Advances and Applications in Mathematical Sciences
Volume 21, Issue 6, April 2022, Pages 3563-3575
© 2022 Mili Publications, India

APPLICATION OF DATA DRIVEN MODEL FOR PREDICTION AND ESTIMATION OF EVAPOTRANSPIRATION – A CASE STUDY

S. V. S. N. D. L. PRASANNA and NAGAVENI THALLAPALLI

Associate Professor
Department of Civil Engineering
University College of Engineering (A)
Osmania University, Hyderabad-500007, India
E-mail: prasanna.s@uceou.edu

Associate Professor
Department of Mechanical Engineering
University College of Engineering (A)
Osmania University, Hyderabad-500007, India

Abstract

Water is becoming a scarce resource as a result of increased demand for hydropower, irrigation, and water supply, among other things. Crop growth simulation models simulate crop development and yield by using weather data such as temperature, solar radiation, and rainfall. The critical process of water balance that is an important component of energy balance is Evapotranspiration. It changes due to a variety of parameters such as wind, temperature, humidity, and the availability of water. Because it accounts for 15% of the water vapour in the atmosphere, evapotranspiration is an important process in the water cycle. Analytically, the FAO Penman-Monteith method is used to calculate reference crop evapotranspiration (ETo). Furthermore, an equation for estimating Evapotranspiration is developed using a reliable method known as Multi Gene Genetic Programming (MGGP). Six different measured weather variables are taken and compared with MGGP results. The present study exhibited that, the expected results of the Evapotranspiration of the arrangement formed by data are favourable and also generalize the testing data. The corresponding R-squared values for random data and continuous data from the year 2009 to 2017 is 0.976 and 0.976 obtained from the training values, while the values for the testing are 0.972 and 0.964 respectively. Based on the

2020 Mathematics Subject Classification: 60G25.

Keywords: Evapotranspiration, Penman, Simulation, Multi objective Optimization, Multi Gene Genetic Programming (MGGP).

*Corresponding author; E-mail: nagaveni.t@uceou.edu

Received December 9, 2021; Accepted December 17, 2021

performance analysis, the efficiency and reliability of the proposed MGGP model are validated and predicted for the year 2021 and found the results are satisfactory.

1. Introduction

Evaluation of Hydrological parameters is critical for interpreting water quality data. Variations in hydrological conditions have a significant impact on water quality. It is critical that personnel involved in hydrological or water quality measurements are generally familiar with the principles and techniques used by one another [1]. The conventional and/or physical measurement methodologies, viz., snow and precipitation using rain gauges, lysimeters along with atmometers for Evapotranspiration, infiltration by means of double ring infiltrometers, electromagnetic, gravimetric, and volumetric methods for soil moisture, and piezometers for groundwater, can all be used to quantify these hydrological parameters. However, all the above mentioned methods provide point-based prediction of different hydrologic parameters [2]. Evaporation occurs from each of the various process mechanisms possible in any soil-plant-atmosphere. The factors affecting Evaporation possible from the soil are soil water content, soil type, the occurrence or absence of surface insulations, and the environmental conditions that are forced on the type of soil [3].

Hatice Citakoglu et al. proposed a new compact method using multi-gene genetic programming (MGGP) to estimate solar radiation in Turkey. 163 stations meteorological data from 1975 to 2015 for 7 areas of Turkey were used for MGGP modelling. It was concluded that the MGGP modelling and empirical equations obtained after calibration, are found to most effective for estimating solar radiation [4]. Multiple Genetic Programming (MGP) technique projected by Ali Danandeh Mehr et al., stated that, this technique shows an improvement in the accuracy of prediction when compared to the standalone Genetic Programming (GP). The measurement of the amount of Rainfall from two weather forecasting stations in the Lake Urmia Basin, Iran, are used to demonstrate the model's 1-month rainfall forecasting capability. They concluded that the MGP model outstrips the standards at both the station points in terms of statistical performance [5]. Anurag Malik et al. investigated the ability of various predictive models, comprising of Support Vector Machine (SVM), Multiple Model-Artificial Neural Network

(MMANN), Multivariate Adaptive Regression Spline (MARS), Multi-Gene Genetic Programming (MGGP), and 'M5Tree,' to estimate pan evaporation on a monthly scale (EPm) at two stations in India (e.g. Ranichauri and Pantnagar). The Gamma test was used to determine which input variables were the most effective for the five different models. According to the findings, MM-ANN-1 and MGGP-1 models, will aid local participants in proper management of water resource [6].

