Geomechanics and Engineering, *Vol. 35, No. 5 (2023) 499-509*
https://doi.org/10.12989/gae.2023.35.5.499

Application of deterministic models for obtaining groundwater level distributions through outlier analysis

Dae-Hong Min[1a], Saheed Mayowa Taiwo[2b], Junghee Park[3c], Sewon Kim[4d] and Hyung-Koo Yoon*[1]

[1]*Department of Construction and Disaster Prevention Engineering, Daejeon University, Daejeon 34520, Republic of Korea*
[2]*Infrastructure Group, Manawatu District Council, 4743 Feilding, Palmerston North, New Zealand*
[3]*Department of Civil, Environmental Engineering, Incheon national University, Incheon, 22012, Republic of Korea*
[4]*Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology, Republic of Korea*

(*Received March 3, 2023, Revised November 10, 2023, Accepted November 16, 2023*)

Abstract. The objective of this study is to perform outlier analysis to obtain the distribution of groundwater levels through the best model. The groundwater levels are measured in 10, 25 and 30 piezometers in Seoul, Daejeon and Suncheon in South Korea. Fifty-eight empirical distribution functions were applied to determine a suitable fit for the measured groundwater levels. The best fitted models based on the measured values are determined as the Generalized Pareto distribution, the Johnson SB distribution and the Normal distribution for Seoul, Daejeon and Suncheon, respectively; the reliability is estimated through the Anderson-Darling method. In this study, to choose the appropriate confidence interval, the relationship between the amount of outlier data and the confidence level is demonstrated, and then the 95% is selected at a reasonable confidence level. The best model shows a smaller error ratio than the GEV while the Mahalanobis distance and outlier labelling methods results are compared and validated. The outlier labelling and Mahalanobis distance based on median shown higher validated error ratios compared to their mean equivalent suggesting, the methods sensitivity to data structure.

Keywords: error ratio; GEV model; goodness of fit; groundwater level; mahalanobis distance; test statistic

1. Introduction

The sustainable development of groundwater resources highly depends on its scientific management (Olabode and San 2023, Rajabian 2023, Shirazi *et al.* 2013). In particular, Interpolation is one of the scientific techniques used in statistical modeling to estimate measured value according to its spatial variability. In other words, it can be used to predict the water levels at unsampled location through the data measured from sampled locations (Chowdhury *et al.* 1991, Saidi *et al.* 2010, Mubashir *et al.* 2017, Ali *et al.* 2018). Although, there is less or absence of reliable deterministic function to describe this data spatial variabilities. However, probabilistic function can be used to provide a solution to this approach (Kitanidis 1997, Yoon *et al.* 2015, Wang *et al.* 2015). Thus, it is important to determine the best function that can describe the measured values accurately. Nevertheless, measured values are inherently included outliers, aberrant and extreme values that should be removed to increase the reliability of the measured data. In past few decades, probabilistic methods have been extensively applied for modeling and quantifying outlier(s) in various fields for obtaining effective friction angle based on Bayesian approach (Chunlai *et al.* 2018, Wang *et al.* 2019), probability-based geotechnical model (Sari 2021, Tian *et al.* 2016), reliable soil properties (Parry *et al.* 2014, Taiwo and 2018) and subsurface stratification (Wang *et al.* 2019). The results of their studies focused mainly on measuring uncertainty associated with the measured data and however, the reliability of each model is often not addressed.

There is a limitation to capture reliable data based on the only statistical analysis (i.e., mean, median, percentiles, error bars) because the error from measurement, instrument fault and environmental conditions, may be included in the raw data (Ahmadi and Sedghamiz 2007, D'Agostino and Stephens 2007). Thus, they should be removed prior to spatial interpolation. The empirical estimation technique can equally be applied to determine the dominant function that properly reflect the trends of measured values. Although, in hydrology, many families of distribution functions applied in this field, is also a good choice for fitting a groundwater level (Fürst *et al.* 2015) thus, the additional distribution functions, which can parametrically model should be considered. Among these functions, the generalized extreme value (GEV) function belong to an advanced distribution family, comprised of bounded and non-negative theoretical functions (i.e., Frechet, Weibull, Gumbel) hence, suitable for extreme values identification (Zalina *et al*. 2002). Although, the GEV function is mostly considered for extreme value analysis however, its regional application is still subjected to critiques. Gorgoso-Varela and Rojo-Alboreca (Gorgoso-Varela and Rojo-Alboreca

*Corresponding author, Associate Professor
E-mail: hyungkoo@dju.ac.kr
[a]Ph.D. Candidate
[b]Three-waters officer
[c]Assistant Professor
[d]Researcher

ISSN: 2005-307X (Print), 2092-6219 (Online)

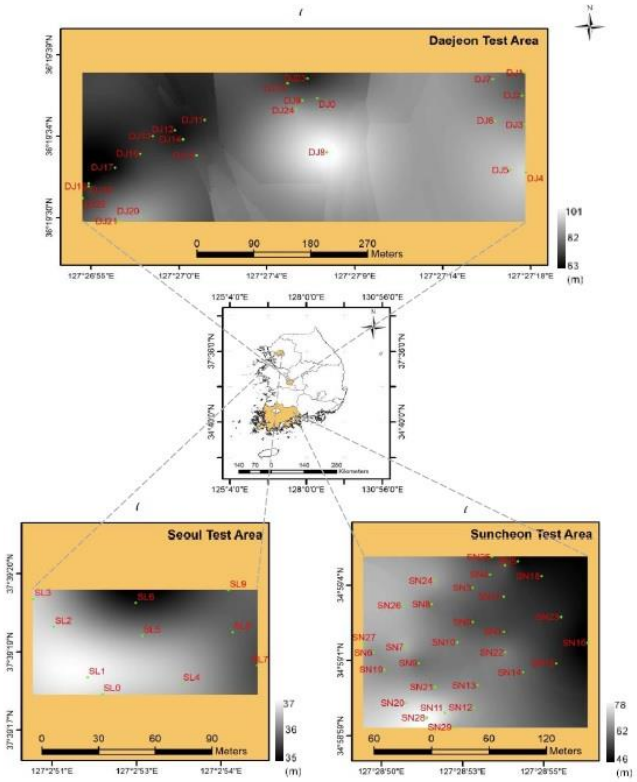

Fig. 1 Site description: (a) Seoul; (b) Daejeon and (c) Suncheon. Note that SL, DJ and SN represent the names of the borehole locations in Seoul, Daejeon and Suncheon, respectively

2014) used only the two functions of Gumbel and Weibull to increase the reliability of the dominate function. Arreyndip and Joseph (Arreyndip and Joseph 2016) tried to identify the consistent function among the Frechet, Weibull, Gumbel and GEV models according to measured values. These results show that the GEV cannot be completely considered as a dominant function for outlier analysis especially where spatial variation is required. Therefore, it is necessary to determine a function that can provide a better correlation with the measured values rather than mere considering an advanced distribution functions such as the GEV. Thus, this study aimed to select a best function that reliably predict the measured groundwater level in the test locations.

