doi.org/10.62683/SINARG2025.116

Research paper

APPLICATION OF THREE REGRESSION METHODS FOR FILLING MISSING VALUES OF ANNUAL MAXIMUM DAILY PRECIPITATION

Nikola Đokić[1]

Abstract

This study examines the effectiveness of three regression methods – multiple linear, random forest, and log-linear (gamma) when applied to annual maximum daily precipitation data sets to fill in missing values. Gridded observations data of extreme daily precipitation, sourced from the Digital Climate Atlas of Serbia platform, were utilized for this study in the area of Niš. The dataset, which is complete for the period 1950–2020, was intentionally modified to simulate missing data. These artificial gaps, or 'holes,' were introduced systematically at the beginning, end, and randomly selected locations within the dataset. The data omission was carried out incrementally at rates of 5%, 10%, 15%, and 20%. The performance of the methods for completing incomplete series was evaluated in terms of standard metrics like the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). The results indicated a commendable performance across all evaluated methods, even when addressing 20% missing data. Notably, multiple linear regression emerged as the most effective technique among those tested.

Key words: *Missing Data, Precipitation, Regression Methods, Gridded Datasets*

[1] M.Sc. Civ. Eng., assistant, University of Niš, Faculty of Civil Engineering and Architecture, Serbia, nikola.djokic@gaf.ni.ac.rs, ORCID 0009-0008-3078-1473

1. INTRODUCTION

Effective planning and management of water resources rely on the availability of reliable and precise precipitation data collected from meteorological stations [1]. Precipitation data is commonly gathered using rainfall gauges, which are regarded as the primary source of precipitation observations. These instruments provide direct measurements of precipitation at specific locations [2, 3]. However, precipitation data is frequently incomplete. The incompleteness of precipitation data may be due to instrument failure, measurement errors, or other factors. The presence of missing data in rainfall time series can significantly impact the accuracy of statistical analyses [4]. A common approach to address this issue is to discard the years or periods that contain missing data. However, this strategy may lead to significantly reduced sample sizes or can obstruct an accurate characterization of the upper tail of the distribution of the random variable [5]. An alternative approach involves supplementing the target station with data from appropriate nearby stations.

Various techniques have been developed to estimate and reconstruct missing data, which can be generally divided into three categories: empirical methods, statistical approaches, and function-fitting techniques [6]. The methods described for addressing missing data aim to enhance the accuracy and reliability of hydrological models. Techniques such as the regional weighting method, linear regression, Kriging, etc., are used in practice to fill in monthly and annual rainfall data [7]. However, due to the significant spatial variability of extreme precipitation, such techniques are not recommended for filling daily and sub-daily precipitation [6]. Since daily and sub-daily precipitation data serve as inputs for hydrological models, selecting an appropriate method for estimating missing values is crucial. Previous research has discussed the use of multiple linear regression [8], simple substitution [9], the Theil method [10], and machine-learning algorithms [11, 12] for this purpose.

This study evaluates the effectiveness of three methods for filling artificially induced gaps in annual maximum daily precipitation data from the period 1950 to 2020. The methodologies under investigation include Multiple Linear Regression (MLR), Random Forest (RF), and Log-Linear Gamma (LLG) regression techniques. Gridded data representing the annual maximum daily precipitation at nine locations across the city of Niš, sourced from the Digital Climate Atlas of Serbia, were utilized.

2. MATERIALS AND METHODS

This chapter defines the specific location of the investigation and provides a detailed explanation of the techniques used to fill the data series. It also describes how the so-called "holes" were intentionally introduced into the complete sequence of daily precipitation records. Validation was conducted by comparing the reconstructed values with the actual omitted precipitation data.

Due to the insufficient density of observation stations with available daily precipitation data, a suitability test for filling in incomplete sequences was conducted using a grid network sourced from the Digital Climate Atlas of Serbia [13].

2.1. Study Area and Data

Niš, one of the oldest cities in the Balkans, is located in southern Serbia and serves as an important economic, cultural, and geographical hub. Positioned at the crossroads of Central

and Southeast Europe, Niš has historically been a gateway between Eastern and Western civilizations. The city lies at 43°19′N latitude and 21°54′E longitude, with an elevation averaging 194 meters above sea level. The following figure (Figure 1) includes the location of the research area on the territory of the Republic of Serbia, with a detailed overview of the positions of 9 significant points from the gridded observations of the Digital Climate Atlas of Serbia, which are located over the city of Niš.

