DOI: 10.14750/ME.2025.025

UNIVERSITY OF MISKOLC
MIKOVINY SÁMUEL DOCTORAL SCHOOL OF EARTH SCIENCES
Head of the Doctoral School:
Prof. Dr. Péter Szűcs

INNOVATIVE HYDROGEOPHYSICAL APPROACH FOR CHARACTERIZING THE QUATERNARY AQUIFER SYSTEM IN DEBRECEN AREA, EASTERN HUNGARY

PhD Thesis

By:
Musaab Adam Abbakar Mohammed
Hydrogeological Engineer

Scientific supervisors:
Prof. Dr. Péter Szűcs
Prof. Dr. Norbert Péter Szabó

Miskolc, 2025
HUNGARY

DOI: 10.14750/ME.2025.025

Statement of the Supervisors

For the PhD Thesis

" INNOVATIVE HYDROGEOPHYSICAL APPROACH FOR CHARACTERIZING THE QUATERNARY AQUIFER SYSTEM IN DEBRECEN AREA, EASTERN HUNGARY," by **Musaab A. A. Mohammed**.

Throughout his doctoral studies, the candidate has demonstrated intellectual rigor, innovative thinking, and a commitment to advancing the use of geophysical methods for solving hydrogeological problems. His research addresses a critical challenge in modern hydrogeology: the accurate characterization of heterogeneous aquifer systems in regions under significant pressure from agricultural and domestic water use. The candidate integrated geophysical data, machine learning techniques, and numerical modeling to develop a comprehensive framework for evaluating groundwater resources in the Quaternary aquifer system in the Debrecen area.

The PhD thesis is structured around several key innovations. The candidate successfully developed a novel approach to process and analyze well-logging data. He applied novel and robust clustering techniques including the most frequent value-assisted cluster analysis (MFV-CA) and self-organizing maps (SOM) to produce geologically realistic models that aligned closely with drilling data. In addition to lithological characterization, the candidate applied the Csókás method, which was previously developed in the Department of Geophysics, University of Miskolc, to estimate hydraulic conductivity across the aquifer. This approach extended the spatial coverage of limited pumping test data, enabling the creation of high-resolution 3D maps and proving effective for detailed aquifer characterization.

The candidate’s innovative use of deep learning further enhances the efficiency and accuracy of geological and hydrogeophysical analysis. He proposed hybrid deep learning models to improve lithological classification and hydraulic conductivity estimation. These models achieved remarkable accuracy for lithology and hydraulic conductivity prediction, while significantly reducing computational demands. Finally, the candidate introduced an innovative approach to simulate 3D steady-state flow conditions using the results of the well-logging data as the primary input. He integrated geophysics-based conceptual models with numerical modeling techniques to develop a stable groundwater flow model. This approach enhanced the efficiency of groundwater modeling and reduced reliance on extensive field measurements.

The findings of the candidate have immediate practical applications for groundwater development and management in the Debrecen area while also providing methodological frameworks that can be adapted for similar investigations worldwide. The work has resulted in several peer-reviewed publications in leading journals, indicating the high quality and relevance of the research findings. In addition to his technical achievements, the candidate has demonstrated outstanding academic and professional qualities. He has consistently demonstrated a strong work ethic, intellectual curiosity, and the ability to approach complex problems with creativity, determination, and innovative solutions. We confidently recommend that he be awarded the PhD degree, as he has successfully fulfilled all the academic and research requirements for this qualification.

MAY 2025, MISKOLC

Scientific Supervisors

Prof. Dr. Péter Szűcs

University Full Professor

Prof. Dr. Norbert Péter Szabó

University Full Professor

DOI: 10.14750/ME.2025.025

Table of Contents

1. INTRODUCTION

1.1 Background and aims

The characterization of the topography-driven aquifer systems represents a major concern in hydrogeological investigations. The primary objective of such investigations is to obtain detailed information about the geometry, lithology, and hydraulic properties of these systems (Tóth 2009). This information is traditionally acquired through pumping tests, core drilling, and laboratory analyses, which allow for the detailed examination of geological, petrophysical, and hydrogeological properties. Understanding the variations in these properties is crucial for predicting groundwater flow, recharge rates, and contamination pathways (Anderson et al. 2015). However, these conventional methods may not adequately capture the aquifer's heterogeneity, as pumping and samples are typically conducted and collected at sparse locations, potentially leading to incomplete representations of the subsurface heterogeneity (Williams and Paillet 2023). Moreover, these methods are often expensive and time-consuming. On the other hand, the use of geophysical data such as electrical, electromagnetic, and well logging provides indirect measurements of subsurface properties by observing variations in the physical properties of the subsurface. These methods can cover larger areas more rapidly and indicate quasi-continuous structural features, lithological variations, and hydrogeological properties with higher resolution (Rubin and Hubbard 2006; Mohammed et al. 2024c).

Geophysical well logs provide a high-resolution perspective of the subsurface, enabling a more detailed and accurate characterization of the aquifer properties (Williams and Paillet 2023). Different types of well logs are used for solving groundwater problems including spontaneous potential (SP), natural gamma ray (NGR), resistivity, and nuclear magnetic resonance (NMR). The SP log measures the natural electrical potential created by differences in ionic concentration between the borehole fluid and the formation water. This log provides valuable information on the lithology and salinity of the formation water. NGR logging measures the natural radioactivity emitted by formations and helps identify lithology and potential groundwater zones. The resistivity logging measures the electrical resistivity of formations surrounding the borehole allowing for reservoir identification and evaluation of fluid saturation (Serra 1983). Geophysical well logs play a vital role in evaluating groundwater aquifers, providing continuous in-situ measurements of subsurface properties. However, due to the complex and often ambiguous nature of log responses, especially in heterogeneous formations, advanced analytical techniques are essential to accurately extract meaningful lithological and hydrogeological information. Different techniques are used to interpret well-logging data including deterministic, multivariate statistics, and inversion methods (Szabó 2015). Deterministic methods rely on established physical relationships between well-log measurements and subsurface properties (Timur 1968). Statistical methods offer a data-driven approach to reveal hidden relationships between different logging measurements and geological and hydrogeological properties of the aquifer (Keys 1990).

Geophysical logging data is widely used for rock typing problems and geological modeling, as it enables a detailed characterization of rock composition, stratification, and variations in lithology (Kobr et al. 2005). Hydrogeologists have relied on visual interpretation of these logs, with sandy aquifers commonly identified by their low NGR responses and higher resistivity compared to surrounding beds. While this approach can be effective for addressing simple problems, such as well designing, it is prone to misinterpretations. Factors such as subjective human judgment in selecting layer boundaries can lead to inaccuracies, particularly in complex geological settings. This subjectivity limits the precision of lithological classification and can affect the overall understanding of subsurface conditions (Mohammed et al. 2024e). To overcome these limitations, there is a growing need for automated methods that ensure consistent and objective determination of lithology and layer boundaries. Recent advances in machine learning have made it possible to apply multivariate unsupervised learning techniques, such as cluster analysis, factor analysis, and self-organizing maps (Kohonen 1982; Asfahani et al. 2018; Szűcs et al. 2021; Mohammed et al. 2023). These methods have proven effective in solving complex rock-typing problems, as they can reduce subjectivity, improve accuracy, and automate the interpretation of geophysical well logs.

Hydrogeological parameters are typically determined through laboratory experiments and pumping tests (Klute and Dirksen 1986). While these methods are effective, they require significant resources for data collection and analysis, especially when conducted over large spatial scales. One of the primary attributes of geophysical logs is their ability to contribute to aquifer analysis through quantitative log interpretation. The relationship between well log data and parameters is established through response functions, though no comprehensive models directly link aquifer parameters to well logging data (Paillet and Crowder 1996; Szabó 2015). As a result, hydrogeologists rely on empirical methods for estimating aquifer properties (Jorgensen 1988; Nelson 1994). While theoretically straightforward, applying this direct method in practice is challenging. This difficulty arises because the parameters in these empirical equations are not well known, and geophysical logs are rarely calibrated for diverse geological settings. In this context, Csókás (1995) developed a methodology to determine hydraulic conductivity. This method effectively addresses the limitations of the empirical methods in which the parameters are derived solely from well-log data. The Csókás method developed Prof. Dr. János Csókás the former Head of the Geophysical Department at the University of Miskolc. This method relies on the interpretation of geophysical logs sensitive to lithology and groundwater saturation and provides a continuous profile of hydraulic conductivity.

Despite the advancements facilitated by deterministic, statistical, and inversion techniques in analyzing geophysical data, their computational demands often prompt the exploration of more efficient and cost-effective techniques. In this context, the machine learning (ML) and deep learning (DL) models, characterized by their capacity for fast clustering, dimensionality reduction, classification, and regression, are suitable for handling large datasets encountered in geophysical investigations (Horrocks et al. 2015). While ML and DL have been extensively applied in well log analysis for characterizing petroleum reservoirs (Wiener et al. 1991; Yang et al. 2004), their application in groundwater system characterization remains limited in the literature (Mohammed et al. 2025a), even though water-bearing formations often exhibit greater heterogeneity and complexity compared to petroleum reservoirs (Paillet and Crowder 1996). DL models can be broadly categorized into two types including semi-supervised and supervised methods. Semi-supervised methods, such as autoencoder neural networks and adversarial neural networks, are designed to work with a combination of labeled and unlabeled data (Dramsch 2020; Mohammed et al. 2024f). These techniques can extract meaningful features from well-logging datasets, which can then be correlated with lithological and hydrogeological parameters (Horrocks et al. 2015; Hussain et al. 2023). The supervised learning models, such as artificial neural networks (ANNs) can accurately learn and model nonlinear relationships between well-logs and lithological and hydrogeological properties (Wiener et al. 1995; Imamverdiyev and Sukhostat 2019). The integration of deep learning into well-log analysis represents a paradigm shift, as it allows for the processing of vast amounts of data with speed and precision previously unattainable through statistical and inversion methods.

The Quaternary aquifer system in the Debrecen area is considered a key source for groundwater extraction to fulfill the supply demands. The growing population and agricultural development in the region have intensified groundwater extraction, resulting in several environmental challenges including compaction, water quality deterioration, and intensified vertical groundwater flow (Bendefy 1968; Szanyi 2004; Mohammed et al. 2025b). The Debrecen area is part of the Nyírség–Hajdúság groundwater body that comprises a complex aquifer system. Previous studies have enhanced the understanding of the hydrogeological setting in the region, though the level of detail and coverage has varied significantly across investigations. For instance, Tóth and Almási (2001) introduced a fluid potential pattern for the GHP, grouping the Quaternary succession, Zagyva Formation, and Újfalu Formation into a single unit termed the Nagyalföld Aquifer. Similarly, studies focusing on geothermal resources (Buday et al. 2015) treated the Quaternary sequences as three layers. Szanyi (2004) conducted a 3D numerical modeling to simulate the vertical spread of pollution within the uppermost layer of the Debrecen area. This simulation provided a detailed understanding of how contaminants migrate vertically through the subsurface system. Research about stratigraphic analysis of the Debrecen area was conducted by Püspöki et al. (2013) and provided compelling evidence for the presence of an incised valley system within the region. This finding prompted the initiation of a comprehensive database, aimed at contextualizing the hypothesized valley system. Recently, Carpio (2024) developed a regional-scale conceptual model of hydrostratigraphical units within the Quaternary aquifer system by analyzing the qualitative characteristics of the geophysical well logs.

Despite numerous hydrogeological studies, there remains a lack of detailed quantitative models that jointly and accurately define the spatial continuity of aquifer layers and their hydrogeological characteristics. This is because previous studies have primarily relied on well-to-well correlations to analyze lithological variations and on pumping test data to estimate aquifer properties. However, as noted, these techniques have their limitations, and the resulting lithological and hydrogeological information fails to fully capture the heterogeneity of the system. To address these limitations, I developed a hydrogeophysical approach that utilizes well-logging data to produce automated, high-resolution, and continuous characterizations of the Quaternary aquifer system in the Debrecen area. Within this framework, I aim to evaluate the performance of various deterministic, unsupervised learning, and deep learning algorithms in analyzing well-log data to identify lithological, petrophysical, and hydraulic variations within the aquifer system. Based on the interpreted well-logs, I construct a conceptual hydrogeological model to demonstrate the potential of geophysical data as effective inputs for groundwater flow modeling. This approach will enhance both the accuracy and efficiency of flow simulations, thereby supporting more informed and sustainable groundwater management.

1.2 Description of the study area

The study area is situated around Debrecen city, encompassing approximately 680 km^2 (Fig. 1). The study area lies between EOVY coordinates 820,000 to 860,000 meters and EOVX coordinates 243,000 to 260,000 meters, as defined by Hungary's national projection system, the Egységes Országos Vetületi Rendszer (EOV – Unified National Projection System). This area is part of the Great Hungarian Plain (GHP) in which substantial variations in land elevation have transpired due to tectonic movements, erosion, and extensive sedimentation processes (Püspöki et al. 2021; Mohammed et al. 2024g). These geological events have influenced the study area, leading to an elevation ranging from 81 to 164 m above sea level (a.s.l). The region's climate is predominantly continental, with annual mean temperatures ranging from 10° to 11° C. Annual precipitation totals vary from 550 to 600 mm, and potential evapotranspiration ranges between 600 and 700 mm/year (NATéR).

Geologically, the investigated depth extends to approximately 220 meters, encompassing the full Quaternary sequence and the upper portion of the underlying Late Miocene to Pliocene sediments. The Pliocene sediments begin with the Újfalu Sandstone Unit, which consists of alternating delta front and delta plain deposits made up of sandstone, siltstone, and clay marl. This is overlain by the Zagyva Unit, composed of fluvial and lacustrine sediments, including medium- to fine-grained sand, silt, clay, and clay marl. A distinct feature within this unit is the Nagyalföld Variegated Clay layer, characterized by interbedded variegated clay, lignite, and pebbly sand beds (Sztanó et al. 2023). The Quaternary sequences date back to the Pleistocene epoch and are primarily composed of fluvial sediments and sandy loess. Additionally, fluviolacustrine and lacustrine deposits are present within the Pleistocene sequence. The evolution of these deposits was significantly influenced by climatic and tectonic changes, which altered the fluvial transport capacity and shaped the regional landscape. The thickness of the Quaternary sediments in the study area generally ranges between 150 and 200 m (Haas 2012) and is stratigraphically divided into three main units: the upper, middle, and lower Pleistocene beds. The upper and lower units are dominated by river channels and overbank deposits, while the middle unit is characterized by fluviolacustrine and lacustrine sediments (Székely et al. 2020).