After the selection of the suitable model, the determination of reliable confidence intervals is an important procedure in performing outlier analysis (Grubbs 1969; Benstock and Cegla 2017). The confidence interval is the range within an acceptable sampling error where the probable value falls, it consists of a minimum and a maximum level. Note that the unbiased values are identified and removed through the selected confidence level. The confidence interval has been generally suggested between 5% - 95% in a normalized probability density function (Hook 1977, Cheung *et al.* 1990) although, Taiwo and Yoon (Taiwo and Yoon 2018) suggested a confidence interval ranges between 10% - 90% with consideration of the number of outlying data. However, researchers ignored prior knowledge about the measured data when choosing a confidence interval. Therefore, selection of confidence interval is mostly subjective and lack methodological approach. Consequently, this study suggests an alternative method to determine the confidence interval thus, suggests a reasonably range to remove extreme values. The existing outlier analysis lacked clear removal criteria and insufficient generalization for large-scale data. Therefore, in this study, we aim to validate a model that provides outlier removal criteria for spatial prediction.

This study begins with an introduction on the background theory, including the goodness of fit test and the methods of outlier analysis. The site descriptions of the measured values in Seoul, Daejeon and Suncheon are explained and the suitable model that best fit the measured values from the fifty-eight applied models are also considered. The identification of the unbiased values, according to proposed confidence intervals, are described and the suitable ranges are suggested. Finally, the performance of the suggested outlier analysis methods was validated and compared.

2. Site description and data acquisition

The ground water level was obtained from the Geotechnical Information Portal System in South Korea (https://www.geoinfo.or.kr), which provided the results of piezometric tests of regions around South Korea. The results performed with 10, 25 and 30 piezometric locations in Seoul, Daejeon and Suncheon, respectively, were selected, and the positions of each borehole were expressed

in Fig. 1. The names of the piezometers were defined according to their boring number in the three test areas. The three study areas: Seoul, Daejeon and Suncheon with an estimated aerial extent of 12,150 m^2, 232,825 m^2 and 57,575 m^2 as well as elevations of 34 - 37 m, 60 - 105 m and 46 - 80 m respectively

3. Methodology

3.1 Interpolation

Interpolation is a technique used to predict the values at unsampled location through the spatial relationship of the measured values from sampled location within a given neighborhood. Interpolation is a commonly used method in spatial prediction, employed for input data used in outlier analysis. Interpolation techniques can be divided into two main groups: deterministic and geostatistical. Deterministic interpolation techniques predict value from sampled locations, based on the degree of similarity or smoothing between sampled locations. An exact interpolator predicts value identical to the measured value at a sampled location while a different value from the measured value is inexact interpolator (Hemeda 2022, Yasser *et al.* 2022, Johnston *et al.* 2001). Although, Geostatistical interpolation techniques utilizes the statistical relationship between the measured locations by using their spatial autocorrelation through a variogram model. Detailed discussions are provided in (Cressie 1993, Chiles and Delfiner 1999). However, in this study, deterministic method, which predicts data through spatial distributions from an iterative probabilistic function is considered. Deterministic methods involve various mathematical approaches, and the most commonly used one is Inverse Distance Weighting (IDW) (Shepard 1968). IDW is an interpolation method that assigns greater weights to sample data points that are closer in distance, and uses these weights to make predictions.

3.2 Goodness of fit test

Goodness of fit (GOF) denotes how well the observed values fit the empirical distribution function, and thus GOF can be apply to measure the difference between the empirical and the theoretical distribution functions (Chowdhury *et al*. 2010). Methods for measuring this discrepancy, includes Anderson-Darling (AD), Kolmogorov-Smirnov (KS) and Chi-square (χ^2), among them, the AD method is selected in this study because it gives a larger weight to the tail of the distribution curve thus, suitable for identifying misfits associated with extreme value. (Anderson and Darling 1954, Damazio and Kelman 1986, Marsaglia and Marsaglia 2004). Thus, AD is more suitable for identifying outliers (i.e. extremely high or low values) (Fürst *et al*. 2015). The test statistic (A^2) from the AD method [25] is mathematically expressed in Eq. (1).

$$A^2 = -\sum_{i=1}^{n}\left[(2i-1)\frac{\left\{InF_x(x_i)+In\left(1-F_x(x_{n+1-i})\right)\right\}}{n}\right]-n \qquad (1)$$

where n denotes the number of data and $In(F_x(x_i))$ denotes the natural logarithm of cumulative distribution values of the I^{th} data in the theoretical distribution function. The proposed distribution is accepted if the test statistic is less than the corresponding critical value (at selected significant level) otherwise it is rejected (Stephens 1986). Nevertheless, a smaller test statistic depicts a reliable model fit.

3.3 Model distribution fits

Fifty-eight empirical distribution functions were applied to find a suitable fit for the measured groundwater levels at the Seoul, Daejeon and Suncheon locations. The GOF based on AD was analyzed through a distribution fitting software, Easy-Fit (version 5.6), and the results were displayed in Fig. 2. Five models depicting lowest test statistic (i.e., good correlation) among the models were selected (Hosking and Wallis 1997). However, if the generalized (GEV) model is not selected among the first five models, then it is shown in the fifth graph (Fig. 2). The reason for considering the GEV model is for comparison (Fig. 3), as it is mostly applied in outlier analysis.

3.4 Outlier labeling

Useful information regarding measured discordancy in a dataset can be obtained by comparing N-sample variables as a point in N-dimensional space (Hosking and Wallis 1997). Point(s) that are far from the center of the space is identified as an outlier. Based on this fundamental technique, outlier labeling consists of two independent unidirectional outlier lines, which are plotted along horizontal and vertical axes, according to a proposed confidence level [8]. The data inside the region where the two boundary lines intersect are considered to be reliable values; however, those left in the remaining areas are regarded as outlier. The values corresponding to a proposed confidence interval are calculated through Eq. (2)

$$Value_{CI} = (X_{mean}) \pm Z_{(\alpha/2)} \cdot \frac{s}{\sqrt{N}} \qquad (2)$$

where X_{mean}, s and N represent the mean value, standard deviation and number of data, respectively (James *et al.* 2013). $Z_{(\alpha/2)}$ is the standardized threshold value, which depends on a selected significant level (α). An alternative approach based on median value was also considered as proposed by (Iglewicz and Martinez 1982, Hampel 1985) in Eq. (3)

$$Value_{CI} = (X_{med}) \pm median\{(|x_i - (X_{med})|,...,|x_N - (X_{med})|)\} \cdot g(N, \alpha_N) \qquad (3)$$

where X_{med} represents the median of the ordered samples $x_i \ldots x_n$ of variable X. $g(N, \alpha_N)$ shows the threshold factor (Hampel 1985) which depends on both the number of data (N) and the chosen significance level (α).

