Assessing Gene Expression Programming as a technique for seasonal streamflow prediction: A case study of NSW

R I Esha¹,², M A Imteaz¹ and A Nazari¹
¹Swinburne University of Technology, Melbourne, Australia
²E-mail: resha@swin.edu.au

Abstract. This research aims to provide long term streamflow forecast models using multiple climate indices as the predictors with the help of an advanced evolutionary method, Gene Expression Programming (GEP) to solve the developed symbolic regression problems as it is found to be superior than other traditional methods. Being a transparent model, GEP is able to provide the relationship between input (climate indices) and output (streamflow) variables with mathematical expressions which help the users to understand the underlying hydrological process between the climate mode and streamflow without having much knowledge about the used software. Two stations of New South Wales (NSW) are chosen based on their longer data record and fewer missing values. Several preliminary researches including single and multiple correlation analyses reveal PDO (Pacific Decadal Oscillation), IPO (Inter Decadal Pacific Oscillation), IOD (Indian Ocean Dipole) and ENSO (El Nino Southern Oscillation) are few among the influential indices on the study region. The resultant models appear to be more efficient with up to 50% higher Pearson correlation (r) values than that of the simple MLR technique adapted in one of our previous studies. Furthermore, the statistical performance analyses including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Willmott index of agreement (d) and Nash-Sutcliffe efficiency (NSE) ensure high predictability of the developed models. The similar correlation values (r) generated from calibration and validation periods which ranges between 0.74 and 0.91 increase the reliability of the resultant models for predicting seasonal streamflow up to three months in advance.

1. Introduction
An efficient forecast model is of immense importance to the water stakeholders as it can influence the low-risk decision making process to a great extent which will enhance the potential economic benefit as well [1,2]. As the remote climate drivers fluctuate at very low frequencies, they have better predictability of streamflow while comparing to initial catchment condition.

Australia is greatly influenced by the climate anomalies originated from the surrounding Pacific, Indian and Southern oceans. Some of the major climate drivers influencing southeast Australia’s climate include but not limited to ENSO, IPO (PDO), SAM (Southern Annular mode) and IOD [3] originated from Pacific Ocean, the Indian Ocean and the Southern Ocean respectively.

The ENSO (El Niño-Southern Oscillation) phenomenon which causes climate variability in many parts of the world is the direct consequence of large-scale interactions between ocean and atmospheric
circulation patterns in the equatorial Pacific Ocean [4]. Researchers have investigated the extent of influences of ENSO on streamflow in Australia [5-10]. The El Niño–Southern Oscillation Modoki (EMI) events are strongly impactful on the climates of various regions across the world which includes Japan, New Zealand, western coast of United States [11], Australia [12], South China [13]. In addition to ENSO indices, eastern Australia is evident to be influenced by IOD [14] along with the interdecadal modulation impact of ENSO originated from the low frequency variability in Pacific Ocean, which is recognized as Pacific Decadal Oscillation or PDO [15]. Past research studies have investigated [16-18] the influence of Interdecadal Pacific Oscillation (IPO) on the decadal to multidecadal variation of rainfall and streamflow and suggested that IPO should be considered as a dominant climate index. IPO and PDO are highly correlated as these are the two indices of the same phenomenon where IPO is defined for the whole Pacific Basin and PDO is defined for the North Pacific (pole ward of 20°N) Basin.

Many hydrologists established the existence of strong correlations between streamflow and large-scale climate drivers, though the nature of the relationship remained a difficult question to deal with. According to [5] the relationship between streamflow and remote climate drivers is more likely to be non-linear, thus a non-linear model is expected to give better solutions than a linear model.

There are five different forecasting methods in practice which are time series analysis, regression analysis, artificial intelligence method (e.g. ANN, fuzzy logic etc), the hybrid and Monte Carlo simulation methods. Some of the existing data-driven techniques include statistical methods like Simple Regression (SR), Multi-Linear Regression (MLR) and Auto Regressive Moving Average (ARMA) in addition to Artificial Intelligence (AI) methods like Artificial Neural Networks (ANNs), Adaptive Neuro-Fuzzy Inference System (ANFIs), Genetic Programming (GP) and Genetic Expression Programming (GEP) [19-27].

Though the difficulties in dealing with artificial intelligence models encourage the users to attempt comparatively simple statistical models, the limitations of these models are evident when data become complex. In addition, one of the major advantages of artificial intelligence-based models like GEP and ANN over regression-based models is, they do not impose any fixed model structure on the data, rather they allow the data itself to identify the model structure by using artificial intelligence [28] which sometimes makes the models more robust.