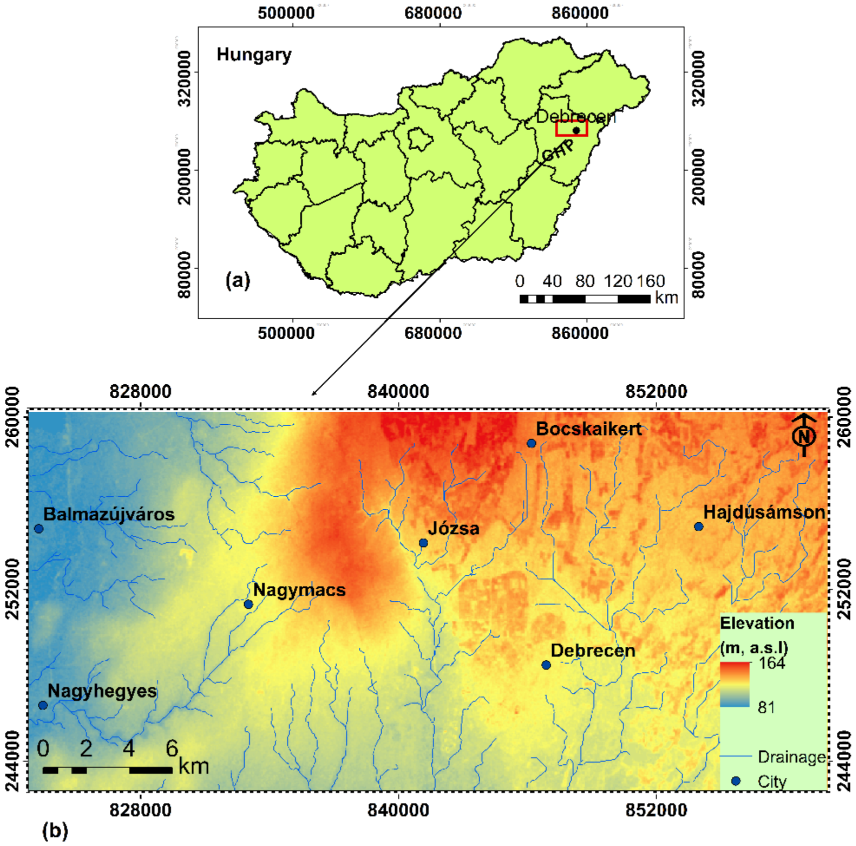


Fig. 1 (a) The location of the study area in eastern Hungary within the GHP. (b) The digital elevation model (DEM) of the study area.

The primary hydrostratigraphic units within the investigated depth comprise both Pre-Quaternary and Quaternary formations, which together constitute the Nagyalföld Aquifer system (Fig. 2). The Nagyalföld Aquifer is identified as the most important non-karstic aquifer in Hungary with a permeability greater than 1,000 mD (Tóth and Almási 2001; Buday and Püspöki 2011). The Pre-Quaternary succession within the Nagyalföld Aquifer, known as the Late Miocene Unit (LMU), is predominantly made up of thick silt layers with occasional thin interbeds of fine sand. This unit represents a low-permeability background that underlies the more productive overlying deposits. The Quaternary sequence is subdivided into three main hydrostratigraphic units, arranged from oldest to youngest (Carpio 2024). The lowermost is the Incised Valley Unit (IVU), consisting of elongated sand and gravel bodies trending in a northeast-southwest direction. This unit is notably deficient in clay with high hydraulic conductivity and efficient groundwater flow. Above the IVU lies the Alluvial Unit (AU), which is characterized by three horizontally extensive sand bodies interlayered with silty clay deposits. This unit displays significant lateral heterogeneity, reflecting the complex depositional dynamics of alluvial environments. The uppermost part of the Quaternary sequence is the Coarsening-Upward Unit (CUU), which features a vertically heterogeneous succession of clay, silt, and sand, representing a typical coarsening-upward trend formed under prograding depositional conditions.

Groundwater flow within the Nagyalföld Aquifer system is primarily driven by topographic gradients, following the natural slope of the land surface from higher elevation recharge zones toward lower elevation discharge areas (Czauner et al. 2024). The Debrecen area serves as a transitional zone between two distinct regions: the Nyírség recharge zone in the northeast, composed predominantly of alluvial fan deposits, and the Hortobágy discharge zone to the southwest (Erdélyi 1976). In terms of hydrochemistry, the groundwater in the shallow porous sediments is typically of the Ca-Mg-HCO_3 type. As groundwater moves along its flow path and interacts with surrounding minerals at greater depths, it evolves into a Na-HCO_3 type. This shift reflects geochemical processes such as ion exchange and mineral dissolution that occur with increasing residence time and depth (Varsányi 1992; Mohammed et al. 2025c).

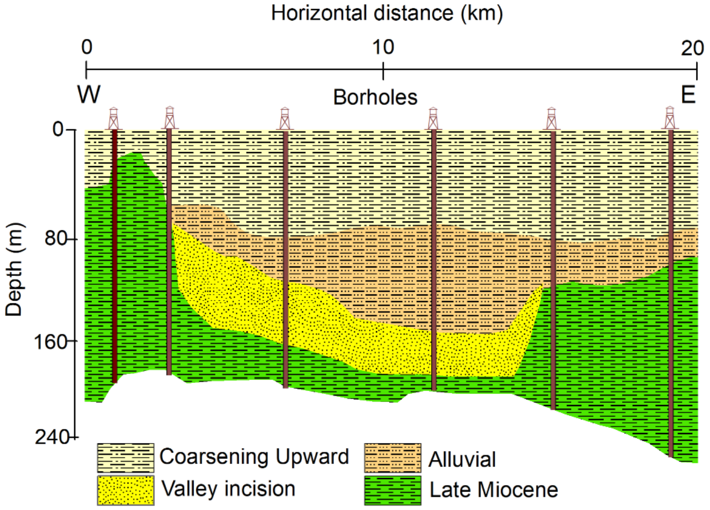


Fig. 2 The hydrostratigraphical units within the Quaternary aquifer system modified after Carpio (2024).

2. MATERIALS AND METHODS

I employed a hydrogeophysical approach using well-logging data to characterize the Quaternary aquifer system in the Debrecen area. The workflow (Fig. 3) begins with well-logging data analysis, where missing data is filled in using deep learning (DL) model. Following this, lithological modeling is conducted using a combination of unsupervised and supervised learning techniques and DL models. Hydraulic conductivity is then estimated based on deterministic and DL models to generate a 3D spatial distribution. Finally, the geological and hydrogeological information derived from the well-logging data analysis is synthesized into a conceptual model, which is subsequently transformed into a numerical MODFLOW-USG model for groundwater flow simulation.

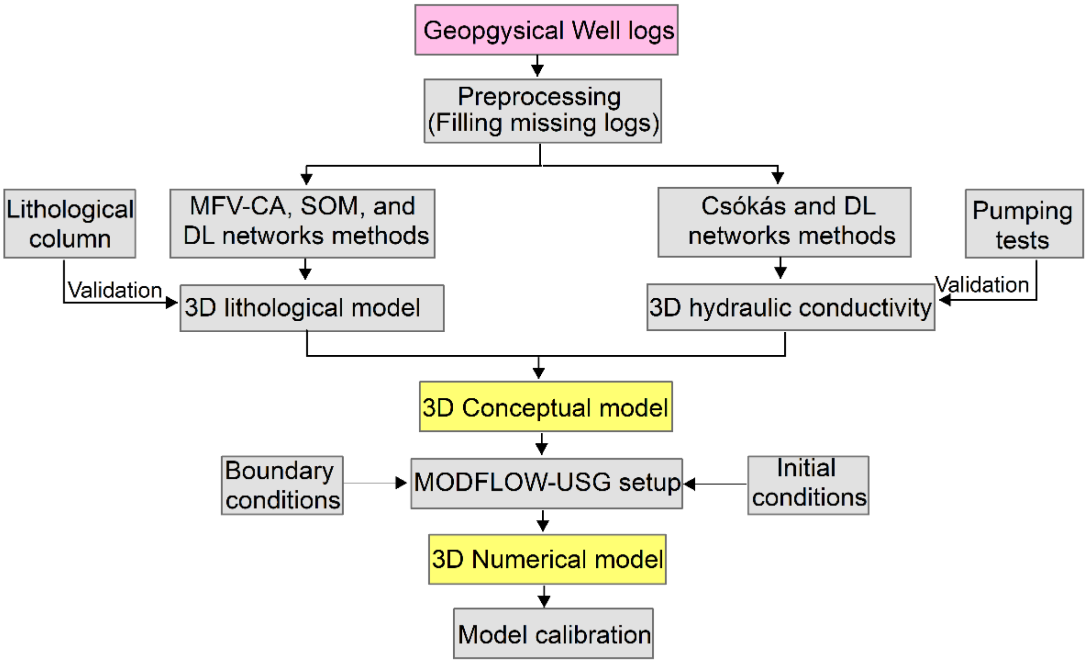


Fig. 3 Summary of the workflow conducted in this PhD thesis

2.1 Hydrogeophysical dataset

The data used in this thesis including geophysical and hydrogeological data is provided by the Supervisory Authority of Regulating Activities (SZTFH), formerly known as the Hungarian Mining and Geological Survey (MBFSZ). Well-logging data were collected from twenty-four boreholes used for groundwater production (Fig. 4). These boreholes have depths ranging from 95 m to over 200 m. The data included multiple well-logs including SP, NGR, and short-normal (RS) and long-normal resistivity log (RD) with the RD log being encountered in a limited number of boreholes (B2, D11, D12, and D14). Before analyzing the well logs, the data were normalized to account for variations in digitization scales. The SP and NGR logs were originally digitized from different sources and exhibited inconsistent value ranges. As a result, their digitization was not standardized across boreholes. To ensure consistency across the dataset, and because these logs are primarily used for qualitative interpretation, the SP log was normalized to a range of -30 to 30 mV, while the NGR log was scaled to 0 to 100 API units. In contrast, the resistivity logs were already digitized with consistent units and values across the dataset. Therefore, the original resistivity values were preserved without modification to maintain their integrity for quantitative analyses.

The hydrogeological data includes groundwater level measurements, water chemistry, and pumping test data. Groundwater level measurements were obtained from 90 wells across the study area while pumping test data were collected from 8 production wells. The water chemistry data were obtained in the form of electrical conductivity (EC) measurements from groundwater samples collected in the logged boreholes. These datasets, in combination with the geophysical data, form the foundation for the hydrogeophysical analysis conducted in this PhD thesis.

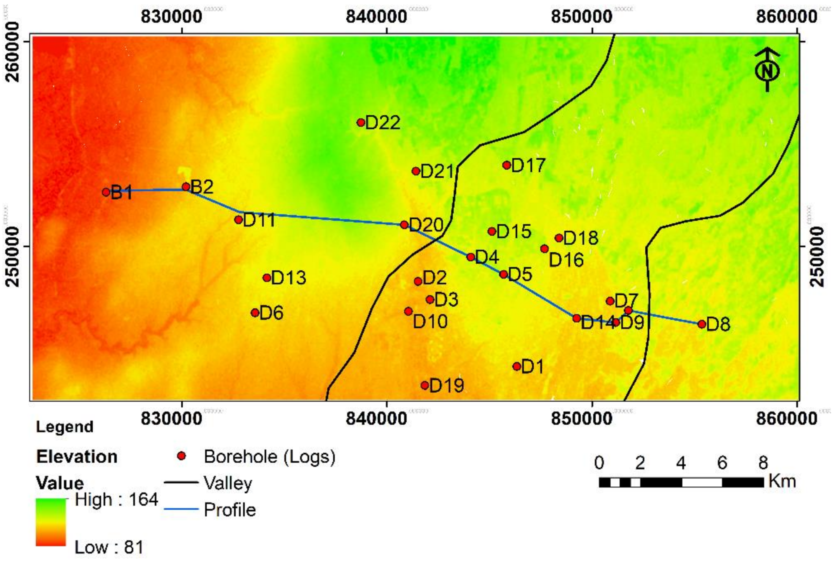


Fig. 4 The distribution of the data collection points and profiles used for the analysis of well-logging data.

2.2 Hydrogeophysical data analysis

Geophysical well-logging data plays a vital role in hydrogeological investigations by providing continuous and high-resolution information about subsurface formations. These logs are crucial for identifying lithological boundaries, evaluating aquifer properties, and informing decisions related to well design and completion (Tselentis 1985). In this thesis, a combination of deterministic methods and machine learning techniques was applied to preprocess and analyze well-log data. These approaches enabled the extraction of key lithological, petrophysical, and hydrogeological parameters necessary for understanding the structure and dynamics of the aquifer system. The following subsections provide detailed descriptions of the various techniques applied for analyzing well-logging data.

