of climate indices were developed and their performances were evaluated. While preparing the model sets, only one climate index from each Indian and Pacific Ocean was selected. Such selection was necessary so that any autocorrelation effect between Indian Ocean indices (i.e., WTIO, DMI) and Pacific Ocean Indices (i.e., SOI, Nino3, and Nino4) is avoided.

##### 4.4.1 GEP model development and selection of best forecast model

GEP model development is not that straight-forward as several trials and errors are required to obtain optimum output from the developed model. This involves setting up and deciding on appropriate parameters (i.e., head size, gene number, and linking function). Head size and gene numbers are often get modified to achieve optimum results. To keep the model equation simple, head size is usually kept limited

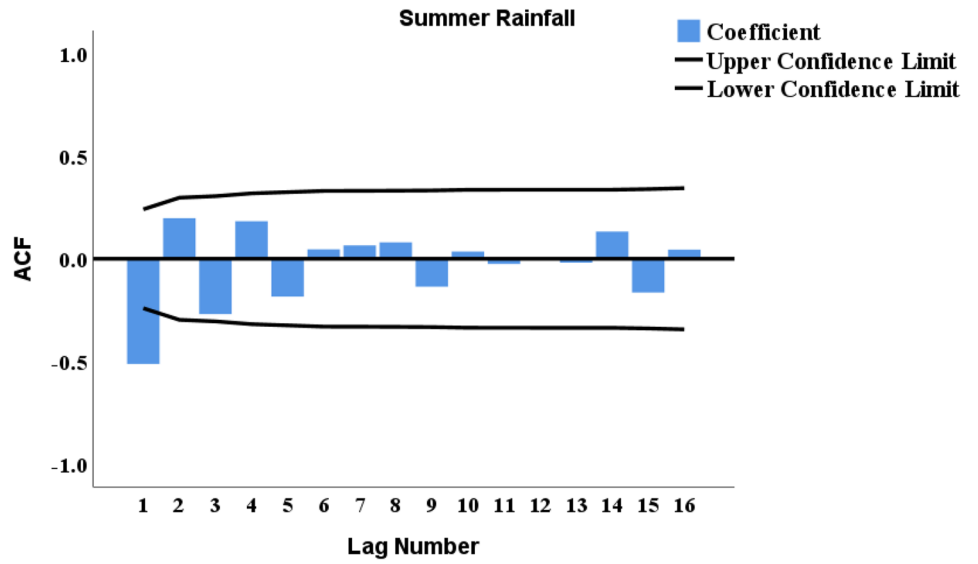


**Fig. 6** Rainfall data for Anna Plains: **a** before differencing, **b** after differencing

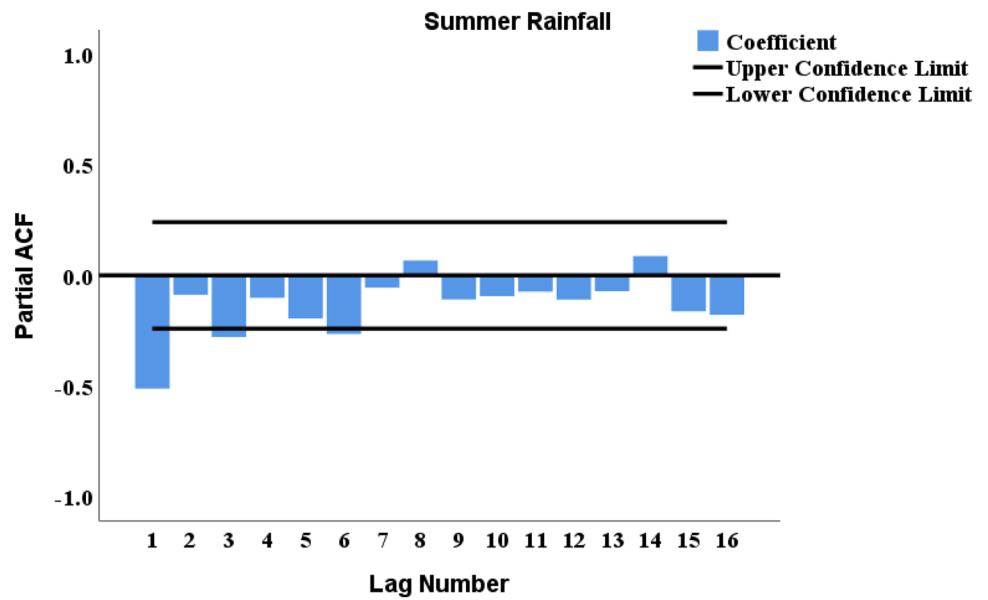
to 7–10 and gene number is set between 3 and 6. For this study, in particular, keeping head size 9 and gene number 5 has been found as most suitable. Furthermore, “addition” is found as best suited as the linking function. Several GEP models were developed using different combinations such

as DMI-Nino3, DMI-SOI, WTIO-SOI, WTIO-Nino3, and WTIO-Nino4 for the Kimberley region. All these analyses were performed using the ‘*GeneXpro-Tools 5*’ software. Table 13 presents different model sets prepared with

**Fig. 7** Rainfall data for Anna Plains: **a** ACF plot and **b** PACF plot



(a)



(b)

**Table 9** ARIMA and transfer function orders

Station name	ARIMA order			Transfer function order		
	Autoregressive (AR)	Difference (D)	Moving average (MA)	Numerator	Denominator	Difference
Anna plains	0	1	1	1	1	1
Bidyadanga	0	1	1	1	1	1
Gogo station	3	1	2	1	1	1
Quanbun downs	3	1	1	1	1	1

**Table 10** Pearson correlation ( $r$ ) results with the different model sets in ARIMAX

Region	Station name	Pearson correlation ( $r$ ) for different model sets				
		DMI-Nino3	DMI-SOI	WTIO-SOI	WTIO-Nino3	WTIO-Nino4
Kimberley	Anna plains	–	–	0.83	–	0.84
	Bidyadanga	–	–	0.68	–	–
	Gogo station	–	–	0.65	–	–
	Quanbun downs	0.47	0.50	0.53	0.44	–

**Table 11** Model description of the selected ARIMAX models in the calibration period

Station name	Model type	Predictors	Lag month	Model fit statistics				Ljung-box Q (18)		
				Pearson correlation ( $r$ )	RMSE	MAE	Normalized BIC	Statistics	DF	Sig ( $p$ )
Anna plains	ARIMAX (0,1,1)	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.83	39.64	29.51	8.17	17.33	17	0.43
Bidyadanga	ARIMAX (0,1,1)	WTIO <sub>Aug</sub> -SOI <sub>May</sub>	4	0.68	57.90	41.97	8.94	8.82	17	0.94
Gogo station	ARIMAX (3,1,2)	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.65	40.01	27.49	20.21	20.21	13	0.09
Quanbun downs	ARIMAX (3,1,1)	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	1	0.53	43.56	36.33	8.48	8.72	14	0.84

influential climate indices using the GEP method and their respective Pearson correlation ( $r$ ) values.