One of the main advantages of GEP models over some other data-driven models is that the resultant model is not a complete “black-box”, rather the relationship between input (climate indices) and output (streamflow) variables can be explained with mathematical expression (combination of basic operators and functions). GEP was found to give better performance than other data-driven methods such as ANN and ANFIS [29-32]. Reference [25] investigated the comparative performance of ANN, ANFIS, GEP and ARMA models to forecast lake levels in Turkey and concluded that GEP was the better performer among all other data driven models. GEP model was suggested as a feasible alternative to ANN, ANFIS and MLR time series when these models were applied to simulate rainfall-runoff transformation process [25,32] successfully generated flow duration curves using non-linear regression equations developed by GEP. ANN and GEP were compared for estimating reference evapotranspiration where GEP provided explicit equations, even though the outcomes from ANN were slightly better than GEP [33]. In the study of reference [34] GEP was efficient to predict one day ahead river flow with high correlation coefficient.

The study intends to provide deterministic forecast as it can play more important roles in solving water management problems by enabling the water stakeholders to take accurate decisions knowing the predicted amount of future streamflow, rather than the probabilistic approaches which have been attempted by many researchers till date [3,5,35-37]. Furthermore, the Bayesian joint probability (BJP) method used by Australian Bureau of Meteorology (http://www.bom.gov.au/water/ssi/index.shtml) to provide futuristic streamflow is again a probabilistic method. Therefore, the present study is an endeavor to explore the non-linear relationship between climate indices and seasonal streamflow with a view to providing deterministic forecast.
2. Study area and data
North-east part of New South Wales is chosen as the current study area considering its agricultural importance, climatic variation and geographic location. Three streamflow stations (Figure 1) Hunter River at Singleton (Station ID 210001), Goulburn River at Coggan (Station ID 210006) and Namoi River at North Cuerindi (Station ID 419005) were selected based on their long data records and fewer missing values (less than 5%).

![Map of the study area showing the locations of the streamflow stations.](image)

Five climate indices: ENSO based SST anomalies NINO3.4, EMI, IPO, PDO, and DMI (IOD) were chosen for the analysis based on the previous research works on streamflow and rainfall in the study area. The oceanic and atmospheric climate indices data are collected from Climate Explorer website (http://climexp.knmi.nl) while the EMI data is obtained from the website of JAMSTEC (http://www.jamstec.go.jp/frcgc/research/dl/iod/modoki) for duration of 102 years (1914-2015). An overview of the used climatic variables is presented in Table 1.

102 years of observed monthly streamflow (in cumec) was collected from the Australian Bureau of Meteorology (BOM) (http://www.bom.gov.au/waterdata/) which ranges from 1914 to 2015. The whole dataset was divided into two segments where 96 years’ (from 1914 to 2009) data was selected for the calibration of the models and the remaining 6 years’ (from 2010 to 2015) data was used for validating the models. Using the collected monthly streamflow data, seasonal mean discharge data was derived for spring (September-October-November).

| Predictors | Predictor Definition | Origin | Data Period | Data Source |
|------------|----------------------|--------|-------------|-------------|
| PDO        | SST anomaly in North Pacific Ocean, (north of 20°N Latitude) | Pacific Ocean | 1914-2015 | ERSST (http://climexp.knmi.nl/) |
| IPO        | SST anomaly in North and South Pacific Ocean (includes South of 20°N Latitude) | Pacific Ocean | 1914-2015 | HadISST1 (http://climexp.knmi.nl/) |
| Nino3.4    | Average SST anomaly over central Pacific Ocean(5°S-5°N,120°-170°W) | Pacific Ocean | 1914-2015 | HadISST1 (http://climexp.knmi.nl/) |
3. Methodology
GEP is a search technique which involves individuals of non-linear structures with different size and shape encoded in linear chromosomes of fixed lengths. The genotype i.e., chromosomes (which are generally consisted of more than one gene of equal length) and the phenotype i.e., expression trees or ETs (expression of the genetic information encoded in chromosomes) are the two important entities of GEP which are different both structurally and functionally. The head contains both functions(nodes) and terminal symbols(leaves) while the tail contains only the terminal symbols. In each problem, the head length (h) is decided by the user and tail length (t) is formulated by the equation:

\[ t = (n-1)h + 1 \]  

where n is the number of arguments of the functions.

