scientific reports

OPEN

Assessing heterogeneous groundwater systems: Geostatistical interpretation of well logging data for estimating essential hydrogeological parameters

Musaab A. A. Mohammed[1,2]✉, Yetzabbel G. Flores[1], Norbert P. Szabó[1] & Péter Szűcs[1]

This research presents an unsupervised learning approach for interpreting well-log data to characterize the hydrostratigraphical units within the Quaternary aquifer system in Debrecen area, Eastern Hungary. The study applied factor analysis (FA) to extract factor logs from spontaneous potential (SP), natural gamma ray (NGR), and resistivity (RS) logs and correlate it to the petrophysical and hydrogeological parameters of shale volume and hydraulic conductivity. This research indicated a significant exponential relationship between the shale volume and the scaled first factor derived through factor analysis. As a result, a universal FA-based equation for shale volume estimation is derived that shows a close agreement with the deterministic shale volume estimation. Furthermore, the first scaled factor is correlated to the decimal logarithm of hydraulic conductivity estimated with the Csókás method. Csókás method is modified from the Kozeny-Carman equation that continuously estimates the hydraulic conductivity. FA and Csókás method-based estimations showed high similarity with a correlation coefficient of 0.84. The use of factor analysis provided a new strategy for geophysical well-logs interpretation that bridges the gap between traditional and data-driven machine learning techniques. This approach is beneficial in characterizing heterogeneous aquifer systems for successful groundwater resource development.

Keywords Debrecen, Pannonain basin, Factor analysis, Csókás method, Hydraulic conductivity

The characterization of heterogeneous groundwater aquifers is one of the challenges that requires the integration of advanced geological and geophysical techniques to address the inherent complexities[1,2]. This spatial heterogeneity complicates the ability to accurately estimate parameters such as hydraulic conductivity and the contaminants distribution[3–6]. Traditional geological and hydrogeological methods can provide valuable insights, but they often fall short of capturing the full lithological variability within the aquifers[7–9]. On the other hand, Geophysical well logging offers a unique opportunity to attain a more comprehensive understanding of the aquifer system as it gives a continuous estimation of the aquifer characteristics[10–13]. The petrophysical and hydraulic parameters can be obtained by analyzing well-logging data using deterministic and inverse modeling[14–17]. However, formulating the mathematical problem that relates the hydrogeological parameters to the measured geophysical data is often complex and associated with a high uncertainty[18,19].

The challenging and ill-posed nature of the hydrogeophysical inverse problem has limited the application of inversion-based models to estimate hydrogeological parameters[20,21]. One critical parameter in hydrogeology, hydraulic conductivity, presents particular difficulties due to its nonlinearity to other petrophysical and fluid properties, making its accurate prediction from geophysical data problematic[22,23]. Consequently, groundwater researchers often resort to pumping experiments[24], which are costly and time-consuming to quantify hydraulic

[1]Faculty of Earth and Environmental Science and Engineering, University of Miskolc, Egyetemváros, Miskolc 3515, Hungary. [2]College of Petroleum Geology and Minerals, University of Bahri, Khartoum, Sudan. ✉email: musab20501@gmail.com

conductivity[25]. In this context, Csókás[26] introduced an improved methodology for estimating hydraulic conductivity in loosely consolidated hydrogeological units, exclusively relying on well-logging data. The successful application of this method necessitates the interpretation of geophysical logs sensitive to lithology and water saturation[27]. The main advantage of well logging-based methods is their ability to give continuous profile estimation for petrophysical and hydrogeological parameters crucial to simulating and understanding the hydrodynamic conditions of the heterogeneous groundwater systems[28,29].

Recent advancements in machine learning have opened up new possibilities for interpreting well-logging data[30–34]. Unsupervised learning methods offer efficient insights into petrophysical and hydrogeological characteristics. Among them, factor analysis, a powerful multivariate statistical approach, allows to reveal the complex interdependencies within well-logging data[35,36] and can be used to extract details from multidimensional datasets that are not immediately observable[37]. Several works on using factor analysis for the interpretation of well logs data are reported in the literature[38,39]. For instance, Li et al.[40] applied factor analysis for the characterization of gas-bearing formation in Sichuan Basin, China, and proved its efficiency in formation evaluation compared to the other conventional methods. A study conducted by Asfahani[41] indicated the effectiveness of factor analysis for lithology identification. Puskarczyk[42] identified the finer-scale variation of the lithofacies in a shale formation within the Carpathian Basin using principal component analysis (PCA). The study indicated that PCA can be successfully employed in mapping the gas-saturated and sandstone-claystone formations.

Multivariate statistical and inverse modeling approaches are mainly applied for characterizing gas and oil reservoirs. While gas and oil reservoirs have been extensively investigated, the heterogeneous groundwater aquifers present a more challenging environment due to their complex lithological nature[18]. The primary objective of this study is to use the exploratory factor analysis integrated with the Csókás method for characterization of the Quaternary aquifer in the Debrecen area. The present research stands out for its innovative interpretation of the well-logging data that bridges the gap between conventional and unsupervised learning data-driven techniques.

Study area
Geography
The research site is situated around the Debrecen area, Eastern Hungary, encompassing approximately 650 km² (Fig. 1). It is integral to the Great Hungarian Plain (GHP) in which substantial variations in land elevation have transpired due to contemporary tectonic movements, erosion, and extensive sedimentation processes[43]. The geological movements have notably influenced the topography in the study area, leading to an elevation ranging from 88 to 160 m above sea level (a.s.l). The region's climate can be characterized as predominantly continental, with annual mean temperatures ranging from 10° to 11 °C. The annual precipitation varies from 550 to 600 mm, and potential evapotranspiration ranges between 600 and 700 mm/year[44].
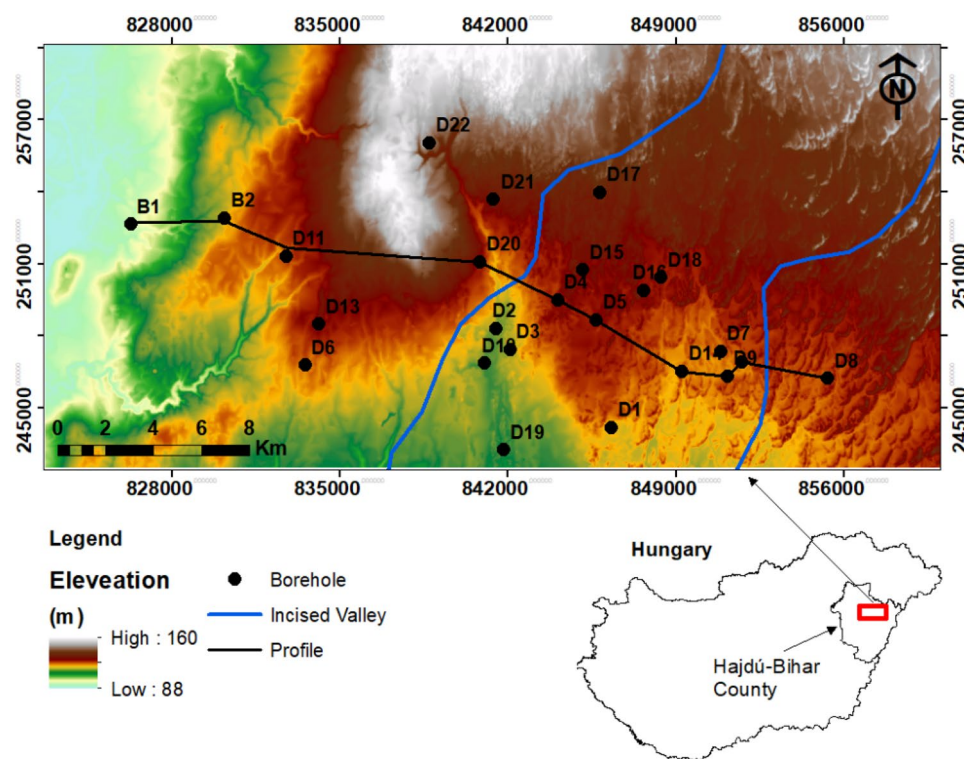
Figure 1. Geographic map created with ArcGIS Desktop v. 10.8[45], showing the location of the study area within the Eastern part of Hungary.
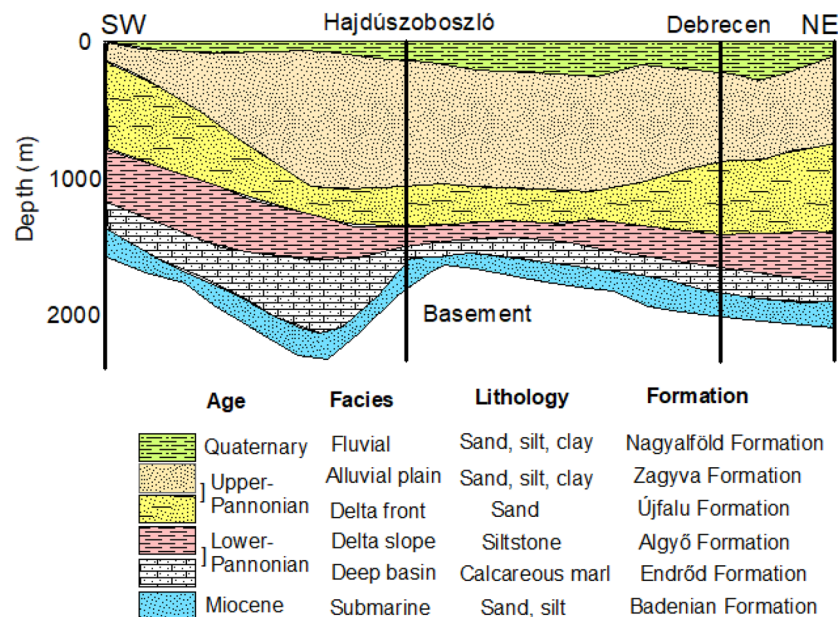
Geology