As illustrated in Table 13, the WTIO-SOI model showed good prediction performance in all four rainfall stations. Apart from the WTIO-SOI model, WTIO-Nino3 also showed good predictability for Quanbun Downs only. However, to obtain a consistent and user-friendly model set for all these stations, only WTIO-SOI models were analyzed further. Table 14 presents the selected best models for each rainfall station with their statistical performances both in the calibration and validation period.

Table 14 shows that the Pearson correlation ( $r$ ) is increased in the validation period for all the selected rainfall stations except for Quanbun Downs. For all these models, relatively low *RMSE*, and *MAE* values also indicate them as good prediction models. Also, the *NSE* values ranged from 0.57 to 0.72 in the calibration period and from 0.45 to 0.70 in the validation period respectively, suggesting a good fit. Overall, for all the

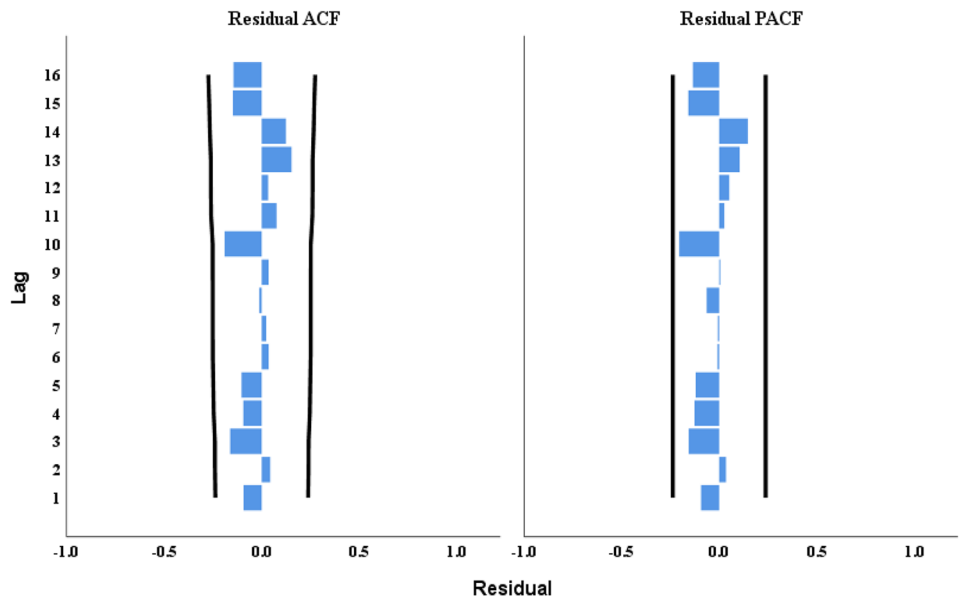
developed models, the *NSE* value has been found as above 0.50 that indicates a good fit for all the models. The refined Willmott index of agreement ( $d_r$ ) value close to or more than 0.70 in both calibration and validation period also indicates the skillfulness of the developed model. All the developed GEP models showed prediction capability of the seasonal summer rainfall four months in advance except for Quanbun Downs, in which the prediction deemed possible only one month in advance. Many other models were found with longer lead times but having lower correlation values, or higher errors for these selected stations, hence they were not chosen as the best model. During the evaluation procedure, it was discovered that as the number of generations increases, so do the correlation values. The correlation value, however, does not increase after reaching an optimum level. The number of generations got too big for a few stations to discover an optimal solution, and the statistical performance of those models appeared to be poor.

**Table 12** Model description of the selected ARIMAX model both in calibration and validation period

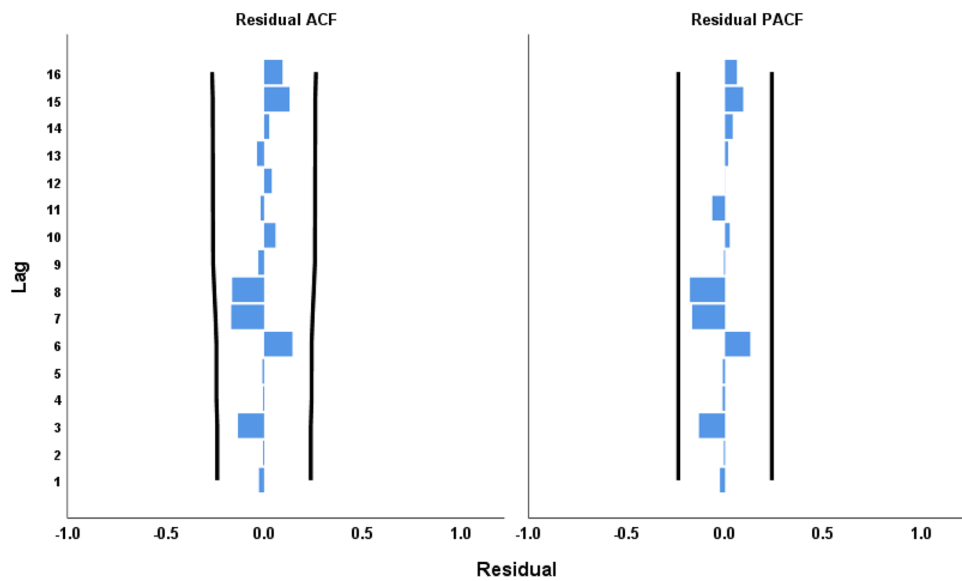
Station name	Model set	Calibration period				Validation period			
		Pearson correlation ( $r$ )	RMSE	MAE	Refined Willmot index of agreement ( $d_r$ )	Pearson correlation ( $r$ )	RMSE	MAE	Refined Willmot index of agreement ( $d_r$ )
Anna plains	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	0.83	39.64	29.51	0.68	0.65	53.54	40.97	0.64
Bidyadanga	WTIO <sub>Aug</sub> -SOI <sub>May</sub>	0.68	57.90	41.97	0.63	0.77	38.77	32.96	0.70
Gogo station	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	0.65	40.01	27.49	0.66	0.74	40.33	28.86	0.70
Quanbun downs	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	0.53	43.56	36.33	0.57	0.59	49.07	39.03	0.63