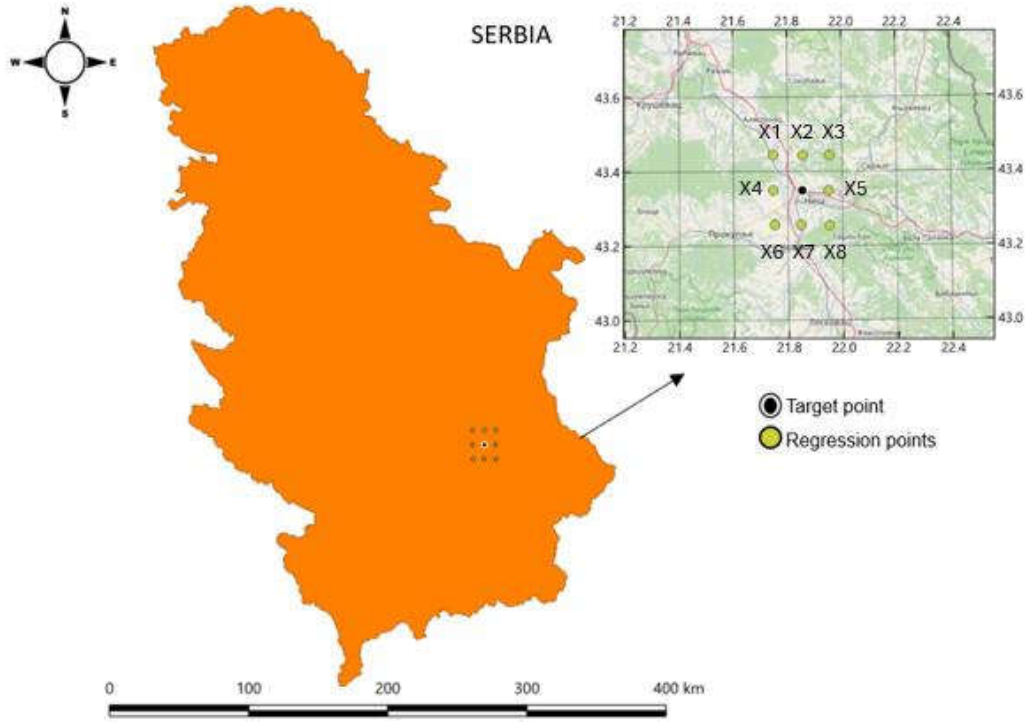

Figure 1. Study area with eight regression and one target point gained from the Digital Climate Atlas of Serbia

Gridded meteorology refers to meteorological data that is organized into a structured grid format, rather than being based solely on individual weather station observations. This approach allows for consistent spatial and temporal coverage, making it useful for climate modeling, weather forecasting, and environmental research [14].

The grid of the atlas was formed based on the measured climatic variables at meteorological and climatological stations in Serbia, which were then interpolated to a regular grid of points, with a horizontal distribution of ~10x10 km. Annual maximum daily precipitation data in the period 1950-2020 were used from the climate atlas of Serbia. An "intervention" was conducted on the complete dataset, resulting in the removal of 5 %, 10%, 15%, and 20% of the data from the beginning, from the end, and in random places. A total of 36 incomplete data series were obtained in this manner. Such an intervention was carried out only at one (central/target) measurement point, to further establish a regression based on intact data from the 8 nearby points. Used regression techniques are detailed below.

2.2. Regression methods

2.2.1. Multiple Linear Regression

MLR is a statistical method used to analyze the relationship between a dependent variable and two or more independent variables. It extends simple linear regression, which only deals with one independent variable, to handle multiple predictors [1, 15]. The missing data (Yi) is estimated from the following equation:

$$Y_i = b_o + \sum_{i=1}^{n}(b_i X_i) \quad (1)$$

where Yi is the estimated rainfall data, Xi is the observed rainfall value of the ith surrounding point, and bi, b1, …, bn are the regression coefficients.

2.2.2. Random Forest Regression

Among its many uses, RF is a reliable technique for regression, prediction, and classification. By sampling the training data at random, it builds several decision trees and uses a bagging process to determine which trees produce the best predictions [16]. In 2001, Breiman presented an expansion of this algorithm [17].

RF works well for producing predictive models for tasks involving both regression and classification. Binary decision trees, specifically Classification and Regression Trees (CART), are the foundation of this approach. Software XLstat was utilized to carry out RF regression [18]. The number of trees, the mtry value, and particular tree parameters like minimum node size, minimum child size, and maximum depth are among the parameters that XLstat needs to know. Longer computation times result from using more trees, even though this improves model stability. To lessen the chance of overfitting, the maximum depth parameter limits the trees' complexity. When building individual decision trees, the mtry parameter indicates how many features are chosen at random at each split.