The research consists of diverse geological formations including Mesozoic basement rocks, Miocene deposits, Pannonian layers, and Quaternary Formation (Fig. 2). The Mesozoic rocks are composed of metamorphic and igneous rocks, and they are primarily associated with the Tisza Mega-Unit[46,47]. These rocks encompass a variety of rock types, such as granites, gabbros, and basalts, alongside schists and phyllites[48]. The Miocene Formation is characterized by an assortment of sedimentary rocks, encompassing marl, sandstones, and claystone[49]. The Pannonian sediments are classified into two distinct parts, namely the Lower and Upper Pannonian[47]. During the early stages of the Lower Pannonian period, the initially deposited coarse-grained sandstone and coastal sandy conglomerates underwent lateral transformations into siltstone, known as the Algyő Formation. Simultaneously, there was the development of calcareous marl and limestone, referred to as the Endrőd Formation[49]. Conversely, the Upper Pannonian era comprises a succession of sedimentary layers, encompassing sandy delta plain and delta front sediments, interspersed with alluvial siltstone, sandstone, clay, marl, and quartz pebbles. These particular deposits are observed within the Újfalu and Zagyva formations[50]. The surface of the GHP predominantly consists of Quaternary deposits. These deposits encompass fluvial sediments, river sediments, and sandy loess. The thickness of Quaternary deposits in the research area varies from 80 to 150 m. These deposits are categorized into three segments: upper, middle, and lower Pleistocene beds[47]. The lower and upper sections predominantly comprise river and overbank sediments, while the middle section predominantly encompasses coarse-grained fluviolacustrine sediments[51].

Hydrogeology

In the Great Hungarian Plain (GHP), five hydrostratigraphic units were identified based on their lithology and chronostratigraphy. These units are the Pre-Neogene impermeable layer, the Pre-Pannonian aquifer, the Endrőd confining layer, the Algyő confining layer, and the Nagyalföld water-bearing stratum[49]. The Nagyalföld Aquifer, which encompasses the Újfalu and Zagyva Formations along with Quaternary sediments, has been recognized as the main aquifer with a permeability exceeding 1000 mD[49,53].

Recently, Flores et al.[44] conducted an extensive regional-level hydrostratigraphical investigation, concentrating on the upper section of the Nagyalföld aquifer. Their findings revealed that the key hydrostratigraphic components in their study encompass the Pre-Quaternary and Quaternary sequences (Fig. 3). The Pre-Quaternary sequence of the Late Miocene is distinguished by substantial layers of silt with occasional intercalated fine sand. In contrast, the Quaternary sequence is characterized by three hydrostratigraphic divisions, ordered from older to younger. The first is an incised valley unit, described as an elongated body of sand and gravel with minimal clay content. Above it, the alluvial unit is depicted as a succession of three consecutive sand bodies with significant horizontal variability and deposits of silty clay. Finally, the coarsening upward unit is described as a sequence displaying pronounced heterogeneity, featuring clay, silt, and sand bodies. The observations have unveiled the existence of two distinct hydraulic systems in the study area. In the upper system, groundwater flow is predominantly governed by gravitational forces, while the lower system experiences overpressure[10]. Hydraulic interaction between these two systems frequently occurs, particularly in areas where low-permeability layers exert outward pressure[52].

Figure 2. Geological cross-section showing the main lithological formations in the study area modified after Juhász[52] and Tóth and Almási[49].
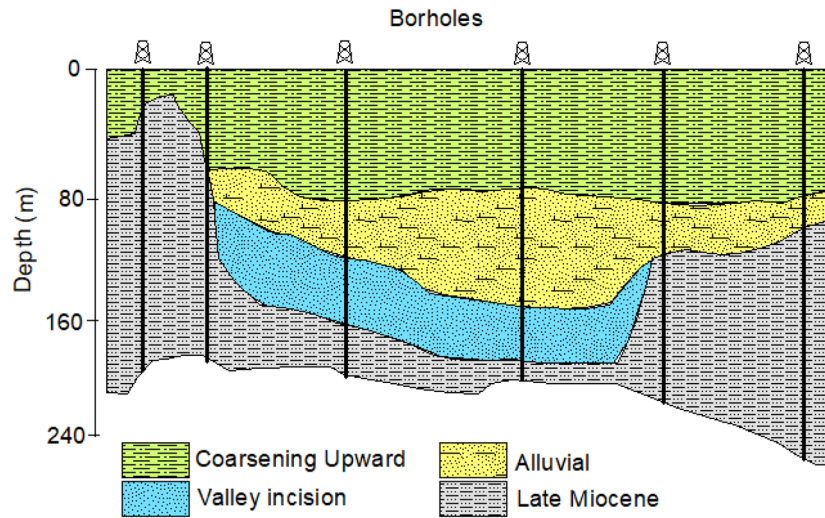
Figure 3. The main hydrostratigraphical units within the Nagyalföld aquifer, modified after Flores et al.[44].

Materials and methods

This study used geophysical well logging data to identify and characterize groundwater aquifers in the Eastern Hungary region surrounding Debrecen. In this work, the aquifer geometry and the petrophysical and hydrogeological parameters of the Quaternary aquifers in the study region are defined utilizing data collected from twenty-four (24) boreholes. This study employed three well logs including spontaneous potential (SP), natural gamma ray (NGR), and deep normal resistivity (RS), and analyzed using Csókás method and factor analysis.

Csókás approach

Csókás[26] model is used for estimating hydraulic conductivity from the well logs data. This method can be seen as an empirically refined version of the equations proposed by Kozeny[54] and Carman[55]. The Kozeny-Carman equation takes into account several key parameters, such as the density of water (ρ_w), viscosity (μ), porosity (φ), the dominant grain size of the aquifer materials (d), and the acceleration due to gravity (g). The Kozeny-Carman-based hydraulic conductivity (K_{KC}) can be estimated using Eq. (1).

$$K_{KC} = \frac{\rho_w\}}{\mu}\frac{d^2}{180}\frac{\varphi^3}{(1-\varphi)^2}. \tag{1}$$

Csókás approach proves to be particularly applicable in situations involving lossy geological formations. This suitability is established through an empirical connection between the effective grain size of water-saturated sediments (d_{10}) and the formation factor ($F = \frac{R_0}{R_w}$) (Eq. 2). Alger[56] investigation revealed that, apart from the porosity (φ), the resistivity of water (R_w) also exerts an influence on the formation factor. In this research, the effective porosity is estimated using Eq. (3)[57], considering the shale volume (V_{sh}) present in the geological formation. The shale volume however, is estimated using Larianov[58] equation (Eqs. 4 and 5). Consequently, the hydraulic conductivity (K, m/s), calculated using the Csókás method can be determined using Eq. (6).

$$d_{10} = C_d \log F, \tag{2}$$

$$\varphi_e = \varphi \times (1 - V_{sh}), \tag{3}$$

$$V_{sh} = 0.33\left(2^{2*I_\gamma} - 1\right), \tag{4}$$

$$I_\gamma = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}}, \tag{5}$$

$$K = C_k \frac{\varphi}{(1-\varphi)^4}\frac{\left(log\frac{R_0}{R_w}\right)^2}{\left(\frac{R_0}{R_w}\varphi\right)^{1.2}}, \tag{6}$$

where φ_e is the effective porosity, I_γ is the gamma-ray intensity, which is calculated using a linear formula that uses the gamma-ray response of the log (GR_{log}), minimum (GR_{min}) and the maximum (GR_{max}) gamma-ray. C_k is the proportionality constant and has the value 855.7 * 5.22 * 10⁻⁴.

Exploratory factor analysis (FA)

Factor analysis is an unsupervised machine learning method that facilitates the reduction of complex datasets into a more manageable set of factors. In this study, factor analysis was employed to extract factor logs representing the largest portions of variance within the dataset from the analysis of the available well logs of SP, NGR, and RS[59]. These factor logs are then linked to shale volume estimated using the Larionov[58] equation and hydraulic conductivity determined by the Csókás[26] method. The correlation of factor logs with these parameters aids in developing site-specific equations that facilitate direct connections between the factor log and aquifer parameters that can be used as alternatives to the existing methods.

During the initial stages, standardization of well logs was necessary, given the use of different probes and, consequently, varying measurement units (Eq. 7), followed by the integration of data into a matrix (D) (Eq. 8), and the application of a factor analysis model (Eq. 9).

$$\check{D}_{il} = \frac{\left(D_{il} - \overline{D}_l\right)}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(D_{il} - \overline{D}_l\right)^2}},$$

(7)

$$D = \begin{bmatrix} SP_1 & GR_1 & RS_1 \\ SP_2 & GR_2 & RS_2 \\ \vdots & \vdots & \vdots \\ SP_i & GR_i & RS_i \\ \vdots & \vdots & \vdots \\ SP_n & GR_n & RS_n \end{bmatrix},$$

(8)

$$\hat{D} = FW^T + E.$$

(9)

In this context, $\hat{D}_{il}$ represents the scaled data for the n-th observation within the l-th well log. $\overline{D}_l$ corresponds to the average value of the unprocessed data from the l-th well log, where L is the total number of borehole geophysical tools, and N is the count of measuring points in the specified depth range. F is the factor score matrix of dimensions N by M, where M is the number of extracted factors, W is the factor loading matrix of dimensions L by M. E is the matrix of residuals with dimensions N by L, and T represents the matrix transpose operator.