Most common traditional methods are adopted for prediction of Evapotranspiration via empirical equations and experimental methods. In the present-day scenario data driven models are widely used in predicting hydrological parameters. Garg et al. have proposed a simulation approach (MGGP) for the relation amongst water content and two input parameters via., soil suction and volumetric crop root content. The projected method employs a statistical stepwise regression approach. The experimental data was compared to the simulated data obtained from the models [7]. It has been observed that the ability of MGGP is to remove unnecessary variables automatically.

In the present study, a total of 144 data points that include 5 climatic variables via., maximum and minimum temperatures (°C), relative humidity (%), wind speed (km/day), and sunshine hours were collected from meteorological centre [8]. In view of this, adopting the hydrological parameters, the objectives of the present study are, to estimate the Evapotranspiration using the traditional analytical method via., Penman-Monteith method. To develop empirical models for prediction of Evapotranspiration, using Multi Gene Genetic Programming (MGGP). An attempt is made to emphasize that, the choice of the training and testing data also can have an impact on the training phenomenon of the methods. The training and testing data was adopted in two ways, firstly, 108 training and 36 testing data points were selected from the year 2009 to 2020 randomly. Secondly, 108 training data points are taken continuously from the year 2009 to 2017 and 36 testing points were from 2018 to 2020. The above two models (Random and Continuous data) are adopted for prediction of evapotranspiration for the year 2021. Further, the study was extended to compare simulated data obtained from MGGP model with analytical results. Despite the fact that there have been numerous applications of data-driven models in hydrological studies, this study is an attempt to incorporate MGGP in order to improve the estimation accuracy of Evapotranspiration.

2. Data Collection

The study area comprises of Nizamabad district. It is one among the 33 districts of Telangana region with a geographical area of 4,288 square kilometres. It is located at 18°41′N 78°6′E. Nizamabad is bordered Nirmalon the North, Jagtial and Rajanna Sircillaon the East, Kamareddy on the South, and Nanded on the West. Figure 1 depicts a map of the research area's location.

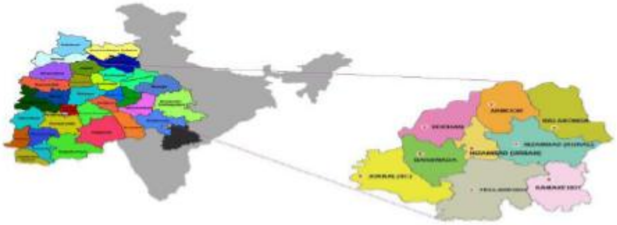

Figure 1. Location of Study Area.

3. Modelling

The present study is carried out using two-way methodology via., analytical calculation and mathematical modelling analysis. The detailed understanding is present the subsequent sub sections.

Analytical Analysis:

For analytical analysis, the Reference Evapotranspiration (ETo) for the study area is evaluated using Penman-Monteith method [9]. In order to estimate ETo various required climate parameters via., meteorological parameters via., max. and min. temperatures, relative humidity, average wind velocity, duration of sunshine were considered on monthly basis for 12 years (144 months) from world weather online. The evocative statistical features of the meteorological variables are presented in Figure 2. The Reference Evapotranspiration (ETo) is calculated using Penman-Monteith detailed in equation (1) and its values for the months January to November for year 2021 as depicted in the Figure 7.

$$ETo = \frac{0.408\Delta(R_n - G) + \gamma(900/(T + 273))U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \tag{1}$$

where, $\Delta =$ slope of vapour pressure curve $(kPa/^{\circ}C)$, $R_n =$ Net radiation $(MJ/m^2.day)$, $G =$ soil heat flux density $(MJ/m^2.day)$, $\gamma =$ psychometric constant $(kPa/^{\circ}C)$, $T =$ Mean daily air temperature $(^{\circ}C)$, $U_2 =$ Average wind speed at 2m height (m/s), $e_s =$ saturation vapour pressure (kPa), $e_a =$ actual vapour pressure function (kPa).

