SPATIAL INTERPOLATION OF RAINFALL AT LAKE TOBA REGION WITH ORDINARY AND EXTERNAL DRIFT KRIGING

Nicholas ABISHA [1], Sri NURDIATI [2*], Endar Hasafah NUGRAHANI [3], Mohamad Khoirun NAJIB [2] & Syukri Arif RAFHIDA [2]

DOI: 10.21163/GT_2026.213.07

ABSTRACT

This study aims to interpolate rainfall in the Lake Toba region using ordinary kriging (OK) and external drift kriging (EDK) with elevation and distance-to-lake as drift variables. Rainfall data from 34 stations (2007-2017) were evaluated across three scenarios: (1) without imputation, (2) sparse data using only complete stations, and (3) all stations using k-nearest-neighbors imputation. Leave-one-out cross-validation revealed that in dense-data scenarios (1 and 3), OK and EDK produced similar accuracy, indicating that dense local observations override the need for auxiliary variables. However, in the sparse-data scenario (2), EDK with distance-to-lake significantly outperformed both OK and EDK with elevation. Geographically, this highlights that standard assumptions of orographic lift fail in the caldera due to complex rain shadows over Samosir Island. Instead, distance-to-lake serves as a reliable proxy for the localized lake-land breeze circulations that drive the basin's microclimate. Ultimately, this study demonstrates that spatial interpolation in complex, data-sparse regions must be grounded in variables that accurately reflect physical hydrometeorological realities.

Keywords: *Rainfall, Kriging, Lake Toba, Elevation, Distance-to-Lake.*

1. INTRODUCTION

Recent studies have reported a concerning decline and fluctuation in Lake Toba's water level, which has drawn increasing attention due to its implications for tourism and hydropower production (Irwandi et al., 2021). Falling water levels may disrupt cruise ship operations, while Irwandi et al. (2021) also noted a reduction in hydroelectric power generation affecting aluminium production at PT Inalum. These changes are likely driven by rising air temperatures and decreasing rainfall, as supported by multiple observations across the region.

Building on this, Irwandi et al. (2023) analyzed climate trends in the Lake Toba region using bias-corrected ERA5-Land data. The utility of ERA5 reanalysis data for characterizing extreme rainfall events has also been demonstrated in other parts of the Indonesian archipelago (Koesuma et al., 2025). The analysis found that 70\% of the area showed increasing rainfall trends, while the rest showed declines. However, further verification using in-situ station data is required to confirm the spatial distribution of these trends. Understanding this spatial variability is essential for assessing its linkage to lake hydrology and the ongoing water level decline.

---

[1] Master's Program in Applied Mathematics, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia, abishanicholas@apps.ipb.ac.id (NA),

[2] Division of Computational Mathematics, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia. [*] Corresponding author nurdiati@apps.ipb.ac.id (SN), mkhoirun_najib@apps.ipb.ac.id (MKN), asrafhida@apps.ipb.ac.id (SAR)

[3] Division of Mathematical Economics, Finance, and Actuarial Science, School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia, e_nugrahani@apps.ipb.ac.id (EHN),

Spatial interpolation methods are frequently employed to estimate rainfall at ungauged locations. Common techniques include inverse distance weighting (IDW), ordinary kriging (OK), and external drift kriging (EDK), among others (Fung et al., 2022; Caloiero et al., 2021). Comprehensive reviews indicate that while geostatistical methods generally outperform deterministic techniques for monthly and annual rainfall, their relative success depends heavily on network density and the availability of secondary variables (Ly et al., 2013). Previous comparative studies show varying results regarding kriging performance. For example, Das & Islam (2021) found that kriging with external drift (KED) using longitude as a covariate significantly outperformed both OK and IDW in mapping annual average rainfall across Bangladesh. Conversely, Frazier et al. (2016) found that OK produced lower error statistics than multivariate methods for rainfall interpolation in the Hawaiian Islands, suggesting that added model complexity does not always guarantee improved accuracy. These findings underscore that the effectiveness of kriging methods depends not only on station density, but also on the suitability and spatial relevance of drift variables (Kyriakidis et al., 2001). Historically, elevation has been the most common auxiliary variable incorporated into multivariate geostatistical algorithms for rainfall interpolation, often yielding significant accuracy improvements in mountainous regions (Goovaerts, 2000).

A critical factor influencing kriging accuracy is the choice and fitting of the variogram model. The spherical variogram model remains widely used due to its suitability for isotropic spatial processes and ease of parameter estimation, as demonstrated in recent applications such as Somayasa et al. (2021) where it was employed to model spatial variability in agricultural land using weighted universal kriging. Empirical variograms are typically estimated using methods such as the Cressie-Hawkins estimator, and parameter fitting is often performed via weighted least squares (Pearse et al., 2024). The Cressie-Hawkins estimator remains a relevant and widely adopted approach for robust variogram estimation, as demonstrated by its implementation in recent geostatistical software such as SciKit-Gstat (Mälicke, 2022) and GstatSim (Mackie et al., 2023), which emphasize reproducibility and flexibility in modern Python-based spatial analysis workflows.

Beyond software adoption, recent literature also affirms the value of the Cressie-Hawkins estimator in applied geostatistics. Gutierrez-Lopez (2021) proposed a modification of the Cressie-Hawkins estimator (CH-GLo) by incorporating the fourth statistical moment, enhancing its performance for mapping extreme hydrological events. The modified estimator effectively reduced the impact of outliers in rainfall data during an extreme storm event in Queretaro, Mexico. Similarly, Mahdi et al. (2020) conducted comparative analyses of empirical variogram estimators including CH. While alternative models like VNN performed better in their groundwater case study, CH remained a robust and valid comparator, especially in small-sample contexts. In the domain of geochemical exploration, Nguyen (2020) employed the Cressie-Hawkins estimator to mitigate the influence of outliers in the spatial analysis of ore-forming elements. Their results demonstrated the estimator's utility in quantifying moderate spatial dependence in skewed geochemical datasets. Collectively, these applications demonstrate that the CH estimator remains a practical and theoretically grounded approach in diverse spatial modeling problems.