3.5 Mahalanobis square distance

Mahalanobis squared distance (MSD) is an alternative method for identifying outlier(s) by considering the distance between each sampled location with respect to their mean

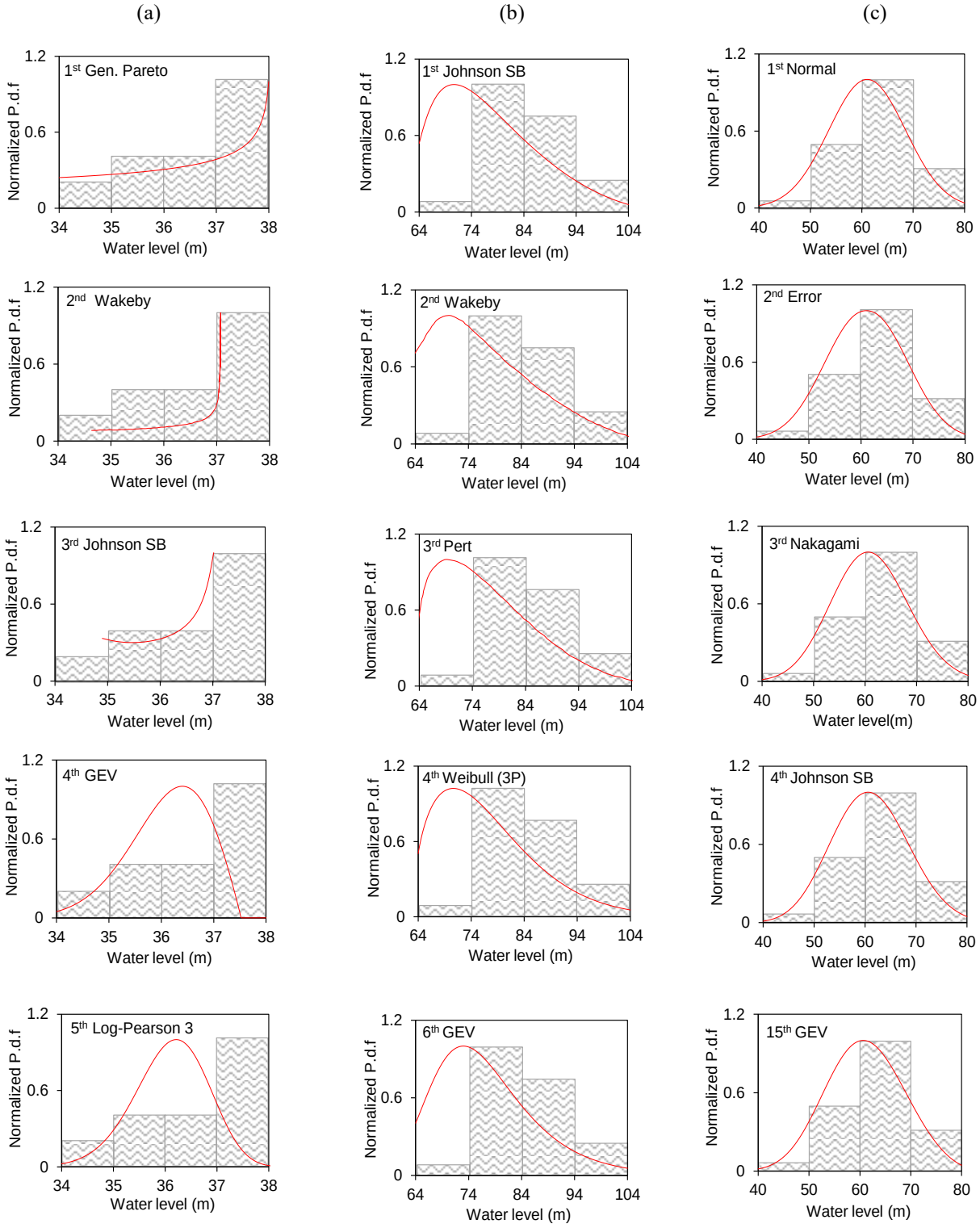

Fig. 2 Model distribution fits according to the Anderson-Darling ranking criterion: (a) Seoul, (b) Daejeon and (c) Suncheon

absolute deviation (MAD), and thus the method shows how far apart is the individual data from mean values (Taguchi and Jugulum 2022). Note that if MAD is 0, the selected data is distributed on the same position as the mean value. The mathematical equation for calculating MSD is addressed in Eq. (4)

$$Vaule_{CI} = (x_i - x_{mean})^T \sum^{-1} (x_i - x_{mean}) \quad (4)$$

where $(xi - x_{mean})^T$ denotes the transpose of the deviation from the mean and $\sum^{-1}$ represents unbiased covariance

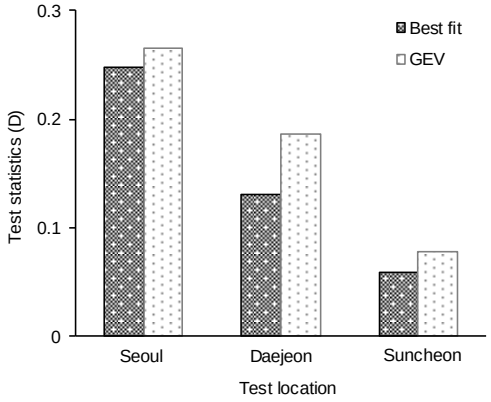

Fig. 3 A Bar-chat showing Comparison in test statistics between the best distribution and the GEV

matrix of the total number of sample N. MSD accounts for the variance of each data and the covariance between them through the covariance matrix ($\sum^{-1}$). In other words, it possesses a good measure of the relationship between the local neighbor data and the referenced (center) data distribution. Note that the distance calculated based on Eq. (4) is reduced to Euclidean distance when the covariance matrix conforms to identity matrix. Theoretically, the covariance matrix ($\sum^{-1}$) in MSD considers the relationship between the datasets as well as the spatial correlation between their nearest neighbor data to the referenced (center) data. This depicts its main difference from outlier labeling procedure where the datasets is only considered independently. However, masking and swamping effects also play a major role in acceptance of MSD as a reliable criterion for outlier analysis. Masking occurred when small cluster of outlying data attracted the mean xmean and increase $\sum^{-1}$ towards the direction of the mean, thus decreasing the MSD therefore, unable to detect outlying data as an outlier(s). This effect refers to a Type-II hypothesis error indicating failure in identifying an outlying data when truly it is an outlier. Contrarily, Type-I hypothesis error indicating identifying non-outlying data as an outlier when truly it is a real data. This equally defined a swamping effect where the MSD is increased as the mean xmean is attracted by small cluster of outlying data and increase $\sum^{-1}$ away from the direction of the mean, leading to an increased in MSD and thus, non-outlying data are detected as an outlier(s). However, Leys *et al*. (2013) recommended a replacement of the mean value in Eq. (4) by a median value because it is less attracted by outlier and sample size (De *et al.* 2000). Thus, a modified Mahalanobis square distance (MMSD) is depicted in Eq. (5) as

$$Vaule_{CI} = (x_i - x_{median})^T \sum^{-1} (x_i - x_{median}) \quad (5)$$

where $(xi - x_{median})^T$ denote the transposition of the deviation from the median, and $\sum^{-1}$ denotes the unbiased covariance matrix.