2.2.1 Most frequent value-assisted cluster analysis (MFV-CA)

The Most frequent value (MFV) method (Steiner 1988) employs a fully automated data weighting process, providing robust estimations by minimizing sensitivity to outliers. In MFV-assisted cluster analysis (MFV-CA), closer data points are assigned higher weights to emphasize their proximity within clusters, whereas data points farther apart receive relatively lower weights. The weighted average is computed using a symmetric weight function ($\emptyset$), with MFV representing the point where the $\emptyset$ weight function attains its peak value. The MFV and $\emptyset$ are calculated as

$$\mathrm{MFV} = \frac{\sum_{i=1}^{n} x_i \emptyset_i}{\sum_{i=1}^{n} \emptyset_i} \tag{1}$$

$$\emptyset_i = \frac{\varepsilon^2}{\varepsilon^2 + (x_i - \mathrm{MFV})^2} \,. \tag{2}$$

The expression (x_i − MFV) represents the distance within the cluster, and ε denotes the dihesion, serving as the shape parameter for the weight function. The MFV is iteratively calculated from the combination of Eq. (1) and (2). In the initial iteration, MFV is replaced by the mean of the data vector and ε is computed as

$$\varepsilon_1 \leq \frac{\sqrt{3}}{2}\big(\max(x_i) - \min(x_i)\big). \tag{3}$$

Initially, a larger value is selected for ε, thereby attributing a relatively significant weight even to outliers. Subsequent iteration steps involve the reciprocal derivation of ε and MFV from each other, as per Eqs. (4) and (5), where j is the iteration step

$$\mathrm{\varepsilon_j} = \frac{3\sum_{i=1}^{n} \frac{\left(x_i - MFV_{j-1}\right)^2}{\left[\varepsilon_{j-1}^2 + \left(x_i - MFV_{j-1}\right)^2\right]^2}}{\sum_{i=1}^{n} \frac{1}{\left[\varepsilon_{j-1}^2 + \left(x_i - MFV_{j-1}\right)^2\right]^2}} \tag{4}$$

$$\mathrm{MFV_j} = \frac{\sum_{i=1}^{n} \left[\frac{\varepsilon^2{}_j}{\varepsilon^2{}_j + \left(x_i - MFV_{j-1}\right)^2}\right] x_i}{\sum_{i=1}^{n} \left[\frac{\varepsilon^2{}_j}{\varepsilon^2{}_j + \left(x_i - MFV_{j-1}\right)^2}\right]}. \tag{5}$$

Upon completion of the maximum iterations (j=50), the utilization of a small ε value ensures that data points closer to MFV carry a greater weight, whereas outliers carry a reduced weight. Initially, the MFV weights (Steiner-Cauchy weights) utilize the Euclidean distance (D_E) (Eq. 6). Subsequently, the weighting is adapted, resulting in the robust MFV-CA using Steiner weighted distance (D_{st}) with c^{MVF} as the cluster centroid (Eq. 7).

$$D_E = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{6}$$

$$D_{st} = \sqrt{\left(\sum_{i=1}^{n} \emptyset_i\right)^{-1} \sum_{i=1}^{n} \emptyset_i (x_i - c^{MFV})^2}, \tag{7}$$

where x_i and y_i are the data vectors (i.e. data objects). The optimal number of clusters (k) is determined using the elbow method. This method utilizes a visual representation of the clustering results, where a line plot of the within-cluster sum of squares (WCSS) is analyzed. The WCSS (Eq. 8) represents the total squared distance between each data point and its assigned cluster's centroid. The elbow method identifies the "elbow" point in this plot, where the rate of decrease in WCSS slows down significantly.

$$\mathrm{WCSS} = \sum_{k=1}^{k}\sum_{i=1}^{n_k} \mathrm{D}^2{}_{st}(\mathrm{c^{MFV}}_{k}, x_{\mathrm{i(k)}}) \,, \tag{8}$$

where D_{st} represents the Steiner distance between the k^{th} vector with n_k elements that belong to the k^{th} cluster with k^{th} centroid ($\mathrm{c^{MFV}}_{k}$).

2.2.2 Self-organizing map (SOM)

Self-organizing map (SOM) (Kohonen 1982) is an unsupervised neural network that is used for clustering and visualizing high-dimensional data by projecting it onto a lower-dimensional space. SOM consists of a grid of neurons, typically organized in a 2D lattice. SOM algorithm operates on the principle of competitive learning in which each neuron in the SOM represents a data vector characterized by a weight vector (W_i) of the same dimension as the input data $\mathbf{X} = (x_1, x_2, \ldots x_n)$. The training process involves iteratively presenting the input data to the network, identifying the best matching unit (BMU) which is the neuron whose weight vector is most like the input data, and updating the weights of the BMU and its neighboring neurons based on Euclidean Distance. The BMU can be found as

$$\mathrm{BMU} = \mathrm{argmin_i} = \|\mathbf{X} - \mathbf{W}_\mathrm{i}(\mathrm{t})\|, \tag{9}$$

where $\mathrm{argmin_i}$ is the argument of the minimum over all indices i, $\|\cdot\|$ represents the Euclidean distance, $W_\mathrm{i}(\mathrm{t})$ is the weight vector of node i at time step, t. The BMU is the neuron with the smallest Euclidean distance to the input vector. In other words, it is the neuron that minimizes Eq. 9. Once the BMU is identified, the weight vectors of the BMU and its neighboring nodes are updated to move closer to the input vector. The weight update can be performed as

$$\mathbf{W}_\mathrm{i}(\mathrm{t}+1) = \mathbf{W}_\mathrm{i}(\mathrm{t}) + \eta(\mathrm{t})\mathrm{h_{i,0}}(\mathrm{t})(\mathbf{X}(\mathrm{t}) - \mathbf{W}_\mathrm{i}(\mathrm{t})), \tag{10}$$

where $\eta(\mathrm{t})$ is the learning rate at time step t and $\mathrm{h_{i,0}}(\mathrm{t})$ is the neighborhood function that determines how much the neighboring nodes of the BMU are adjusted. The neighborhood function, typically a Gaussian function, ensures that the BMU and its surrounding nodes are updated, but nodes farther away are updated less as

$$h_{i,BMU}(t) = \exp\left(-\frac{\left\|r_i - r_{i,BMU}\right\|^2}{2\sigma^2(t)}\right), \quad (11)$$

where $r_i - r_{i,BMU}$ is the distance between node i and the BMU in the grid and σ is the neighborhood radius, which decreases over time. The choice of SOM architecture, including the number of neurons and the grid topology, was determined through a series of experiments aimed at optimizing the balance between resolution and generalization.

2.2.3 Csókás method

The Csókás method is an empirically modified version of the Kozeny (1927) and Carman (1937) equations for estimating the hydraulic conductivity based solely on geophysical well logs. Kozeny-Carman equation considers the water density (ρ_w (g/cm^3)), viscosity (μ (Pa·s)), porosity (φ (fraction)), dominant grain size of spherical particles (d (cm)), and normal acceleration of gravity (g (cm/s^2)). By considering a rock as an assembly of capillaries, the hydraulic conductivity (K (cm/s)) can be calculated by Eq. (12). As the Kozeny-Carman approach is based on rock analysis, it is possible to use Eq. (13) to obtain the dominant grain size (d) from the grain size distribution curve as

$$K = \frac{\rho_w g}{\mu}\frac{d^2}{180}\frac{\varphi^3}{(1-\varphi)^2}, \quad (12)$$

$$d = \frac{d_{10} + d_{60}}{2}\sqrt{\frac{d_{10}}{d_{60}}}, \quad (13)$$

where d_{10} and d_{60} in Eq. 6 refer to the grain diameter at 10% and 60% of the cumulative frequency of the grain size distribution curve produced by sieve analysis. The Csókás method is valid in loose and saturated formations with a formation factor of less than 10. The formation factor can be derived using Eq. (14), in which the R_0 is the resistivity of the fully saturated rock and R_w is the resistivity of formation water as

$$F = \frac{R_0}{R_w}. \quad (14)$$

In the laboratory, Alger (1971) found a direct relationship between the grain size of water-saturated sediments and the formation factor. The formation factor of loose sediments was experimentally matched to the effective grain size (d_{10}) obtained by sieve analysis as

$$d_{10} = C_d \log F, \tag{15}$$

where C_d is the site constant and is proposed to be $5.22 \cdot 10^{-4}$ for medium to well-sorted sediments with F of less than 10. The shale volume in the aquifer materials impacts the hydraulic conductivity and effective porosity. To estimate porosity, Archie (1942) proposed an empirical equation as

$$F = a\,\varphi^{-m}, \tag{16}$$

where m is the cementation exponent, which ranges practically from 1.4 to 1.8 for loose sediments, and a is the coefficient of tortuosity ($a \approx 1$). The effective porosity (φ_e) is calculated using Eq. (17) (Schlumberger 1991), while the shale volume (V_{sh}) is calculated using Larionov (1969) equation (Eq. 18) in which the natural gamma-ray intensity (I_γ) is calculated using the Schlumberger (1984) linear formula (Eq. 19)

$$\varphi_e = \varphi * (1 - V_{sh}), \tag{17}$$

$$V_{sh} = 0.083 \left(2^{3.7 I_\gamma} - 1\right), \tag{18}$$

$$I_\gamma = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}}. \tag{19}$$

Consequently, by incorporating Alger's (1971) empirical equation into the Kozeny-Carman equation, Csókás derived Eq. (20) for the estimation of hydraulic conductivity (m/s)

$$K = C_k \frac{\varphi^3}{(1 - \varphi)^4} \frac{\left(\log \frac{R_0}{R_w}\right)^2}{\left(\frac{R_0}{R_w}\,\varphi\right)^{1.2}}, \tag{20}$$

where C_k is the proportionality constant and has a value of $855.7\, C_t C_d^2$. Quantity C_t is a constant that varies with formation temperatures, C_t is equal to $1+3.37*10^{-2}\,T + 2.21*10^{-4}\,T^2$, where T is the formation temperature (in Celsius degrees).

The Logan (1964) method is employed to validate the hydraulic conductivity estimated by the Csókás (1995) model. This method is grounded in the Thiem (1906) solution, applicable to wells pumped at a steady and constant rate within an infinite, homogeneous, and isotropic aquifer. This estimation is reliant on the available data, including discharge (Q), drawdown (s), and aquifer thickness (b). The hydraulic conductivity can be estimated using the Logan method by Eq. (21) and Eq. (22) for unconfined and confined aquifers, respectively

$$\mathrm{K} = 2.43\,\frac{\mathrm{Q}}{\mathrm{s}\,(2\mathrm{b}-\mathrm{s})}, \tag{21}$$

$$\mathrm{K} = 1.22\,\frac{\mathrm{Q}}{\mathrm{s} * \mathrm{b}}\ . \tag{22}$$

2.2.4 Deep Learning (DL) models

Semi-supervised and supervised DL models are used to analyze well-logging data. Semi-supervised models included autoencoder neural networks (AE-NN), while supervised models, such as gated recurrent units (GRU), multilayer perceptron neural networks (MLPNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) networks, were used for both classification and regression tasks.

AE-NN is a class of semi-supervised learning techniques used for dimensionality reduction and feature learning (Géron 2022). The architecture of AE-NN consists of an encoder network, which maps the input data into a lower-dimensional latent space, and a decoder network, which reconstructs the original data from the latent space representation log (LS) (Wang et al. 2014). The encoder network comprises multiple layers of neurons implemented as fully connected feedforward neural networks (Fig. 5). Each layer applies a linear transformation followed by a nonlinear activation function (ReLU) to capture complex relationships within the data. The output layer of the encoder represents the LS, where each neuron encodes a distinct feature of the input data. The output of the encoder (h) can be obtained as

$$\mathbf{h} = \mathrm{f}(\mathbf{W}\mathbf{x} + \mathbf{b}), \tag{23}$$

where **x** represents the vector of input data, **W** is the weight matrix, **b** is the bias vector, and f represents the activation function.

The decoder network mirrors the encoder architecture in reverse to reconstruct the original input data from the LS representation. Similar to the encoder, the decoder comprises multiple layers of neurons with corresponding weights and biases. The output layer of the decoder attempts to reconstruct the input data by applying linear transformations and non-linear activation functions to the LS representation. The output of the decoder (y) can be obtained as

$$\mathbf{y} = \mathrm{g}(\mathbf{W}'\mathbf{h} + \mathbf{b}'), \tag{24}$$

where $\mathbf{W}'$ denotes the weight matrix of the decoder, $\mathbf{b}'$ is the bias vector of the decoder, and g represents the activation function applied element-wise.

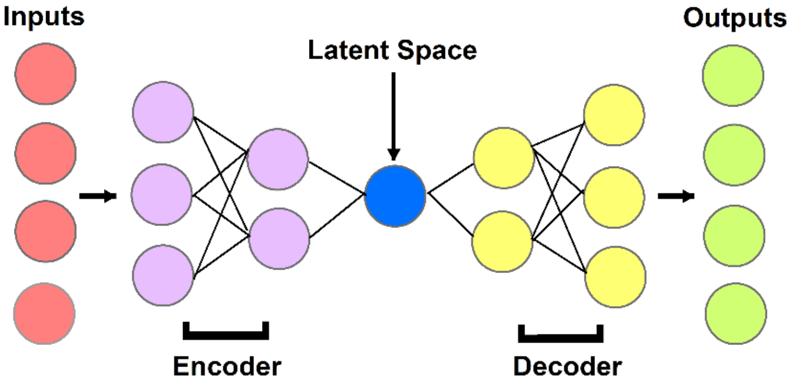
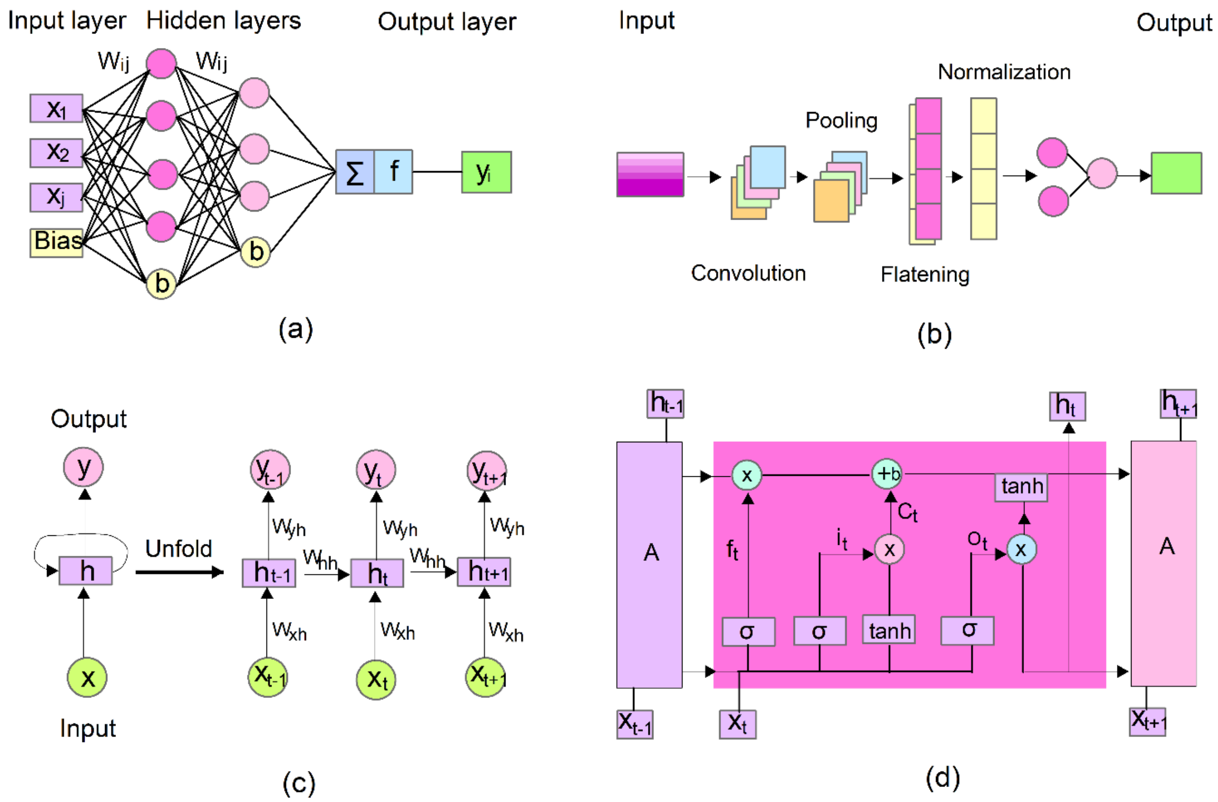


Fig. 5 The structure of the deep autoencoder neural network.