While previous geostatistical studies have evaluated kriging methods globally, there is a distinct lack of comparative research tailored to the unique topographical and microclimatic conditions of Indonesia's major caldera lakes. The novelty of this study lies in explicitly testing whether standard auxiliary variables like elevation—which are often successful in temperate or continental mountainous regions—remain effective in the equatorial, lake-driven microclimate of the Lake Toba basin, particularly when compared against a localized distance-to-water metric. This study aims to map the spatial distribution of rainfall across the Lake Toba region using and comparing OK and EDK methods, incorporating robust spherical variogram models estimated via the Cressie-Hawkins method. By filling this critical regional knowledge gap with a rigorously validated spatial rainfall model, this research provides the essential hydrometeorological baseline needed to investigate and mitigate Lake Toba's declining water levels.

2. MATERIALS AND METHODS

2.1. Study Area and Data

The study area is the Lake Toba region, Indonesia, covering a geographical grid 86 from 98.27◦E-99.32◦E longitude and 2.05◦N-3.05◦N latitude. The locations of the 34 meteorological stations used in this study are shown in **Fig. 1**. The primary data consisted of monthly rainfall (mm) time series from 34 stations, collected from January 1971 to December 2017, and obtained from the Indonesian Agency for Meteorology, Climatology, and Geophysics (BMKG). Two external variables were used for EDK: (1) Elevation (Elev), obtained from the public Open-Elevation API, depicting the caldera's physical terrain (**Fig. 1A**), and (2) Distance from the Lake (Dist), defined as the shortest distance from each station to the shoreline, calculated using GADM administrative area shapefiles, representing the proximity to the primary moisture source (**Fig. 1B**).

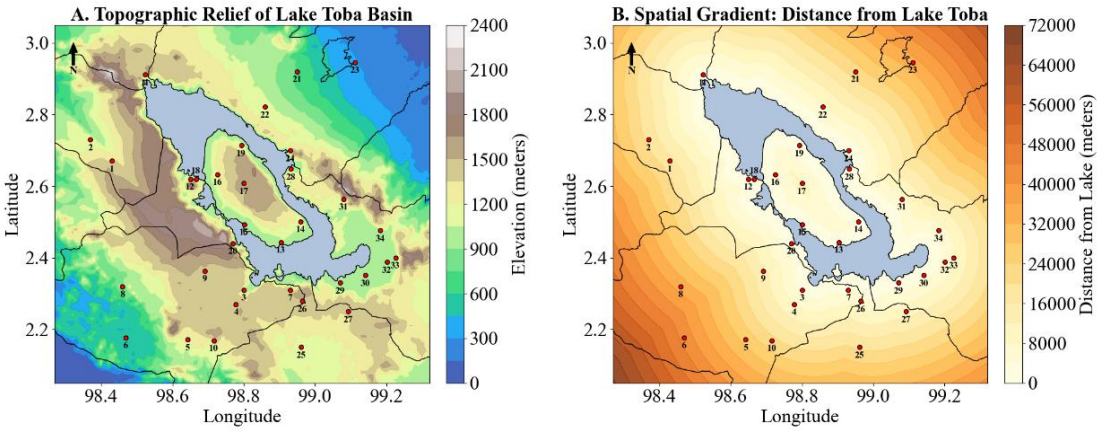

Fig. 1. Location of the 34 meteorological stations in the Lake Toba region. Panel (A) displays the topographic relief (elevation) of the basin, while Panel (B) illustrates the spatial gradient representing the shortest distance from each station to the lake shoreline.

2.2. K-Nearest Neighbors Imputation

Imputation is a class of procedures performed with the aim of estimating missing data values (Batista and Monard 2002). One imputation method that can be applied is k-nearest-neighbors (KNN). The KNN imputation method applies the concept of estimation based on the nearest neighbors. The KNN method works by predicting the value of a missing observation using information from the k nearest samples. The proximity between two points is calculated using a distance metric. The most commonly used metric is the Euclidean distance (Kuhn and Johnson 2013):

$$d(a,b) = \sqrt{\sum_{j=1}^{p} (x_{aj} - x_{bj})^2}$$

This method was selected for its computational efficiency and its effectiveness in preserving the local spatial correlation structures of the rainfall data.

2.3. Empirical Variogram

The experimental variogram, or empirical variogram, is a function computed from the observed data $z(x_i), i = 1, 2, \ldots$, to describe the spatial correlation before a theoretical model is fitted (Webster and Oliver 2007). The classical empirical variogram estimator, also known as the method of moments estimator, was introduced by Matheron (Cressie 1993; Webster and Oliver 2007). It is defined as:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \left(z(x_i) - z\left(x_j\right)\right)^2,$$

where $N(h)$ is the set of all pairs of observation points (i, j) separated by a lag vector h.

While common, the Matheron estimator is known to be sensitive to outliers in the data (Cressie and Hawkins 1980; Oliver and Webster 2015). To address this, a more robust empirical variogram estimator was proposed by Cressie and Hawkins (1980), which is given by

$$\hat{\gamma}(h) = \frac{\left(\frac{1}{|N(h)|}\sum_{(i,j)\in N(h)} \left|z(x_i) - z\left(x_j\right)^{0.5}\right|\right)^4}{2\left(0.457 + \frac{0.494}{|N(h)|} + \frac{0.045}{|N(h)|^2}\right)}.$$

In this study, the empirical variogram was modeled by applying this robust Cressie-Hawkins estimator (Cressie 1985; Al-Mofleh et al. 2016; Mohebzadeh 2018).