The primary factor explains the majority of the variation in the dataset, while the subsequent factors contribute to a relatively smaller portion of the variance. The factor loading matrix, which measures the degree of association between the factors and the actual data, offers precise weights for each data category. Because the factors are statistically uncorrelated, the correlation matrix of the observed data can be indicated using Eq. (10) as

$$R = \frac{1}{N} D^T D = L L^T + \Psi.$$

(10)

In this context, Ψ represents a diagonal matrix containing specific variances. When Ψ takes on a value of 0, the issue can be resolved through the solution of an eigenvalue problem. If Ψ differs from 0, the factor scores are determined using the maximum likelihood method, and the subsequent objective function is then optimized to simultaneously estimate both L and Ψ[60] (Eq. 11).

$$\Omega(L, \Psi) = tr(R, LL^T - \Psi)^2 = min.$$

(11)

Factor loadings are usually subjected to an orthogonal transformation to enhance the interpretability of factors, as proposed by Ref.[37]. In this study, factor rotation was carried out using the varimax technique, following Kaiser[61] approach. Factor scores can be derived by applying a linear approach with the assumption of linearity[62] (Eq. 12).

$$F^T = (L^T \Psi^{-1} L)^{-1} L^T \Psi^{-1} D^T.$$

(12)

The Pearson[63] (R) and Spearman[64] (ρ) correlation coefficients are utilized to assess the relationships between the extracted factor logs, well logs, and petrophysical and hydrogeological parameters. Pearson correlation coefficient evaluates the strength and direction of the linear relationship between the continuous variables while Spearman rank correlation coefficient measures the strength and direction of the monotonic relationship. Both coefficients range from −1 to 1, with 1 indicating a perfect positive relationship, 0 indicating no relationship, and −1 indicating a perfect negative relationship. These coefficients provided simple sensitivity analysis to evaluate the associations between well logs and the extracted factor logs.

Results

This research introduces factor analysis for the interpretation of well logs for estimation of shale volume (V_{sh}), effective porosity (φ_e), and hydraulic conductivity (K) of the Quaternary aquifers in the Debrecen area. The data is analyzed in 1D along the boreholes, and the obtained results are interpolated in 2D along a profile. The distribution of the borehole along the profile is illustrated in Fig. 4, with the stratigraphic bounding surfaces described by Flores et al.[44]. These surfaces are created with the geometrical convergence interpolation[65] of the identified well tops following the sequence stratigraphical principles[66]. The hydrostratigraphic units in the area from the bottom to top are the Late Miocene, incised valley, alluvial, and coarsening upward units (Fig. 4). The Late Miocene
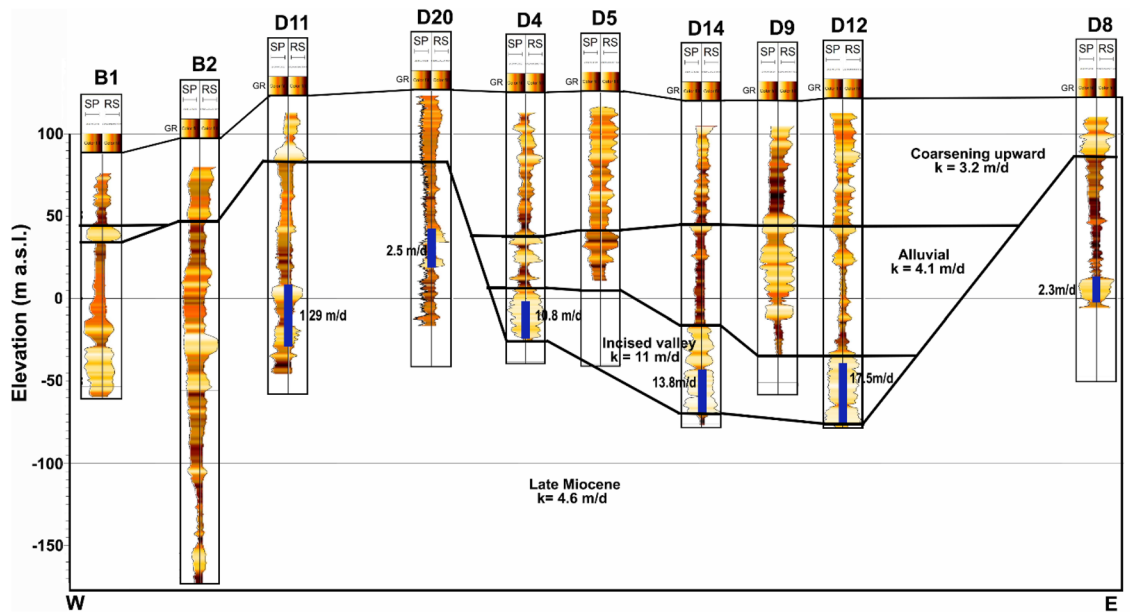
Figure 4. 2D geophysical profile showing the pattern of the SP, GR, and SP logs and the distribution of the main hydrostratigraphical units in the study area.

unit is characterized by low occurrence of silty sand lithologies embedded in thick silty clay sequences while the incised valley unit is dominated by a thick sequence of gravel and sand deposits. Over them, the alluvial unit is characterized by the occurrence of two sandy channel deposits, embedded into a thick clayed floodplain deposit. The coarsening upward unit is characterized by coarsening upward facies that are made up of a successive intercalation between clay, silt, and sand. Several aquifer units are developed within these hydrostratigraphical units with the incised valley deposits hosting the main aquifer in the study area[51].

FA-based shale volume

The well logging data comprises a total of 34,328 data points along 24 boreholes and is divided into two parts in which 60% of the data is used for correlation and 40% of the data for testing the resulting relationship. The first factor explained 81.7% of the total variance, indicating its robust representation of underlying features in the dataset. A higher positive loading is given to NGR (0.70) and medium negative loading to RS (− 0.57).

The scores of the first factor of the 60% of the data are correlated to shale volume estimated from the Larionov equation and yielded a strong exponential relationship with a Spearman correlation coefficient of 0.91 (Fig. 5a). This relationship underscores the importance of the first factor as a powerful proxy for shale volume[59]. Accordingly, a site-specific equation is obtained that linked shale volume (V_{sh}) to the scaled first factor (F_1) and written as
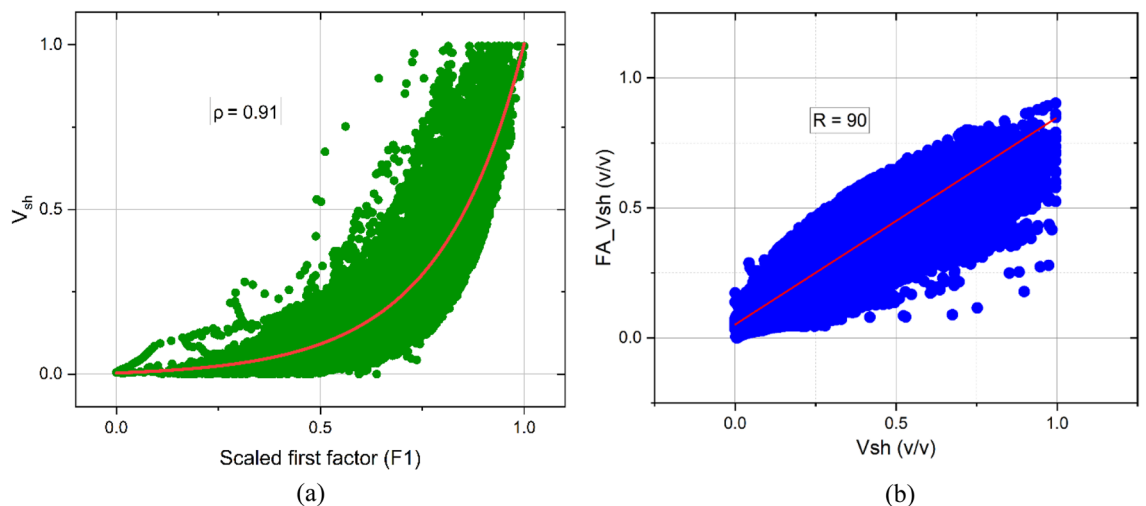
Figure 5. (**a**) The relationship between the scaled first factor and shale volume and (**b**) the correlation between the factor analysis-based and Larionov shale volume.

$$V_{sh} = ae^{bF_1},$$

where a and b are site-specific constants from the local regression that is given with 95% confidence. The average values of a = 0.0153 [0.0067, 0.281] and b = 4.2276 [3.736, 5.2244]. To evaluate the practical utility of the relationship between the first factor and shale volume, 40% of the data is used. Accordingly, the correlation between shale volume obtained from factor analysis and Larionov[58] method is illustrated in Fig. 5b. The promising results obtained from this validation process, where the correlation coefficient reached 0.90, underscores the applicability of the factor analysis-based shale volume estimation.

Based on the obtained relationships, the FA-based shale volume is estimated in 1D (Fig. 6) and 2D along the profile (Fig. 7). The 2D spatial variation of the Larionov[58] equation-based shale volume (Fig. 7a) is compared to FA-shale volume (Fig. 7b). The comparison between the two approaches showed a close agreement. The descriptive statistics of the FA-based shale volume are illustrated in Fig. 8. The shale volume values are then compared to lithofacies proportion calculations based on the analysis of the well-logs data assuming 2 m layer thickness (Fig. 9). The FA-based shale volume of the coarsening upward unit exhibited significant variability, ranging from 0.05 to 50%, with a mean value of 20%. The lithofacies proportion (Fig. 9a) indicated that this unit consists of 37.7% clay, 42.5% silt, and 19.8% sand. The alluvial unit displayed almost similar variability in shale volume, ranging from 0.07 to 72%, with a mean of 34%. This unit consists of 41.9% clay, 26.9% silt, and 3.2% sand (Fig. 9b). The valley incision unit exhibited a relatively uniform distribution of shale volume, varying from almost zero (0.5%) to 9%. Consequently, the facies analysis indicated that this unit is composed of 78.7% sand (Fig. 9c). The Late Miocene unit displayed shale volume variations from 0.05 to 77%, with a mean value of 26%. This unit is dominated by clay and silt layers that make up more than 80% of the unit (Fig. 9d).