Climate variables	Period	Max	Min	Continuous Data			Random Data		
				mean	Standard Deviation	skewness	mean	Standard Deviation	skewness
Max Temp $(T_{\max})$ $(^{\circ}C)$	Training	43	27	33.4259	4.4684	0.7937	33.61	4.7150	0.7270
	Testing	43	28	32.5556	5.0687	0.8730	32.571	4.2630	1.601
Mix Temp $(T_{\max})$ $(^{\circ}C)$	Training	33	16	23.1111	4.1239	0.3089	23.4655	4.0714	0.3644
	Testing	31	15	23.9444	3.9507	0.6711	22.7143	4.1478	1.7170
Humidity (%)	Training	88	15	49.999	21.2666	0.1012	51.000	22.1636	0.0507
	Testing	86	21	53.0833	23.3659	-0.0004	51.143	20.5143	0.1645
Max Wind Speed (kmph)	Training	24.50	9.0	14.4306	3.5903	0.8983	14.8250	3.7074	0.7503
	Testing	24.70	9.9	15.8611	4.1918	0.4970	14.636	4.1608	0.9921
ET (mm/day	Training	10.35	3.88	5.9826	1.5626	0.8478	6.0050	1.6619	0.8421
	Testing	10.14	4.03	5.9053	1.9420	0.9272	5.7904	1.6714	1.0532

Figure 2. Statistical Parameters of Meteorological Variables.

Multi Gene Genetic Programming (MGGP):

Genetic programming (GP) is a technique for generation of genetic mathematical models introduced by Koza [10, 11]. Koza proposed GP as a generalisation of genetic algorithm. The GP method is mainly applied to non-linear regression problems to create mathematical expressions that associate among a given set of independent and the dependent quantities to provide a good fit [12]. Multi Gene Genetic Programming (MGGP) is a newly developed method of GP [13]. It mainly helps in the improvement of precision of GP.

MGGP mainly differ from GP in the number of trees that can be used in MGGP. GP mainly uses single trees whereas MGGP uses combination of trees. A typical MGGP model is shown in Figure 3. The advantage of the MGGP process, is that it is a data-driven procedure that builds models using experimental data. In MGGP the final structure of the target model is made in a linear form by combining two or more non-linear gene. The general form of a MGGP model is shown in equation (2).

$$y = a_0 + a_1 Gene_1 + a_2 Gene_2 + \ldots + a_n Gene_n \tag{2}$$

where, $y =$output, a_0 bias term, $Gene_i$ non-linear genes, $a_i =$coefficient of related genes, $n =$ number of genes.

Each case has a terminal and a function set. The basic mathematical operators (+, -, /), Boolean algebra operators (e.g., AND and OR), or other user-defined mathematical symbols form the elements of function set. The problem with the input variables represents the terminal set [13].

$$y = a_0 + a_1 \times [5 \times (\sin(x)) + \left(\frac{x}{9}\right) + a_2 \times [2 + \left(\frac{x}{3}\right)]$$

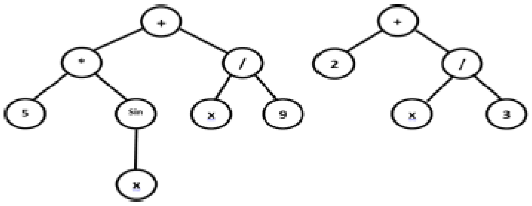

Figure 3. Typical MGGP Model.