**Fig. 8** Residual ACF and PACF plots for **a** Anna Plains, **b** Bidyadanga, **c** Gogo Station and **d** Quانبun Downs



(a)



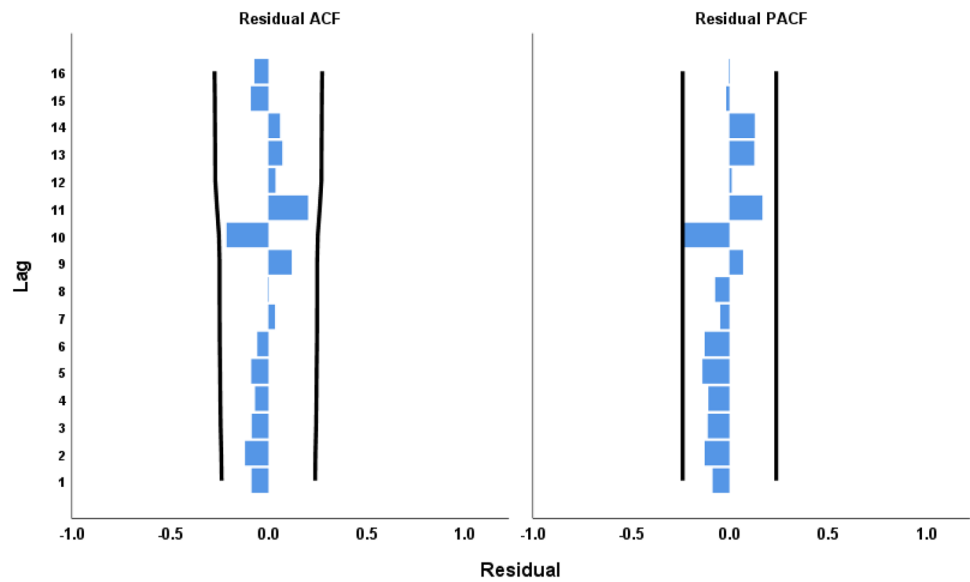
(b)

GEP models offer a unique feature of the model expression structure in terms of Expression Trees (ETs). The procedure of obtaining the output equation from ET has already been discussed in earlier Sect. 3.3.1. As an example Fig. 9 shows the ETs of Anna Plains. Furthermore, Table 15 presents the output equation only for the best model from each station considering higher correlation values with lower errors.

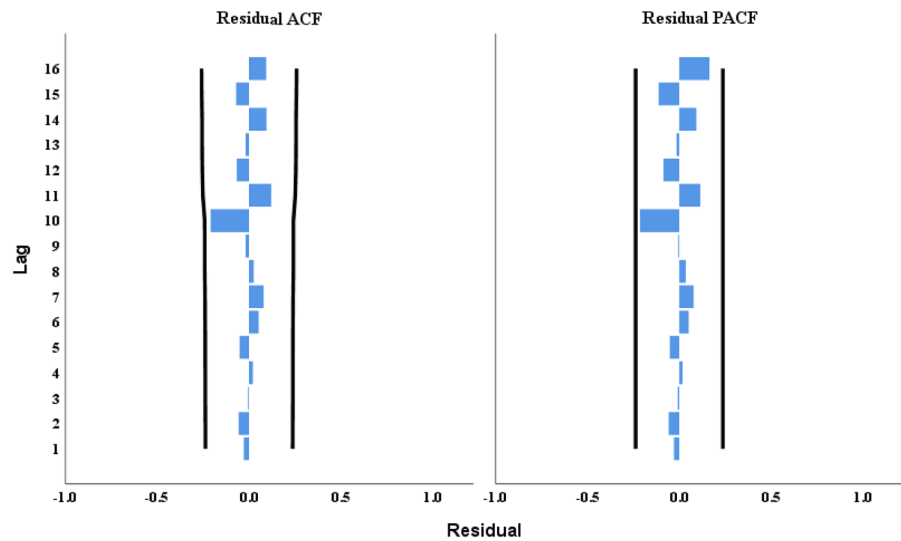
#### 4.5 Comparisons between MLR, ARIMAX, and GEP model

The statistical performance indicators of the best models developed using different forecast methods are presented in Table 16, where indicators like Pearson correlation ( $r$ ), refined Willmot Index of Agreement ( $d_r$ ), RMSE and MAE values for the selected model sets are presented. A schematic diagram of how the best performing models were selected for each stations are presented in Fig. 10.

Fig. 8 (continued)



(c)



(d)

For, Anna Plains, the  $WTIO_{April}-SOI_{Mar}$  model developed using MLR showed its capability to forecast summer rainfall eight months in advance with a correlation of 0.36 and 0.42 in calibration and validation periods respectively. However, for the ARIMAX model a different model set,

$WTIO_{Aug}-SOI_{Mar}$  showed its capability of forecasting four months before the event with a correlation of 0.83 and 0.65 in calibration and validation periods, respectively. The same model set, i.e., the  $WTIO_{Aug}-SOI_{Mar}$  model developed using GEP showed similar forecasting capability as ARIMAX

**Table 13** Pearson correlation ( $r$ ) results with the different model sets in GEP

Region	Station name	Pearson correlation ( $r$ ) for different model sets				
		DMI-Nino3	DMI-SOI	WTIO-SOI	WTIO-Nino3	WTIO-Nino4
Kimberley	Anna Plains	–	–	0.82	–	0.74
	Bidyadanga	–	–	0.85	–	–
	Gogo Station	–	–	0.76	–	–
	Quanbun Downs	0.74	0.70	0.76	0.58	–

**Table 14** Model description and selection of best GEP model in calibration and validation period

Station name	Model Set	Calibration period					Validation period				
		Pearson cor-	RMSE	MAE	NSE	Refined Willmot index of agreement ( $d_r$ )	Pearson cor- relation ( $r$ )	RMSE	MAE	NSE	Refined Willmot index of agreement ( $d_r$ )
		relation ( $r$ )									
Anna plains	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	0.82	36.74	27.52	0.66	0.70	0.82	39.63	28.96	0.67	0.75
Bidyadanga	WTIO <sub>Aug</sub> -SOI <sub>May</sub>	0.85	36.92	27.20	0.72	0.75	0.87	34.99	28.20	0.70	0.74
Gogo station	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	0.76	30.58	23.95	0.58	0.69	0.78	37.01	31.06	0.59	0.68
Quanbun downs	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	0.76	35.82	27.56	0.57	0.69	0.74	47.06	35.76	0.45	0.65

(i.e., four months in advance), with a significant correlation of 0.82 in both calibration and validation period.