3.6 Performance comparison of models and proposed outlier methods

Generally, error-based or correlation-based procedures are used to measure the goodness-of-fits of models. The error-based comparison considered in this study is through calculated error ratio from a cross-validation technique involving five-fold procedure. The five-fold is selected due to the relative lower number of data in the test locations. The predicted data, with respect to each removed fold division were compared with the corresponding measured values to estimate the percentage error ratio in Eq. (6) as

$$\%Err = \left(\frac{|V_m - V_p|}{V_m} \right) \times 100 \quad (6)$$

where V_m and V_p represent the measured and predicted values respectively.

4. Results

4.1 Outlier analysis based on GOF

The outlier analysis can identify the biased data, which hinder the spatial continuity of the measured values; thus, they should be removed to generate a reasonable distribution curve. The test statistic according to the proposed distribution function is less than the corresponding critical value 1.9286 (0.05 significant level), thus it is regarded as the best distribution function. The groundwater level was digitized through each best model of the Generalized Pareto, Johnson SB and Normal distributions for Seoul, Daejeon and Suncheon, respectively. The test statistics based on AD ranking criteria are demonstrated in Fig. 3, and the respective values estimated by the GEV model are also plotted for comparison. The calculated test statistics of the best models show lower values than those estimated by the GEV model; Table 1 summarized the test statistics according to each model in detail. The Generalized Pareto, Johnson SB and Normal models show the highest correspondence with the measured values for Seoul, Daejeon and Suncheon locations, respectively. However, the GEV is ranked 4th, 6th and 15th as a low correlation in the Seoul, Daejeon and Suncheon test locations, respectively.

The 2-dimensional outlier labeling plots for the groundwater levels and error ratios are summarized in Figs. 4(a), 5(a) and 6(a) for Seoul, Daejeon and Suncheon regions, respectively. The outlier labeling plots based on the GEV model are also demonstrated in the same figures for comparison. A reliable confidence interval is necessary to identify the extreme values. Thus, the various confidence intervals are applied at 50%, 60%, 70%, 80%, 90%, and 99.97% (almost 100%) with 10% increments, and 6 further sub-increments are applied with 1% increments between 94% and 99.97%, including 94%, 95%, 96%, 97%, 98%, 99%, and 99.9%. According to the selected confidence interval (50-99.9%), the groundwater levels in Seoul, Daejeon and Suncheon were determined to be 35.19-37.01 m, 70.38-84.35 m and 56.19-65.77 m respectively, with the error ratios of 0.08-0.65%, 0.38-1.39% and 0.28-1.02% in the same locations. However, the groundwater levels based

Table 1 Ranking of the best fits according to the Anderson-Darling distribution test
*Boldly highlighted values show the test statistics for the best and the Generalized Extreme Value (GEV) fit with their respective rankings. Lower test statistics values depict a better distribution fit

Test locations	Model test	Anderson-Darling Distribution	
	Model fit	Ranking	Test statistics
Seoul	**Gen. Pareto**	**1**	**0.24726**
	Wakeby	2	0.24726
	Johnson SB	3	0.25446
	GEV	**4**	**0.26472**
	Log-Pearson 3	5	0.31792
Daejeon	**Johnson SB**	**1**	**0.12959**
	Wakeby	2	0.13409
	Pert	3	0.15895
	Weibull (3P)	4	0.16833
	GEV	**6**	**0.18673**
Suncheon	**Normal**	**1**	**0.09632**
	Error	2	0.09758
	Nakagami	3	0.0989
	Johnson SB	4	0.09903
	GEV	**15**	**0.10927**

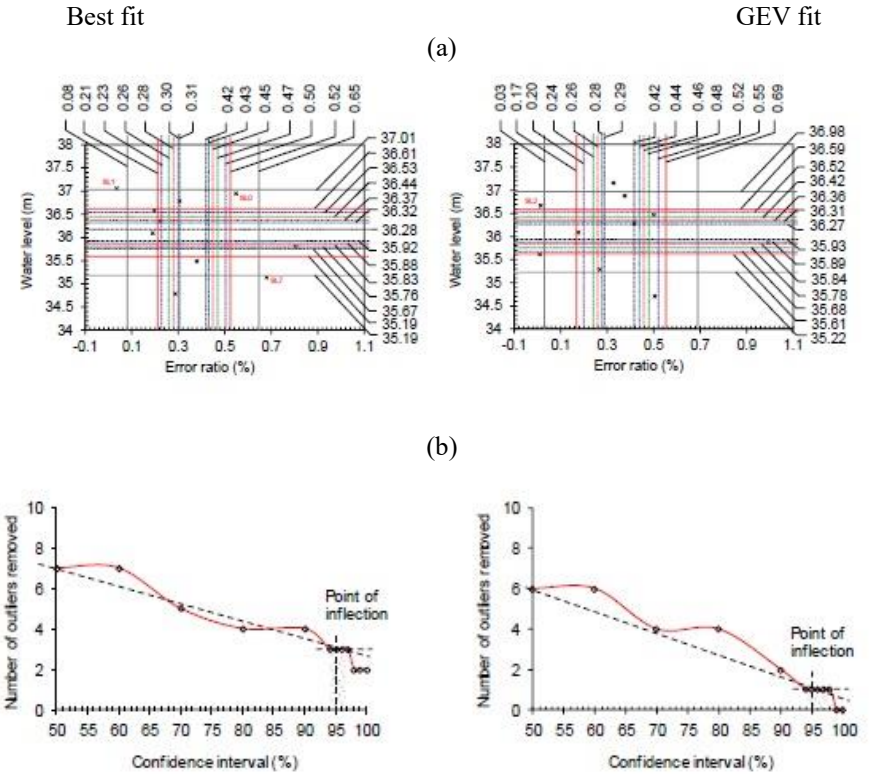

Fig. 4 Outlier analyzed results in Seoul: (a) Outliers labeling at 7 stages including 50%, 60% 70%, 80%, 90%, 95% and 99.97% of the confidence levels and (b) the number of outlier data based on various confidence levels

on the GEV model show 35.22-36.98 m, 70.30-81.40 m and 56.06-65.88 m in the Seoul, Daejeon and Suncheon test locations respectively, and the ranges are slightly different with those estimated by best models in the three test locations. The error ratios derived from the GEV model were 0.03-0.69%, 0.50-1.82% and 0.28-1.16%, respectively, in the same test locations, and they show a relatively wide range values.