The steps to predict streamflow are as follows:
- Initially random population is generated which is consisted of individual chromosomes of fixed length.
- Each chromosome in the initial population is expressed by the expression trees and evaluated using the predicted-observed data pairs of the training period as well as an appropriate fitness function.
- The next step is to determine the set of terminals T and the set of functions F to create the chromosomes.
- The next major step is to select the chromosomal architecture which is composed of length of head (h), the number of genes per chromosome and genetic operators.
- The proper linking function (addition, multiplication, subtraction or division) needs to be chosen to connect the algebraic sub-trees.
- Finally, the default values of the genetic operators need to be selected for the GeneXpro program[38]. An overview of the used parameters is described in Table 2.

| Parameter | Value |
|-----------|-------|
| Function set | +, -, *, /, Pow (x, y), \( \exp(x) \), \( \ln(x) \), \( x^2 \), \( x^3 \), \( x^4 \), \( x^5 \) |
| Genetic operator | Optimal evolution |
| Chromosomes | 30 |
| Head size | 7-10 |
| Number of genes | 3-9 |
| Linking Function | Addition, Multiplication |
| Fitness Function error type | R-square, R, RMSE, RRSE, RAE |
| Mutation rate | 0.00138 |
| Inversion rate | 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

The development of GEP models and all the relevant statistical calculations are performed using the “GeneXpro tools 5.0” software.

4. Results and discussion

In one of our preliminary studies [39] single concurrent and lagged correlation analyses had been carried out to identify the most influential climate indices and their extent of impact with the increase of lagged months on spring streamflow of the study region. These analyses revealed the suitable combination sets of multiple indices for Multiple Linear Regression (MLR) analysis. The same combination sets of different lagged indices were used in the present study for GEP analysis while the best models were selected based on their statistical performances in terms of higher Pearson correlation values and lower errors i.e., Root Relative Squared Error (RRSE), Relative Absolute Error (RAE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Nash-Sutcliffe Efficiency (NSE).

Table 3. Performance test of the best GEP models for calibration and validation period.

| Station Name       | Model         | Calibration | Validation |
|--------------------|---------------|-------------|------------|
|                    |               | r   RRSE  RAE  RMSE  MAE  NSE  r   RRSE  RAE  RMSE  MAE  NSE |
| Singleton          | PDO<sub>Apr</sub> NINO3.4<sub>May</sub> | 0.74  0.67  0.67  13.85  10.62  0.55  0.79  0.78  0.70  10.60  7.47  0.39 |
| Coggan             | PDO<sub>Mar</sub> NINO3.4<sub>Jun</sub> | 0.87  0.24  0.65  1.07  0.82  0.76  0.91  0.92  0.69  1.68  1.18  0.15 |
| North Cuerindi     | PDO<sub>May</sub> NINO3.4<sub>May</sub> | 0.76  0.64  0.72  5.82  4.86  0.58  0.82  0.90  0.80  9.68  7.99  0.19 |

From Table 3 it is evident that the developed models have high correlation (r) values with very low errors which satisfies the statistical performance of the GEP models.

The output equations from the best developed GEP models for all 3 stations are as follows:

\[
Q_{\text{Singleton}} = \left[-\frac{4.76d_0(d_0+7.87)}{2.69 \times 10^8 \times d_1}\right]^5 + \frac{d_1^2}{d_1+0.21} + 14.40 - \frac{0.21}{d_0} - 10.93d_1 \times 0.78d_1 - \frac{d_0d_1}{0.77} 
\] (2)

where, \(d_0 = \text{PDO}_{\text{Apr}}\) and \(d_1 = \text{NINO3.4}_{\text{May}}\)

\[
Q_{\text{North Cuerindi}} = d_0d_1 - 4.75d_1 + d_1^4 + \frac{d_1^2(0.91-d_1)}{(d_1+1.59)(d_0-0.44)} - \frac{0.07d_1^2d_0-0.90}{d_3} + 6.47
\] (3)

where, \(d_0 = \text{PDO}_{\text{May}}\) and \(d_1 = \text{NINO3.4}_{\text{May}}\)

\[
Q_{\text{Coggan}} = \frac{d_0^2d_0}{1.55(0.88+2d_0)} + \frac{d_1^2d_1+6.44}{20.73+(-1.55)d_1} - \frac{0.47}{d_1(-0.88)^d_1}
\] (4)

where, \(d_0 = \text{PDO}_{\text{Mar}}\) and \(d_1 = \text{NINO3.4}_{\text{Jun}}\).