Effective porosity

The effective porosity is essential for assessing the rate of groundwater flow within the aquifer[67]. In this study, the FA-based shale volume is substituted into Schlumberger[57] formula for a more practical estimation of effective porosity. The obtained parameters from the FA approach are compared to those of conventional approaches and showed a close agreement with a 0.93 correlation coefficient (Fig. 10). Figure 11 shows the 2D interpolation of the effective porosity obtained from the empirical method (Fig. 11a) and factor analysis (Fig. 11b), in which a close agreement between the two approaches is indicated. As a result, the obtained effective porosity for the hydrostratigraphical units is illustrated using a Box plot (Fig. 12). The effective porosity of the coarsening upward unit exhibited notable variability, ranging from nearly impermeable conditions at 0.005% to highly permeable conditions at 47%, with an average of 18%. The effective porosity of the alluvial unit displayed a similar pattern ranging from 0.004 to 44%. The valley incision unit demonstrated a more uniform distribution of effective porosity, varying from 16 to 33%, with a mean of 25% while the Late Miocene unit exhibited effective porosity values ranging from almost zero to 50%, with a mean of 16%.

Hydraulic conductivity

In the sedimentary clastic formations, the hydraulic conductivity and the amount of shale are generally inversely correlated[68]. In this research, the hydraulic conductivity values obtained from the Csókás method are correlated
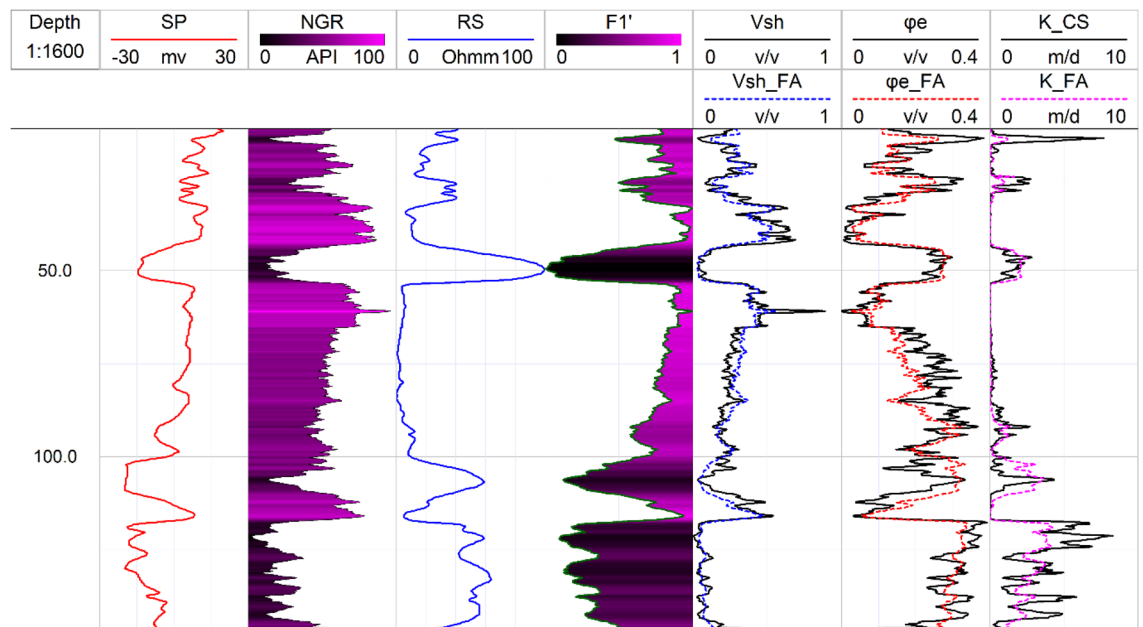
Figure 6. 1D analysis and interpretation of the well logs data in borehole B1 showing the estimated petrophysical and hydrogeological using conventional and factor analysis methods.

(a)

(b)
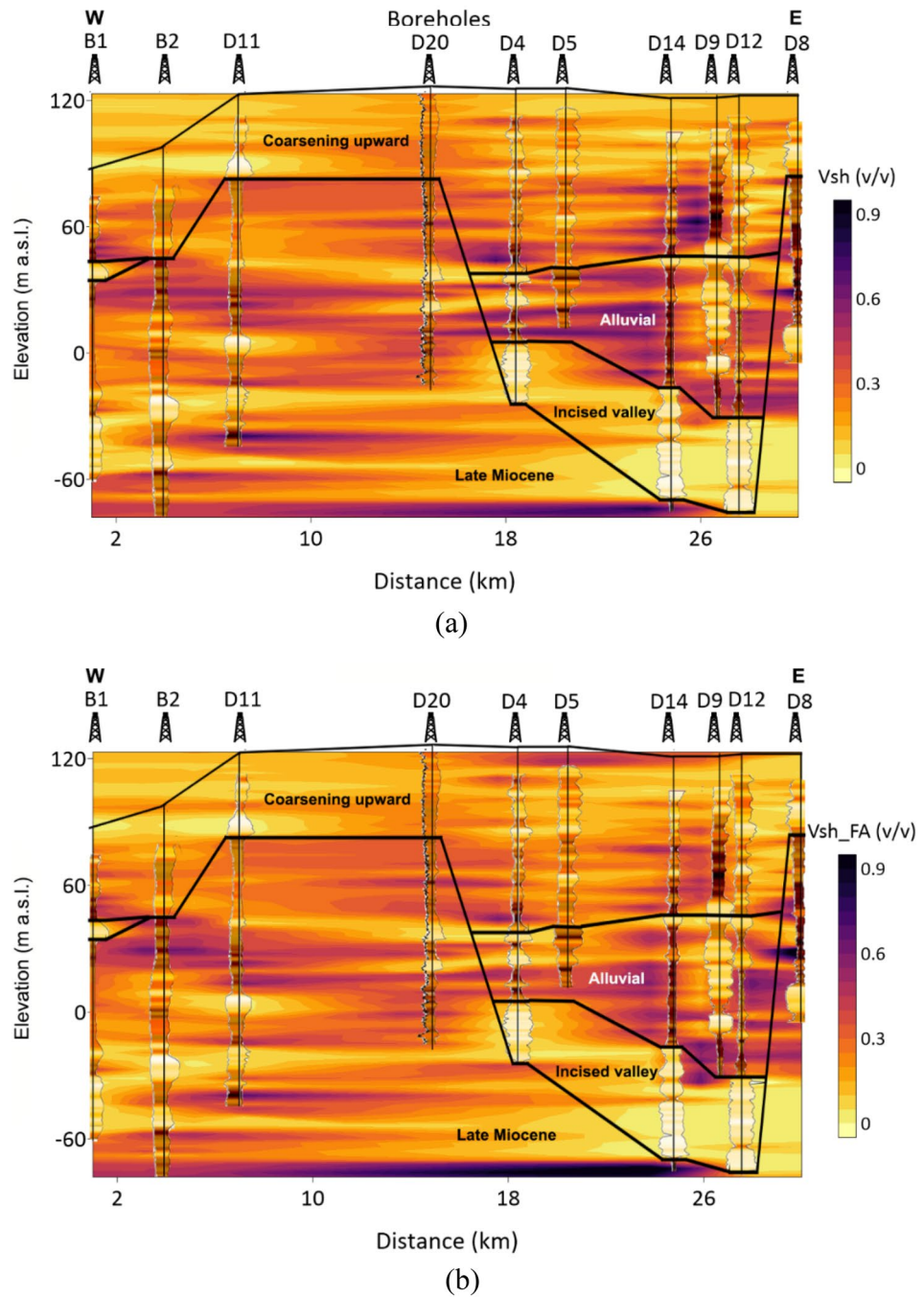
Figure 7. The estimated shale volume along the profile based on the (**a**) Larionov equation and (**b**) factor analysis.

to the first factor. Accordingly, a strong negative nonlinear relationship with a correlation coefficient of − 0.84 is detected[69] (Fig. 13a) that takes the following form,