The boundary lines according to each confidence interval estimated through Eq. (2) were aligned accordingly

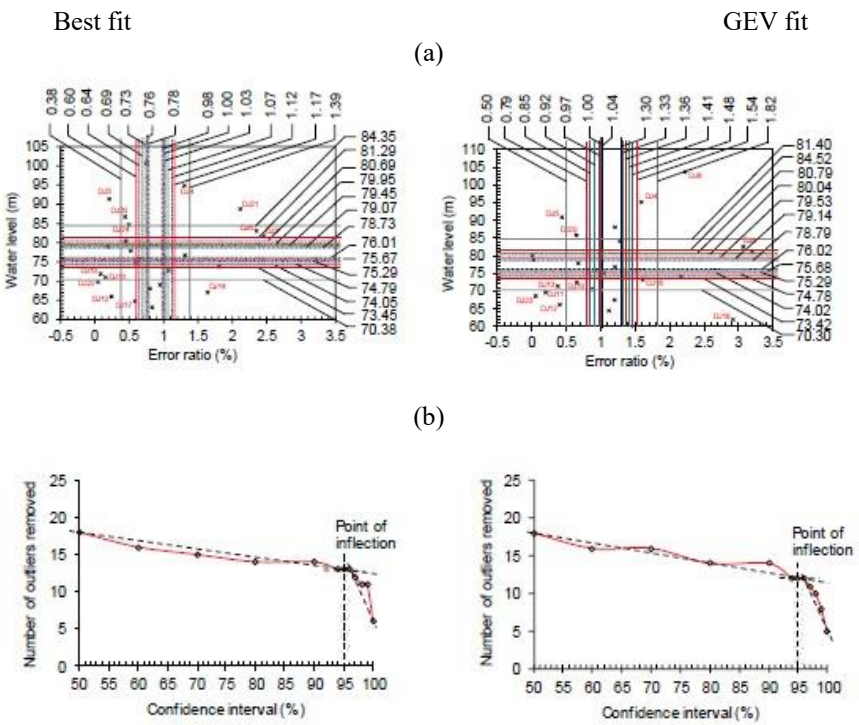

Fig. 5 Outlier analyzed results in Daejeon: (a) Outliers labeling at 7 stages including 50%, 60% 70%, 80%, 90%, 95% and 99.97% of the confidence levels and (b) the number of outlier data based on various confidence levels

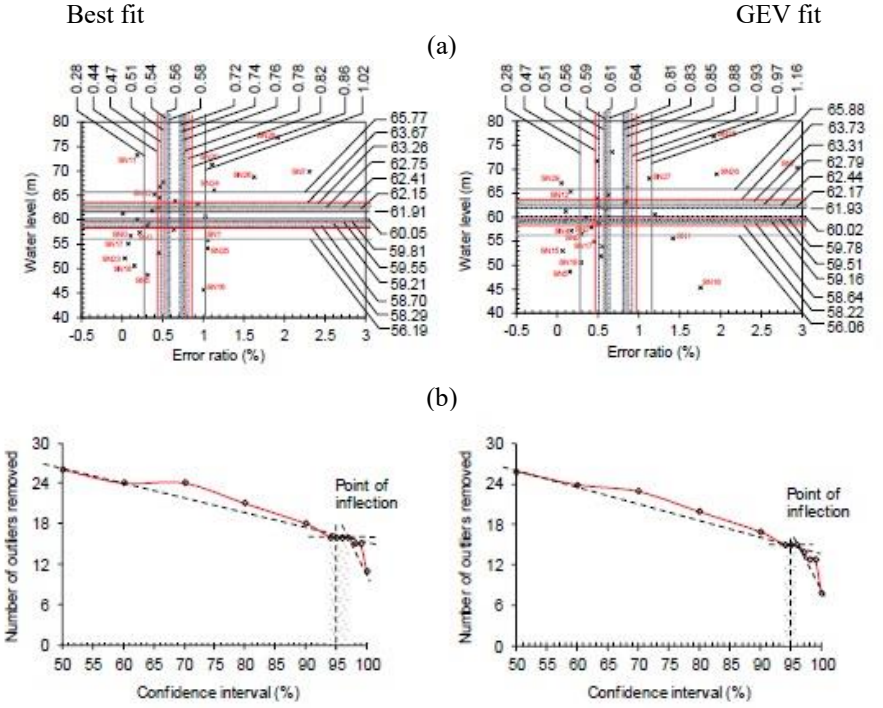

Fig. 6 Outlier analyzed results in Suncheon: (a) Outliers labeling at 7 stages including 50%, 60% 70%, 80%, 90%, 95% and 99.97% of the confidence levels and (b) the number of outlier data based on various confidence levels

with their intersection as shown in Figs. 4(a), 5(a) and 6(a).

The number of data removed as outlier for each confidence interval are represented in Figs. 4(b), 5(b) and 6(b); the number of outliers gradually increase as the confidence interval become wider. Thus, the number of data to be removed increases as the confidence interval is close to 50% indicating decrease in the reliability of the acceptance region. The figures show the inflection point where two different slopes meet near a confidence level of 95%, which means the 95% is a reasonable criterion to remove biased data because the number of removed values are almost constant after the 95% confidence interval. A

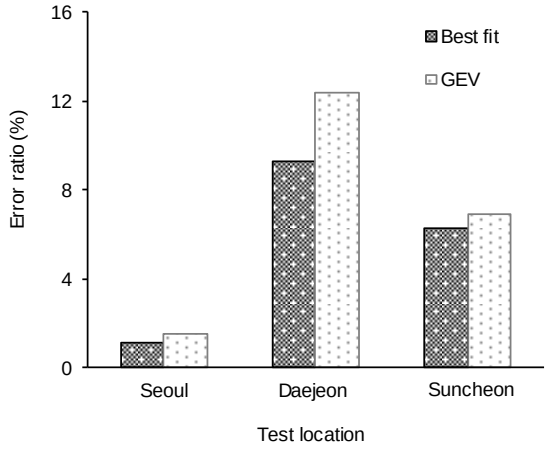

Fig. 7 A Bar-chat showing Comparison between Error ratios reduced after validation of the best function and Generalized Extreme Value function at 95% confidence level for Seoul, Daejeon and Suncheon

similar confidence interval was also demonstrated by Thode 2002. The 3, 13 and 16 data were identified as outliers based on best models at Seoul, Daejeon and Suncheon, respectively. However, the GEV method shows a different number of data with 1, 12 and 15 for Seoul, Daejeon and Suncheon locations, respectively.

To compare the reliability of the best models and the GEV, the calculated percentage error ratios were demonstrated in Fig. 7. The average values are 1.11%, 9.24% and 6.29% for best models, and 1.47%, 12.39% and 6.93% for the GEV distribution fits in the Seoul, Daejeon and Suncheon test locations, respectively. The error ratios based on the best model were much smaller than those calculated by the GEV which also corroborate with the test statistic values estimated for the models.

5. Discussion

The outlier analysis according to mean value (Eq. (2)) and median value (Eq. (3)) is affected by swamping (i.e., non-outlying data is identified as outlier) and masking (i.e., unable to identified all outlying data) respectively. These effects resulted from high variance between the groundwater level and error ratio. The effects also show that the detection of outliers in a bivariate (spatial) distribution by considering the variables might lead to high error and misleading interpretation as recommended by D'Agostino and Stephens 1986.