2.4. Theoretical Variogram

A theoretical variogram is a model obtained by fitting a mathematical function to the experimental (empirical) variogram (Webster and Oliver 2007). One form of a theoretical variogram is the spherical model, which is given by:

$$\gamma(h;\theta) = f(x) = \begin{cases} c_0 + c\left[\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right], & 0 \le h < a \\ c_0 + c, & h \ge a \end{cases},$$

where $\theta = (c_0, c, a)$. The spherical variogram function is designed to be bounded above.

This is because bounded semivariance is more commonly encountered in geostatistics (Cressie 1985; Webster and Oliver 2007). The parameters θ are defined as follows:

- Total Sill $(c_0 + c)$: The maximum value of the semivariance.
- Range (a): The value of h at which the theoretical variogram reaches the total sill.
- Nugget Effect (c_0): The component of the total sill representing the semivariance at lag 0, often interpreted as micro-scale variation.
- Partial Sill (c): The other component of the total sill, representing the spatially correlated portion of the variance.

A robust method for fitting the spherical theoretical variogram, according to Cressie (1985) as cited in Al-Mofleh (2016), is the Weighted Least Squares (WLS) method. This method applies the weights:

$$w_i = \frac{|N(h_i)|}{\gamma(h_i;\theta)^2}, \quad i = 1,2,\ldots,n,$$

where $|N(h_i)|$ denotes the number of pairs represented by the lag h_i, and $\gamma(h_i;\theta)$ is the value of the theoretical semivariance at lag h_i.

2.5. Ordinary Kriging

Ordinary Kriging (OK) is a geostatistical estimation method that predicts the value of a random variable, Z, at an unobserved point, x_0, using a weighted linear combination of sample points $z(x_1),\ldots,z(x_n)$ (Webster and Oliver 2007). The OK estimator is:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i),$$

where λ_i are the weights assigned to each observation.

The weights are the solutions to the system of equations $A\lambda = b$, where A is the covariance matrix between all observations, b is the covariance vector between all observations and the target location, and λ is the weight vector.

2.6. External Drift Kriging

The Ordinary Kriging method assumes that the mean of the process is stationary (constant). However, some spatial processes contain a trend, also known as a drift, where the expected value is not stationary (Webster and Oliver 2007). When the trend is a function of K external variables, $y_1, y_2, \ldots, y_K$, which are known at all locations, the model is modified to incorporate these auxiliary predictors (Hengl et al., 2003):

$$Z(x) = \sum_{k=0}^{K} \beta_k y_k(x) + \varepsilon(x),$$

where β_k are unknown coefficients.

This process is known as Kriging with External Drift or External Drift Kriging (EDK). The EDK estimator is:

$$\hat{Z}_{KED}(x_0) = \sum_{i=1}^{n} \lambda_i^{KED} z(x_i).$$

The weights λ_i^{KED} are solved using the extended matrices: $\lambda^{KED} = \left\{\lambda_1^{KED}, \ldots, \lambda_n^{KED}, \varphi_0, \ldots, \varphi_p\right\}^{\top}$, where λ^{KED} is the vector of solved weights and φ_p are the Lagrange multipliers.

3. EXPERIMENTAL DESIGN

The analysis was performed using data from the 2007--2017 period. To assess the impact of missing data (NaN values) on interpolation accuracy, three distinct data-handling scenarios were designed:

- Scenario 1 (S1): All 34 stations were used. Missing values (288 cells, 6.41% of data) were left as NaN, with the risk of bias from an unbalanced dataset.
- Scenario 2 (S2): The dataset was reduced to only the 15 stations that had complete data records for the entire period. This ensures minimal bias but reduces spatial coverage.
- Scenario 3 (S3): All 34 stations were used. The 288 missing values were imputed using KNN method, balancing data completeness and spatial coverage.

For Scenario 3, a Monte-Carlo simulation (50 iterations) was performed to determine the optimal imputation hyperparameters. This simulation, which involved randomly masking known values, found that $K = 16$ neighbors with distance-based weighting yielded the lowest RMSE. Finally, for all three scenarios, the 11-year monthly time-series data was aggregated into four quarterly averages to minimize seasonal variance: Jan-Mar (JFM), Apr-Jun (AMJ), Jul-Sep (JAS), and Oct-Dec (OND).

The kriging interpolation was implemented using the Python GSTools library. The interpolation was performed on a 100×100 grid covering the study area. The full experimental plan involved 36 unique interpolation models:

- 12 Ordinary Kriging (OK) models (4 quarters × 3 scenarios).
- 24 External Drift Kriging (EDK) models (4 quarters × 3 scenarios × 2 drift variables [Elev and Dist]).

Model accuracy was evaluated using Leave-One-Out Cross-Validation (LOOCV). In this study, for each iteration ($i = 1, \ldots, n$), a single station (x_i) was removed from the dataset. The remaining $n - 1$ stations were used to re-calculate the variogram parameters and perform kriging to predict the value at x_i. This process was repeated for all n stations. The resulting prediction errors ($e_i = \hat{z}(x_i) - z(x_i)$) were compiled, and RMSE, MAE, and MAPE were calculated to compare model performance.

4. RESULTS AND DISCUSSIONS

4.1. Data Validation, Drift Selection, and Variogram Modeling

The analysis began with an exploratory data analysis to validate the three data-handling scenarios. As shown in **Table 1**, a comparison of Scenario 1 (original data) and Scenario 3 (imputed data) reveals that their descriptive statistics are highly comparable.