For Bidyadanga, the WTIO<sub>Aug</sub>-SOI<sub>Nov</sub> model set developed using MLR showed a significant correlation of 0.44 and 0.51 in the calibration and validation periods, respectively. It showed a prediction capability of one-month in advance. In both ARIMAX and GEP methods, the WTIO<sub>Aug</sub>-SOI<sub>May</sub> model set showed promising performance pertaining high correlations in both calibration and validation periods. In calibration and validation periods, the ARIMAX model returned a correlation value of 0.68 and 0.77 respectively, wherein for the GEP model, the correlation values obtained were as high as 0.85 and 0.87 respectively. Both the model sets showed their prediction capability four months in advance.

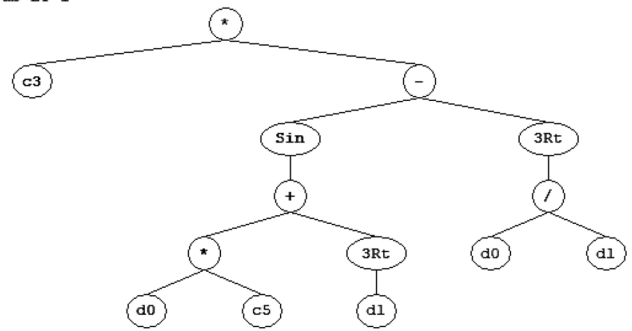
For Gogo Station, the WTIO<sub>Aug</sub>-SOI<sub>Mar</sub> model set showed the best correlation in MLR, ARIMAX, and GEP methods, whereas, WTIO<sub>Oct</sub>-SOI<sub>Nov</sub> model set was outstanding in performance for Quanbun Downs. For Gogo Station, Pearson correlation values for calibration and validation periods were found as 0.36 and 0.35 (MLR), 0.65 and 0.74 (ARIMAX), and 0.76 and 0.78 (GEP). Similarly, for Quanbun Downs, correlation values were reported as 0.32 and 0.36 (MLR), 0.53 and 0.59 (ARIMAX), and 0.76 and 0.74 (GEP) in calibration and validation periods, respectively. For Gogo Station, the selected model has been found as capable to predict four months in advance, where, for Quanbun Downs, the selected model can predict just a month earlier than the actual event.

While the performance of the forecasting methods is compared in terms of Pearson correlation ( $r$ ) values, it is evident that GEP models have shown impressive performance than the other two methods. In calibration and validation periods, the correlation ( $r$ ) value in the GEP model ranged from 0.76 to 0.82 and 0.74 to 0.87, respectively. These correlation values are quite outstanding compared to MLR model (0.32 to 0.36 in calibration and 0.35 to 0.51 in validation period) and ARIMAX model (0.53 to 0.83 in calibration and 0.59 to 0.77 in validation period). This is also depicting a greater underlying contribution of the predictors is evident in the developed models, as a Pearson correlation ( $r$ ) value of more than 0.5 is considered as a large effect (Field 2013).

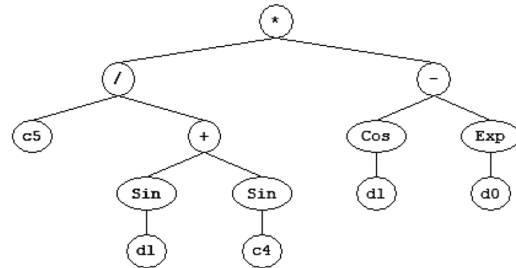
A similar observation about the GEP model's superiority over others was made in terms of the refined Willmot index of agreement ( $d_r$ ) values. The refined Willmot index of agreement ( $d_r$ ) value is an indicator of the model fitness, where, a relatively high positive value indicates a good fit (Willmott et al. 2012). In the GEP model, the ' $d_r$ ' value in calibration and validation period ranged from 0.69 to 0.75 and 0.65 to 0.75 respectively. Such positive ' $d_r$ ' values are