The following parameter values were used:

- Forest parameters: Sampling (Random with replacement); Method (Random Input); Sample size (71); Number of trees (300),
- Stop conditions: Construction time (300); Convergence (50),
- Tree parameters: minimum node size (2); minimum son size (1); maximum depth (30); mtry (3); CP (0,0001).

2.2.3. Log-Linear (Gamma) Regression

Log-linear regression is one of the specialized cases of generalized linear models (GLM). The 'linear' in the name means the model's presumption of a linear relationship between the input and output variables. And 'log' refers to the model's use of a logarithmic transformation of the input data before fitting it into a linear equation. GLM offers considerable flexibility in assuming the distribution of ‘errors’ about the mean response. They handle response variables that follow different distributions, and in this case, the gamma-distributed data is used. Applying a logarithmic transformation to variables in a regression model is a widely embraced technique for addressing scenarios where a non-linear relationship intricately intertwines the independent and dependent variables. This method allows for a more coherent interpretation of the data, enabling analysts to uncover patterns that might otherwise remain obscured. By re-scaling the variables, one can often reveal a clearer, more linear relationship, transforming complex interactions into more manageable and visually

interpretable forms [19]. This transformation accommodates nonlinear relationships in the data, resulting in a more accurate prediction of those relationships. In the log-linear model (2), the literal interpretation of the estimated coefficient b_i is that a one-unit increase in *X* will produce an expected increase in log *Y* of b_i units. In terms of *Y* itself, this means that the expected value of *Y* is multiplied by e^{bi}.

$$logY_i = b_o + \sum_{i=1}^{n}(b_i logX_i) \quad (2)$$

2.3. Methods Performance

The efficiency of the filling data was compared using three different error indices: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination R^2. The error measures were used to compare the estimations with the observed values. The three error indices are given as follows:

- Mean absolute error (**MAE**)

The Mean Absolute Error (MAE) stands out as a crucial metric in the realm of model evaluations, revered for its clarity in quantifying estimation errors. This powerful measure provides straightforward insight into the magnitude of errors, allowing researchers and analysts to gauge the precision of their predictions. As highlighted by Willmott and colleagues in 2009, the MAE is a recommended method for assessing accuracy [20]. The MAE is computed using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_i - Y_i\right| \quad (3)$$

where $\hat{Y}_i$ is the observed value of the rainfall data from the target point.

- Root mean square error (**RMSE**)

RMSE measures the difference between the estimated and observed values. The best method gives the lowest computed value of the RMSE. The RMSE value varies from 0 to +∞. The RMSE is computed using the following equation:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{Y}_I - Y_I)^2}{n}} \quad (4)$$

- Coefficient of determination (**R^2**)

The **coefficient of determination**, denoted as **R^2**, is a statistical measure used in regression analysis to assess how well a model explains the variability of the dependent variable. The R^2 is computed using the following equation:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\cdot(Y_i - \bar{Y}_i)}{\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} \quad (5)$$

where $\bar{Y}_i$ and $\bar{Y}$ are the average precipitation values of estimated and observed data, respectively.

3. RESULTS AND DISCUSSION

The following table (Table 1) summarizes the values of error indicators (RMSE, MAE, and R2) for all three applied regression techniques and all considered incomplete sets of precipitation data, with omission from the beginning (B), end (E), and in random (R) places along strings.

Table 1. Values of different error indices: mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination R^2 for three different regression methods and different types of omission datasets

		Regression method								
		MLR			RF			LLG		
Rate of omission	Perf. metrics	Omission								
		B	E	R	B	E	R	B	E	R
5%	RMSE	1.42	0.82	0.89	1.51	2.50	4.92	1.45	2.33	1.70
	MAE	1.16	0.68	0.70	1.35	2.27	3.19	1.14	2.26	1.56
	R^2	0.94	1.00	0.99	0.90	0.91	0.99	0.92	0.99	0.98
10%	RMSE	1.31	1.18	0.79	5.32	4.41	1.14	10.68	5.23	10.37
	MAE	1.01	1.00	0.61	3.33	3.91	0.81	4.92	3.66	5.35
	R^2	0.99	1.00	0.99	0.94	0.92	0.93	0.98	0.94	0.94
15%	RMSE	1.19	1.01	0.90	4.62	4.15	3.06	7.87	4.12	4.38
	MAE	0.96	0.84	0.74	2.98	3.20	2.16	3.78	2.49	2.52
	R^2	0.99	0.99	0.99	0.94	0.90	0.96	0.96	0.94	0.96
20%	RMSE	1.03	0.89	0.66	4.75	3.59	3.31	9.85	3.78	7.05
	MAE	0.74	0.72	0.59	3.33	2.62	2.15	5.13	2.38	2.90
	R^2	1.00	0.99	0.99	0.95	0.91	0.93	0.89	0.94	0.95

From the previous table, it can be seen following:

- The largest errors are obtained using the LLG model, where RMSE goes up to 10.68 mm,
- The smallest errors are obtained using MLR for all considered cases,
- RF regression gave satisfactory results, with the biggest RMSE value of 5.32 mm,
- The error values did not strictly increase as the number of missing data in the series increased, moreover, the smallest errors were registered in the case when multiple linear regression was used over the series with 20% of missing data.