The predictability of the best GEP models is explained through the time series and scatter plots of observed and simulated flows in Figure 2.
Figure 2. Comparison of time series analyses and scatter plots between observed and simulated streamflow during the calibration and validation periods.

In Table 4 the performances of the GEP models are compared with that of the MLR models developed in one of our previous studies (Esha and Imteaz, 2018). It is evident from both analyses (MLR and GEP) that the selected best models have statistically significant correlation (r) values with lower errors. While comparing the best models from both analyses, it is found for every station GEP models have outperformed MLR models with their much higher correlation values (almost twice of MLR) and much lower errors.

Table 4. Comparison of MLR and GEP models in terms of Pearson correlation (r) values.

| Stations Name | Best Models | GEP Pearson Correlation (r) | MLR Pearson Correlation (r) |
|---------------|-------------|-----------------------------|-----------------------------|
|               | Calibration| Validation                  | Calibration                 | Validation                 |
| Singleton     | PDO_{Apr}  | 0.74                        | PDO_{Mar}                   | 0.41                       |
|               | NINO3.4_{May} | 0.79                  | NINO3.4_{Jun}               | 0.65                       |
5. Conclusion

The study is conducted to assess the ability of GEP technique to predict seasonal streamflow of north-east NSW region using combination of lagged climate indices as predictors. The performance of the developed GEP models were evaluated based on several statistical parameters which includes Pearson correlation values (r), RRSE, RAE, RMSE, MAE and NSE. One best models from each of the three stations were selected considering their higher correlation values and lower errors. The outcomes of this study were compared with the results obtained from MLR analysis which was a part of one of our previous studies. The comparison revealed that for every station GEP models outperformed MLR models with their much higher correlation values which ranges between 0.74 and 0.91, whereas for MLR models the values varied between 0.41 and 0.65. The developed models are efficient models with their much higher correlation values which ranges between 0.74 and 0.91, whereas for MLR models the values varied between 0.41 and 0.65. The developed models are efficient to predict spring streamflow up to three months ahead, thus enable the water stake holders to take low risk decision at the earlier stage of the crop period. The predictability of the GEP models on seasonal streamflow of other regions of NSW will be assessed in our future studies.

References

[1] White W B et al. 2004 Forecasting Australian drought using Southern Hemisphere modes of sea - surface temperature variability International Journal of Climatology: A Journal of the Royal Meteorological Society 24 1911-27

[2] Abawi Y, Dutta S, Zhang X and McClymont D 2005 ENSO-based streamflow forecasting and its application to water allocation and cropping decisions—an Australian experience” Regional Hydrological Impacts of Climate Change–Impact Assessment and Decision Making, Brazil (Oxfordshire, UK: IAHS Publication) pp 346-354

[3] Duc H N, et al. 2017 Association of climate drivers with rainfall in New South Wales, Australia, using Bayesian model averaging Theor. Appl. Climatol. 127 169-85

[4] Kiladis G N and Henry F D 1989 Global climatic anomalies associated with extremes in the Southern Oscillation J. Climate 21069-90

[5] Piechota T C et al. 2001 Development of exceedance probability streamflow forecast J. Hydrol. Eng. 6 20-8

[6] Chiiew F H S et al. 1998 El Nino/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting J. Hydrol. 204 138-149

[7] Chiiew F H S and Leahy M J 2003 Inter-decadal Pacific oscillation modulation of the impact of El Niño/Southern Oscillation on Australian rainfall and streamflow MODSIM 2003 International Congress on Modelling and Simulation vol 1 (Townsville, Australia: Modelling and Simulation Society of Australia and New Zealand)

[8] Detttinger M D and Henry F D 2000 Global characteristics of stream flow seasonality and variability J. Hydrometeorology 1 289-310

[9] Dutta S C et al. 2006 Rainfall and streamflow response to El Niño Southern Oscillation: a case study in a semiarid catchment, Australia Hydrol. Sci. J. 51 1006-20

[10] Sharma A 2000 Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3—A nonparametric probabilistic forecast model J. Hydrol. 239 249-58

[11] Ashok K, Behera S K, Rao S A, Weng H and Yamagata T 2007 El Niño Modoki and its possible teleconnection J. Geophys. Res. Ocean. 112 C11007