**Fig. 9** Expression tree of anna plains

Sub-ET 1



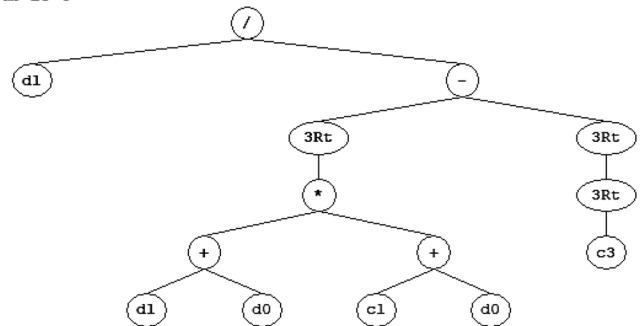
Sub-ET 2



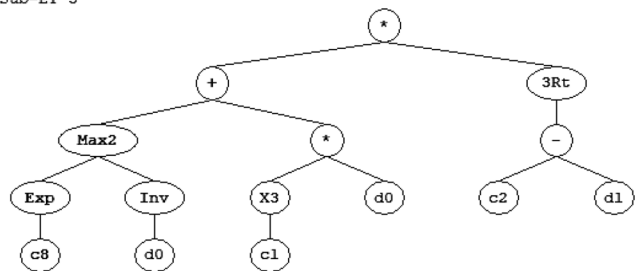
Sub-ET 3



Sub-ET 4



Sub-ET 5



**Table 15** Output equation of best developed GEP model for the selected rainfall stations

Rainfall station	Output equation
Anna plains	$25 * \left[ \sin(6.93d_0 + \sqrt[3]{d_1}) - \sqrt[3]{\frac{d_0}{d_1}} \right] + \left[ \frac{3.65}{\sin(d_1) - \sin(5.25)} * \{\cos(d_1) - \exp(d_0)\} \right] + \exp(4.78) + \left[ \frac{d_1}{\sqrt[3]{(d_1+d_0)^*(d_0-2.23)}} - \sqrt[3]{2.38} \right] + \left[ \text{Max} \left\{ \exp(3.49), \frac{1}{d_0} \right\} \right]$ $+ \{(-3.14)^3 * d_0\} * \sqrt[3]{-1.81 - d_1}$ <p>Here, <math>d_0 = WTIO_{Aug}</math> and <math>d_1 = SOI_{Mar}</math></p>
Bidyadanga	$\frac{1}{\cos \left[ \frac{2.46}{\text{Min}(d_1, 1.00d_1)} \right]^2} + \frac{1}{\cos \left[ \frac{5.96}{\text{Min}(d_0+5.90d_1, d_1)} \right]^2} + \frac{1}{\cos(\text{Avg}(d_1 * (8.1-d_0), d_0 * (78.03+d_0)))} + \left[ \text{Min} \left( 3.94, \frac{1}{-7*d_0} \right) + 2.92 \right]^2 * d_1 + \sqrt{\exp\{\text{Min}(d_0, \cos d_1^2)\}} * 9.74^2$ <p>Here, <math>d_0 = WTIO_{Aug}</math> and <math>d_1 = SOI_{May}</math></p>
Gogo station	$\left[ \text{Avg} \left\{ -14.79 * (d_1 - 4.35), \text{Max} \left( -\frac{2.96}{d_1}, -4.5 \right) \right\} * \sqrt[3]{d_0} \right] + \left[ \text{Avg} \left\{ -11.22, \sin(-6.25 * d_1^2) * \exp(1.35)^2 \right\}^2 + \frac{\text{Log}(\frac{d_0}{2.5}, \sqrt[3]{d_1})}{\frac{d_1+3.65}{\text{Avg}(d_1, 2.19)}} \right] * d_1$ $+ \left[ \{-7.52 + \text{Min}(0.49, d_1)\}^2 - \{\sqrt[3]{\sin(d_1)} * (-18.19 * d_1)\} \right]$
Quantun downs	<p>Here, <math>d_0 = WTIO_{Aug}</math> and <math>d_1 = SOI_{Mar}</math></p> $\text{Avg} \left\{ \left\{ \sin(d_1), d_1^2 \right\} - 3.91 \right\} * d_1^4 + \frac{1}{\text{Avg} \left\{ d_1^2, (d_0 - 3.02) \right\} * \text{Avg}(d_1^2, \sqrt[3]{(-0.94)})} + \left[ d_1^3 + \left\{ \text{Avg}(7.64, d_0) \right\}^3 \right] * 2.26 + \left[ \text{Max} \left[ \text{Avg} \left\{ (-25.83 * d_1), \frac{1}{d_0} \right\} * \frac{1}{d_1}, \frac{1}{\text{Avg}(d_1, 0.58)} \right] \right] +$ $\text{Min} \left[ \left[ -21.93, [-21.93 * 4.22 * \{\text{Avg}(d_1, d_0) + d_0\}] \right] \right]$ <p>Here, <math>d_0 = WTIO_{Oct}</math> and <math>d_1 = SOI_{Nov}</math></p>

**Table 16** Comparison of Pearson's correlation ( $r$ ), refined Willmot index of agreement ( $d_r$ ), RMSE and MAE between MLR, ARIMAX, and GEP model

Station name	Model type	Predictors	Lag month	Pearson's correlation ( $r$ )		Refined Willmot index of agreement ( $d_r$ )		RMSE		MAE	
				Calibration	Validation	Calibration	Validation	Calibration	Validation	Calibration	Validation
Anna plains	MLR	WTIO <sub>April</sub> -SOI <sub>Mar</sub>	8	0.36	0.42	0.53	0.56	59.07	62.68	43.06	50.50
	ARIMAX	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.83	0.65	0.68	0.64	39.94	53.54	29.51	40.97
	GEP	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.82	0.82	0.70	0.75	36.74	39.63	27.52	28.96
Bidyardanga	MLR	WTIO <sub>Aug</sub> -SOI <sub>Nov</sub>	1	0.44	0.51	0.54	0.56	61.93	59.72	47.31	49.10
	ARIMAX	WTIO <sub>Aug</sub> -SOI <sub>May</sub>	4	0.68	0.77	0.63	0.70	57.90	38.77	41.97	32.96
	GEP	WTIO <sub>Aug</sub> -SOI <sub>May</sub>	4	0.85	0.87	0.75	0.74	36.92	34.99	27.20	31.06
Gogo station	MLR	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.36	0.35	0.54	0.52	43.95	54.29	35.88	46.47
	ARIMAX	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.65	0.74	0.6	0.70	40.01	40.33	27.49	28.86
	GEP	WTIO <sub>Aug</sub> -SOI <sub>Mar</sub>	4	0.76	0.78	0.69	0.68	30.58	37.01	23.95	31.06
Quanbun downs	MLR	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	1	0.32	0.36	0.52	0.54	51.89	57.33	41.80	48.81
	ARIMAX	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	1	0.53	0.59	0.57	0.63	43.56	49.07	36.33	39.03
	GEP	WTIO <sub>Oct</sub> -SOI <sub>Nov</sub>	1	0.76	0.74	0.69	0.65	35.82	47.06	27.56	35.76

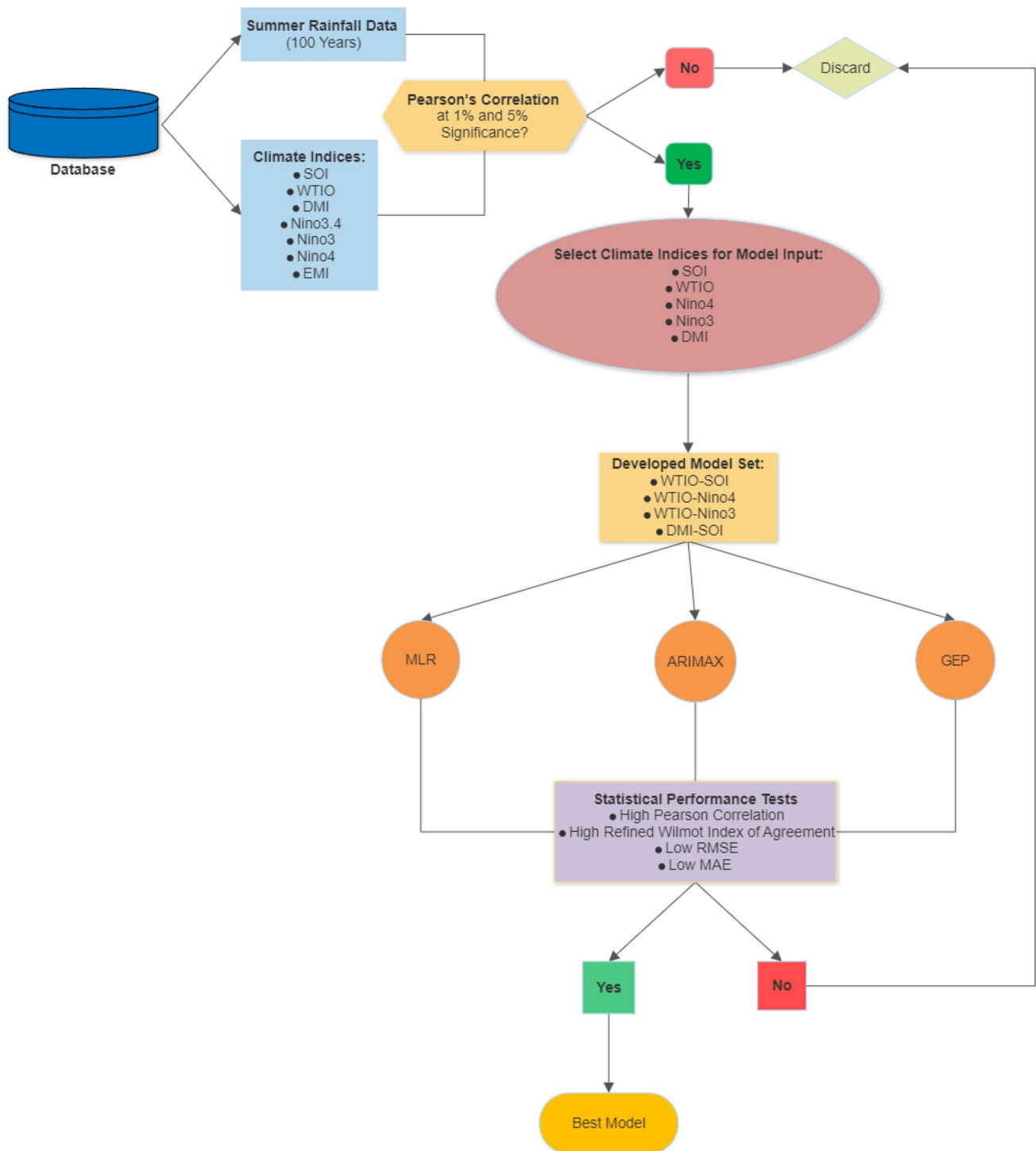
a good indicator for GEP model's prediction capability over MLR and ARIMAX models.