In this work the software tool GPTIPS is used to develop the equations for the prediction of evapotranspiration [14]. This software is a new code written by "Genetics and Regression Symbolic Programming" on the basis of multi gene GP used with MATLAB 2021a. The standards of the maximum number and the depth of the genes determine the extent and different systems of the model for the space of universal solutions. In the current study, the Performance assessment is carried out. For this, out of the 144 data points obtained from 12 years of data for the first trial, 108 random points were selected for training and 36 points for testing. Further, 108 training data points are taken continuously from the year 2009 to 2017 and

36 testing points were from 2018 to 2020. The meteorological parameters considered for the development of MGGP model are maximum $-x_1$; minimum temperatures $(^{\circ}C) - x_2$; relative humidity $(\%) - x_3$; wind speed $(kmph) - x_4$; sunshine hours $-x_5$ and Evapotranspiration $- y$. The individual performance fitness functions are evaluated via. R^2, Mean Square Error (RMSE), MAPE and MAE are given as follows:

$$R^2 = \left[\frac{\sum_{i=1}^{n} (a_i - \overline{a}_i)(p_i - \overline{p}_i)}{\sqrt{\sum_{i=1}^{n} (a_i - \overline{a}_i)^2 \sum_{i=1}^{n} (p_i - \overline{p}_i)^2}} \right]^2 \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - p_i)^2} \tag{4}$$

$$MAPE = \frac{\sum_{i=1}^{n} | a_i - p_i |}{\sum_{i=1}^{n} a_i} \times 100 \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{n} | a_i - p_i |^2}{n} \tag{6}$$

Where, $n =$ total number of data points, a_i and $p_i =$ actual output values, $\overline{a}_i$ and $\overline{p}_i =$ average predicted and actual outputs.

The higher value of R^2 and lower values of MAE, MAPE and RMSE are the signs of developed equations having better performance. Other initial parameters were established for the implementation of GPTIPS is given in the Table 1.

Table 1. Parameter Settings for MGGP.

Run Parameter	Value
Population size	250
Max generations	150
Generation elapsed	14
Input variables	5
Training instances	108

Tournament size	25
Elite fraction	0.7

4. Results and Discussions

The developed MGGP model for two various datasets was implemented in predicting the Evapotranspiration, and the empirical method (Penman-Monteith) using training and testing datasets. The present sections highlight the results evaluated after processing the data.

MGGP model for Prediction of ETo:

The MGGP model was developed for different datasets via., Random and Continuous data points. For the random data, total of 108 (75%) data points for the training and 36 (25%) points for testing out of 144 were selected randomly. Subsequently, for continuous datasets with 108 data points from the years 2009 – 2017 for training and 2018-2020 for testing the data. Number of models on front was 16 with total models as 2250 respectively. The following equation (7) was evolved as the best model for Evapotranspiration for random dataset. Further, another dataset of continuous data point's one best fit model was developed and is shown by equation (8).

$$ET_o = 0.0199x_3 - 0.0618x_2 + 0.0101x_4 - 0.0619x_5 - 1.81e^{-5}\tanh(x_2) + 0.00807x_2x_4 + 0.00103x_1x_5^2 - 1.81e^{-5}x_2x_3x_4 - 1.41e - 6x_3^2x_4x_5\tanh(x_3) - 1.41e^{-6}x_2x_3x_4\tanh(x_3) + 0.76 \tag{7}$$

$$ET_o = 0.596x_2 - 0.263x_1 + 0.037x_3 + 0.233x_4 + 0.02990x_1x_5 - 2.7e^{-5}x_3x_4 (2.0x_2 + x_3) - 2.7e^{-5}x_2x_3x_5 - 1.35 \tag{8}$$

For any model development, the accuracy of the equation can be elaborated by plotting the measurement obtained from training and test data. Figure 4 depicts the measured Evapotranspiration values (expected output) and the one calculated by MGGP (predicted output), demonstrating an excellent correlation between the multi gene GP predictions and the measured data. This graph shows that an acceptable agreement between predicted values and experimental data can be obtained with a linear correlation coefficient of 0.97597 for Random and 0.96406 for continuous dataset. The Figure 5 depicts the comparison between the actual and predicted values of ET_o for both the datasets. Bias term in the selected cross validated mathematical model was estimated as +0.76 for random data and -1.35 for continuous dataset respectively, that was included in the expression.