Table 1.

Mean and standard deviation of rainfall data across all scenarios and partitions.

Scenario		Mean (mm)				Std. Dev. (mm)			
		JFM	AMJ	JAS	OND	JFM	AMJ	JAS	OND
S1	No imp.	172.034	174.171	152.627	247.280	46.311	50.837	41.322	58.152
S2	$n = 15$	173.196	174.585	151.693	250.506	49.090	42.064	29.317	44.889
S3	Imputed	172.778	173.132	154.041	244.665	46.417	49.671	40.997	54.535

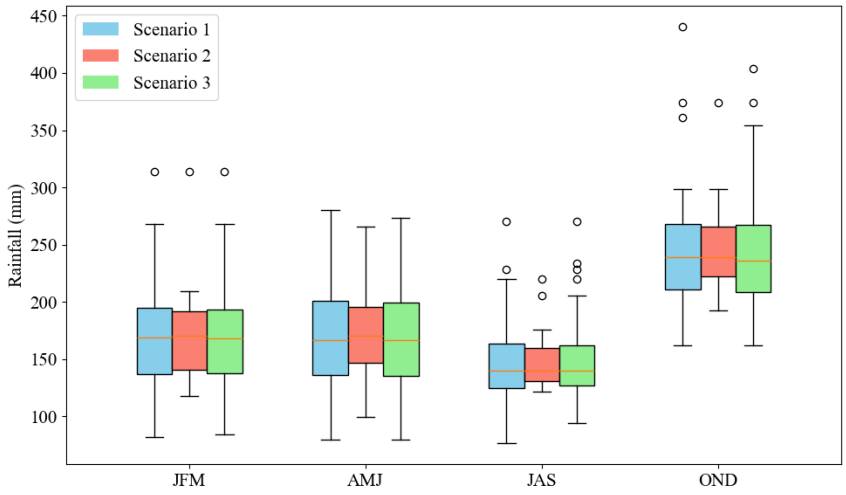

Fig. 2. Boxplot of rainfall in all partitions across all scenarios.

The minor differences in mean and standard deviation (e.g., S1 OND Mean 247.28 mm vs. S3 OND Mean 244.67 mm) are the expected result of including the 6.41% of imputed data points. This lack of significant statistical distortion validates the S3 dataset as a complete and reliable representation of the original data. Conversely, Scenario 2 (reduced stations) showed notable deviations, such as a significantly lower variance in the JAS quarter (std. dev. 29.32 mm), confirming it as a biased sample.

Beyond central tendency, the data's distribution (**Fig. 2**) revealed the presence of significant outliers in all datasets except for AMJ in all scenarios. Shapiro-Wilk normality tests (**Table 2**) confirmed that for both S1 and S3, the driest (JAS) and wettest (OND) quarters were non-normally distributed ($p < 0.05$). This is a critical finding, as standard variogram estimators (like Matheron's) are highly sensitive to such extreme values in the data. Therefore, the data's non-normality and outliers justify the use of the robust Cressie-Hawkins estimator for variogram modeling.

Table 2.

Shapiro-Wilk normality test results across all scenarios and partitions.

Scenario		Quarter	*W* Statistic	*p*-value	Decision
S1	**No imp.**	**JFM**	0.9534	0.1548	Normal
		AMJ	0.9614	0.2678	Normal
		JAS	0.9164	0.0128	Not normal
		OND	0.8901	0.0025	Not normal
S2	$n = 15$	**JFM**	0.8557	0.0209	Not normal
		AMJ	0.9675	0.8197	Normal
		JAS	0.8585	0.0230	Not normal
		OND	0.8818	0.0505	Normal
S3	**Imputed**	**JFM**	0.9538	0.1592	Normal
		AMJ	0.9631	0.2980	Normal
		JAS	0.8958	0.0036	Not normal
		OND	0.9144	0.0113	Not normal

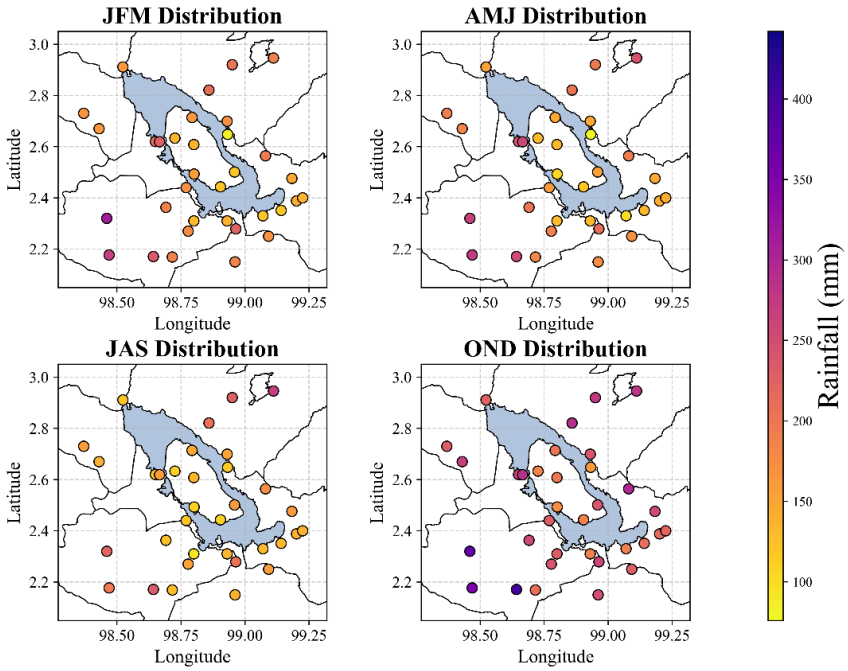

Fig. 3. Exploratory spatial distribution of observed seasonal rainfall across the Lake Toba basin for the JFM, AMJ, JAS, and OND quarters.