The MLR, ARIMAX, and GEP models were further compared considering their error measurement statistics. As presented in Table 16, the developed GEP model returned relatively low RMSE and MAE values in both calibration and validation periods, while compared to the MLR and ARIMAX models. As a lower value of RMSE and MAE indicates a better predictability performance of the developed models, the GEP model can be considered as best compared to the rest of the developed models (Saigal and Mehrotra 2012; Singh et al. 2005).

To understand the predictability performance of the developed models at different rainfall stations, observed versus predicted plots were drawn for selected MLR, ARIMAX, and GEP models. As presented in Fig. 11, it is evident that the GEP model showed prominent prediction performance as almost mimicking the trend of the naturally occurred rainfall events. GEP model's trend also demonstrated its capability of capturing extreme events such as heavy rainfalls and droughts. As observed, the ARIMAX model showed moderate prediction performance as it successfully captured some of the extreme events while been failed in other instances. In contrast, the MLR model showed relatively poor performance among all the three techniques, capturing none of the extreme events as well as demonstrating instances of underestimation and overestimation of the events. Both MLR and ARIMAX models, being a linear and time-series approach respectively were outperformed by the non-linear GEP model in which the existing non-linear relationship between rainfall and climate indices was considered to better explain the underlying variability.

### 5 Conclusion

In this study, the rainfall forecasting capability of an artificial intelligence method (i.e., GEP) was evaluated against the conventional linear method (i.e. MLR) and time-series technique (i.e., ARIMAX), where lagged climate indices were used as input variables. Monthly summer rainfall events for four stations located in the Kimberley region of NWWA were analyzed, and the investigation resulted in the identification of influential climate indices and their interactions responsible for the region's summer rainfall variability. Climate indices namely WTIO and SOI were found as the most dominant factor contributing to the rainfall events for the NWWA region. Thus, several model sets were developed using the combination of WTIO-SOI at different lagged months and used as the input sets for



**Fig. 10** A schematic diagram of selection of best performing model

MLR, ARIMAX, and GEP models. To achieve the best prediction results, a different combination of model sets were analyzed for different techniques that returned different prediction performances for different models at different lead times.

Overall, the prediction model developed using the GEP technique showed good predictability compared to MLR and ARIMAX techniques for all four rainfall stations. The WTIO-SOI model set for the GEP model showed a high correlation ranging from 0.76 to 0.85 in the calibration period and 0.74 to 0.84 in the validation period.