$$\log(K) = a\left(1 - F_1^b\right)^c + d,$$

where a, b, c, and d represent site-specific regression coefficients. These coefficients showed values of 19.2, 4.27, 0.2, and − 19, respectively. The correlation between the hydraulic conductivity of the factor analysis and the Csókás method is shown in Fig. 13b, in which a close agreement (R = 0.88) is indicated. Accordingly, the hydraulic conductivity is mapped into 2D to reveal the vertical and horizontal variation (Fig. 14). The descriptive statistics of the FA-based hydraulic conductivity are illustrated in Fig. 15. The hydraulic conductivity of the coarsening upward unit ranged from nearly impermeable conditions at 0.005 m/d to more conductive zones with values of up to 2.3 m/d. The mean hydraulic conductivity for this unit was approximately 0.5 m/d. In the alluvial unit,
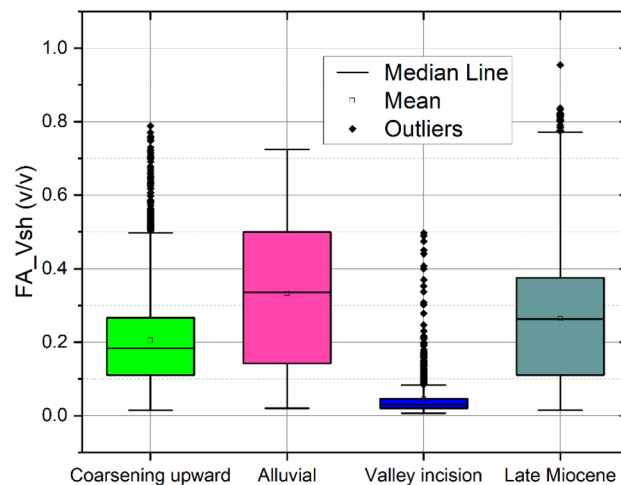
Figure 8. Box plot showing the statistical summary of the shale volume for the main hydrostratigraphical unit.
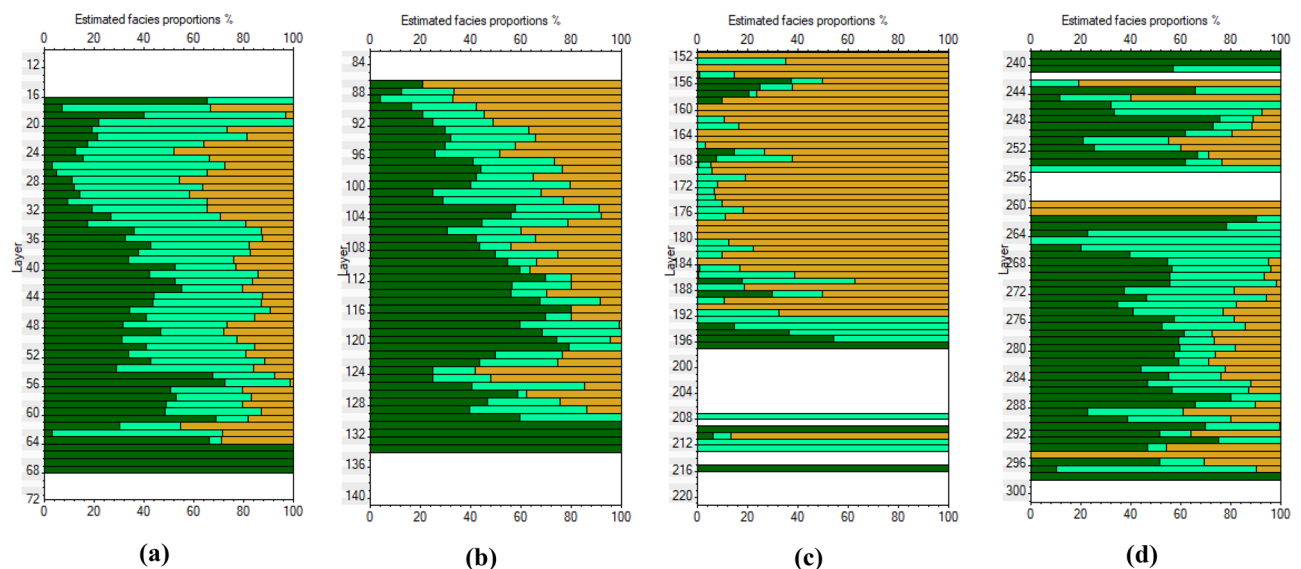

Figure 9. The estimated facies proportion for (**a**) coarsening upward, (**b**) alluvial unit, (**c**) valley incision, and (**d**) Late Miocene unit.

it varied from 0.004 to 0.7 m/d. Notably, the valley incision unit demonstrated a more uniform distribution of hydraulic conductivity, with values ranging from 0.05 to 3.6 m/d and a mean value of approximately 1.3 m/d. The Late Miocene unit exhibited hydraulic conductivity values ranging from almost zero to 0.8 m/d, averaging approximately 0.01 m/d.

Discussion

Factor analysis allowed the extraction of factor log that captured a significant portion of the data variance. Simple sensitivity analysis is conducted using Pearson and Spearman correlation coefficients (Fig. 16). These coefficients assist in understanding the relationship between well logs and the resulting factor logs and identifying which logs have the most significant impact on the outcome. The Pearson correlation coefficients, assuming linearity, displayed values of 0.43, 0.90, − 0.92, 0.81, − 0.67, and 0.38 between the extracted first factor and SP, NGR, RS, shale volume, effective porosity, and hydraulic conductivity, respectively (Fig. 16a). On the other hand, the Spearman rank correlation coefficients revealed stronger associations (0.41, 0.91, − 0.89, 0.91, − 0.75, and − 0.84). GR and RS logs exhibited higher correlations with the first factor because these logs are primarily sensitive to clay content, serving as indicators of lithological variation[70]. On the other hand, the SP log showed a lesser correlation with the extracted factor indicating its lower influence on the resulting factor log. This observation aligns with the initial hypothesis and underlines the dominant role of lithological characteristics in shaping the variability captured by the first-factor log.

Accordingly, the analysis of well-log data provided crucial implications for understanding the aquifer system in the study area. For instance, the variability in shale volume across the hydrostratigraphical units underscored
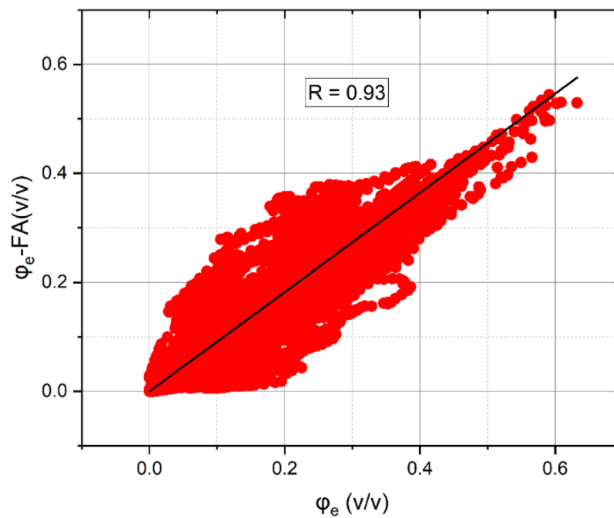
Figure 10. The correlation between the effective porosity using conventional methods and factor analysis.

the horizontal and vertical heterogeneity of subsurface geology[49,71]. The broad range of the estimated parameters depicted the heterogeneous nature of coarsening upward, alluvial, and Late Miocene units in which highly permeable materials coexisted within the less permeable zones[43]. The presence of low permeability shaly layers can act as barriers to flow, influencing the direction and velocity of groundwater movement. In contrast, the highly permeable sandy and gravely layers can facilitate rapid groundwater flow, potentially serving as potential aquifer zones[44]. The incised valley deposits, on the other hand, showed a more uniform distribution for the aquifer parameters with lower shale volume and higher effective porosity and hydraulic conductivity. The uniformity of this unit suggests a relative homogeneity, making it a potentially promising groundwater source[44,72].

Factor analysis has proven to be a successful method for characterizing the main hydrostratigraphical units in the Debrecen area, considering the limited number of available well logs which is a notable limitation in recent investigations. Given this constraint, factor analysis emerged as a suitable method for the estimation of key petrophysical and hydrogeological parameters and facilitating the characterization of groundwater systems[38]. However, in the petroleum industry, where more comprehensive reservoir characterization is required, more sophisticated machine learning methods such as neural networks are commonly employed[73,74]. These methods offer high accuracy and flexibility in handling complex relationships between well-log data and target parameters[75]. However, they require larger datasets and computational resources for training and optimization. The factor analysis approach demonstrated a higher generalization ability in which the obtained practical equations can be safely used for estimating the characteristics of the clastic heterogeneous aquifers, especially within the Pannonian Basin. The shared geological history and lithological composition of these aquifers suggest favorable conditions for employing this factor analysis-based approach, However, slight fluctuations in the regression coefficient are expected due to the variation in saturation and degree of cementation[76].

Conclusion

The main aim of this research is to detect the vertical and horizontal distribution of the petrophysical and hydrogeological parameters within the main hydrostratigraphical units of the Quaternary system. This research demonstrated the potential of factor analysis in redefining the interpretation of well-log data. The conclusions of this research can be summarized as follows:

- The first factor extracted from the data matrix containing SP, NGR, and RS logs explained 81.7% of the data variance that showed a solid exponential relationship with the shale volume determined by the Larionov equation. This relation allowed the development of a universal equation that can be used independently for shale volume estimation. The shale volume estimated using this practical equation closely agrees with the deterministic approach.
- Based on the FA-based shale volume, the effective porosity is estimated and showed a close agreement with that of the deterministic approach. Moreover, a nonlinear relationship is obtained between the first scaled factor and the hydraulic conductivity. The FA-based hydraulic conductivity estimation revealed a significant correlation with the Csókás-based hydraulic conductivity, showing high variations within the hydrostratigraphical units. However, the distribution of hydraulic conductivity within the valley incision unit showed a more uniform pattern, making this unit a promising groundwater aquifer.
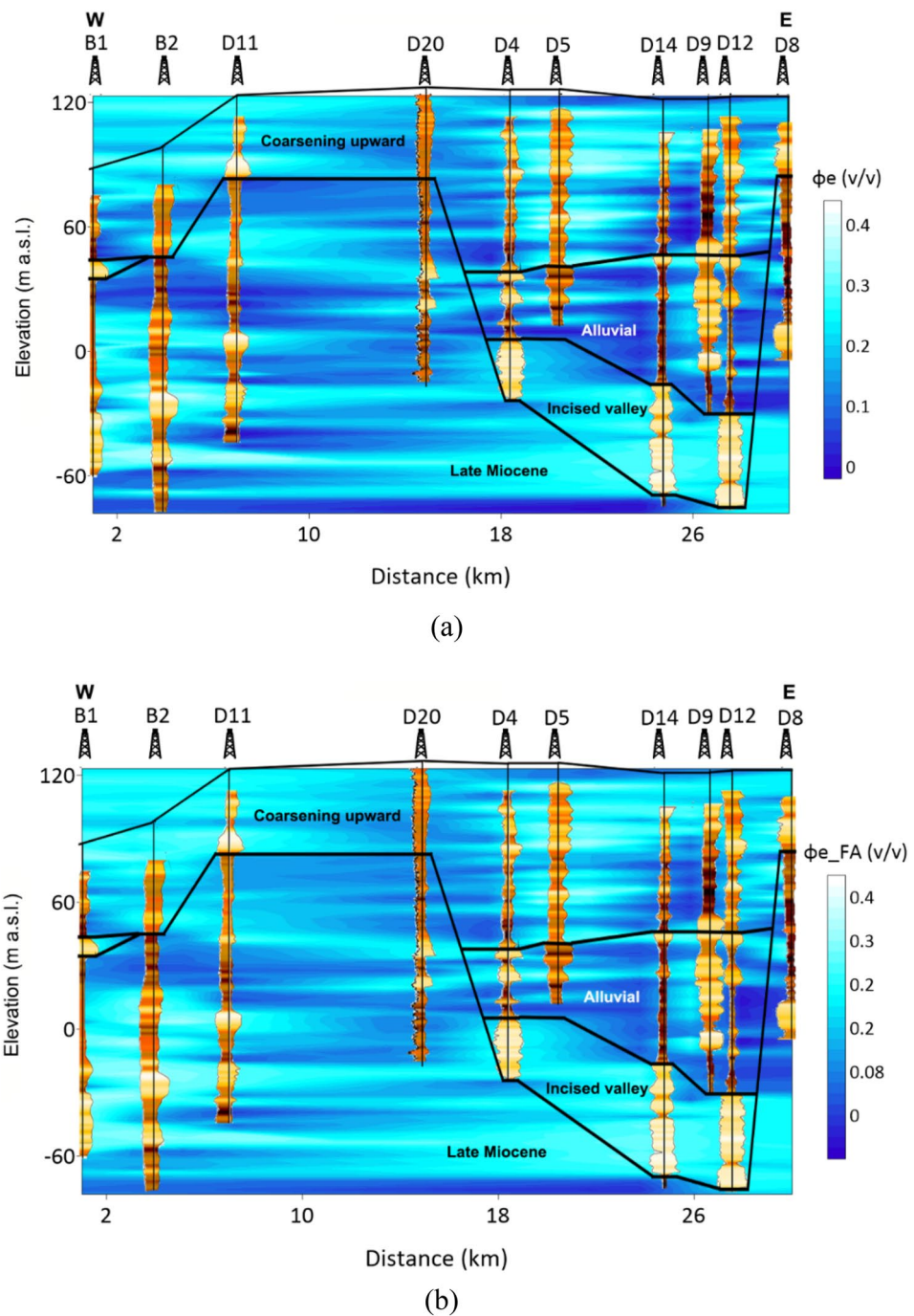
Figure 11. The estimated effective porosity along the profile based on the (**a**) empirical method and (**b**) factor analysis.