In addition to examining the statistical distribution, an exploratory spatial data analysis (ESDA) was conducted to visualize the raw seasonal precipitation patterns across the basin. As illustrated in **Fig. 3**, the observed rainfall data exhibits clear spatial heterogeneity. Throughout the year, higher precipitation volumes are consistently concentrated in the southwestern and northeastern regions of the outer basin. In contrast, a distinct dry zone persists over Samosir Island and the central caldera, particularly evident during the drier JAS quarter. This preliminary spatial mapping visually corroborates the complex microclimatic dynamics of the region, highlighting the need to carefully select and incorporate physical geographical variables into the interpolation model.

To evaluate potential drift variables for EDK, a Pearson correlation analysis was performed (**Table 3**). The analysis revealed that elevation (Elev) has a weak and mostly non-significant correlation with rainfall. In stark contrast, the distance from the lake (Dist) exhibited a strong, statistically significant positive correlation with rainfall across all quarters and scenarios ($r > 0.59$, $p < 0.01$). This confirms that Dist is a highly relevant and promising auxiliary variable, while Elev is expected to perform poorly.

Finally, spherical variogram models were fitted (**Table 4**). The fitted parameters for S1 and S3 were highly comparable, further demonstrating that the KNN imputation successfully preserved the underlying spatial correlation structure. In contrast, S2 (reduced stations) produced markedly different parameters, including a nugget effect of zero for most seasons. This suggests the sparse data led to an over-simplified model, unlike the more realistic micro-scale variability (noise) captured in S1 and S3.

Table 3.

Pearson correlation coefficients (r) and p-values for drift variables against quarterly rainfall.

Drift var.	Quarter	S1		S2		S3	
		r	p-value	r	p-value	r	p-value
Elev	JFM	-0.08	0.636	-0.02	0.940	-0.10	0.576
	AMJ	-0.11	0.541	0.17	0.554	-0.12	0.482
	JAS	-0.40	0.020	-0.22	0.427	-0.39	0.024
	OND	-0.19	0.271	-0.07	0.802	-0.20	0.245
Dist	JFM	0.65	0.000	0.75	0.001	0.66	0.000
	AMJ	0.70	0.000	0.65	0.009	0.68	0.000
	JAS	0.66	0.000	0.59	0.019	0.68	0.000
	OND	0.68	0.000	0.77	0.001	0.69	0.000

To show the variogram visually, the plot for the S3 models was shown (**Fig. 4**). The plot shows a clear spatial dependence structure that provided a reliable basis for interpolation. However, a clear full sill is not visible in the empirical plots. This is likely because the spatial extent of the study area (approx. 1.05 degrees) is comparable to the fitted spatial correlation range (0.75–1.00 degrees), meaning the maximum lag distance is insufficient to capture the full decorrelation. Additionally, outliers appear at greater distances. This is why weighted least squares was chosen for theoretical variogram fitting, it reduces the effect of these outliers at greater distances. These outliers may also show that the CH estimator has its limits.

Table 4.

Parameters of the fitted spherical variogram model.

Scenario		Parameter	JFM	AMJ	JAS	OND
S1	No imp.	Range (deg)	0.91	0.75	1.00	1.00
		Total Sill (mm^2)	3044.3	3667.86	2573.82	4590.05
		Nugget (mm^2)	592.58	400.59	195.93	645.97
S2	$n = 15$	Range (deg)	0.87	0.87	0.87	0.87
		Total Sill (mm^2)	4306.53	3595.28	1359.12	6019.41
		Nugget (mm^2)	0.00	0.00	183.08	0.00
S3	Imputed	Range (deg)	0.92	1.00	1.00	1.00
		Total Sill (mm^2)	3174.820	4076.37	2531.76	4224.46
		Nugget (mm^2)	460.27	665.67	167.23	719.72

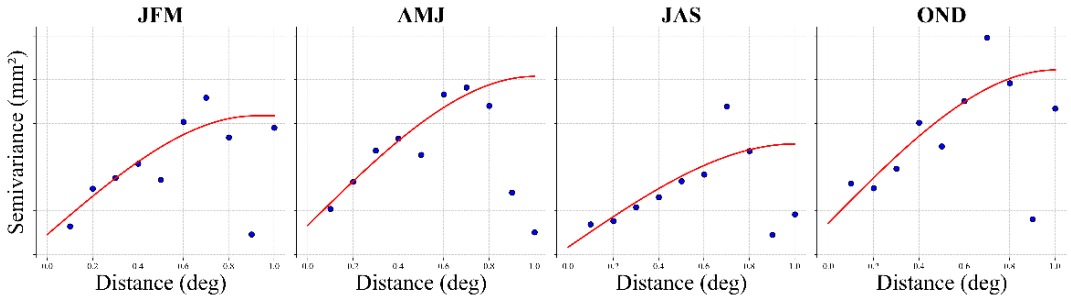

Fig. 4. Visual representation of the fitted theoretical variogram models in Scenario 3.

4.2. Kriging Performance and Spatial Interpolation

The 36 kriging models were evaluated using Leave-One-Out Cross-Validation (LOOCV), with results presented in **Table 5** (for OK) and **Table 6** (for EDK). Kriging was performed on untransformed data, as preliminary tests with log, square-root, and Box-Cox transformations did not improve evaluation metrics.

Table 5.

LOOCV evaluation metrics for Ordinary Kriging (OK).