[12] Taschetto A S et al. 2009 Effect of anomalous warming in the central Pacific on the Australian
monsoon Geophys. Res. Lett. 36 L12704
[13] Feng J and Li J 2011 Influence of El Niño Modoki on spring rainfall over south China J. Geophys. Res. Atmosph. 116 D13102
[14] Verdon D C and Stewart W F 2005 Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia Water Res. Res. 41 W09413
[15] Westra S and Ashish S 2009 Probabilistic estimation of multivariate streamflow using independent component analysis and climate information J. Hydrometeorol. 10 1479-92
[16] Power S et al. 1999 Inter-decadal modulation of the impact of ENSO on Australia Clim. Dynam. 15 319-24
[17] Kiem A S, Stewart W F, and George K 2003 Multi - decadal variability of flood risk Geophys. Res. Lett. 30 1035
[18] Kiem A S and Stewart W F 2004 Multi - decadal variability of drought risk, eastern Australia Hydrolog. Proc. 18 2039-50
[19] Maier H R and Graeme C D 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications Environ. Model. Softw. 15 101-24
[20] Govindaraju R S 2000 Artificial neural networks in hydrology. I: Preliminary concepts J. Hydro. Eng. 5 115-23
[21] Coulibaly P, Anctil F and Bobee B 2000 Daily reservoir inflow forecasting using artificial neural networks with stopped training approach J. Hydrol. 230 244-57
[22] Khu S T, et al. 2001 Genetic programming and its application in real - time runoff forecasting I JAWRA 37 439-51
[23] Nourani V, Kisi Ö and Komasi M 011 Two hybrid artificial intelligence approaches for modeling rainfall–runoff process J. Hydrol. 402 41-59
[24] Traore S and Guven A 2012 Regional-specific numerical models of evapotranspiration using gene-expression programming interface in Sahel Water Resour. Manag. 26 4367-80
[25] Kisi O et al. 2012 Intermittent streamflow forecasting by using several data driven techniques Water Resour. Manag. 26 457-74
[26] Kiafar H et al. 2017 Evaluating the generalizability of GEP models for estimating reference evapotranspiration in distant humid and arid locations Theor. Appl. Climatol. 130 377-89
[27] Sattar A M A, Gharabaghi B and McBean E A 2016 Prediction of timing of watermain failure using gene expression models Water Resour. Manag. 30 1635-51
[28] Aziz K et al. 2013 Regional flood estimation in Australia: Application of gene expression programming and artificial neural network techniques Adapting to Change: the Multiple Roles of Modelling: Proc. of the 20th Int. Congress on Modelling and Simulation (MODSIM2013), 1-6 December 2013, Adelaide, South Australia
[29] Aytek A, Asce M and Alp M 2008 An application of artificial intelligence for rainfall-runoff modeling J. Earth Syst. Sci. 117 145-55
[30] Azamathulla H M and Ghani A A 2011 Genetic programming for predicting longitudinal dispersion coefficients in streams Water Resour. Manag. 25 1537-44
[31] Guven A and Aytek A 2009 New approach for stage–discharge relationship: gene-expression programming J. Hydrol. Eng. 14. 812-20
[32] Razaq S A et al. 2016 Prediction of flow duration curve in ungauged catchments using genetic expression programming Procedia Engineering 154 1431-8
[33] Yassin M A, Alazba A A, and Mohamed A M 2016 Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate Agric. Water Manag. 163 110-24
[34] Fernando A, Shamseldin A and Abrahart R 2012 River flow forecasting using gene expression programming models 10th Int. Conf. on Hydroinformatics HIC 2012, Hamburg, Germany (Hamburg, Germany: Tu Tech Innovation)
[35] Ruiz J E, Cordery I, and Sharma A 2007 Forecasting streamflows in Australia using the tropical
Indo-Pacific thermocline as predictor *J. Hydrol.*, 341 156-64

[36] Robertson D E and Wang Q J 2009 A Bayesian joint probability approach to seasonal prediction of streamflows: Predictor selection and skill assessment *H2009: 32nd Hydrology and Water Resources Symposium, Newcastle: Adapting to Change* Engineers Australia1545

[37] Wang Q J and Robertson D E 2011 Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences *Water Resour. Res.* 47 W02546

[38] Shiri J and Kişl Ö 2011 Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations *Comput. Geosci.* 37 1692-701

[39] Esha R I., and Imteaz M A 2018 Assessing the predictability of MLR models for long-term streamflow using lagged climate indices as predictors: A case study of NSW (Australia) *Hydrol. Res.* 50 nh2018171