Nevertheless, Davies and Gather 1993 recommended that the outlier removal rate (i.e., the number of outliers/total number of data) should be less than or equal to 50% to produce an enhanced resolution map after outlier analysis. The calculated outlier removal rates based on the mean values were shown to be 30% (3/10), 52% (13/25) and 53% (16/30) for Seoul, Daejeon and Suncheon, respectively. In other words, the resolution of the generated maps at Daejeon and Suncheon may decrease after the outlier analysis due to the high removal rate. Moreover, the removed data tend to be concentrated in the local areas because the outlier analysis is considered solely on statistical basis and therefore, hindered the resolution of the generated map. Thus, the following problems should be considered to increase the map resolution: 1) the removal rate and 2) localized outlier data.

To overcome the weaknesses, the mean values are replaced with a median value to reflect the ground water levels and error ratios midpoints as recommended by Iglewics and Martinez 1982 in Eq. (2). Mean values is replaced by median value because the median value is immune to extreme values Hampel 1985. The mean and median values are estimated to be 36.09 and 36.22 for Seoul, 77.37 and 75.71 for Daejeon, and 60.98 and 60.98 for Suncheon, respectively. Note that the differences between the mean and median values show that the distributions have nonsymmetrical characteristics in the sites in Seoul and Daejeon, while the distributions in Suncheon demonstrate a symmetrical characterization due to the same mean and median values. The outlier analysis was performed based at the 95% confidence interval, according to Eqs. (2) and (3), and the result is plotted in Fig. 8 through the outlier labeling method. The number of outlier data estimated by the median value is zero and thus, the calculated outlier removal rate decreased to 0% in all regions. Even though, the method based on the median value, decreased the removal rate (swamping effect), it has no consideration for the spatial distribution of outlier data and thus, additional methods for the satisfaction of both the removal rate and localized outlier data is necessary.

In other to reflect the spatial distribution, the Mahalanobis distance (MD), which considered localized spatial data distribution is considered. In this study, both the mean and median values are also used as criteria. Note that, if the calculated MD values based on Eqs. (4) and (5) are 0, the selected data are located at the same position as the mean or median values. A two dimensional ellipses based on the Mahalanobis distance are demonstrated in Figure 8 with a confidence level of 95% and thus, the data outside the ellipse are regarded as outliers. The distances based on the mean and median from the center to the vertex are calculated as 0.92, 11.76 and 8.89 for Seoul, Daejeon and Suncheon, respectively. The distance of the co-vertices are also deduced to be 0.26, 0.84 and 0.61 for Seoul, Daejeon and Suncheon, respectively. Note that the calculated distances on the basis of the mean and median are the same, and the only shift of an ellipse occurred according to the reference value. The highest vertex value in Daejeon demonstrates the huge spatial variance, while Suncheon and Seoul show relatively low spatial data variance due to small values of the distances. The calculated outlier removal rates based on Eqs. (7) and (8) were equal to 0% (0/10) and 10% (3/30) for Seoul and Suncheon, respectively. However, the different removal rates are deduced to 24% (6/25) and 32% (8/25) based on the mean and median values, respectively, in Daejeon. The reason why the calculated removal rate in Daejeon is different between the mean and the median is because of the high dispersion in the measured values with the same results for the vertex and co-vertex values. Ultimately, the outlier removal rate is less than 50%, and

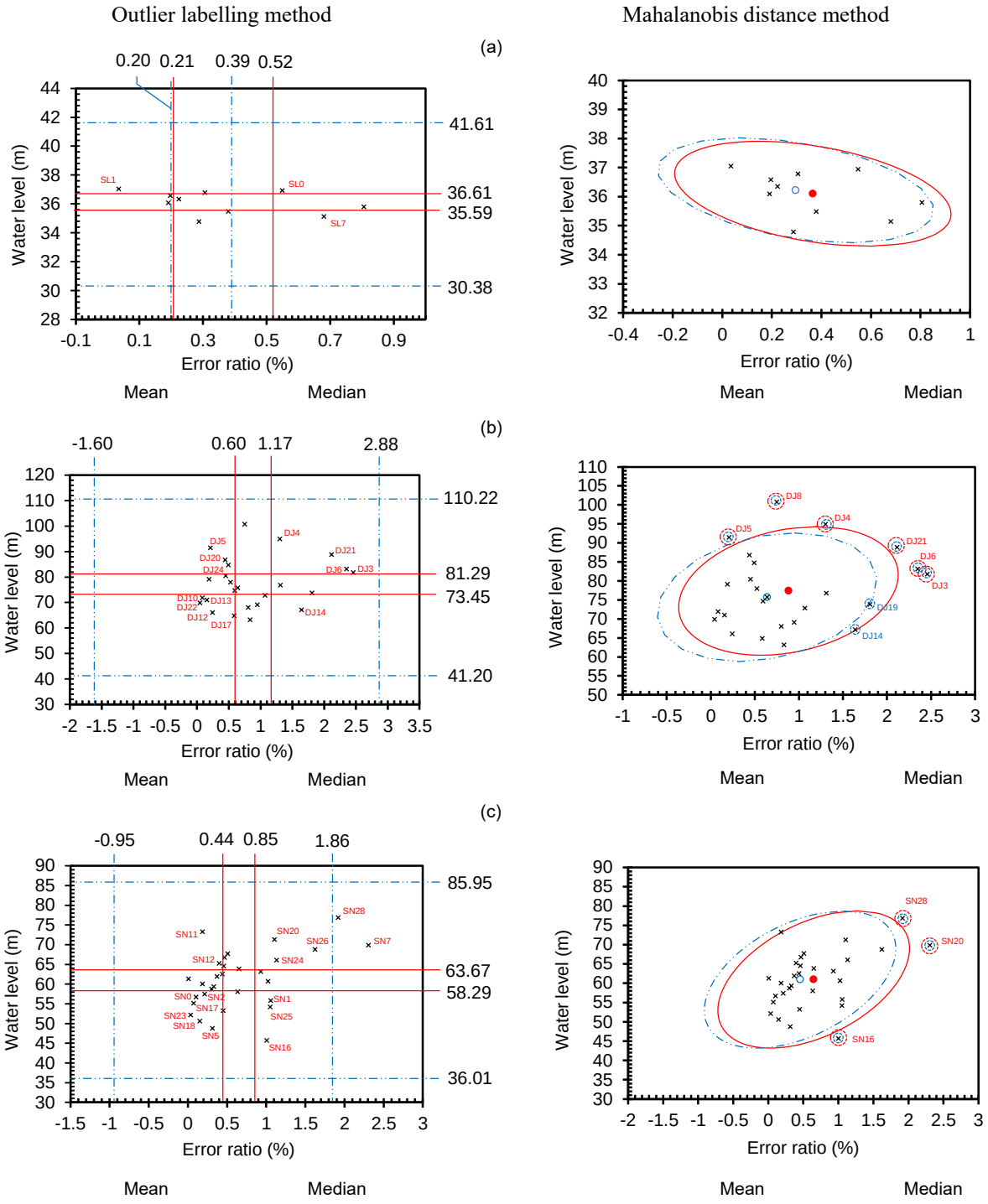

Fig. 8 Comparisons of outlier labeling and Mahalanobis distance methods at a 95% confidence level: (a) Seoul; (b) Daejeon and (c) Suncheon

thus, the swamping effect is highly minimized regarding the spatial distribution.