The computational results showed that classical statistical methods such as MLR performed excellently, as also noted by Sattari M.T. and Kusiak A. in their work [1], as well as Hasanpour Kashani, M., and Dinpashoh, Y. in their study [21].

In the following Table 2, a correlation matrix between the included regression (X1-X8) and target points (T) is shown, for the case where the smaller errors are found.

Table 2. Correlation matrix of the nine considered locations

	X1	X2	X3	X4	X5	X6	X7	X8	T
X1	**1**	0.981	0.963	0.977	0.874	0.964	0.926	0.241	0.885
X2	0.981	**1**	0.994	0.992	0.939	0.976	0.966	0.274	0.945
X3	0.963	0.994	**1**	0.983	0.951	0.973	0.970	0.270	0.951
X4	0.977	0.992	0.983	**1**	0.941	0.988	0.979	0.256	0.952
X5	0.874	0.939	0.951	0.941	**1**	0.932	0.969	0.269	0.993
X6	0.964	0.976	0.973	0.988	0.932	**1**	0.985	0.255	0.940
X7	0.926	0.966	0.970	0.979	0.969	0.985	**1**	0.268	0.973
X8	0.241	0.274	0.270	0.256	0.269	0.255	0.268	**1**	0.260
T	0.885	0.945	0.951	0.952	0.993	0.940	0.973	0.260	**1**

It can be seen that all but the last eighth point have a strong correlation with each other.

Standardized coefficients, also known as beta coefficients, are used in regression analysis to compare the relative importance of predictor variables. They are calculated by standardizing the data so that all variables have a mean of 0 and a standard deviation of 1. This allows for direct comparison of the effects of different variables, regardless of their original units of measurement. In multiple regression, standardized coefficients help determine which independent variables have the greatest impact on the dependent variable. A higher absolute value of a standardized coefficient indicates a stronger influence on the outcome [22]. The following figure (Figure 2) shows the values of the standardized coefficients for the eight predictors used with a 95% confidence interval. Blocks shaded in a lighter blue represent locations with a more significant influence in the regression model, while darker shades indicate a smaller influence.

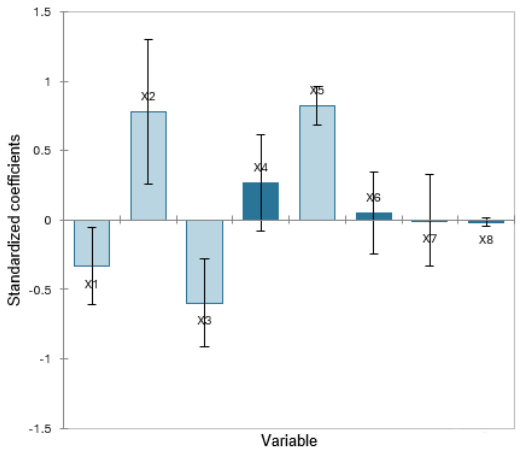

Figure 2. Values of the standardized coefficients of predictors with a 95% confidence interval

From the figure above, it can be seen that X5 has the strongest influence (0.78), because it has the highest standardized coefficient, meaning it contributes the most to predicting the dependent variable. Conversely, X7 and X8 have the lowest coefficients - their impact on the prediction is minimal.

4. CONCLUSION

In the study reported in this paper, the annual maximum daily precipitation data at the nine locations from a grid network sourced from the Digital Climate Atlas of Serbia are considered. Each point/location contains complete data strings, with intentional holes introduced at the central point in various locations along the string, resulting in 36 new strings for this point. Three different methods were applied to fill artificially induced missing data. The results demonstrated that the MLR method is well-suited for estimating missing precipitation data, even in cases where 20% of the data is missing. RF had satisfactory but not as good results as MLR, while LLG indicated that in some cases, it can give a weaker forecast of potential values of extreme precipitation. This work represents a respectable example of how incomplete precipitation datasets can be supplemented, indicating that a more conservative approach is potentially better than newer techniques involving the application of artificial intelligence and machine learning to solve such problems.