The determination of the coefficient values of R^2 – goodness of fit, RMSE - root mean squared error, MAE - mean absolute error, SSE - sum of squared errors, and MSE - mean squared errors are highlighted for both the datasets in Table 4. In summary, the MGGP model predictions are optimal when $R^2, R, MAE,$ and RMSE are close to $1, 1, 0,$ and 0, respectively. As highlighted in Table 2, the present methodology can predict Evapotranspiration that is comparable to the actual core measurements. It can be seen that, the MGGP model has a high degree of accuracy in predicting objectives.

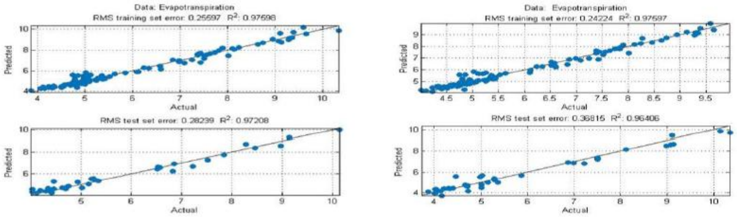

Figure 4. Graphs of Measured and Predicted Values of Evapotranspiration by MGGP model for Random (left) and Continuous (Right) Dataset.

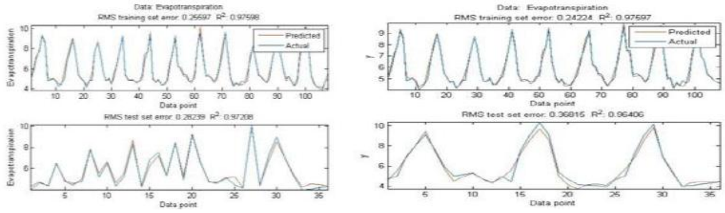

Figure 5. Graphs of RMS values of Evapotranspiration by MGGP model for Random (left) and Continuous (Right) Dataset.

The Figure 6 portrays how the best values (log) and mean fitness change with the number of generations for both the datasets. The graph also highlights that as the number of generations increases, so does the fitness value. Furthermore, as shown in Figure 6, the computational complexity of this model is the lowest in absolute terms, equal to 144 for random datasets and 80 for Continuous dataset respectively, which shows the Pareto front curve.

Table 2. Metrics for Evaluation for Random and Continuous Datasets.

Metric	Random Datasets		Continuous Datasets	
	Training Data	Test Data	Training Data	Test Data
R^2	0.97598	0.97208	0.97597	0.96406
RMSE	0.25597	0.28239	0.24224	0.36815
MAE	0.1834	0.21615	0.1688	0.28191
SSE	7.0762	2.8707	6.3372	4.8792
Max. abs. error	0.83471	0.81054	0.82436	1.1093
MSE	0.06552	0.079742	0.058678	0.13553

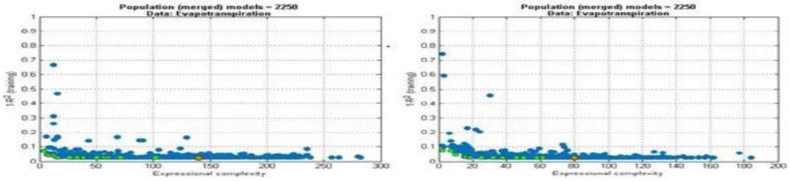

Figure 6. Variation of fitness with the number of generations (left-Random and right-Continuous datasets).

Analytical results

The annual ET_o for the present study area are calculated analytically using Penman-Monteith method for 11 months from January to November for the year 2021. Moreover, the values of Evapotranspiration adopted in the model development using MGGP were also analytically estimated using equation (1). The comparative results that evaluated from the two models and from the empirical equation are highlighted in Figure 7.

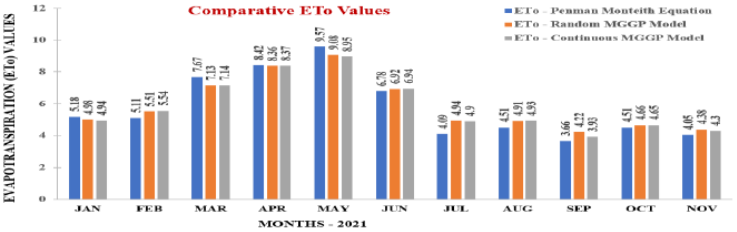

Figure 7. Evapotranspiration Results for Year-2021.