Scenario		Quarter	RMSE	MAE	MAPE (%)
S1	**No imp.**	**JFM**	36.9534	26.6975	16.8347
		AMJ	35.5512	28.8207	18.8534
		JAS	34.9528	27.8754	19.2172
		OND	50.1285	34.9198	14.3593
S2	$n = 15$	**JFM**	45.4856	35.0459	19.2639
		AMJ	39.0001	31.7394	18.5105
		JAS	32.9532	25.1486	15.8229
		OND	41.1272	29.1662	10.819
S3	**Imputed**	**JFM**	35.4235	26.3160	16.3979
		AMJ	35.5623	29.4276	19.1713
		JAS	34.4482	27.7322	18.3211
		OND	45.6477	32.9888	13.7802

The validation metrics reveal two critical findings. First, for the dense-data scenarios (S1 and S3), the performance of OK and EDK was nearly identical. In S3, for example, the OND (wettest quarter) RMSE for OK was 45.65 mm, while EDK-Elev was 46.06 mm and EDK-Dist was 45.39 mm. These differences are negligible, which leads to a significant conclusion: with a sufficiently dense station network, the added complexity of EDK provides no tangible benefit. The kriging estimate is dominated by the wealth of nearby observations, rendering the influence of the auxiliary drift variable, whether strongly correlated or not, statistically insignificant.

The central finding of this study emerged from the sparse-data Scenario (S2). Here, the EDK-Dist model was the undisputed best-performing model, particularly in the JAS (RMSE 23.39) and OND (RMSE 26.97) quarters. This shift in performance aligns with findings by Camera et al. (2014), who observed that the ranking of interpolation techniques is directly dependent on observation density. They noted that higher station density leads to greater stability in method ranking, whereas sparse networks often necessitate specific techniques that can better capture spatial variability. This contrast is evident in our results: while dense scenarios showed little difference between methods, the sparse scenario (S2) revealed a stark divergence between the successful EDK-Dist and the poor-performing S2-OK (OND RMSE 26.97) and S2-EDK-Elev (OND RMSE 50.42).

Table 6.

LOOCV evaluation metrics for External Drift Kriging (EDK).

Scenario		Drift	Quarter	RMSE	MAE	MAPE (%)
S1	No imp.	Elev	JFM	35.8549	26.7138	16.6096
			AMJ	34.5249	27.0139	16.9901
			JAS	35.1094	27.5594	18.8782
			OND	50.5462	34.8214	14.2939
		Dist	JFM	36.3958	27.7241	17.3533
			AMJ	34.2258	27.6334	18.2767
			JAS	33.5745	26.3913	18.2241
			OND	50.4825	35.7599	14.5359
S2	$n = 15$	Elev	JFM	59.2182	41.8769	22.4403
			AMJ	49.171	38.2899	21.7479
			JAS	37.242	29.5129	18.6888
			OND	50.4235	34.2488	12.7015
		Dist	JFM	42.7698	31.7958	17.8031
			AMJ	37.1501	28.1417	16.3599
			JAS	23.3909	18.3329	11.5188
			OND	26.965	20.6127	7.694
S3	Imputed	Elev	JFM	34.6249	26.1632	16.1201
			AMJ	35.2068	28.0403	17.7584
			JAS	34.5546	27.397	17.9709
			OND	46.0575	33.0014	13.7706
		Dist	JFM	35.2718	27.5913	17.0083
			AMJ	34.0632	27.9627	18.4094
			JAS	33.1907	26.129	17.3101
			OND	45.3904	33.5958	13.9203

The failure of the EDK-Elevation model to outperform Ordinary Kriging is consistent with geostatistical theory; while Goovaerts (2000) demonstrated that external drift variables can significantly reduce error, this is contingent on a strong correlation. Similarly, Wagner et al. (2012) found that in monsoon-driven mountainous regions, elevation may show no significant correlation with rainfall. They demonstrated that variables representing the distance from a moisture source or orographic barrier (such as the Western Ghats escarpment) can be far superior predictors than local elevation. In our study, the 'Dist' variable, reflecting the lake's influence, acted as a powerful guide for the interpolation precisely because the observation data was too sparse to capture the pattern on its own.

The superior performance of the EDK-Dist model in the sparse-data scenario is deeply rooted in the physical microclimatology of the Lake Toba caldera. As a massive inland water body covering over 1,100 square kilometers, Lake Toba acts as the primary local moisture source for the immediate surrounding basin. The thermal contrast between the lake surface and the surrounding landmass generates strong, localized diurnal lake-land breeze circulations. Moisture evaporated from the lake is transported inland by these breezes, resulting in a precipitation gradient that is heavily dependent on proximity to the shoreline. In a sparse observation network where macro-scale synoptic weather patterns are under-sampled, this distance-to-lake variable effectively captures the dominant microclimatic driver of regional rainfall, providing a highly reliable spatial guide for the interpolation algorithm.

Conversely, the failure of elevation to improve interpolation accuracy is a direct result of the caldera's complex topography and resulting rain shadow effects. While geostatistical theory often assumes a positive linear correlation between elevation and rainfall due to orographic lift, the Lake Toba basin is characterized by exceptionally steep caldera walls and the central presence of Samosir Island.

Prevailing regional winds deposit the majority of their moisture on the windward slopes of the outer basin. By the time air masses descend into the caldera or cross over Samosir Island, they have lost significant moisture, creating localized rain shadows. Consequently, high-elevation stations on leeward slopes may record significantly lower rainfall than lower-elevation stations on windward slopes. Because a simple elevation variable cannot distinguish between windward and leeward positioning, its correlation with rainfall breaks down, rendering it an ineffective auxiliary drift variable for this specific geographical context.