- The proposed methodology demonstrated potential for characterizing heterogeneous aquifer systems, and the findings can be directly applied to aquifers within the transboundary Pannonian Basin and other regions sharing similar geological and hydrogeological characteristics.
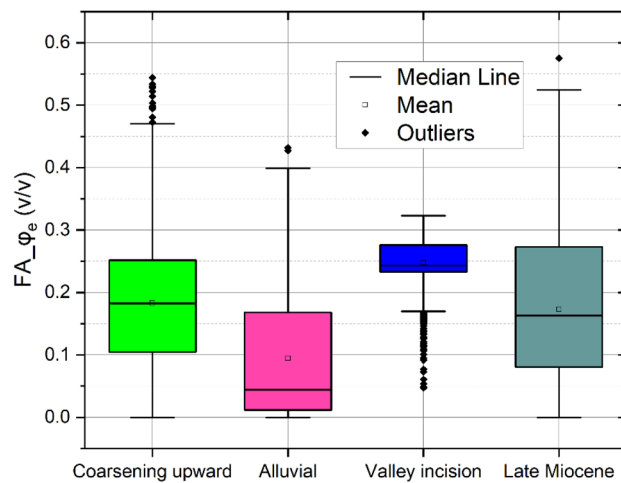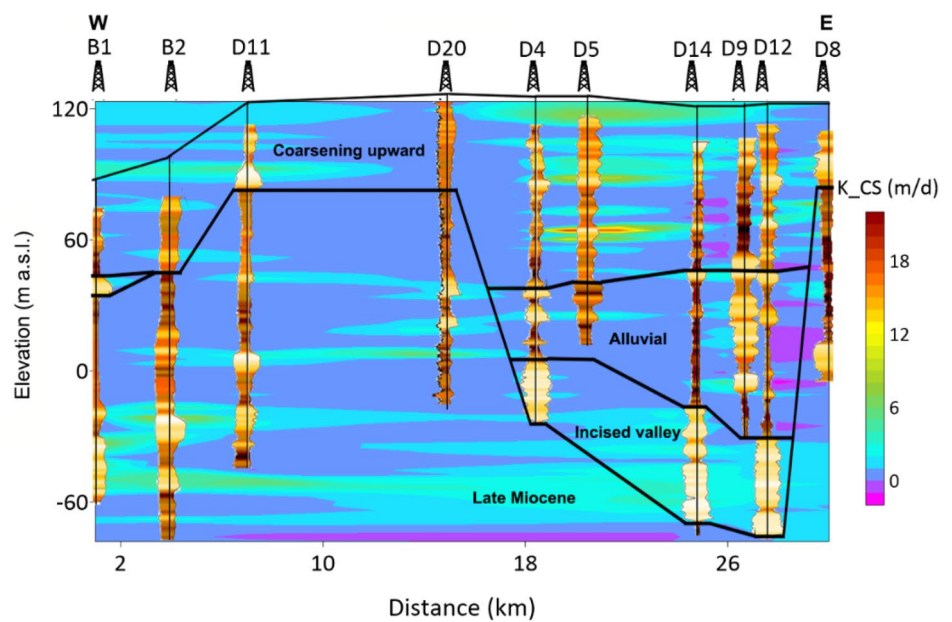
Figure 12. Box plot showing the statistical summary of the calculated effective porosity for the main hydrostratigraphical unit.
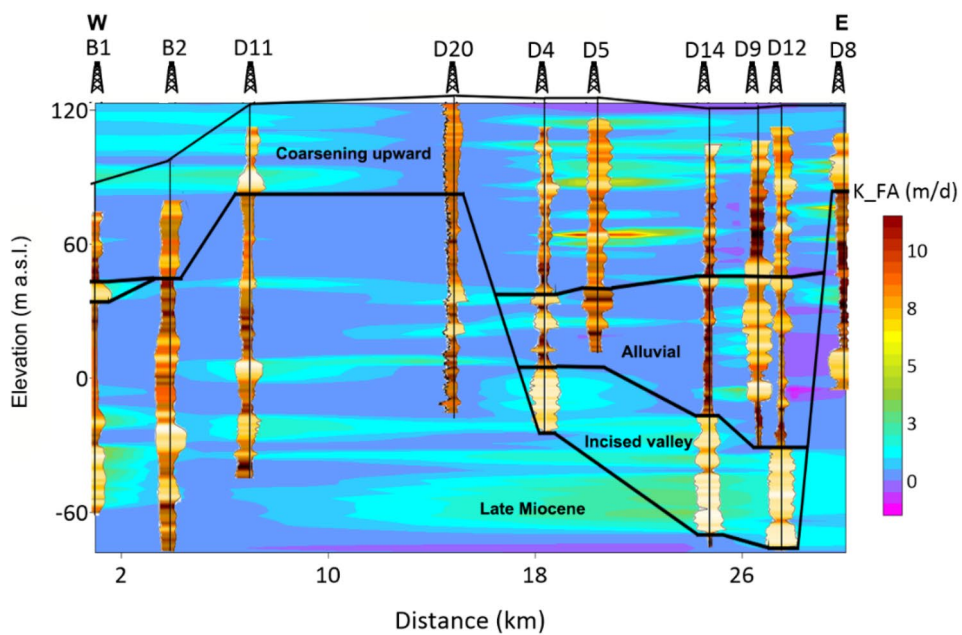

(a)   (b)
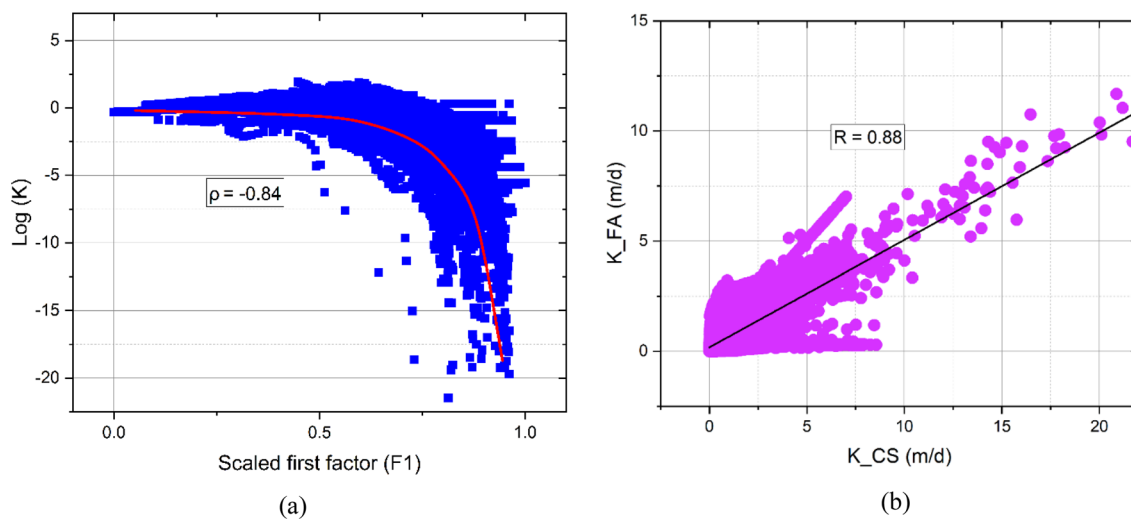
Figure 13. (**a**) The relationship between the scaled first factor and hydraulic conductivity and (**b**) the correlation between the factor analysis-based and Csókás method shale hydraulic conductivity.
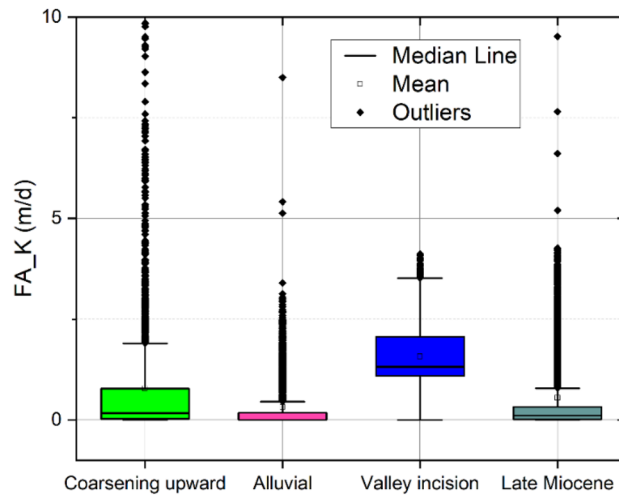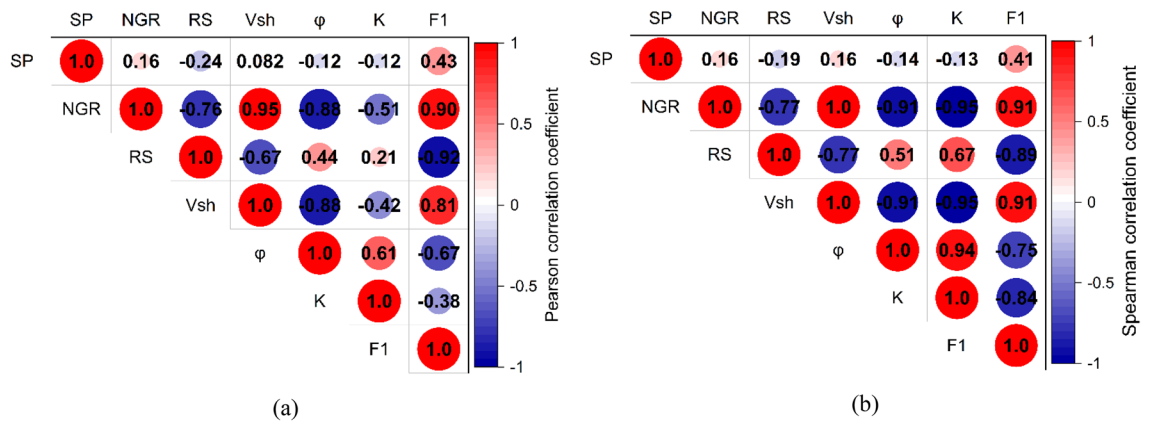
(a)