**Funding** Open Access funding enabled and organized by CAUL and its Member Institutions.

## Declarations

**Conflicts of interest** No conflict of interest to be reported.

**Data availability** Data and materials will be available upon request to the first author.

**Code availability** Code will be available upon request to the first author.

**Ethics approval** Not required as no animal or human were involved in the study

**Consent to participate** We agree to participate any relevant survey or other activity.

**Consent for publication** We agree Springer to publish the article with proper copyright.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abbot J, Marohasy J (2012) Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv Atmos Sci* 29:717–730
- 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
- Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. 2012. Comparison of multiple linear and non-linear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research*, 48.
- Adhikary SK, Muttill N, Gokhan Yilmaz A (2016a) Ordinary kriging and genetic programming for spatial estimation of rainfall in the Middle Yarra River catchment, Australia. *Hydrol Res* 47:1182–1197
- Adhikary SK, Muttill N, Yilmaz AG (2016b) Genetic programming-based ordinary kriging for spatial interpolation of rainfall. *J Hydrol Eng* 21:04015062
- Akhtar M, Corzo G, Van Andel S, Jonoski A (2009) River flow forecasting with artificial neural networks using satellite observed precipitation pre-processed with flow length and travel time information: case study of the Ganges river basin. *Hydrol Earth Syst Sci* 13:1607–1618
- Alavi AH, Gandomi AH (2011) A robust data mining approach for formulation of geotechnical engineering systems. *Int J for Computer-Aided Engineering* 28:242–274
- Ashok K, Guan Z, Yamagata T (2003b) A Look at the Relationship between the ENSO and the Indian Ocean Dipole. *J Meteorol Soc Jpn* 81:41–56
- Ashok, K., Guan, Z. & Yamagata, T. 2003a. Influence of the Indian Ocean Dipole on the Australian winter rainfall. *Geophysical Research Letters*, 30.
- Azamathulla HM (2012) Gene expression programming for prediction of scour depth downstream of sills. *J Hydrol* 460:156–159
- Azamathulla HM, Ghani AA (2010) Genetic programming to predict river pipeline scour. *Journal of Pipeline Systems Engineering* 1:127–132
- Azamathulla HM, Ghani AA, Zakaria NA, Guven A (2010) Genetic programming to predict bridge pier scour. *J Hydraul Eng* 136:165–169
- Box, G. E. & Jenkins, G. 1976. *Time series analysis: Forecasting and control*
- Cai W, Van Rensch P, Cowan T, Hendon HH (2011) Teleconnection pathways of ENSO and the IOD and the mechanisms for impacts on Australian rainfall. *J Clim* 24:3910–3923
- Cai, W. & Cowan, T. 2006. SAM and regional rainfall in IPCC AR4 models: Can anthropogenic forcing account for southwest Western Australian winter rainfall reduction? *Geophysical Research Letters*, 33.
- Chiang YM, Chang FJ (2009) Integrating hydrometeorological information for rainfall-runoff modelling by artificial neural networks. *Hydrol Process* 23:1650–1659
- Chowdhury RK, Beecham S (2013) Influence of SOI, DMI and Niño3. 4 on South Australian rainfall. *Stoch Env Res Risk Assess* 27:1909–1920
- Corporation, I. 2013. IBM SPSS Forecasting 22. Armonk, NY, USA.
- Cryer, J. D. & Chan, K.-S. 2008. *Time series analysis: with applications in R*, Springer Science & Business Media.
- Drecourt J-P (1999) Application of neural networks and genetic programming to rainfall runoff modeling. *Water Resour Manage* 13:219–231
- Drosowsky W, Chambers LE (2001) Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J Clim* 14:1677–1687
- England MH, Ummerhofer CC, Santoso A (2006) Interannual rainfall extremes over southwest Western Australia linked to Indian Ocean climate variability. *J Clim* 19:1948–1969
- Esha RI, Imteaz MA (2020) Pioneer use of Gene Expression Programming for predicting seasonal streamflow in Australia using large scale climate drivers. *Ecohydrology*. <https://doi.org/10.1002/eco.2242>
- Evans FH, Guthrie MM, Foster I (2020) Accuracy of six years of operational statistical seasonal forecasts of rainfall in Western Australia (2013 to 2018). *Atmos Res* 233:104697
- De Falco, I., Della Cioppa, A. & Tarantino, E. 2005. A genetic programming system for time series prediction and its application to El Niño forecast. *Soft Computing: Methodologies and Applications*. Springer.
- Feng J, Li J, Li Y (2010) Is there a relationship between the SAM and southwest Western Australian winter rainfall? *J Clim* 23:6082–6089
- Feng J, Li J, Xu H (2013) Increased summer rainfall in northwest Australia linked to southern Indian Ocean climate variability. *Journal of Geophysical Research: Atmospheres* 118:467–480
- Feng J, Li J, Li Y, Zhu J, Xie F (2015) Relationships among the monsoon-like southwest Australian circulation, the southern annular mode, and winter rainfall over southwest western Australia. *Adv Atmos Sci* 32:1063–1076
- Fernando AK, Shamseldin AY, Abrahart RJ (2012b) Use of gene expression programming for multimodel combination of rainfall-runoff models. *J Hydrol Eng* 17:975–985

- Fernando A, Shamseldin A, Abrahart R (2012a) River flow forecasting using gene expression programming models.
- Ferranti L (2012) Calibration and validation of seasonal forecasts. ECMWF Seminar on Seasonal Prediction, 3–7 September, 2012.
- Ferreira C (2001) Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems*.
- Field A (2013) *Discovering statistics using IBM SPSS statistics*, Sage.
- Fierro AO, Leslie LM (2013) Links between central west Western Australian rainfall variability and large-scale climate drivers. *J Clim* 26:2222–2246
- Foootan E, Awange J, Schumacher M, Anyah R, van Dijk A, Kusche J (2016) Quantifying the impacts of ENSO and IOD on rain gauge and remotely sensed precipitation products over Australia. *Remote Sens Environ* 172:50–66
- Gandomi A.H, Alavi, AH (2013) *Expression programming techniques for formulation of structural engineering systems*, Chapter.
- Ghamariadyan M, Imteaz MA (2020) A Wavelet Artificial Neural Network method for medium-term rainfall prediction in Queensland (Australia) and the comparisons with conventional methods. *Int J Climatol* 41(S1):E1396–E1416. <https://doi.org/10.1002/JOC.6775>
- Ghamariadyan M, Imteaz MA (2021a) Prediction of seasonal rainfall with one-year lead time using climate indices: a wavelet neural network scheme. *Water Resour Manag* 35:5347–5365. <https://doi.org/10.1007/s11269-021-03007-x>
- Ghamariadyan M, Imteaz MA (2021b) Monthly rainfall forecasting using temperature and climate indices through a hybrid method in Queensland, Australia. *J Hydrometeorol* 22(5):1259–1273. <https://doi.org/10.1175/JHM-D-20-0169.1>
- Goddard L, Mason SJ, Zebiak SE, Ropelewski CF, Basher R, Cane MA (2001) Current approaches to seasonal to interannual climate predictions. *Int J Climatol* 21:1111–1152
- Gupta HV, Kling H (2011) On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. *Water Resources Research*, 47.
- Güven A, Aytekin A (2009) New approach for stage–discharge relationship: gene-expression programming. *J Hydrol Eng* 14:812–820
- Güven A, Günal M, Engineering D (2008) Genetic programming approach for prediction of local scour downstream of hydraulic structures. *Journal of Irrigation* 134:241–249
- Hashmi MZ, Shamseldin AY, Melville BW (2011) Statistical downscaling of watershed precipitation using Gene Expression Programming (GEP). *Environmental Modelling Software* 26:1639–1646
- Hossain I, Esha RI, Imteaz MA (2018a) An attempt to use non-linear regression modelling technique in long-term seasonal rainfall forecasting for Australian Capital Territory. *Geosciences* 8(8):1–12. <https://doi.org/10.3390/geosciences8080282>
- Hossain I, Rasel H, Imteaz MA, Mekanik F (2018b) Long-term seasonal rainfall forecasting: efficiency of linear modelling technique. *Environmental Earth Sciences* 77:280
- Hossain I, Rasel HM, Mekanik F, Imteaz MA (2020) Artificial Neural Network modelling technique in predicting Western Australian seasonal rainfall. *Int J Water* 14(1):14–28
- Islam F, Imteaz MA (2019) Development of prediction model for forecasting rainfall in Western Australia using lagged climate indices. *Int J Water* 13:248–268
- Islam F, Imteaz MA (2020) Use of Teleconnections to Predict Western Australian Seasonal Rainfall Using ARIMAX Model. *Hydrology* 7:52
- Kamruzzaman M, Beecham S, Metcalfe A (2013) Climatic influences on rainfall and runoff variability in the southeast region of the Murray-Darling Basin. *Int J Climatol* 33:291–311
- Kamruzzaman M, Beecham S, Metcalfe AV (2017) Changing patterns in rainfall extremes in South Australia. *Theoret Appl Climatol* 127:793–813
- Khu ST, Liang SY, Babovic V, Madsen H, Muttil N (2001) Genetic programming and its application in real-time runoff forecasting. *J Am Water Resour Assoc* 37:439–451
- Kirono DG, Chiew FH, Kent DM (2010) Identification of best predictors for forecasting seasonal rainfall and runoff in Australia. *Hydrol Process* 24:1237–1247
- Koza JR (1994) Genetic programming as a means for programming computers by natural selection. *Stat Comput* 4:87–112
- Lin Z, Li Y (2012) Remote influence of the tropical Atlantic on the variability and trend in North West Australia summer rainfall. *J Clim* 25:2408–2420
- Ljung GM, Box GE (1978) On a measure of lack of fit in time series models. *Biometrika* 65:297–303
- Marshall A, Hendon H (2014) Impacts of the MJO in the Indian Ocean and on the Western Australian coast. *Clim Dyn* 42:579–595
- McBride JL, Nicholls N (1983) Seasonal relationships between Australian rainfall and the Southern Oscillation. *Mon Weather Rev* 111:1998–2004
- McCuen RH, Knight Z, Cutter AG (2006) Evaluation of the Nash-Sutcliffe efficiency index. *J Hydrol Eng* 11:597–602
- 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
- Mekanik F, Imteaz M, Talei A (2016) Seasonal rainfall forecasting by adaptive network-based fuzzy inference system (ANFIS) using large scale climate signals. *Clim Dyn* 46:3097–3111
- Montazerolghaem M, Vervoort W, Minasny B, McBratney A (2016) Long-term variability of the leading seasonal modes of rainfall in south-eastern Australia. *Weather and Climate Extremes* 13:1–14
- Murphy BF, Timbal B (2008) A review of recent climate variability and climate change in southeastern Australia. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 28:859–879
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. *J Hydrol* 10:282–290
- Nicholls N (2010) Local and remote causes of the southern Australian autumn-winter rainfall decline, 1958–2007. *Clim Dyn* 34:835–845
- Rasel H, Imteaz M, Mekanik F (2016) Investigating the influence of Remote Climate Drivers as the Predictors in Forecasting South Australian spring rainfall. *International Journal of Environmental Research* 10:1–12
- Risbey JS, Pook MJ, McIntosh PC, Wheeler MC, Hendon HH (2009) On the remote drivers of rainfall variability in Australia. *Mon Weather Rev* 137:3233–3253
- Rotstayn L, Jeffrey SJ, Collier MA, Dravitzki S, Hirst A, Syktus J, Wong K, Physics, (2012) Aerosol-and greenhouse gas-induced changes in summer rainfall and circulation in the Australasian region: a study using single-forcing climate simulations. *Atmospheric Chemistry* 12:6377
- Saigal S, Mehrotra D (2012) Performance comparison of time series data using predictive data mining techniques. *Advances in Information Mining* 4:57–66
- Savic DA, Walters GA, Davidson JW (1999) A genetic programming approach to rainfall-runoff modelling. *Water Resour Manage* 13:219–231
- Schepen A, Wang Q, Robertson D (2012) Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *J Clim* 25:1230–1246
- Shabani S, Candelieri A, Archetti F, Naser G (2018) Gene expression programming coupled with unsupervised learning: a two-stage