MLPNN employs a feedforward network structure consisting of three key components including the input, hidden, and output layers (Fig. 6a) (Popescu et al. 2009). The input layer acts as the receptor for dataset features, with each neuron ($\mathbf{x}$) representing a distinct feature. The hidden layers are responsible for nonlinear transformations of input data with activation functions (f). The output layer synthesizes the computations of the network, resulting in a single neuron for regression or multiple neurons for classification ($\mathbf{y}$). The MLP processes the information through forward propagation and adapts through backpropagation (Dayhoff 1990). Forward propagation involves the passing of data from the input layer to the output layer, while backpropagation computes the error between predicted and actual outputs using loss functions. The weighting coefficients ($\mathbf{W}$) and biases ($\mathbf{b}$) of neurons are then adjusted iteratively to minimize this error. The output is obtained as

$$\mathbf{y}_{\mathrm{i}} = \mathrm{f}_{\mathrm{i}}\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{x}_{\mathrm{i}}\mathbf{W}_{\mathrm{i}}\right) + \mathbf{b}. \tag{25}$$



(a)
(b)
(c)
(d)

Fig. 6 The typical structure of the deep learning networks of (a) MLP, (b) CNN, (c) RNN, and (d) LSTM.

CNN is a specialized neural network designed for processing grid-like data structures (LeCun et al. 1989). The fundamental operations within CNNs include convolution and pooling (Fig. 6b). The convolutional layers apply filters to input data, to detect local features and patterns. These layers are followed by pooling layers, which reduce the dimensionality of the data, preserving the most critical information while reducing computational complexity. After several convolutional and pooling layers, the output is flattened into a one-dimensional vector and passed to fully connected layers. These layers are like traditional neural networks, where each neuron is connected to every neuron in the previous layer. These fully connected layers interpret the high-level features extracted by the convolutional layers and make final predictions (Zhu et al. 2018).

CNN employs activation functions to introduce nonlinearity that facilitates the modeling of complex relationships within the data.

RNN is used for the analysis and comprehension of sequential data (Rumelhart et al. 1986). They possess a unique architecture that allows them to retain memory and capture temporal dependencies within sequential data. The architecture of RNN involves recurrent connections that help to maintain a form of memory by passing information from one time step to the next (Fig. 6c). At each time step (t), RNN computes an output ($\mathbf{h}_t$) based on the current input ($\mathbf{x}_t$) and the hidden state from the previous step ($\mathbf{h}_{t-1}$). This hidden state sums up information learned from past inputs and facilitates the understanding of context and sequential patterns. Training RNN involves a backpropagation through time (BPTT) algorithm (Lillicrap and Santoro 2019), which is an extension of the backpropagation algorithm adapted for sequential data. BPTT unfolds the network across time steps and facilitates the computation of gradients and subsequent weight updates (W_{hh}, W_{xh}). The output is calculated as

$$\mathbf{h}_t = f(\mathbf{W}_{hh} * \mathbf{h}_{t-1} + \mathbf{W}_{xh} * \mathbf{x}_t + \mathbf{b}). \tag{26}$$

LSTM network is designed to address the limitations of standard RNN in retaining long-range dependencies within sequential data (Hochreiter 1997). The architecture of LSTM networks introduces memory cells that maintain and update information over extended sequences, enabling them to handle vanishing and exploding gradients during training (Fig. 6d). At each time step (t), an LSTM cell operates on three main gates: the input (i_t), forget (f_t), and output gates (o_t), each of which is controlled by sigmoid activation functions (σ). Additionally, the cell state (C_t) and hidden state (h_t) are updated based on these gates and the input (x_t) at a time step. The input gate controls how much new information is allowed into the cell state (Eq. 27). The forgetting gate decides how much of the previous cell state should be forgotten (Eq. 28). The cell state is updated by combining the previous cell state (C_{t-1}) weighted (W) by the forget gate with the candidate values weighted by the input gate (Eqs. 29 and 30). The output gate determines how much of the cell state should be exposed as the hidden state (Eq. 31). Finally, the hidden state is computed by applying the output gate to the cell state (Eq. 32). LSTM networks use sigmoid (σ) and tanh activation functions to regulate information flow through memory cells and hidden states. The sigmoid function is applied in the forget, input, and output gates, determining how much past information is retained, discarded, or passed forward. The tanh function is used in the candidate memory cell state to introduce non-linearity and ensure smooth updates, as well as in scaling the final hidden state.

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi} * \mathbf{x}_t + \mathbf{W}_{hi} * \mathbf{h}_{t-1} + \mathbf{b}_i), \tag{27}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xi} * \mathbf{x}_f + \mathbf{W}_{hf} * \mathbf{h}_{t-1} + \mathbf{b}_f), \tag{28}$$

$$\hat{C}_t = \tanh(\mathbf{W}_{xc} * \mathbf{x}_{\mathbf{t}} + \mathbf{W}_{hc} * \mathbf{h}_{t-1} + \mathbf{b}_c), \tag{29}$$

$$\mathbf{C}_t = f_t * \mathbf{C}_{t-1} + \mathbf{i}_t * \hat{C}_t, \tag{30}$$

$$o_t = \sigma(\mathbf{W}_{x0} * \mathbf{x}_t + \mathbf{W}_{ho} * \mathbf{h}_{t-1} + \mathbf{b}_o), \tag{31}$$

$$\mathbf{h}_t = \mathbf{o}_t * \tanh(\mathbf{C}_t). \tag{32}$$

GRU (Cho et al. 2014) is also a variant of RNNs specifically designed to capture long-range dependencies within sequential data. The GRU comprises a series of interconnected gated units, including input, forget, and output gates (Fig. 7). This gating mechanism allows the GRU to selectively update (z_t) and reset (r_t) its internal state based on the input data and prior knowledge, enabling it to effectively learn complex relationships. The update gate acts as a filter, controlling the flow of past information to the future, while the reset gate selectively prunes unnecessary memories (Chung et al. 2014). By iteratively adjusting the model parameters through backpropagation and gradient descent optimization, the predictive performance of the GRU model is optimized and the trained model is then utilized to predict the output (h_t). The GRU utilizes two main activation functions: sigmoid and tanh, which regulate information flow and enable the network to capture sequential dependencies effectively. The sigmoid activation function (σ) is applied in the update gate and reset gate, outputting values between 0 and 1 to control how much past information is retained or discarded. The tanh activation function is used in computing the candidate hidden state, introducing non-linearity to model complex relationships in sequential data.

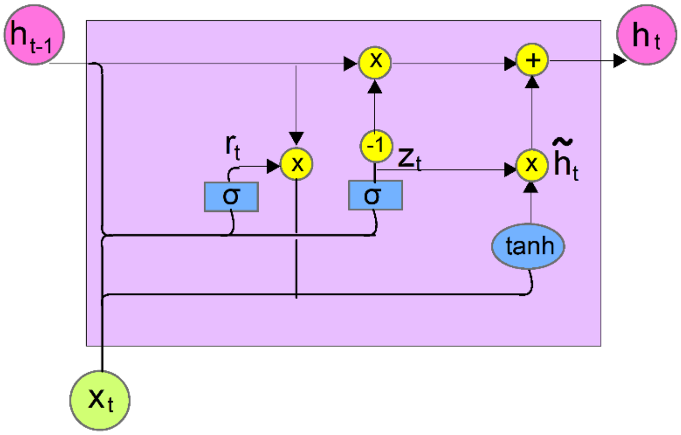


Fig. 7 The structure of the GRU neural network models

The DL models were implemented using Python version 3.7 with Keras and TensorFlow frameworks. The input data was split into an 80:20 ratio for training and validation. The training was optimized using the Adam optimizer. It is an advanced optimization algorithm commonly used in deep learning modeling. It combines the benefits of momentum (which accelerates convergence) and RMSProp (which adapts learning rates) to efficiently update model weights during training (Kingma and Ba 2014). Adam calculates adaptive learning rates for each parameter based on the first moment (mean) and second moment (uncentered variance) of gradients, making it robust for sparse gradients and non-stationary objectives. For classification tasks, the SoftMax activation function was applied in the output layer to generate probability distributions, while ReLU was used in regression tasks to ensure stable gradient propagation. Sparse categorical cross-entropy (SCCE) was used as the loss function for classification (Mannor et al. 2005), effectively handling integer-labeled multi-class outputs. For regression tasks, mean squared error (MSE) was employed to minimize the difference between predicted and actual values, ensuring smooth gradient updates for accurate parameter estimation.

Different Performance evaluation metrics are used to assess the accuracy of models. For classification tasks, accuracy, precision, and recall are used. Accuracy offers a straightforward measure of the overall correctness by quantifying the ratio of correctly predicted instances to the total cases in the dataset. However, its reliability can diminish in scenarios where classes within the dataset are imbalanced. Precision assesses the accuracy of positive predictions, emphasizing the ability of the model to correctly identify positive instances among all instances it predicted as positive. Conversely, recall measures the capacity of the model to capture all relevant positive instances from the entire set of actual positive instances. These metrics are calculated using as

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\sum \mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \cdot 100, \tag{33}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \cdot 100, \tag{34}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot 100, \tag{35}$$

where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative instances, respectively.

The performance of the models for regression tasks is evaluated using mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). RMSE quantifies the average deviation between predicted and actual values, giving larger weight to larger errors due to the squaring operation. Meanwhile, MAE provides a more straightforward measure of average prediction errors, being less sensitive to outliers compared to RMSE. For understanding the explanatory power of regression models, the R^2 reveals the proportion of variance in the target that is explained by the features. These metrics are calculated using as

$$\mathrm{RMSE} = \frac{1}{n} \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{36}$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|, \tag{37}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})^2}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}} \right)^2, \tag{38}$$

where x_i and y_i are the ith values of the actual and predicted values, respectively, n is the number of observations, and $\bar{y}$ and $\bar{x}$ are the mean of the predicted and actual values, respectively.

2.2.5 Particle swarm optimization (PSO)

Particle swarm optimization (PSO) suggested by Kennedy and Eberhart (1995) is a population-based optimization algorithm inspired by the social behavior of bird flocking and fish schooling. PSO is known for its simplicity and ability to handle optimization problems with continuous and discrete variables (Holland 1992). In PSO, a population of candidate solutions (particles) moves through the search space to find the optimal solution. Each particle represents a potential solution to the optimization problem and maintains its position and velocity in the search space. At the start of the optimization process, particles are initialized with random positions and velocities within the search space. The update of particle position and velocity is guided by personal best (pbest) and global best (gbest). The pbest represents the best solution that each particle has individually discovered during its search. The swarm keeps track of the gbest solution found by any particle in the entire population. During each iteration, the velocity of each particle $\boldsymbol{v}_i = [v_{i1}, v_{i2}, \dots\ v_{id}]$ is updated based on its current velocity, its distance to its personal best, and its distance to the global best. This update encourages particles to move towards promising regions of the search space. After updating velocities, particle positions $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \dots\ x_{id}]$ are adjusted accordingly. This process is repeated for a predefined number of iterations or until a termination condition is met. The movement of each particle is governed by its velocity (Eq. 39) and location (Eq. 40) as

$$\mathbf{v}_\mathrm{i}(\mathrm{t}+1) = \mathrm{w}\mathbf{v}_\mathrm{i}(\mathrm{t}) + \mathrm{c}_1\mathrm{r}_1(\mathbf{p}\mathrm{best}_\mathrm{i} - \mathbf{x}_\mathrm{i}) + \mathrm{c}_2\mathrm{r}_2(\mathbf{g}\mathrm{best}_\mathrm{i} - \mathbf{x}_\mathrm{i}), \tag{39}$$

$$\mathbf{x}_\mathrm{i}(\mathrm{t}+1) = \mathbf{x}_\mathrm{i}(\mathrm{t}) + \mathbf{v}_\mathrm{i}(\mathrm{t}+1), \tag{40}$$

where w is the inertia weight that controls the trade-off between global and local search (Shi and Eberhart 1998), c_1 and c_2 are acceleration coefficients representing the cognitive and social components of the movement and set as 2 (Kennedy and Eberhart 1995), and r_1 and r_2 are random values sampled from the uniform distribution in the range [0, 1]. PSO is employed to refine the regression parameters governing the correlation between the LS representations obtained from the AE-NN and shale volume and hydraulic conductivity. PSO aims to optimize the fit of the AE-NN-based model to the observed data, minimizing L_1 norm objective (Eq. 37)

2.3 Numerical flow simulation

A steady-state numerical flow model is constructed to simulate the groundwater flow within the Quaternary aquifer system in which the inputs for the model are solely derived from the analysis of the well logging data. The groundwater modeling is performed in three stages including conceptualization, formulation of the mathematical model, and the translation of the conceptual model into a mathematical model (Anderson et al. 2015). The conceptualization of the system involved characterizing lithological and hydraulic properties. The objective is to create a robust framework that effectively represents the real-world complexities of the groundwater system (Bear and Verruijt 2012). Consequently, the flow simulation was performed within the groundwater modeling system (GMS) program operating under modular groundwater flow model - unstructured grid (MODFLOW-USG) (Panday et al. 2013) developed by the U.S. Geological Survey (USGS). This program allows simulation of groundwater flow and contaminant transport with control volume finite difference (CVFD) unstructured grids that divide the model domain into discrete control volumes.

The conceptual model is translated into mathematical models by assigning boundary conditions and hydraulic stresses. Boundary conditions were defined to simulate the interaction between the aquifer system and the surrounding environment for accurately simulating hydrological processes (Mehl and Hill 2006). Different types of boundary conditions are available for groundwater flow modeling including specified head boundary (Dirichlet boundary), specified flow boundary (Neumann boundary), and specific head and flow boundary (Cauchy boundary) (Jazayeri and Werner 2019). Dirichlet boundary is employed to constrain the simulation of the hydraulic head along the boundary of the model domain while the Neumann boundary is used to represent the inflow through recharge and outflow through discharge zones using the appropriate packages. The calibration of the model is conducted using the parameter estimation (PEST) program within the MODFLOW-USG. The calibration involved adjusting model parameters to minimize the difference between simulated and observed groundwater levels.