REFERENCES

[1] Sattari, M. T., Rezazadeh-Joudi, A., & Kusiak, A.: **Assessment of different methods for estimation of missing data in precipitation studies.** *Hydrology Research*, *48*(4), 1032-1044, 2017.

[2] Duarte, L. V., Formiga, K. T. M., & Costa, V. A. F.: **Comparison of methods for filling daily and monthly rainfall missing data: statistical models or imputation of satellite retrievals?.** *Water*, *14*(19), 3144, 2022.

[3] Falck, A. S., Maggioni, V., Tomasella, J., Vila, D. A., & Diniz, F. L.: **Propagation of satellite precipitation uncertainties through a distributed hydrologic model: A case study in the Tocantins–Araguaia basin in Brazil**. *Journal of Hydrology*, *527*, 943-957, 2015.

[4] Koutsoyiannis, D.: **Advances in stochastics of hydroclimatic extremes**. *L'Acqua*, 23-32, 2021.

[5] Xia, Y., Fabian, P., Stohl, A., & Winterhalter, M.: **Forest climatology: estimation of missing values for Bavaria, Germany.** *Agricultural and forest meteorology*, 96(1-3), 131-144, 1999.

[6] de Oliveira, L. F., Fioreze, A. P., Medeiros, A. M., & Silva, M. A.: **Comparison of gap filling methodologies of annual historical series of rainfall**. *Revista Brasileira de Engenharia Agrícola e Ambiental*, *14*, 1186-1192, 2010.

[7] Simolo, C., Brunetti, M., Maugeri, M., & Nanni, T.: **Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach**. *International Journal of Climatology*, *30*(10), 1564-1576, 2010.

[8] Nathans, L., Oswald, F., & Nimon, K.: **Interpreting multiple linear regression: A guidebook of variable importance.** *Practical Assessment Research & Evaluation, 17 (9), 19*, 2012.

[9] Pappas, C., Papalexiou, S. M., & Koutsoyiannis, D.: **A quick gap filling of missing hydrometeorological data**. *Journal of Geophysical Research: Atmospheres*, *119*(15), 9290-9300, 2014.
[10] Egigu, M.: **Techniques of filling missing values of daily and monthly rain fall data: a review**. *SF J. Environ. Earth Sci*, *3*(1), 2020.
[11] Portuguez-Maurtua, M., Arumi, J. L., Lagos, O., Stehr, A., & Montalvo Arquinigo, N.: **Filling gaps in daily precipitation series using regression and machine learning in Inter-Andean Watersheds**. *Water*, *14*(11), 1799, 2022.
[12] Körner, P., Kronenberg, R., Genzel, S., & Bernhofer, C.: **Introducing Gradient Boosting as a universal gap filling tool for meteorological time series**. *Meteorologische Zeitschrift*, *27*(5), 369-376, 2018.
[13] *Ministarstvo zaštite životne sredine*, 2022, Digitalni atlas klime i klimatskih promena Republike Srbije. Projekat „Unapređenje srednjoročnog i dugoročnog planiranja mera prilagođavanja na izmenjene klimatske uslove u republici Srbiji“, https://atlas-klime.eko.gov.rs.
[14] Nieminen, P.: **Application of standardized regression coefficient in meta-analysis**. *BioMedInformatics*, *2*(3), 434-458, 2022.
[15] Gofa, F., Mamara, A., Anadranistakis, M., & Flocas, H.: **Developing gridded climate data sets of precipitation for Greece based on homogenized time series.** *Climate*, *7*(5), 68, 2019.
[16] Breiman, L.: **Random forests**. *Machine learning*, 45, 5-32, 2001.
[17] Dureh, N., Ueranantasan, A., & Eso, M.: **A comparison of multiple linear regression and random forest for community concern of youth and young adults survey.** *Methods*, 44, 481-487, 2018.
[18] *Addinsoft*. (2025). **XLSTAT statistical software** (Version 2025). [Computer software]. https://www.xlstat.com.
[19] Benoit, K.: **Linear regression models with logarithmic transformations**. *London School of Economics*, London, 22(1), 23-36, 2011.
[20] Tranmer, M., & Elliot, M.: **Multiple linear regression**. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 5(5), 1-5, 2008.
[21] Hasanpour Kashani, M., & Dinpashoh, Y.: **Evaluation of efficiency of different estimation methods for missing climatological data**. *Stochastic environmental research and risk assessment*, *26*, 59-71, 2012.
[22] Willmott, C. J., Matsuura, K., & Robeson, S. M.: **Ambiguities inherent in sums-of-squares-based error statistics**. *Atmospheric Environment*, 43(3), 749-752, 2009.