The resulting interpolation maps (**Fig. 5** and **Fig. 6**) visually confirm these statistical findings. The OK maps (**Fig. 5**) show a consistent spatial trend with lower rainfall concentrated around Samosir Island. The EDK maps (**Fig. 6**) using 'Dist' reinforce this same trend but with a smoother, more regular pattern, which the validation metrics in **Table 6** proved to be a more accurate representation. The EDK-Elev maps, in contrast, appear coarser and less refined, reflecting the irrelevant spatial pattern of the elevation data.

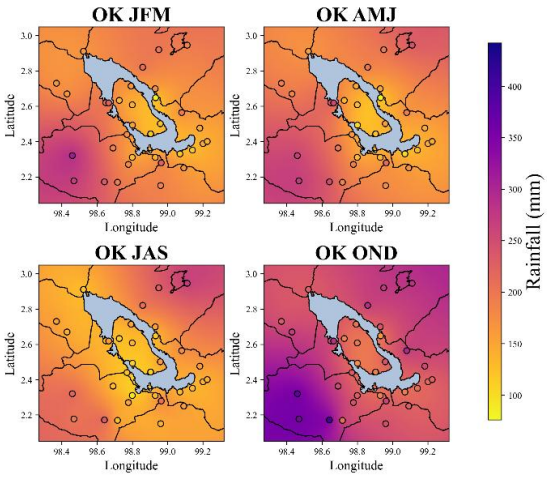

Fig. 5. Interpolated rainfall maps using ordinary kriging for Scenario 3.

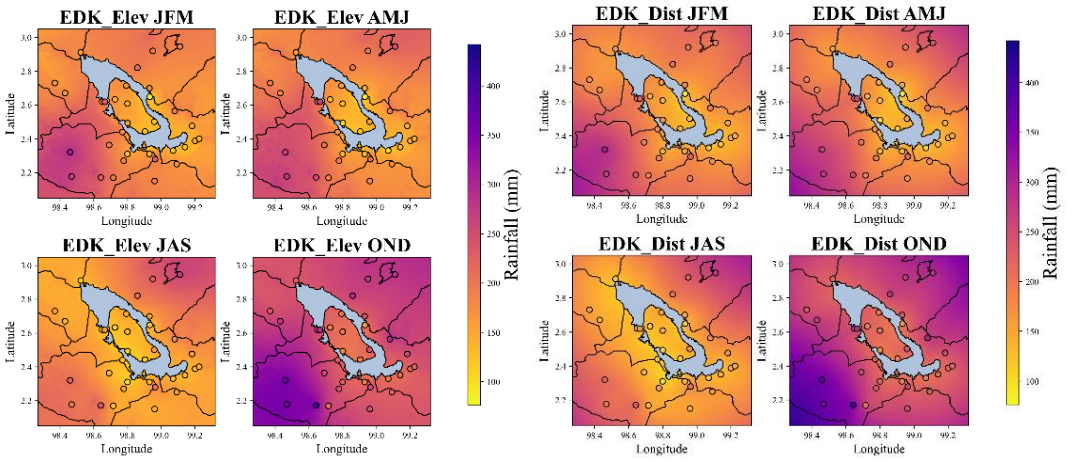

Fig. 6. Interpolated rainfall maps using external drift kriging for Scenario 3 with Elev (left) and Dist (right).

5. CONCLUSION

This study evaluated the spatial interpolation of rainfall in the Lake Toba region by comparing Ordinary Kriging (OK) and External Drift Kriging (EDK) across varied data-handling scenarios. Geographically, the spatial mapping confirmed a consistent precipitation pattern characterized by higher rainfall concentrations in the southwest and northeast, contrasting with a distinct dry zone over Samosir Island and the central caldera. Statistically, the application of k-nearest-neighbors (KNN) imputation proved highly robust, preserving the dataset's underlying spatial correlation structure without introducing the bias inherent in discarding incomplete station records.

The central finding of this research demonstrates that the mathematical effectiveness of interpolation models is deeply governed by the physical geography of the study area. In scenarios with a dense observation network, the abundance of local data points dominates the estimation, rendering the added complexity of EDK largely unnecessary. However, in data-sparse regions, accurate interpolation requires integrating the physical drivers of the local microclimate.

The stark contrast in EDK performance between the two auxiliary variables highlights this geographical dependency. The failure of the elevation variable (EDK-Elev) illustrates that standard assumptions of orographic lift do not uniformly apply in complex terrain; the steep caldera walls and Samosir Island create localized rain shadows that disrupt linear elevation-rainfall correlations. Conversely, the dramatic success of the distance-to-lake variable (EDK-Dist) mathematically confirms Lake Toba's role as the dominant moisture source driving the region's microclimate. By functioning as a proxy for localized lake-land breeze circulations, the distance variable successfully guided the interpolation where raw data was insufficient. Ultimately, this study concludes that spatial interpolation in complex topographical basins cannot rely on statistical optimization alone; it must be grounded in variables that accurately reflect the physical and climatic realities of the landscape.

REFERENCES

Al-Mofleh, H., Daniels, J., & Mckean, J. (2016). Robust variogram fitting using non-linear rank-based estimators. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, *10*(2), 68–77. https://www.researchgate.net/publication/299098198

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. *Soft Computing Systems: Design, Management and Applications*.

Caloiero, T., Pellicone, G., Modica, G., & Guagliardi, I. (2021). Comparative analysis of different spatial interpolation methods applied to monthly rainfall as support for landscape management. *Applied Sciences (Switzerland)*, *11*(20). https://doi.org/10.3390/app11209566

Camera, C., Bruggeman, A., Hadjinicolaou, P., Pashiardis, S., & Lange, M. A. (2014). Evaluation of interpolation techniques for the creation of gridded daily precipitation (1 × 1 km2); Cyprus, 1980–2010. *Journal of Geophysical Research*, *119*(2), 693–712. https://doi.org/10.1002/2013JD020611

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, *17*(5). https://doi.org/10.1007/BF01032109

Cressie, N. A. C. (1993). *Statistics for Spatial Data* (Revised Edition). John Wiley & Sons.