3. RESULTS AND DISCUSSION

3.1 Analysis of well-logging data

3.1.1 Preprocessing of well logging data

The geophysical well logging data used in this thesis (Fig. 4) excludes the deep resistivity (RD) log, as the logging was primarily conducted for well-design purposes rather than for comprehensive subsurface characterization. However, this log provides information about the intrinsic properties of subsurface materials beyond the invaded zone, making it important for the estimation of hydrogeological properties. The RD log is predicted based on the available SP, NGR, and RS logs using the GRU neural network. Initially, the probability density functions (PDF) of the input logs were analyzed (Fig. 8) to assess their variability. The distribution of SP and NGR logs is symmetrically distributed around the mean, while the RS log is right skewed. The symmetric distribution of SP and NGR logs ensures balanced feature representation, while the right-skewed RS log may introduce bias in the GRU model. This requires normalization to ensure balanced representation and improve the model performance. The correlation analysis between well logs used for predicting the RD log revealed important interdependencies (Fig. 9a). The RS log showed a weak to moderate negative correlation (-0.38) with the SP log, while a stronger negative correlation (-0.75) was observed between RS and the NGR log. Additionally, the correlation between SP and NGR was moderate (0.41). Higher correlations between input logs improve prediction accuracy, while lower correlations can limit model performance. Due to the presence of moderate to weak correlations in the dataset, a GRU neural network was used for its ability to model complex, nonlinear relationships, enhancing the prediction of the RD log.

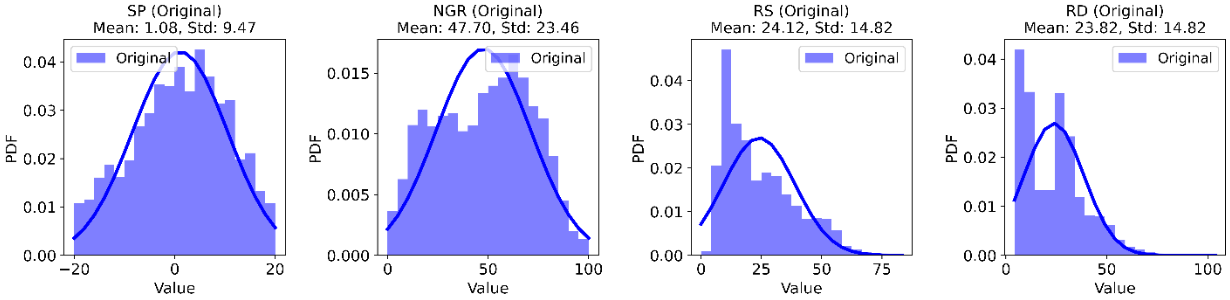
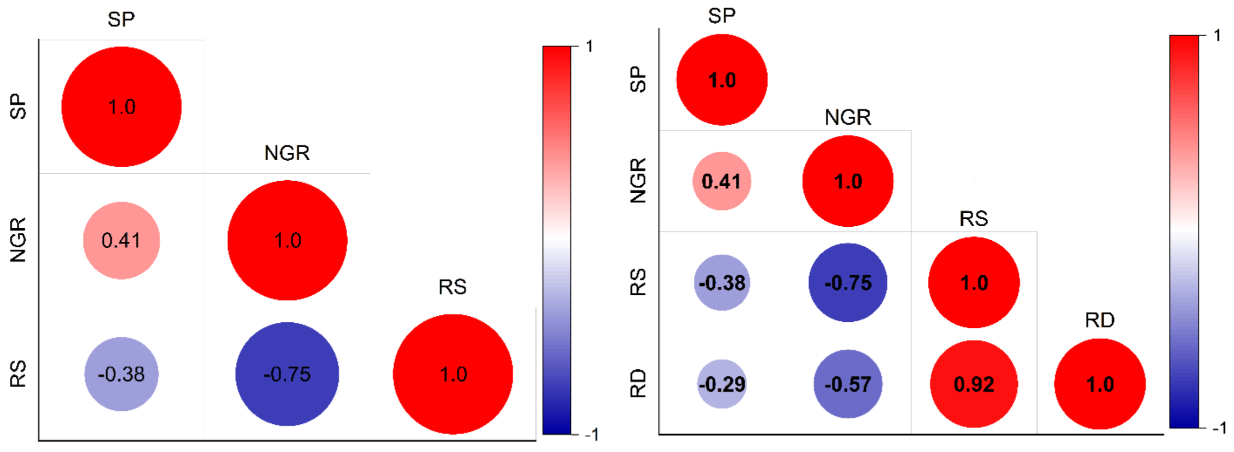


Fig. 8 The probability distribution functions of the well-logs. The y-axis represents the likelihood of log response (integral over the entire range sums to 1).



(a)

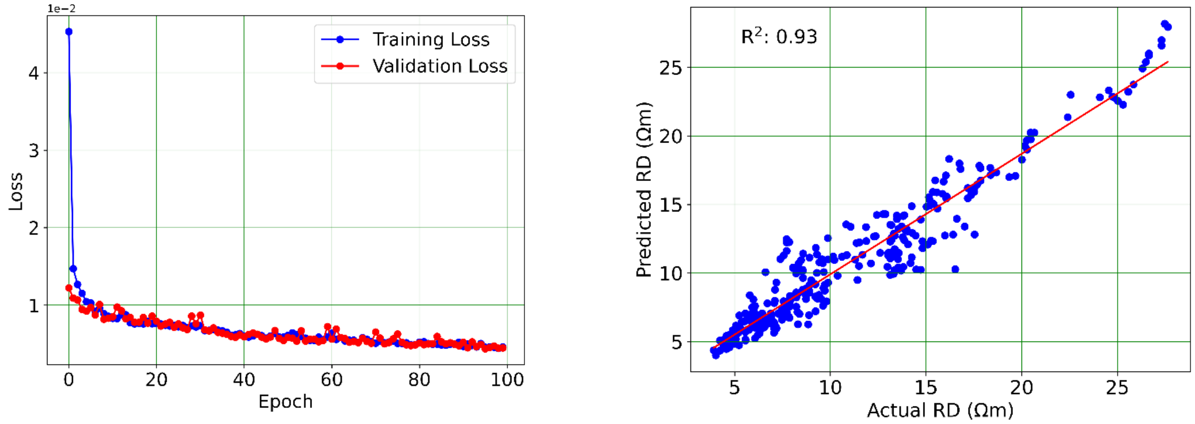
(b)

Fig. 9 Correlation matrix showing the strenght of correlation between (a) the input logs used for the prediction of RD log and (b) the correlation between the input logs and predicted RD log

The GRU model was trained on a borehole with a complete log suite (D14 borehole) that is located in the central part of the study area (Fig. 4). This borehole is selected to allow the GRU model learning from a representative dataset that captures the typical geophysical characteristics of the Quaternary aquifer system. The trained model was then applied to other wells across the study area to predict missing RD logs. The architecture of the GRU model, including the number of hidden layers and units, was determined through a trial-and-error approach to identify the most effective configuration for capturing complex dependencies within the data. Multiple architectures were tested, including variations in the number of GRU layers (ranging from one to three) and different unit sizes. While deeper architectures with additional layers showed marginal improvements in representation learning, they increased computational complexity and the risk of overfitting. Similarly, smaller unit sizes led to underfitting, reducing the model's ability to extract features. After evaluating various configurations, the optimal structure was found to consist of two GRU layers with 100 and 50 units, respectively. To mitigate overfitting, a dropout rate of 0.1 was applied, randomly omitting a fraction of neurons during training to improve model robustness. A learning rate of 0.0002 was selected to ensure stable and efficient convergence. The GRU model was trained and validated over 100 epochs, ensuring sufficient learning without excessive training cycles that could lead to overfitting.

The representation of the training and validation is shown in Fig. 10a. The typical loss pattern for training and validation throughout the epochs indicated that the model is learning effectively, generalizing well to new data, and not overfitting. Accordingly, the logs generated with GRU showed a strong correlation with actual RD measurements in validation wells, with an average determination coefficient (R^2) of 0.93 (Fig. 10b), a root mean square error (RMSE) of 0.06 Ωm, and a mean absolute error of 1.07 Ωm. The PDF of the predicted RD log demonstrated almost similar behavior to that of the RS log (Fig. 8). The correlation between the input logs (SP, NGR, and RS) and the predicted RD log is illustrated in Fig. 9b. The RS and RD exhibit a strong correlation of 0.92, while SP has a weaker negative correlation with RD (-0.29). NGR, on the other hand, shows a moderate negative correlation with RD (-0.57). The strongest relationship between RS and RD demonstrates that shallow resistivity is the most critical predictor of deep resistivity. This high correlation is expected because both logs measure formation resistivity, with RD capturing deeper, less invaded zones. The SP has a low correlation with RD because its response is influenced by electrochemical effects. Therefore, its signal is often affected by noise from borehole conditions and shale effects making its relationship with resistivity indirect (Doll 1949).

An example of the 1D comparison between the actual and predicted RD log is illustrated in Fig. 11 where a strong correlation is indicated. However, some limitations were observed, particularly where abrupt increases in the RS log values occurred, resulting in the peaks of the RD log. These areas exhibited relatively higher fluctuations between the actual and predicted resistivity values, suggesting that the model struggled to accurately represent sudden changes in the RD log, likely due to the inherent complexity of these transitions. Despite these fluctuations, the overall performance of the GRU network can be considered successful for RD log prediction to be used for further analysis. Additional data is required to handle these extreme variations more accurately. This hybrid model is valuable in scenarios where direct measurements of certain well logs are sparse, difficult to obtain, or costly. The ability to predict missing logs from other well-log data offers a cost-effective and time-efficient solution, making it especially beneficial for large-scale regional studies where extensive well-logging campaigns may be impractical. In this thesis, the accurate prediction of RD logs will facilitate the characterization of the subsurface, allowing for a reliable estimation of aquifer properties beyond the invaded zone.



(a) (b)

Fig. 10 (a) Representation of the training and validation for the GRU model throughout the epochs and (b) the correlation between the actual (observed) and predicted deep resistivity (RD) log.

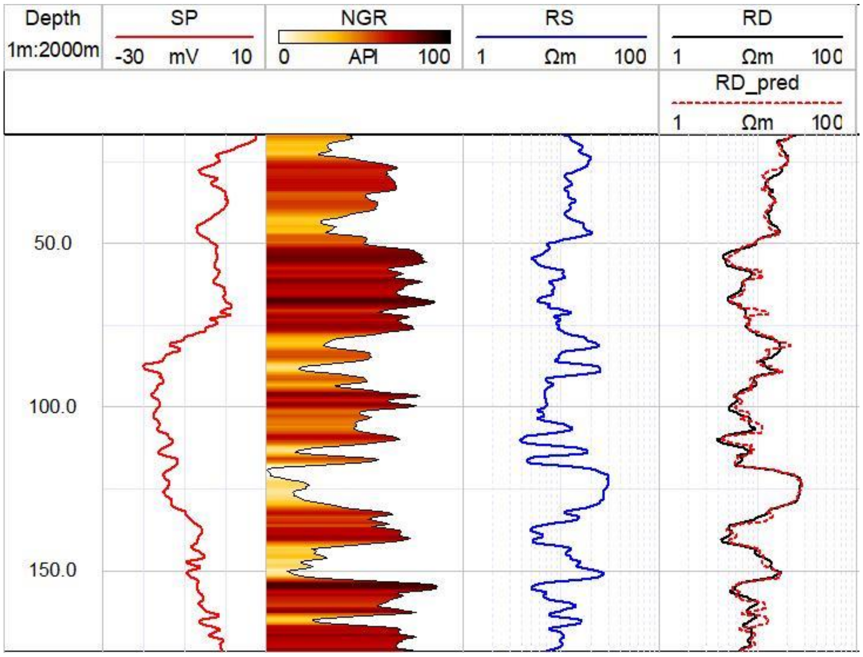


Fig. 11 The results of deep resistivity (RD) log prediction along the B2 borehole (see Fig. 4).

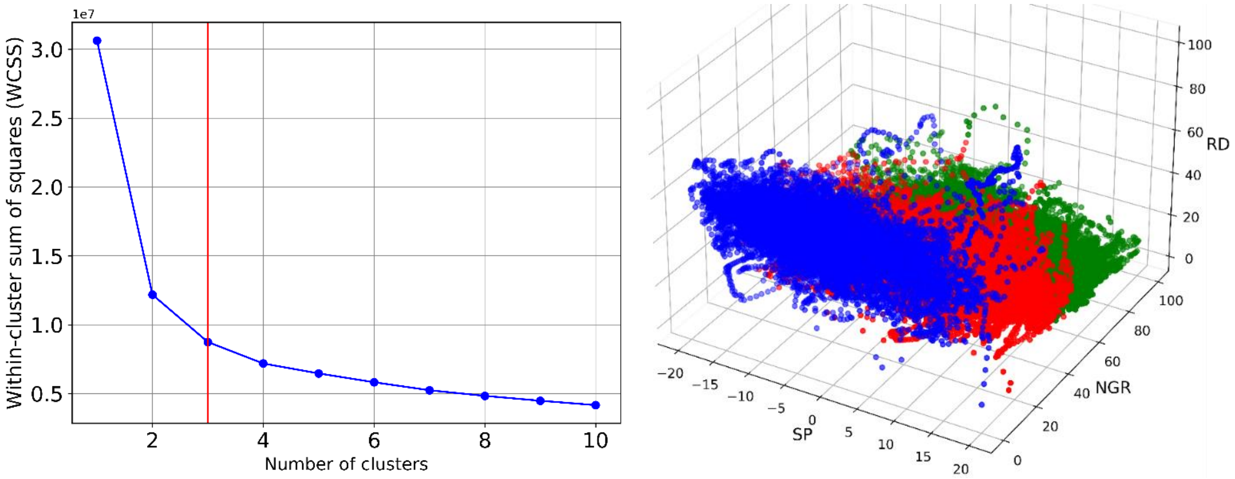
3.1.2 Geological modeling

3.1.2.1 Lithology by clustering methods

The identification of lithological variation is crucial for understanding the complexities within the groundwater systems. For this purpose, the well logs including SP, NGR, RS, and RD logs are analyzed using unsupervised machine learning methods including MFV-CA and SOM, to characterize the main hydrostratigraphical units of the Quaternary aquifer system. Direct visual interpretation of well logs can effectively identify permeable zones in nearly homogeneous systems. However, interpreting lithological variations in heterogeneous systems is challenging, leading to potential biases and inconsistencies. These arise from the subjective nature of manually identifying layer properties and continuity. Moreover, identifying lithology from individual well logs can often lead to misinterpretations. Therefore, I propose an automated approach that jointly analyzes multiple well logs using clustering methods to infer lithological information. This approach detects lithology based on the physical responses of the formations recorded in the logs allowing more consistent identification of the lithological variations.