(b)

Figure 14. The calculated hydraulic conductivity along the profile based on the (**a**) Csókás method and (**b**) factor analysis.

Figure 15. Box plot showing the statistical summary of the FA-based hydraulic conductivity for the main hydrostratigraphical unit.

Figure 16. The correlation between the well logs, hydraulic parameters, and the extracted factor log using (**a**) Pearson and (**b**) Spearman correlation coefficients.

Data availability

The data that support the findings of this study are available from the Supervisory Authority for Regulatory Affairs (SARA), Hungary, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the corresponding author (Musaab A. A. Mohammed) upon reasonable request and with permission of Supervisory Authority for Regulatory Affairs (SARA), Hungary.

References

1. Avci, C. B., Ciftci, E. & Sahin, A. U. Identification of aquifer and well parameters from step-drawdown tests. *Hydrogeol. J.* **18**, 1591–1601 (2010).
2. Lin, H. T. *et al.* Estimation of effective hydrogeological parameters in heterogeneous and anisotropic aquifers. *J. Hydrol.* **389**, 57–68 (2010).
3. Michael, H. A. & Khan, M. R. Impacts of physical and chemical aquifer heterogeneity on basin-scale solute transport: Vulnerability of deep groundwater to arsenic contamination in Bangladesh. *Adv. Water Resour.* **98**, 147–158 (2016).
4. Tran, D. H., Wang, S. J. & Nguyen, Q. C. Uncertainty of heterogeneous hydrogeological models in groundwater flow and land subsidence simulations—A case study in Huwei Town, Taiwan. *Eng. Geol.* **298**, 106543 (2022).
5. Aldana, C., Isch, A., Bruand, A., Azaroual, M. & Coquet, Y. Relationship between hydraulic properties and material features in a heterogeneous vadose zone of a vulnerable limestone aquifer. *Vadose Zone J.* **20**, 1–22 (2021).
6. Møller, I., Karan, S., Gravesen, P. & Rosenbom, A. E. On the representability of soil water samples in space and time: Impact of heterogeneous solute transport pathways underneath a sandy field. *Sci. Total Environ.* **856**, 159039 (2023).

7. Maples, S. R., Foglia, L., Fogg, G. E. & Maxwell, R. M. Sensitivity of hydrologic and geologic parameters on recharge processes in a highly heterogeneous, semi-confined aquifer system. *Hydrol. Earth Syst. Sci.* **24**, 2437–2456 (2020).
8. Bennett, G., Van Camp, M., Shemsanga, C., Kervyn, M. & Walraevens, K. Delineation of the aquifer structure and estimation of hydraulic properties on the flanks of Mount Meru, Northern Tanzania. *J. Afr. Earth Sci.* **196**, 104673 (2022).
9. Díaz-Curiel, J., Arévalo-Lomas, L., Biosca, B., Miguel, M. J. & Caparrini, N. Correlation of boreholes through well logs: Application to the western sector of Madrid. *Sensors* **23**, 1–17 (2023).
10. Szűcs, P., Szabó, N. P., Zubair, M. & Szalai, S. Innovative hydrogeophysical approaches as aids to assess hungarian groundwater bodies. *Appl. Sci.* **11**, 2099 (2021).
11. Hossain, M. I. *et al.* Hydrogeological characterization of saline water aquifer deploying multiple well logs at Khulna, a coastal region of Bangladesh. *Arab. J. Geosci.* https://doi.org/10.1007/s12517-021-06770-8 (2021).
12. Navarro, J., Teramoto, E. H., Engelbrecht, B. Z. & Kiang, C. H. Assessing hydrofacies and hydraulic properties of basaltic aquifers derived from geophysical logging. *Braz. J. Geol.* https://doi.org/10.1590/2317-4889202020200013 (2020).
13. Farrag, A. A., Ebraheem, M. O., Sawires, R., Ibrahim, H. A. & Khalil, A. L. Petrophysical and aquifer parameters estimation using geophysical well logging and hydrogeological data, Wadi El-Assiuoti, Eastern Desert, Egypt. *J. Afr. Earth Sci.* **149**, 42–54 (2019).
14. Aliou, A. S., Dzikunoo, E. A., Yidana, S. M., Loh, Y. & Chegbeleh, L. P. Investigation of geophysical signatures for successful exploration of groundwater in highly indurated sedimentary basins: A look at the Nasia Basin, NE Ghana. *Nat. Resour. Res.* **31**, 3223–3251 (2022).
15. Jiang, W. *et al.* Application of audio-frequency magnetotelluric data to cover characterisation–validation against borehole petrophysics in the East Tennant region, Northern Australia. *Explor. Geophys.* https://doi.org/10.1080/08123985.2023.2246492 (2023).
16. Anomohanran, O., Oseme, J. I., Iserhien-Emekeme, R. E. & Ofomola, M. O. Determination of groundwater potential and aquifer hydraulic characteristics in Agbor, Nigeria using geo-electric, geophysical well logging and pumping test techniques. *Model. Earth Syst. Environ.* **7**, 1639–1649 (2021).
17. Oladele, S., Salami, R. & Dauda, R. S. Petrophysical and hydrogeological characterization of coastal aquifer using geophysical logs in Lekki Peninsula, Lagos, Nigeria. *Groundw. Sustain. Dev.* **22**, 100971 (2023).
18. Paillet, F. L. & Crowder, R. E. A generalized approach for the interpretation of geophysical well logs in ground-water studies—theory and application. *Groundwater* **34**, 883–898 (1996).
19. Jardani, A., Revil, A., Bolève, A. & Dupont, J.-P. Three-dimensional inversion of self-potential data used to constrain the pattern of groundwater flow in geothermal fields. *J. Geophys. Res. Solid Earth* https://doi.org/10.1029/2007JB005302 (2008).
20. Karahan, H. & Ayvaz, M. T. Simultaneous parameter identification of a heterogeneous aquifer system using artificial neural networks. *Hydrogeol. J.* **16**, 817–827 (2008).
21. Kowalsky, M. B., Chen, J. & Hubbard, S. S. Joint inversion of geophysical and hydrological data for improved subsurface characterization. *Leading Edge* **25**, 730–734 (2006).
22. Kobr, M., Mareš, S. & Paillet, F. Geophysical well logging: Borehole geophysics for hydrogeological studies: Principles and applications. *Hydrogeophysics* https://doi.org/10.1007/1-4020-3102-5_10 (2005).
23. Niwas, S. & De Lima, O. A. L. Aquifer parameter estimation from surface resistivity data. *Groundwater* **41**, 94–99. https://doi.org/10.1111/j.1745-6584.2003.tb02572.x (2003).
24. Pliakas, F. & Petalas, C. Determination of hydraulic conductivity of unconsolidated river alluvium from permeameter tests, empirical formulas and statistical parameters effect analysis. *Water Resour. Manag.* **25**, 2877–2899 (2011).
25. Mohammed, M. A. A., Szabó, N. P. & Szűcs, P. Assessment of the Nubian aquifer characteristics by combining geoelectrical and pumping test methods in the Omdurman area, Sudan. *Model. Earth Syst. Environ.* **9**, 4363–4383 (2023).
26. Csókás, J. Determination of yield and water quality of aquifers based on geophysical well logs. *Magyar Geofizika* **35**, 176–203 (1995).
27. Mohammed, M. A. A., Abdelrahman, M. M. G., Szabó, N. P. & Szűcs, P. Innovative hydrogeophysical approach for detecting the spatial distribution of hydraulic conductivity in Bahri city, Sudan: A comparative study of Csókás and Heigold methods. *Sustain. Water Resour. Manag.* **9**, 1–16 (2023).
28. Dudash, L. W., Morgan, T. & Kennedy, J. Integrated geophysical investigation of pauma groundwater basin, California. *Proc. Symp. Appl. Geophy. Eng. Environ. Probl. SAGEEP* **1**, 280–290 (2009).
29. Alfy, M. E. *et al.* Quantitative hydro-geophysical analysis of a complex structural karst aquifer in Eastern Saudi Arabia. *Sci. Rep.* **9**, 1–18 (2019).
30. Dramsch, J. S. 70 years of machine learning in geoscience in review. *Adv. Geophys.* **61**, 1–55 (2020).
31. Caté, A., Perozzi, L., Gloaguen, E. & Blouin, M. Machine learning as a tool for geologists. *Leading Edge* **36**, 215–219 (2017).
32. Joshi, D. *et al.* Prediction of sonic log and correlation of lithology by comparing geophysical well log data using machine learning principles. *GeoJournal* **88**, 1–22 (2021).
33. Urang, J. G., Ebong, E. D., Akpan, A. E. & Akaerue, E. I. A new approach for porosity and permeability prediction from well logs using artificial neural network and curve fitting techniques: A case study of Niger Delta, Nigeria. *J. Appl. Geophys.* **183**, 104207 (2020).
34. Maxwell, K., Rajabi, M. & Esterle, J. Automated classification of metamorphosed coal from geophysical log data using supervised machine learning techniques. *Int. J. Coal Geol.* **214**, 103284 (2019).
35. Puskarczyk, E., Jarzyna, J. A., Wawrzyniak-Guz, K., Krakowska, P. I. & Zych, M. Improved recognition of rock formation on the basis of well logging and laboratory experiments results using factor analysis. *Acta Geophys.* **67**, 1809–1822 (2019).
36. Abordán, A. & Szabó, N. P. Uncertainty reduction of interval inversion estimation results using a factor analysis approach. *GEM Int. J. Geomath.* **11**, 1–17 (2020).
37. Lawley, D. N. & Maxwell, A. E. Factor analysis as a statistical method. *J. R. Stat. Soc. Ser. D* **12**, 209–229 (1962).
38. Szabó, N. P. Hydraulic conductivity explored by factor analysis of borehole geophysical data. *Hydrogeol. J.* **23**, 869–882 (2015).
39. Bueno Buoro, A. & Silva, J. B. C. Ambiguity analysis of well-log data. *Geophysics* **59**, 336–344 (1994).
40. Li, M. *et al.* Application of mathematical statistics to shale gas-bearing property evaluation and main controlling factor analysis. *Sci. Rep.* **12**, 1–15 (2022).
41. Asfahani, J. Statistical factor analysis technique for characterizing basalt through interpreting nuclear and electrical well logging data (case study from Southern Syria). *Appl. Radiat. Isot.* **84**, 33–39 (2014).
42. Puskarczyk, E. Application of multivariate statistical methods and artificial neural network for facies analysis from well logs data: An example of Miocene deposits. *Energies* **13**, 1–18 (2020).
43. Püspöki, Z. *et al.* High-resolution stratigraphy of quaternary fluvial deposits in the Makó Trough and the Danube-Tisza Interfluve, Hungary, based on magnetic susceptibility data. *Boreas* **50**, 205–223 (2021).
44. Flores, Y. G. *et al.* Integration of geological, geochemical modelling and hydrodynamic condition for understanding the geometry and flow pattern of the aquifer system, Southern Nyírség-Hajdúság, Hungary. *Water* **15**, 2888 (2023).
45. ESRI. ArcGIS. at https://www.esri.com/en-us/arcgis/products/arcgis-desktop/resources (2020).
46. Fülöp, J. *Bevezetés Magyarország geológiájába* (Akadémiai Kiadó, 1989).
47. Buday, T. *et al.* Sustainability aspects of thermal water production in the region of Hajdúszoboszló-Debrecen, Hungary. *Environ. Earth Sci.* **74**, 7511–7521 (2015).
48. Mádl-Szőnyi, J. *et al.* Confined carbonates—Regional scale hydraulic interaction or isolation?. *Mar. Petrol. Geol.* **107**, 591–612 (2019).