The error ratios based on cross-validation are calculated in Eq. (6), and the results are demonstrated in Fig. 9 to show the performance of the methods. The error ratios are calculated according to Eq. (6) and shown in Fig. 7. The average error ratios based on the mean values are calculated to be 1.11%, 9.24% and 6.29% in Seoul, Daejeon and Suncheon, respectively; however, the error ratios from the median values are deduced to be 1.09%, 12.76% and 6.83%, respectively, at the same locations. Additionally, both the mean and median values methods i.e. Eqs. (4) and (5) respectively, considering spatial distribution, show the same error ratios of 1.09%, 5.62% and 5.78% for the locations of Seoul, Daejeon and Suncheon, respectively. Although the error ratios estimated by considering spatial distribution are slightly reduced in the Seoul and Suncheon regions, the value is highly decreased in the Daejeon region

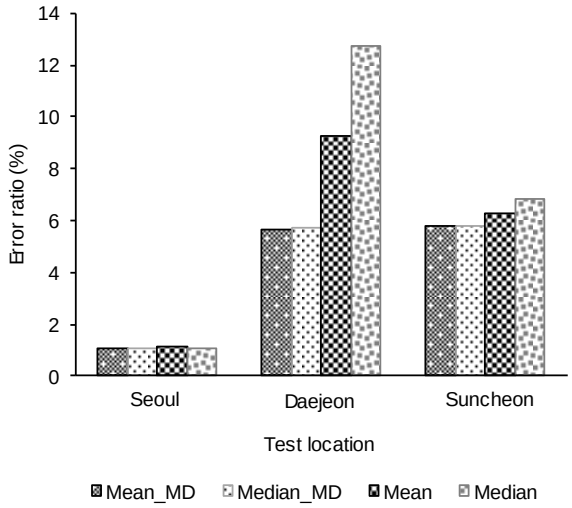

Fig. 9 A Bar-chat showing the performance of the methods for the test locations
Note: Outlier analysis according to Mean, Median, Mean and Median based on Mahalanobis distance is represented by Mean, Median, Mean_MD and Median_MD respectively

due to the scattered distributed data recorded in this test location. The results suggest that if there is minimal or no swamping effect, the mean value method in Eq. (2) i.e., outlier labeling method can be considered for outlier analysis as the method is less rigorous in computation. Otherwise, the Mahalanobis distance methods i.e., considered spatial data distribution with their location distances in Eqs. (4) and (5) should be considered to produce a reliable groundwater level prediction as it shown relatively lower prediction error across the three test locations.

6. Conclusions

The generalized extreme value (GEV) model has been reported as one of the prominent function to identify outlying values however, this study underlined the importance of identifying a suitable distribution function to estimate the measured values. In addition, a comparison of outlier analysis methods was studied and a reliable method was proposed. The detailed conclusions according to this study are illustrated as follows.

The 58 distribution models are fitted to find the best distribution of the groundwater levels, and the Generalized Pareto, Johnson SB and Normal distributions are shown to be the suitable model in the locations of Seoul, Daejeon and Suncheon, respectively. The error ratios calculated by the best model are smaller than those predicted by the GEV model.

The confidence intervals were computed between 50% - 99.97% to find a reasonable values, and the technique exhibited an inflection point at 95%. Thus, the confidence interval is determined to be 95%.

Outlier analysis is performed through mean and median values in the distribution function, and the Mahalanobis distance for the mean and median values are also applied to reflect spatial variability. Although, the method based on mean values is suggested due to its simplicity in computation however, the Mahalanobis distance method outperformed other methods as it shown a better result especially in areas where huge spatial variation exists between the measured values.

Acknowledgments

This research was supported by the Daejeon University Research Grants (2022).

References

Aalianvari, A., Soltani-Mohammadi, S. and Rahemi, Z. (2018), "Estimation of geomechanical parameters of tunnel route using geostatistical methods", *Geomech. Eng*., **14**(5), 453-466. https://doi.org/10.12989/gae.2018.14.5.453.

Ahmadi, S.H. and Sedghamiz, A. (2007), "Geostatistical analysis of spatial and temporal variations of groundwater level", *Environ. Monit. Assess*., **129**(1-3), 277-294. https://doi.org/10.1007/s10661-006-9361-z.

Anderson, T.W. and Darling, D.A. (1954), "A test of goodness of fit", *JASA*, **49**(268), 765-769. https://doi.org/10.1080/01621459.2022.2054816

Arreyndip, N.A. and Joseph, E. (2016), "Generalized extreme value distribution models for the assessment of seasonal wind energy potential of Debuncha. Cameroon", *Renew. Energy*, https:// doi.org/10.1155/2016/9357812.

Aziz, M., Khan, T.A. and Ahmed, T. (2017), "Spatial interpolation of geotechnical data: A case study for Multan City, Pakistan", *Geomech. Eng*., **13**(3), 475-488. https://doi.org/10.12989/gae.2017.13.3.475.

Benstock, D. and Cegla, F. (2017), "Extreme value analysis (EVA) of inspection data and its uncertainties", *NDT & E Int*., **87**, 68-77. https:// doi.org/10.1016/j.ndteint.2017.01.008.

Cheung, S.W., Spitznagel, E., Featherstone, T. and Crane, J.P. (1990), "Exclusion of chromosomal mosaicism in amniotic fluid cultures: efficacy of in situ versus flask techniques", *Prenatal diagnosis*, **10**(1), 41-57. https://doi.org/10.1002/pd.1970100108.

Chiles, J. and Delfiner, P. (1999), "Geostatistics. Modeling spatial uncertainty", John Wiley & Sons, 497. https://doi.org/10.1080/02664763.2012.750474.

Chowdhury, A., Jha, M.K. and Chowdary, V.M. (2010), "Delineation of groundwater recharge zones and identification of artificial recharge sites in West Medinipur district, West Bengal, using RS, GIS and MCDM techniques", *Environ. Earth Sci*., **59**(6), 1209. https://doi.org/10.1007/s12665-009-0110-9.

Chowdhury, J.U., Stedinger, J.R. and Lu, L.H. (1991), "Goodness-of-fit tests for regional generalized extreme value flood distributions". *Water Resour. Res*., **27**(7), 1765-1776. https://doi.org/10.1029/91WR00077.

Cressie, N. (1993), Statistics for spatial data. John Wiley & Sons. https://doi.org/10.1002/119115151.

D'Agostino, R.B. and Stephens, M.A. (1986), "Goodness of fit techniques", Marcel Dekker, https://doi.org/10.1201/9780203753064.

Damazio, J.M. and Kelman, J. (1986), "Use of historical data in flood-frequency analysis", In Hydrologic Frequency Modeling, 487-497. https://doi.org/10.1016/0022-1694(87)90150-8.