Before starting the clustering analysis, the selection of the optimal number of the clusters (i.e. lithology) is essential because a higher number than the optimal gives lithologies that do not exist in the aquifer system. On the other hand, fewer clusters neglect lithological units during the discretization, resulting in low vertical resolution (Szabó et al. 2021). Assigning lithologies to the clusters based solely on mathematical principles can be speculative without additional context about the geological characteristics of the aquifer system. As a result, the optimal number of clusters is determined based on the elbow method and previous geological information. The elbow method identified three clusters as the optimal number, as it resulted in a low within-cluster sum of squares (Fig. 12a). Geologically, the Quaternary sediments in eastern Hungary are described as a complex fluvial system predominated by clastic sedimentary rock. The variations in the fluvial transport capacity impacted the homogeneity and the spatial continuity of the fluvial sediments, resulting in wide variation in the textures of the rock from clay to gravel (Tóth and Almási 2001). Accordingly, three clusters representing three clastic lithologies are accepted as the optimal cluster number.

The formation of the clusters using MFV-CA is illustrated in Fig. 12b. Cluster 1 displayed high NGR responses, along with low RS and RD values. This indicates more conductive materials with higher NGR readings due to the presence of radioactive minerals. Cluster 2 exhibited moderate NGR, RS, and RD responses, suggesting a balanced combination of resistivity and radioactivity. Cluster 3 exhibited low NGR responses with high RS and RD, typically associated with more resistive, non-radioactive formations. Clustering is less dependent on the SP log because its response is primarily influenced by electrochemical potential rather than consistently reflecting lithology. This variability leads to less distinct clustering patterns when compared to logs with clearer physical significance for differentiating lithologies. Based on these log responses, Cluster 1 was assigned to clay, Cluster 2 to clayey sand, and Cluster 3 to sand and gravel. Clay, known for its fine-grained, impermeable nature, displayed high NGR readings due to its higher radioactive element content. Clayey sand, as a transitional lithology, showed intermediate NGR values influenced by grain size and porosity. The sand and gravel lithology, with coarser grain size and higher permeability, exhibited lower NGR values compared to clay. These clusters reflect distinct geological characteristics, including variations in mineral composition and grain size (Mohammed et al. 2024d).



(a) (b)

Fig. 12 (a) Selection of the optimal number of clusters using the Elbow method. (b) Clusters represented in the data space show the physical properties of clay (green), clayey sand (red), and sand and gravel (blue), respectively.

The SOM is further applied to map lithological variations within the aquifer system to reduce the ambiguities of the visual interpretation. The SOM consists of a 1x3 hexagonal grid of neurons (Clusters) with weight associated with each neuron. The hexagonal structure in SOM analysis is preferred over others as it allows for better preservation of neighborhood relationships between data points. The hit counts of neuron 1 are 4953 (34%), for neuron 2 is 6265 (44%), and 3127 (22%) hits for neuron 3 (Fig. 13). The hit counts suggest that more data points are being mapped to neuron 2, indicating that the majority of the input data shares characteristics that are best represented by neuron 2. The weight planes of SP, NGR, RS, and RD contribute to the clustering pattern across the three neurons (Fig. 14). Neuron 1 displays low SP values, low RS and RD, and NGR readings, indicating conductive, clay-rich formations with elevated radioactivity consistent with clay. Neuron 2 shows moderately increasing SP and resistivity values alongside intermediate NGR, reflecting a mix of clay and sand typical of clayey sand. Neuron 3 further elevates SP and resistivity while reducing NGR, suggesting sand-dominant facies with residual clay influence. This indicates a strong correspondence between the SOM neurons and the MFV-CA. Neurons 1, 2, and 3 of the SOM align with Cluster 1, 2, and 3 of the MFV-CA, representing clay, clayey sand, and sand and gravel, respectively.

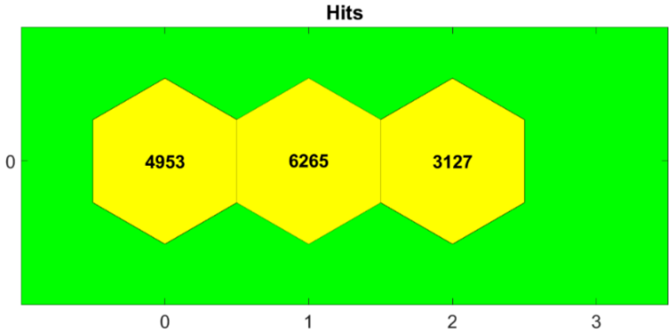


Fig. 13 The hits of each neuron used to perform the mapping

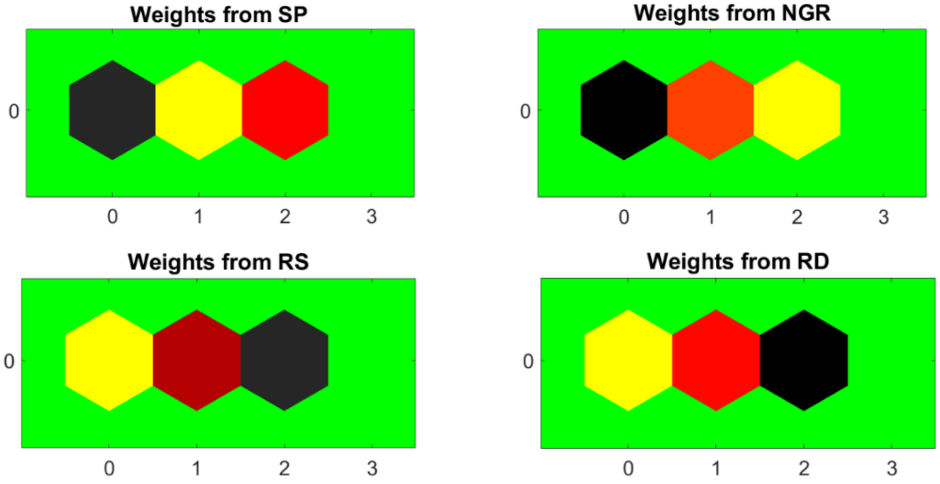


Fig. 14 The weight planes of three neurons for SP, NGR, RS, and RD logs.

Accordingly, the well-logging data are analyzed with MFV-CA and SOM and the results obtained are then compared to those derived from the traditional k-means CA. The results of the 1D analysis showed that traditional k-means clustering exhibited limitations in well-log analysis (Fig. 15). The sensitivity of the k-means CA to data noises affected the discrimination between clay and clayey sand formations, especially in noisy NGR and low RS responses (Mohammed et al. 2024e). This limitation arises because the k-means algorithm relies on a least-squares procedure, which assumes a Gaussian error distribution. Consequently, its performance reduces when facing heavy-tailed error distributions (Szűcs et al. 2006). This results in delayed identification of layer boundaries and inaccurate estimations of effective layer thickness. The MFV-CA demonstrated marked improvements over k-means clustering. It provided enhanced robustness against various error distributions, showing high noise rejection capabilities. SOM also generated a more geologically realistic result. The method captured the logical variations in well-log responses, particularly in zones with complex lithology.

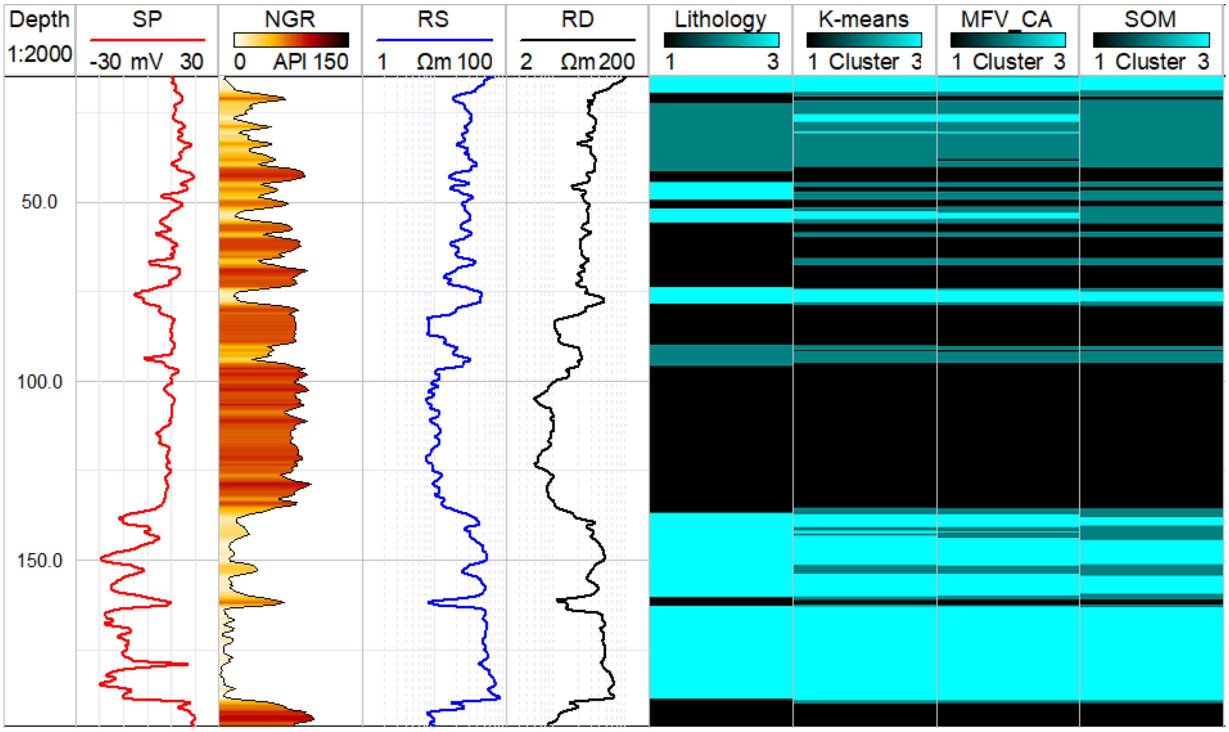
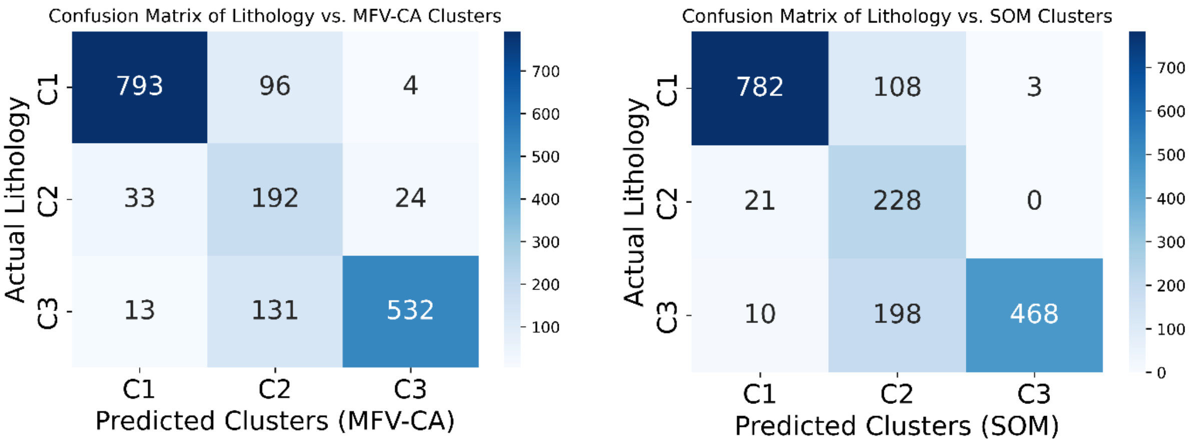


Fig. 15 An example of the 1D distribution of the lithological clusters using k-means cluster analysis (6th track), MFV-CA (7th track) SOM (8th track), and obtained from the interpretation of well logs (1-4 track). The actual lithology obtained from sampling is shown in the 5th track.

The performance of the MFV-CA and SOM models in inferring the actual lithology was evaluated using confusion matrices against the drilling-based lithologies obtained from limited boreholes (Fig. 16). The confusion matrices illustrate the relationship between predicted and actual classifications, displaying both correct and incorrect predictions. Each row represents the actual lithology obtained from drilling, while each column represents the predicted lithology from clustering techniques. The MFV-CA model exhibited strong predictive performance, correctly classifying 793 clayey samples, 192 clayey sand samples, and 532 sand and gravel samples. However, 96 clayey samples were misclassified as clayey sand, and 4 as sand and gravel. Similarly, 33 clayey sand samples were misclassified as clay, while 24 were mistaken for sand and gravel. For the sand and gravel unit, 131 samples were misclassified as clayey sand, and 13 as clay. The SOM model showed comparable performance, with 782 clay samples, 228 clayey sand samples, and 468 sand and gravel samples correctly identified. However, 108 clay samples were misclassified as clayey sand, and 3 as sand and gravel. Additionally, 21 clayey sand samples were misidentified as clay, while 198 sand and gravel samples were mistaken for clayey sand, and 10 were misclassified as clay. Both methods demonstrated a high capability in distinguishing between lithologies with an accuracy exceeding 80%. The misclassifications primarily occur in the transitional clayey sand unit. The MFV-CA method showed better performance in classifying sand and gravel, while the SOM model demonstrated a stronger ability to identify clayey sand correctly.