5. Conclusion

In the present paper, Evapotranspiration values are estimated for Nizamabad district of Telangana State. The meteorological data adopted for the study was collected on monthly basis for 12 years. Further, different datasets via., Random and Continuous data points. For the random data, total of 108 (75%) data points for the training and 36 (25%) points for testing out of 144 were selected randomly. Subsequently, for Continuous datasets 108 data points from the years 2009-2017 for training and 2018-2020 for testing the data. The three different equations via., empirical method and two model developed equations are compared based on the statistical parameters via., mean standard deviation and skewness. Further, the commonly adopted performance criteria of MAE, MSE, SSE, RMSE, and R2 statistics were also compared. The following conclusions have been are drawn after the critical observation of the results:

- The estimation performance of the MGGP model developed equations, for two various datasets via., Random and Continuous datasets are compared and found that both the datasets exhibited good consonance.
- The higher values of multiple regression coefficients evaluated for both the datasets, indicate that the developed equation can predict the physical phenomena more accurately inconsonance with the conventional methods.
- From the analysis, based on the R^2 value it can concluded that, Random or Continuous data points can be adopted for estimation of Evapotranspiration values.
- The MGGP predicted values of ET_o from both the variation of dataset

points showed fair consistency with the traditional empirical Penman-Monteith method.

Thus, the proposed work emphasises the estimation of Evapotranspiration to choose optimal values depending on their needs. The proposed model MGGP effectively integrates the effects of various parameters. The ability of MGGP to generate powerful compact models can be concluded as a significant advantage. The model can be extended with the additions of a greater number of parameters in order to further check the consistency and applicability in the prediction of Evapotranspiration. This estimation can also act as a useful tool for the agricultural purpose.

References

[1] J. Bartram and R. Ballance, Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmers, 1996: CRC Press.

[2] P. K. Thakur, et al., Hydrological parameters estimation using remote sensing and GIS for Indian region: A review. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 87(4) (2017), 641-659.

[3] C. M. Burt, et al., Evaporation research: Review and interpretation, Journal of irrigation and drainage engineering 131(1) (2005), 37-58.

[4] H. Citakoglu, B. Babayigit and N. A. Haktanir, Solar radiation prediction using multi-gene genetic programming approach, Theoretical and Applied Climatology 142(3) (2020), 885-897.

[5] A. D. Mehr and M. J. S. Safari, Multiple genetic programming: a new approach to improve genetic-based month ahead rainfall forecasts, Environmental Monitoring and assessment 192(1) (2020), 1-12.

[6] A. Malik, et al., Modeling monthly pan evaporation process over the Indian central Himalayas: application of multiple learning artificial intelligence model, Engineering Applications of Computational Fluid Mechanics 14(1) (2020), 323-338.

[7] A. Garg, L. Rachmawati and K. Tai, Classification-driven model selection approach of genetic programming in modelling of turning process, The International Journal of Advanced Manufacturing Technology 69(5-8) (2013), 1137-1151.

[8] A. Angstrom, Solar and terrestrial radiation, Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation, Quarterly Journal of the Royal Meteorological Society 50(210) (1924), 121-126.

[9] R. G. Allen, et al., Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome, 300(9) (1998), D05109.

[10] J. R. Koza, Genetic programming: on the programming of computers by means of natural selection 1 (1992): MIT press.

[11] J. R. Koza, Architecture-altering operations for evolving the architecture of a multipart program in genetic programming, 1994: Stanford University.

[12] J. R. Koza, et al., Genetic programming III: Darwinian invention and problem solving, 3 (1999), Morgan Kaufmann.

[13] A. Garg, K. Tai and M. Savalani, Formulation of bead width model of an SLM prototype using modified multi-gene genetic programming approach, The International Journal of Advanced Manufacturing Technology 73(1-4) (2014), 375-388.

[14] D. P. Searson, D. E. Leahy and M. J. Willis, GPTIPS: an open source genetic programming toolbox for multigene symbolic regression, in Proceedings of the International multiconference of engineers and computer scientists Citeseer, (2010).