49. Tóth, J. & Almási, I. Interpretation of observed fluid potential patterns in a deep sedimentary basin under tectonic compression: Hungarian Great Plain, Pannonian Basin. *Geofluids* **1**, 11–36 (2001).
50. Kronome, B. *et al.* Geological model of the Danube Basin; transboundary correlation of geological and geophysical data. *Slovak Geol. Mag.* **14**, 17–35 (2014).
51. Püspöki, Z. *et al.* Tectonically controlled quaternary intracontinental fluvial sequence development in the Nyírség-Pannonian Basin, Hungary. *Sediment. Geol.* **283**, 34–56 (2013).
52. Juhász, G. Lithostratigraphical and sedimentological framework of the Pannonian (sl) sedimentary sequence in the Hungarian Plain (Alföld), Eastern Hungary. *Acta Geol. Hung.* **34**, 53–72 (1991).
53. Mohammed, M. A. A., Szabó, N. P. & Szűcs, P. Joint interpretation and modeling of potential field data for mapping groundwater potential zones around Debrecen. *Acta Geodaet. Geophys.* https://doi.org/10.1007/s40328-023-00433-8 (2024).
54. Kozeny, J. Uber kapillare leitung der wasser in boden. *R. Acad. Sci. Vienna Proc. Class I* **136**, 271–306 (1927).
55. Carman, P. C. Fluid flow through granular beds. *Trans. Inst. Chem. Eng.* **15**, 150–166 (1937).
56. Alger, R. P. Interpretation of electric logs in fresh water wells in unconsolidated formations. *SPE Reprint Ser.* **1**, 255 (1971).
57. Schlumberger,. *Log Interpretation Principles/Applications* (Schlumberger Educational Services, 1991).
58. Larionov, V. V. *Radiometry of Boreholes* 127 (Nedra, 1969).
59. Szabó, N. P. Shale volume estimation based on the factor analysis of well-logging data. *Acta Geophys.* **59**, 935–953 (2011).
60. Jöreskog, K. G. Factor analysis and its extensions. *Fact. Anal.* **100**, 47–77 (2007).
61. Kaiser, H. F. The varimax criterion for analytical rotation in factor analysis. *Psychometrika* **23**, P187-200 (1958).
62. Bartlett, M. S. The statistical conception of mental factors. *Br. J. Psychol.* **28**, 97 (1937).
63. Isaaks, E. H. & Srivastava, R. M. *An Introduction to Applied Geostatistics* (Oxford University Press, 1989).
64. Spearman, C. *The Proof and Measurement of Association Between Two Things* (Appleton-Century-Crofts, 1961).
65. De Boor, C., Höllig, K. & Sabin, M. High accuracy geometric Hermite interpolation. *Comput. Aided Geom. Des.* **4**, 269–278 (1987).
66. Catuneanu, O. *Principles of Sequence Stratigraphy* (Newnes, 2022).
67. Fejes, Z., Szűcs, P., Turai, E., Zákányi, B. & Szabó, N. P. Regional hydrogeology of the tokaj mountains world heritage site, northeast hungary. *Acta Montan. Slov.* **26**, 18–34 (2021).
68. Schon, J. H. & Georgi, D. Dispersed shale, shaly-sand permeability-a hydraulic analog to the Waxman-Smits equation. In *SPWLA Annual Logging Symposium* SPWLA--2003 (2003).
69. Szabó, N. P., Abordán, A. & Dobróka, M. Permeability extraction from multiple well logs using particle swarm optimization based factor analysis. *GEM Int. J. Geomath.* **13**, 1–27 (2022).
70. Szabó, N. P., Dobróka, M. & Drahos, D. Factor analysis of engineering geophysical sounding data for water-saturation estimation in shallow formations. *Geophysics* **77**, WA35–WA44 (2012).
71. Püspöki, Z. *et al.* Obliquity-driven mountain permafrost-related fluvial magnetic susceptibility cycles in the Quaternary mid-latitude long-term (2.5 Ma) fluvial Maros Fan in the Pannonian Basin. *Boreas* **52**, 402–426 (2023).
72. Mohammed, M. A., Szabó, N. P., Flores, Y. G. & Szűcs, P. Multi-well clustering and inverse modeling-based approaches for exploring geometry, petrophysical, and hydrogeological parameters of the Quaternary aquifer system around Debrecen area, Hungary. *Groundw. Sustain. Dev.* https://doi.org/10.1016/j.gsd.2024.101086 (2024).
73. Shehata, A. A., Osman, O. A. & Nabawy, B. S. Neural network application to petrophysical and lithofacies analysis based on multiscale data: An integrated study using conventional well log, core and borehole image data. *J. Nat. Gas Sci. Eng.* **93**, 104015 (2021).
74. Leisi, A. & Saberi, M. R. Petrophysical parameters estimation of a reservoir using integration of wells and seismic data: A sandstone case study. *Earth Sci. Inform.* **16**, 637–652 (2023).
75. Pan, W., Torres-Verdin, C., Duncan, I. J. & Pyrcz, M. J. Improving multiwell petrophysical interpretation from well logs via machine learning and statistical models. *Geophysics* **88**, D159–D175 (2023).
76. Szabó, N. P., Kiss, A. & Halmágyi, A. Hydrogeophysical characterization of groundwater formations based on well logs: Case study on cenozoic clastic aquifers in East Hungary. *Geosci. Eng.* **4**, 45–71 (2015).

Acknowledgements

Author contributions
Musaab A. A. Mohammed: Conceptualization; Data curation; Methodology; Formal analysis; Writing—original draft, Yetzabbel G. Floresa: Writing—review & editing; Validation, Norbert P. Szabó: Methodology; Validation; Supervision, Péter Szűcs: Project administration; Funding acquisition; Supervision.

Funding

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.A.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.