	C1	C2	C3
C1	793	96	4
C2	33	192	24
C3	13	131	532

(a)

	C1	C2	C3
C1	782	108	3
C2	21	228	0
C3	10	198	468

(b)

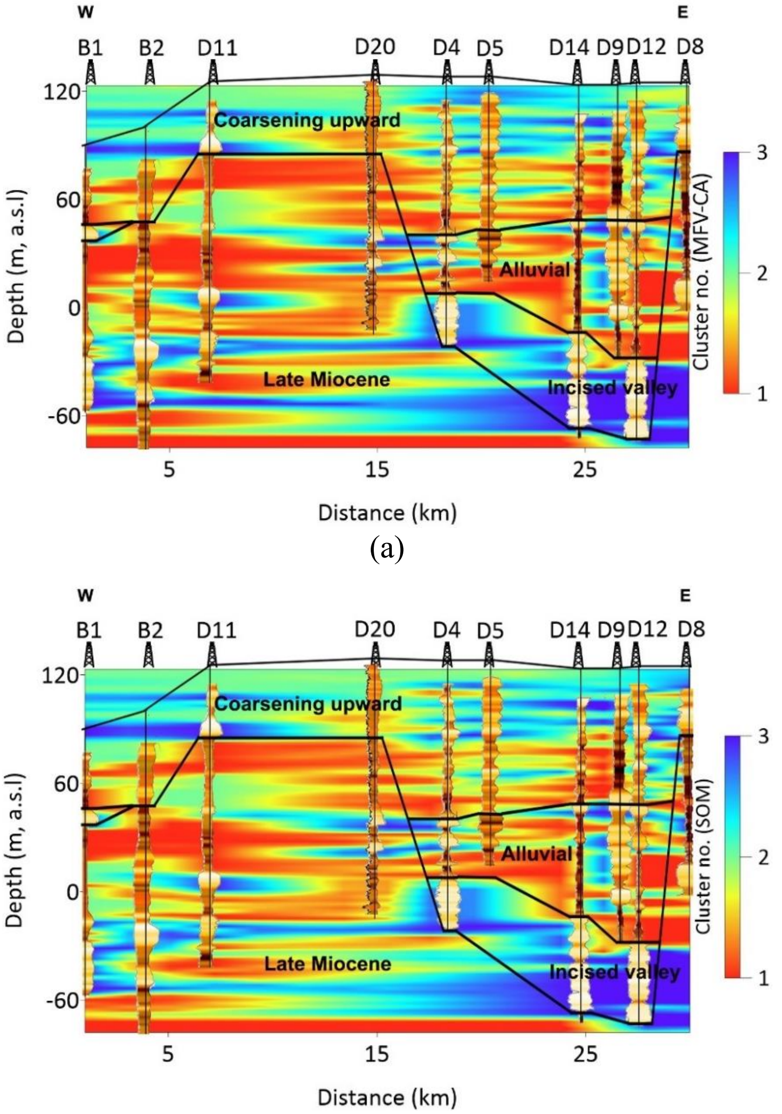
Fig. 16 The confusion matrices between the actual lithology from sampling and predicted clusters using (a) MFV-CA and (b) SOM.

A comparative 1D analysis between the clustering results and the lithology acquired during drilling revealed a generally good agreement in delineating layer boundaries (Fig. 15). However, some disparities were identified. For instance, within the clayey and sandy layers, clustering methods showed a more detailed lithological variation within each layer than indicated by the lithological logs. Nonetheless, in some cases, the lithological logs gave a more detailed description of the lithology along the boreholes. The correlation between the lithology obtained from the drilling and their assigned clusters is illustrated in Table 1. The clustering-based identification of clayey sand lithology encompasses a combination of lithologies, including silty sand, silty clay, sandy clay, clayey sand, and loess. The cluster-based sand and gravel lithology also included silty sand and silty clay of the sampling-based lithology. Therefore, it can be indicated that the limitation of clustering lies in its potential challenge in depicting subtle variations in lithology, especially in distinguishing between closely related compositions such as silty sand and clayey sand lithologies. Despite this limitation, clustering methods demonstrate their effectiveness in simplifying lithology by grouping similar lithological compositions. This highlights the utility of clustering in categorizing lithologies with shared characteristics into broader lithological units for enhanced interpretation and analysis.

Table 1 Comparison and assignment of the lithological logs to the obtained clusters.

Lithology code	Lithology	Cluster code	Cluster litho
1	Clay	1	Clay
2	Sand and gravel	3	Sand and gravel
3	Silty sand	3	Sand and gravel
4	Silty clay	2,3	Sand and gravel and clayey sand
5	Sandy clay	2	Clayey sand
6	Clayey sand	2	Clayey sand
7	Loess	2	Clayey sand

The clustering-based lithology is firstly obtained in 1D along a borehole and the results are interpolated using the ordinary kriging method (Oliver and Webster 1990) to reveal the 2D distribution of the lithological unit along a profile. To preserve the sharp boundaries between the three clustered lithological classes and mitigate artifacts from the inherent spatial continuity assumption in the kriging method, I optimized the visualization by applying a color relief approach. This adjustment emphasizes discrete transitions between clusters, ensuring the final representation aligns with the categorical nature of the clustered data. However, the 2D interpolation of discrete data using the kriging method remains subjective and requires careful interpretation. The 2D analysis demonstrated a nearly similar representation of the groundwater system between the MFV-CA and SOM. The distribution of the extracted lithological clusters along the hydrostratigraphical units is presented in Fig. 17. In this profile, the uppermost part of the profile at a depth between 120 to 45 m a.s.l consists mostly of clayey and clayey sand layers, which are gradually transitioned into a more significant sand layer following an intercalation pattern. However, the sand layers are intermitted and disappear in the central part of the study area. These lithologies make the upper hydrostratigraphic unit of the study area referred to as the coarsening upward unit (CUU), which typically exhibits certain log patterns. In the RS log, a gradual increase in resistivity values upward is observed, while in the NGR and SP logs, a gradual decrease upward is detected, indicating the transition from finer-grained to coarser-grained sediments. CUU hosts the upper unconfined to semi-confined top aquifer unit in the study area (Carpio 2024).



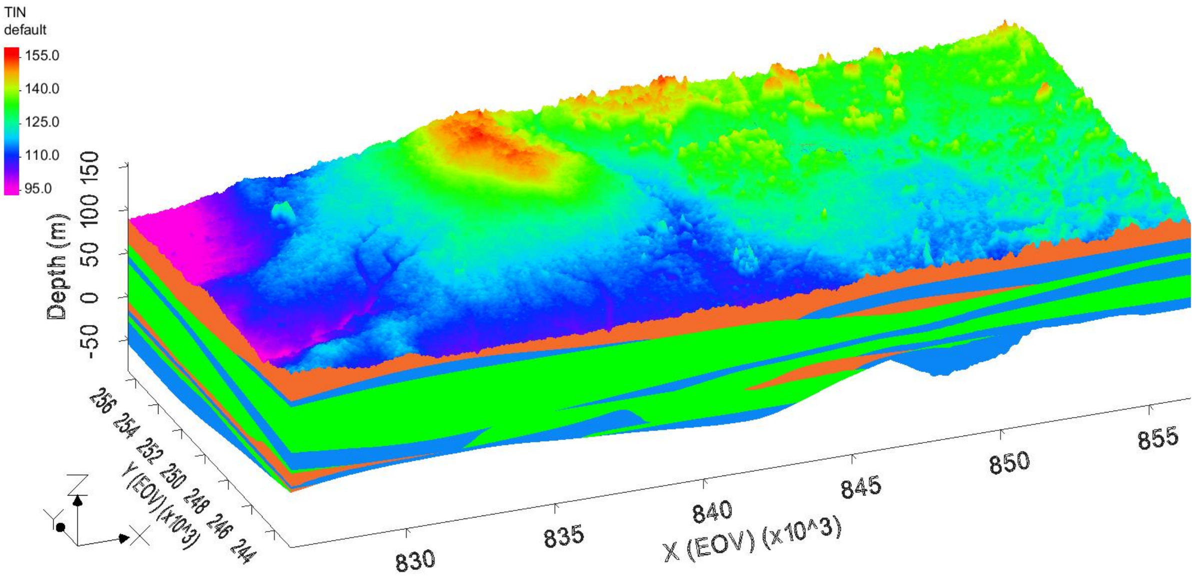
(a)

(b)

Fig. 17 Results of 2D clustering along the profile using (a) MFV-CA, (b) SOM. Clusters 1, 2, and 3 are assigned to clay, clayey sand, and sand and gravel.

The eastern part of the study area is characterized by alternating layers of sand, clayey sand, and clay; however, it is dominated by clays (Mohammed et al. 2024e). These lithologies showed discontinuous patterns as they disappeared in the central and western parts of the area and were described as alluvial units (AU). The AU is deposited in the Lower Pleistocene, and its thickness varies between 25 to more than 100 m. The boundary between the CUU and the AU can be identified by the change in the thickness of the finer sediments. As we go deeper (25 – 70 m b.s.l), an almost continuous unit of sand and gravel layers of an average thickness of 70 m is observed in the eastern parts of the study area. Previous research indicated that this unit is also of Lower Pleistocene and fills up the incised valley unit (IVU) eroding the basal unit in the eastern part of the Debrecen area. IVU is mostly composed of coarse sand and gravel and when it exists it is considered the main aquiferous unit in the region (Püspöki et al. 2016). The basal unit is the Late Miocene unit (LMU), which is encountered mostly in limited boreholes in the western parts of the area. The surface of this unit coincides with the surface of the Zagyva Formation, which is composed of alluvial plain deposits of fine sand, silt, and clay (Buday et al. 2015).

The results of the 1D MFV-clustering were then converted into a 3D geological model (Fig. 18). The analysis delineated 13 layers. The upper part of the system is composed of CUU and AU while the bottom of the system comprises continuous coarse grain of IVU. The heterogeneity in the distribution of the lithological units within the hydrostrsigtaphical units is observed. These units are primarily non-continuous, with the CUU, AU, and IVU concentrated in the eastern parts of the area and the LMU in the western parts. The observed discontinuity in sedimentation patterns is a result of varied depositional characteristics influenced by a range of geological and environmental factors over time (Varsányi et al. 2011). Key processes such as tectonic uplift and subsidence play a crucial role in shaping the landscape by creating valleys and basins, which in turn affect the sedimentation dynamics of alluvial deposits, valley incision, and Late Miocene formations (Tóth and Almási 2001). The validity of the constructed 3D geological model based on the MFV-CA is further confirmed by the alignment of screen intervals within the wells with the clayey sand and sand and gravel intervals identified in the model. The advantage of using clustering techniques lies in its full automation, which yields more consistent information compared to sampling methods in capturing the broader lithological variation in the groundwater system. The resulting classifications served as a robust foundation for developing a conceptual model that accurately captures the lithological variability of the aquifer system, ensuring a realistic representation for further groundwater flow analysis.



(b)

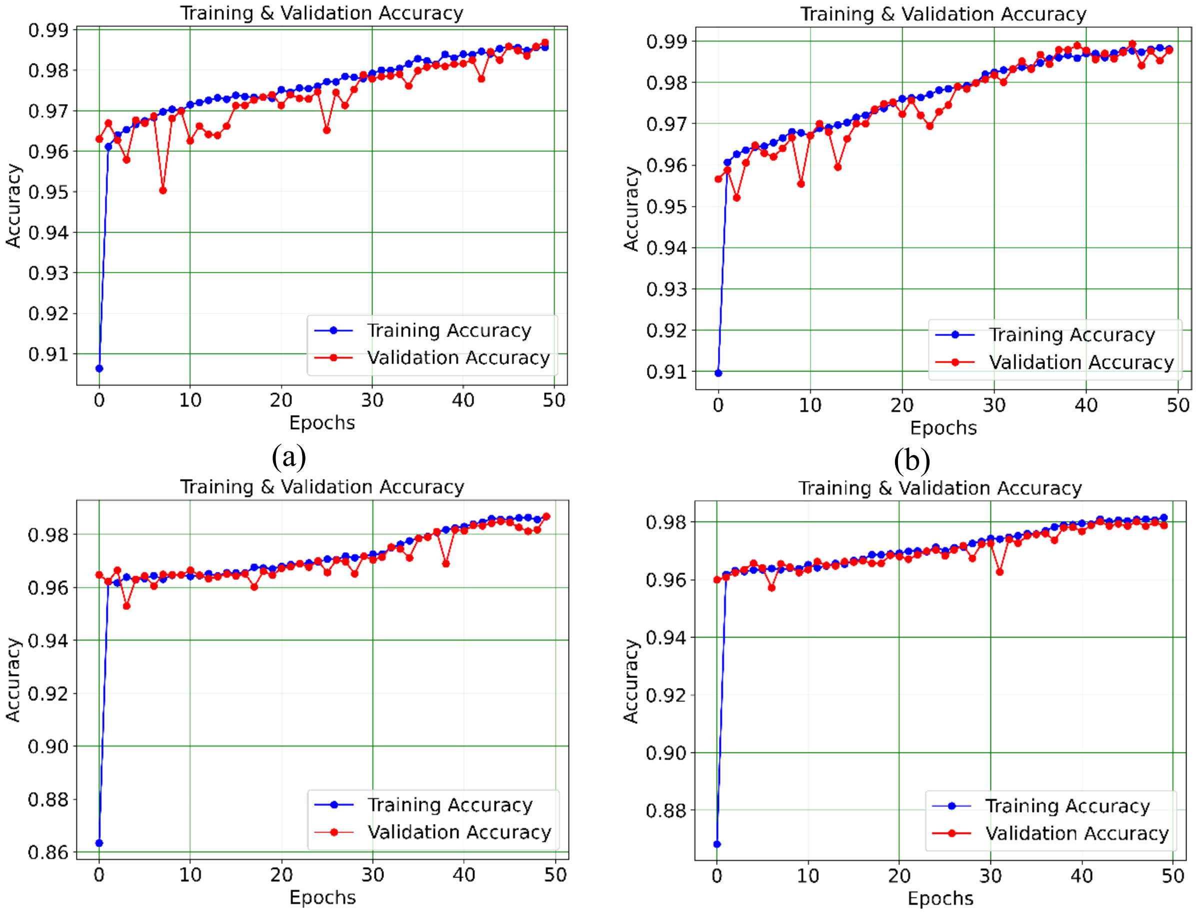
Fig. 18 The 3D geological model obtained from the 3D clustering shows the vertical and horizontal lithological variation in the groundwater system. Clay, clayey sand, and sand and gravel lithologies are represented by green, orange, and blue colors, respectively. The depth is above sea level.

3.1.2.2 Lithology by DL models

The MFV-CA and SOM methods successfully characterized the main hydrostratigraphical units within the aquifer system. However, these methods require considerable computational resources and extended processing times when applied to big datasets. The MFV-CA required over 10 minutes to cluster the dataset, while the SOM completed clustering in approximately 5 minutes. However, this processing time can increase significantly, especially when hyperparameter optimization is necessary and the procedure must be repeated. This limitation underscores the need for more time-efficient methods. To ensure rapid and accurate lithological characterization of groundwater systems, I tested various DL models, including MLPNN, CNN, RNN, and LSTM. Each model was trained to predict lithological clusters. Initially, a series of hyperparameter tuning experiments were conducted to identify the optimal architecture and training configuration for each model. The variations in the number of units, dropout rates, and learning rates across different models are due to the distinct characteristics and learning dynamics of each model type, as well as their specific requirements for optimal performance.