Davies, L. and Gather, U. (1993), "The identification of multiple outliers", *JASA*, **88**(423), 782-792.

https://doi.org/10.1080/01621459.1993.10476339.
De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. (2000), "The Mahalanobis distance", Chemometrics and intelligent laboratory systems, **50**(1), 1-18. https://doi.org/10.1016/s0169-7439(99)00047-7.
Fürst, J., Bichler, A. and Konecny, F. (2015), "Regional frequency analysis of extreme groundwater levels", *Groundwater*, **53**(3), 414-423. https://doi.org/10.1111/gwat.12223.
Gorgoso-Varela, J.J. and Rojo-Alboreca, A. (2014), "Use of Gumbel and Weibull functions to model extreme values of diameter distributions in forest stands", *Annal. Forest Sci*., **71**(7), 741-750. https://doi.org/10.1007/s13595-014-0369-1.
Grubbs, F.E. (1969), "Procedures for detecting outlying observations in samples", *Technometrics*, **11**(1), 1-21. https://doi.org/10.1080/00401706.1969.10490657.
Hampel, F.R. (1985), "The breakdown points of the mean combined with some rejection rules", *Technometrics*, **27**(2), 95-107. https://doi.org/10.1080/01621459.1981.10477698.
Hemeda, S. (2022), Geotechnical modelling and subsurface analysis of complex underground structures using PLAXIS 3D", *Int. J. Geo-Eng*., **13**(1), 9.
Hook, E.B. (1977), "Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use", *Am. J. Human Genetics*, **29**(1), 94. https://doi.org/PMID: 835578.
Hosking, J.R.M. and Wallis, J.R. (1997), "Regional frequency analysis: An approach based on L-moments", Cambridge University Press, 1233. https://doi.org/10.1017/CBO9780511529443.
Iglewicz, B. and Martinez, J. (1982), "Outlier detection using robust measures of scale", *J. Stat. Comput. Simul*., **15**(4), 285-293. https://doi.org/10.1080/00658208810595.
James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), "An introduction to statistical learning", Springer, **12**, 187-190.
Johnston, K., Ver Hoef, J.M., Krivoruchko, K. and Lucas, N. (2001), "Using ArcGIS geostatistical analyst", ESRI. Redlands, https://doi.org/10.1002/ep.10223.
Kitanidis, P.K. (1997), "Introduction to geostatistics: applications in hydrogeology", Cambridge University Press, https://doi.org/10.1007/978-1-4020-5729-8_4.
Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L. (2013), "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median", *J. Exp. Social Psychol*., **49**(4), 764-766. https://doi.org/10.1016/j.jesp.2013.03.013.
Marsaglia, G. and Marsaglia, J. (2004), "Evaluating the Anderson-darling distribution", *J. Stat. Software*, **9**(2), 1-5. https://doi.org/10.18637/jss.v009.i02.
Olabode, O.P. and San, L.H. (2023), "Analysis of soil electrical resistivity and hydraulic conductivity relationship for characterisation of lithology inducing slope instability in residual soil", *Int. J. Geo-Eng*., **14**(1), 7.
Parry, S., Baynes, F.J., Culshaw, M.G., Eggers, M., Keaton, J.F., Lentfer, K., Novotny, J. and Paul, D. (2014), "Engineering geological models: an introduction: IAEG commission 25", *Bull. Eng. Geol. Environ*., **73**(3), 689-706. https://doi.org/10.1007/s10064-014-0576-x.
Rajabian, A. (2023), "Effect of initial failure geometry on the progress of a retrogressive seepage-induced landslide", *Int. J. Geo-Eng*., **14**(1), 11.
Saidi, S., Bouri, S. and Dhia, H.B. (2010), "Groundwater vulnerability and risk mapping of the Hajeb-jelma aquifer (Central Tunisia) using a GIS-based DRASTIC model", *Environ. Earth Sci*., **59**(7), 1579-1588. https://doi.org/10.1007/s12665-009-0143-0.
Sari, M. (2021), "Determination of representative elementary volume (REV) for jointed rock masses exhibiting scale-dependent behavior: a numerical investigation", *Int. J. Geo-Eng*., 12(1), 34.
Shepard, D. (1968), "A two-dimensional interpolation function for irregularly spaced data", *Proceedings of the 1968 23rd ACM National Conference. New York*, NY, USA.
Shirazi, S.M., Imran, H.M., Akib, S., Yusop, Z. and Harun, Z.B. (2013), "Groundwater vulnerability assessment in the Melaka State of Malaysia using DRASTIC and GIS techniques", *Environ. Earth Sci.*, **70**(5), 2293-2304. https://doi.org/10.1007/s12665-013-2360-9
Stephens, M.A. (1986), "Tests based on EDF statistics. Goodness-of-fit techniques", **68**, 97-193. https://doi.org/10.1016/0304-3800(93)E0074-D.
Taguchi, G. and Jugulum, R. (2002), "The Mahalanobis-Taguchi strategy: A pattern technology system", John Wiley & Sons. https://doi.org/10.1002/9780470172247.
Taiwo, S. M. and Yoon, H. K. (2018), "Estimation of elastic wave velocity and DCPI distributions using outlier analysis", *Eng. Geol*., **247**, 129-144. https://doi.org/10.1016/j.enggeo.2018.10.027.
Thode, H.C. (2002), "Testing for normality" 164, CRC press, https://doi.org/10.2307/2332434.
Tian, M., Li, D.Q., Cao, Z.J., Phoon, K.K. and Wang, Y. (2016), "Bayesian identification of random field model using indirect test data", *Eng. Geol*., 210, 197-211. https://doi.org/ 10.1016/j.enggeo.2016.05.013.
Wang, C., Chuai, X., Shi, F., Gao, A. and Bao, T. (2018), "Experimental investigation of predicting rockburst using Bayesian model", *Geomech. Eng*., **15**(6), 1153-1160. https://doi.org/12989.2018/gae.15.6.1153.
Wang, J.B., Liu, X.R., Huang, Y.X. and Zhang, X.C. (2015), "Prediction model of surface subsidence for salt rock storage based on logistic function", *Geomech. Eng*., **9**(1). 25-37. https://doi.org/ 10.12989 / gae.2015.9.1.025.
Wang, X., Wang, H., Liang, R.Y. and Liu, Y. (2019), "A semi-supervised clustering-based approach for stratification identification using borehole and cone penetration test data", *Eng. Geol*., **248**, 102-116. https://doi.org/10.1016/j.enggeo.2018.11.014.
Yasser, F., Altahrany, A. and Elmeligy, M. (2022), "Numerical investigation of the settlement behavior of hybrid system of floating stone columns and granular mattress in soft clay soil", *Int. J. Geo-Eng*., **13**(1), 12.
Yoon, S., Lee, S.R., Kim, Y.T. and Go, G.H. (2015), "Estimation of saturated hydraulic conductivity of Korean weathered granite soils using a regression analysis", *Geomech. Eng*., **9**(1), 101-113. https://doi.org/10.12989/gae.2015.9.1.101.
Zalina, M.D., Desa, M.N.M., Nguyen, V.T.A. and Kassim, A.H. M. (2002), "Selecting a probability distribution for extreme rainfall series in Malaysia", *Water Sci. Technol*., **45**(2), 63-68. https://doi.org/10.2166/wst.2002.0028.

IC