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Mathematical Geology*, *12*(2), 115–125.

Das, S., & Islam, A. R. M. T. (2021). Assessment of mapping of annual average rainfall in a tropical country like Bangladesh: remotely sensed output vs. kriging estimate. *Theoretical and Applied Climatology*, *146*(1–2), 111–123. https://doi.org/10.1007/s00704-021-03729-3

Frazier, A. G., Giambelluca, T. W., Diaz, H. F., & Needham, H. L. (2016). Comparison of geostatistical approaches to spatially interpolate month-year rainfall for the Hawaiian Islands. *International Journal of Climatology*, *36*(3), 1459–1470. https://doi.org/10.1002/joc.4437

Fung, K. F., Chew, K. S., Huang, Y. F., Ahmed, A. N., Teo, F. Y., Ng, J. L., & Elshafie, A. (2022). Evaluation of spatial interpolation methods and spatiotemporal modeling of rainfall distribution in Peninsular Malaysia. *Ain Shams Engineering Journal*, *13*(2). https://doi.org/10.1016/j.asej.2021.09.001

Goovaerts, P. (2000). *Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall*. www.elsevier.com/locate/jhydrol

Gutierrez-Lopez, A. (2021). A Robust Gaussian variogram estimator for cartography of hydrological extreme events. *Natural Hazards*, *107*(2), 1469–1488. https://doi.org/10.1007/s11069-021-04641-9

Hengl, T., Heuvelink, G. B. M., & Stein, A. (2003). *Comparison of kriging with external drift and regression-kriging* *. http://www.itc.nl/library/Academic

Irwandi, H., Rosid, M. S., & Mart, T. (2021). The effects of ENSO, climate change and human activities on the water level of Lake Toba, Indonesia: a critical literature review. In *Geoscience Letters* (Vol. 8, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40562-021-00191-x

Irwandi, H., Rosid, M. S., & Mart, T. (2023). Effects of Climate change on temperature and precipitation in the Lake Toba region, Indonesia, based on ERA5-land data with quantile mapping bias correction. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-29592-y

Koesuma, S., Faqih, A., Hendri, H., Listyarini, J., Kusuma, D. A., Hanggara, A. Z., & Gernowo, R. (2025). Analysis of extreme rainfall of tropical cyclone using solar radiation management and ERA5 data in eastern part of Indonesia. *Geographia Technica*, *20*(1), 298–312. https://doi.org/10.21163/GT_2025.201.20

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (1st ed.). Springer.

Kyriakidis, P. C., Kim, J., & Miller, N. L. (2001). Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. *Journal of Applied Meteorology and Climatology*, *40*(1), 1855–1877.

Ly, S., Charles, C., & Degré, A. (2013). Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale. A review. In *Biotechnol. Agron. Soc. Environ* (Vol. 17, Issue 2).

Mackie, E. J., Field, M., Wang, L., Yin, Z., Schoedl, N., Hibbs, M., & Zhang, A. (2023). GStatSim V1.0: a Python package for geostatistical interpolation and conditional simulation. *Geoscientific Model Development*, *16*(13), 3765–3783. https://doi.org/10.5194/gmd-16-3765-2023

Mahdi, E., Abuzaid, A. H., & Atta, A. M. A. (2020). Empirical variogram for achieving the best valid variogram. *Communications for Statistical Applications and Methods*, *27*(5), 547–568. https://doi.org/10.29220/CSAM.2020.27.5.547

Mälicke, M. (2022). SciKit-GStat 1.0: A SciPy-flavored geostatistical variogram estimation toolbox written in Python. *Geoscientific Model Development*, *15*(6), 2505–2532. https://doi.org/10.5194/gmd-15-2505-2022

Mohebzadeh, H. (2018). *Comparison of Methods for Fitting the Theoretical Variogram to the Experimental Variogram for Estimation of Depth to Groundwater and its Temporal and Spatial Variations*. https://doi.org/10.5829/idosi.aejaes.2018.64.76

Nguyen, T. T. (2020). Analysis of Spatial Dependence of Ore-Forming Elements Using Geostatistics and Moran Correlogram. *Asian Review of Environmental and Earth Sciences*, *7*(1), 47–54. https://doi.org/10.20448/journal.506.2020.71.47.54

Oliver, M. A., & Webster, R. (2015). *Basic Steps in Geostatistics: The Variogram and Kriging*. Springer. http://www.springer.com/series/10183

Pearse, A. R., Cressie, N., & Gunawan, D. (2024). *Optimal prediction of positive-valued spatial processes: asymmetric power-divergence loss*.

Somayasa, W., Sutiari, D. K., & Sutisna, W. (2021). Optimal prediction in isotropic spatial process under spherical type variogram model with application to corn plant data. *Journal of Physics: Conference Series*, *1940*(1), 012003. https://doi.org/10.1088/1742-6596/1940/1/012003

Wagner, P. D., Fiener, P., Wilken, F., Kumar, S., & Schneider, K. (2012). Comparison and evaluation of spatial interpolation schemes for daily rainfall in data scarce regions. *Journal of Hydrology*, *464–465*, 388–400. https://doi.org/10.1016/j.jhydrol.2012.07.026

Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists Second Edition* (S. Senn, M. Scott, & V. Barnett, Eds.; 2nd ed.). John Wiley & Sons Ltd.