- learning process in multi-scale, short-term water demand forecasts. *Water* 10:142
- Shi G, Cai W, Cowan T, Ribbe J, Rotstayn L, Dix M (2008) Variability and trend of North West Australia rainfall: observations and coupled climate modeling. *J Clim* 21:2938–2959
- Shiri J, Kisi O, Yoon H, Lee K-K, Nazemi AH (2013) Predicting groundwater level fluctuations with meteorological effect implications-A comparative study among soft computing techniques. *Comput Geosci* 56:32–44
- Shiri J, Marti P, Singh VP (2014a) Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. *J Hydrological Processes* 28:1215–1225
- Shiri J, Sadraddini AA, Nazemi AH, Kisi O, Landaras G, Fard AF, Marti P (2014b) Generalizability of gene expression programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. *J Hydrol* 508:1–11
- Singh J, Knapp HV, Arnold J, Demissie M (2005) Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *J Am Water Resour Assoc* 41:343–360
- Smith I, McIntosh P, Ansell T, Reason C, McInnes K (2000) Southwest Western Australian winter rainfall and its association with Indian Ocean climate variability. *Int J Climatol* 20:1913–1930
- Taschetto AS, England MH (2009) El Niño Modoki impacts on Australian rainfall. *J Clim* 22:3167–3174
- Thirumalaiah K, Deo MC (2000) Hydrological forecasting using neural networks. *J Hydrol Eng* 5:180–189
- Tibaldi S, Tosi E, Navarra A, Pedulli L (1994) Northern and Southern Hemisphere seasonal variability of blocking frequency and predictability. *Mon Weather Rev* 122:1971–2003
- Tozer, C. R. 2014. *Utilising Insights into Rainfall Patterns and Climate Drivers to Inform Seasonal Rainfall Forecasting in South Australia*. PhD thesis.
- Tularam G (2010) Relationship between El Niño southern oscillation index and rainfall (Queensland, Australia). *Int J Sustain Dev Plan* 5:378–391
- Ummenhofer CC, Sen Gupta A, Pook MJ, England MH (2008) Anomalous rainfall over southwest Western Australia forced by Indian Ocean sea surface temperatures. *J Clim* 21:5113–5134
- Vaze, J., Jordan, P., Beecham, R., Frost, A. & Summerell, G. 2012. Guidelines for rainfall-runoff modelling. *Australian Government Department of Innovation, Industry, science and Research*.
- Verdon, D. C., Wyatt, A. M., Kiem, A. S. & Franks, S. W. 2004. Multi-decadal variability of rainfall and streamflow: Eastern Australia. *Water Resources Research*, 40.
- Willmott CJ, Robeson SM, Matsuura K (2012) A refined index of model performance. *Int J Climatol* 32:2088–2094
- Yilmaz A, Imteaz M, Jenkins G (2011) Catchment flow estimation using Artificial Neural Networks in the mountainous Euphrates Basin. *J Hydrol* 410:134–140
- Zorn CR, Shamseldin AY (2015) Peak flood estimation using gene expression programming. *J Hydrol* 531:1122–1128

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.