For the MLPNN, the hidden layers were configured with 120 and 60 neurons, respectively. A dropout rate of 0.2 was applied to regularize the model, reducing over-reliance on individual neurons and enhancing generalization. The learning rate was set at 0.001 to ensure stable convergence, avoiding the risk of overshooting the optimal solution. In the case of the CNN, 64 filters with a kernel size of 3 were selected to effectively capture fine-grained patterns within the well-log sequences. The fully connected layers were configured with 128 and 64 neurons, providing adequate capacity for extracting high-level features. Given the tendency of CNNs to overfit on small datasets, a dropout rate of 0.3 was applied for regularization. To capitalize on CNN's inherently faster convergence, a slightly higher learning rate of 0.005 was used. For the RNN architecture, which is prone to the vanishing gradient problem (Hochreiter 1997; Mohammed et al. 2024h), 150 and 75 units were selected to balance model complexity and stability. A dropout rate of 0.15 was used to prevent overfitting, and a learning rate of 0.003 was chosen to accelerate adaptation in sequential learning tasks. Lastly, the LSTM networks, known for effectively capturing long-term dependencies, were configured with 100 and 50 units. A dropout rate of 0.25 was applied to improve generalization, while a learning rate of 0.001 enabled fine-tuned updates during training, ensuring both stability and accuracy in learning temporal relationships.

The representation of training and validation for the models is illustrated in Fig. 19. The accuracy metric defined in Eq. 33 is used as a loss function. In the case of MLPNN, the accuracy of the model during training and validation exhibited fluctuations across 50 epochs which showed higher accuracy in training compared to validation. However, there were instances where validation accuracy approached training accuracy. Evaluation metrics for the MLPNN on validation data of accuracy, precision, and recall are 98.69%, 98.78%, and 98.62%, respectively (Table 2). Similarly, the CNN model demonstrated dynamic variations in accuracy between training and validation sets throughout the epochs. The CNN model presented validation metrics of accuracy, precision, and recall of 98.78%, 98.79%, and 98.77%, respectively. The RNN model displayed consistently close accuracy levels between training and validation sets across the epochs. The validation metrics for RNN exhibited robust performance, with accuracy metrics consistently hovering at 98.71%. In the case of LSTM, both training and validation phases showed nearly identical accuracy, resulting in a performance metric of 97.87%. In general, the results for all models indicated that the training dataset sufficiently covers the variability present in the validation set and the models successfully generalized to unseen data.



(a)

(b)

(c)

(d)

Fig. 19 Representation of the training and validation accuracy of the classification through 50 epochs for (a) MLPNN, (b) CNN, (c) RNN, and (d) LSTM-NN models.

Table 2 The performance metrics of the DL classification models during validation with the on-site geological column

Model/metric	MLPNN	CNN	RNN	LSTM
Accuracy (%)	98.69	98.78	98.71	97.87
Precision (%)	98.78	98.79	98.71	97.87
Recall (%)	98.62	98.77	98.71	97.87

Comparing the results of the original model of MFV-CA with the DL models, a close agreement emerged (Fig. 20), signifying the efficiency of the models in approximating lithological variations within the aquifer systems (Mohammed et al. 2024h). Given the small measuring interval of the well logs, which is 1 cm, any misclassified data will have a minimal impact on the overall resolution of the model. The developed models represent a significant advancement in automated hydrogeophysical characterization. Their validated architecture enables accurate subsurface lithology classification based solely on well-log responses, thereby reducing both the time required for clustering analysis and the potential for human bias. Once tuned, these models require only about 3 minutes to converge and characterize lithological units, offering substantial efficiency gains compared to clustering. Although hyperparameter tuning remains a primary challenge in model development, future integration of automated optimization techniques could further enhance efficiency. Moreover, the success of these 1D modeling approaches opens the possibility of extending them to 2D and 3D frameworks. Such extensions would support more comprehensive groundwater system characterization while also reducing the overall computational burden for clustering in larger and more complex datasets. The demonstrated reliability of these models highlights their applicability in other regions with similar hydrogeological and depositional environments, offering a scalable solution for data-driven groundwater studies.

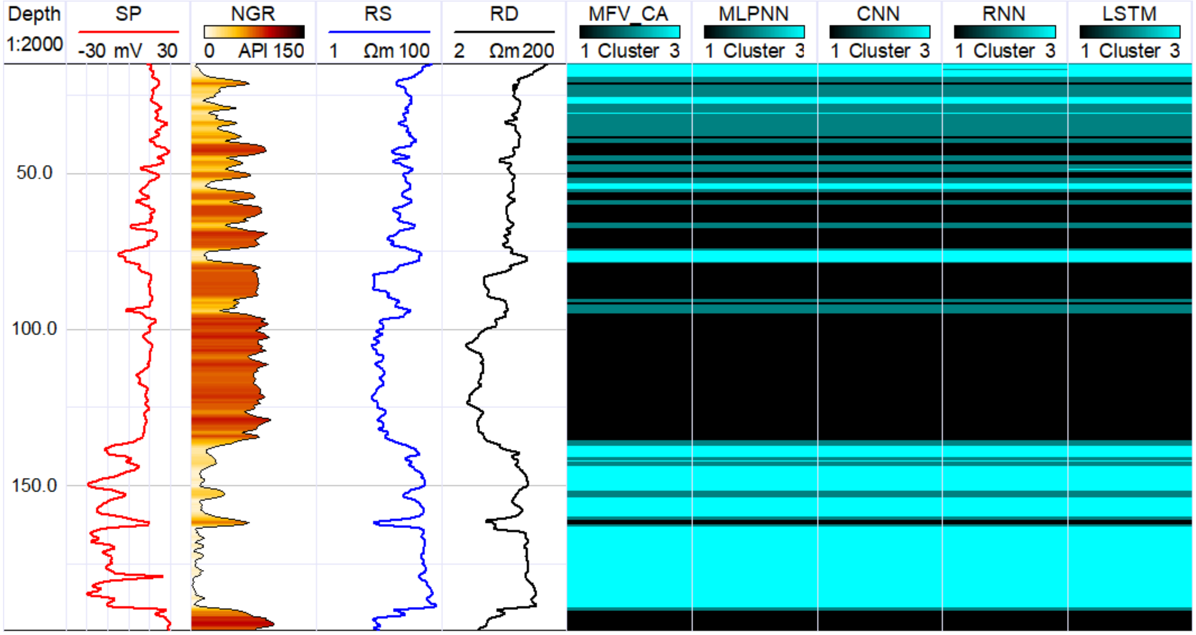


Fig. 20 Comparison between the 1D results of MFV-CA and DL methods along the B14 borehole (see Fig. 4).

3.1.3 Petrophysical and hydrogeological parameters

3.1.3.1 Shale volume and porosity estimation

The shale volume is calculated using Eq. (18) along the D18 borehole (Figs. 4 and 21) and along an E-W profile (Fig. 22a). The descriptive statistics of the volumetric shale volume for the main hydrostratigraphical units are illustrated in Fig. 23a in which the min, max, and mean is shown right to box plot. The shale volume of the CUU exhibited significant variability, ranging from 0.015 to 0.5, with a mean value of 20%. The lithofacies proportion indicated that this unit consists of 37.7% clay, 42.5% silt, and 19.8% sand. The AU also showed high variability in shale volume, ranging from 0.02 to 0.73, with a mean of 0.33. This unit consists of 41.9% clay, 26.9% silt, and 3.2% sand. The IVU exhibited a relatively uniform distribution of shale volume, varying from 0.007 to 0.08. Consequently, the facies analysis indicated that this unit is composed of 78.7% sand. The LMU displayed shale volume variations from 0.015 to 0.77, with a mean value of 0.26%. This unit is dominated by clay and silt layers that make up more than 80% of the unit (Carpio 2024).

The effective porosity is essential for assessing the rate of groundwater flow within the aquifer. Figure 22b shows the 2D interpolation of the effective porosity. The descriptive statistics of the obtained volumetric effective porosity for the hydrostratigraphical units are illustrated using a box plot (Fig. 23b). The minimum, maximum, and mean effective porosity are shown right to the plot for each unit. The effective porosity of the CUU exhibited variability, ranging from nearly impermeable conditions at $6.92*10^{-11}$ to highly permeable conditions at 0.47, with an average of 0.18. The effective porosity of the AU displayed a similar pattern ranging from $2.27*10^{-10}$ to 0.4. The IVU demonstrated a more uniform distribution of effective porosity, varying from 0.17 to 0.32, with a mean of 0.25 while the LMU exhibited effective porosity values ranging from $7*10^{-11}$ to 0.53, with a mean of 0.17.

The described shale volume and effective porosity are considered excluding outliers within each unit, as anomalous measurements may not truly reflect the actual subsurface conditions. Such deviations are likely caused by noise in well-log responses, which can lead to inconsistencies in estimating these parameters. The estimated shale volume and effective porosity hold significant importance for both direct and indirect lithological and hydrogeological parameter estimation.

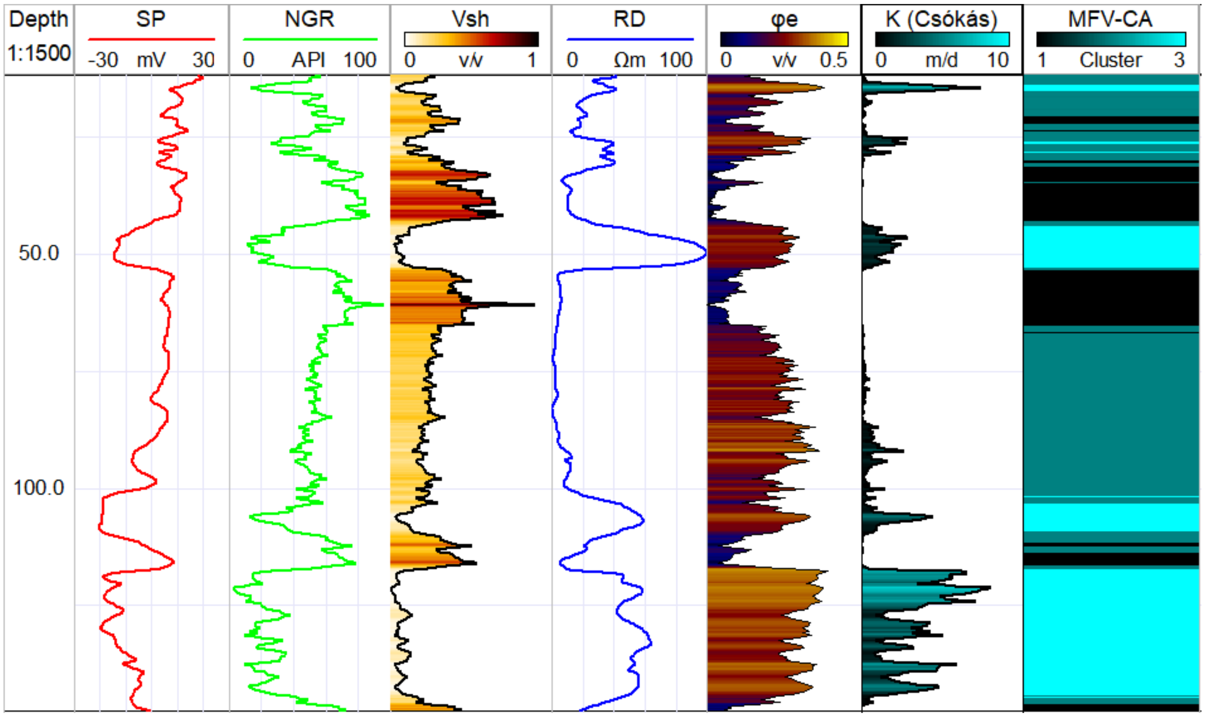
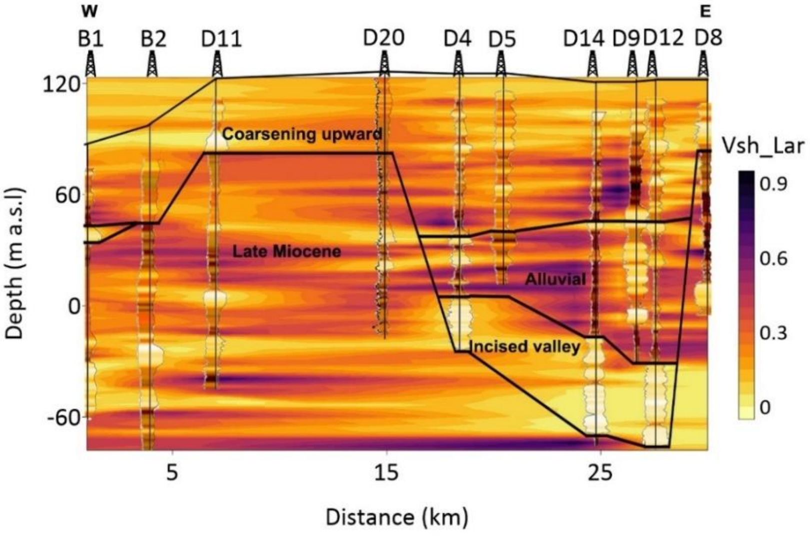
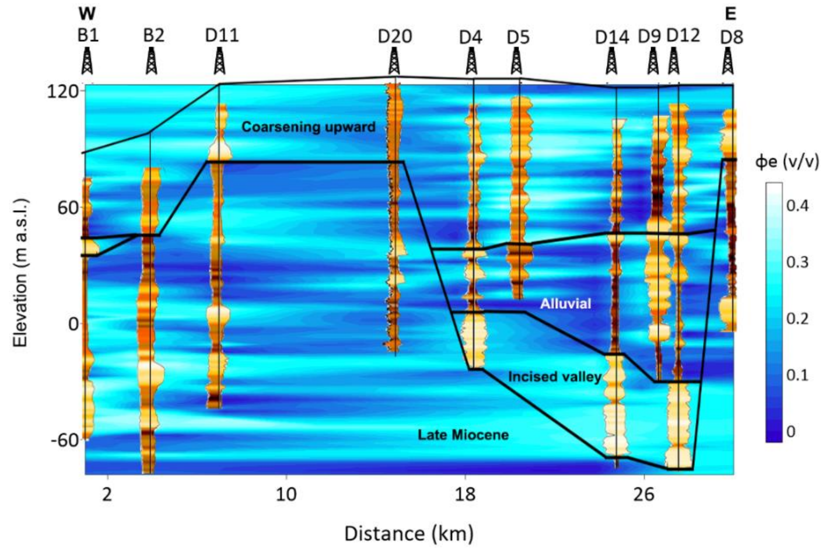


Fig. 21 The 1D interpretation of the petrophysical and hydrogeological parameters in borehole B1. The 1st track represents spontaneous potential (SP), 2nd natural gamma ray (NGR), 3rd is deep resistivity (RD), 4th is shale volume (Vsh), 5th is effective porosity, 6th is hydraulic conductivity, and 7th is the result of the cluster analysis.

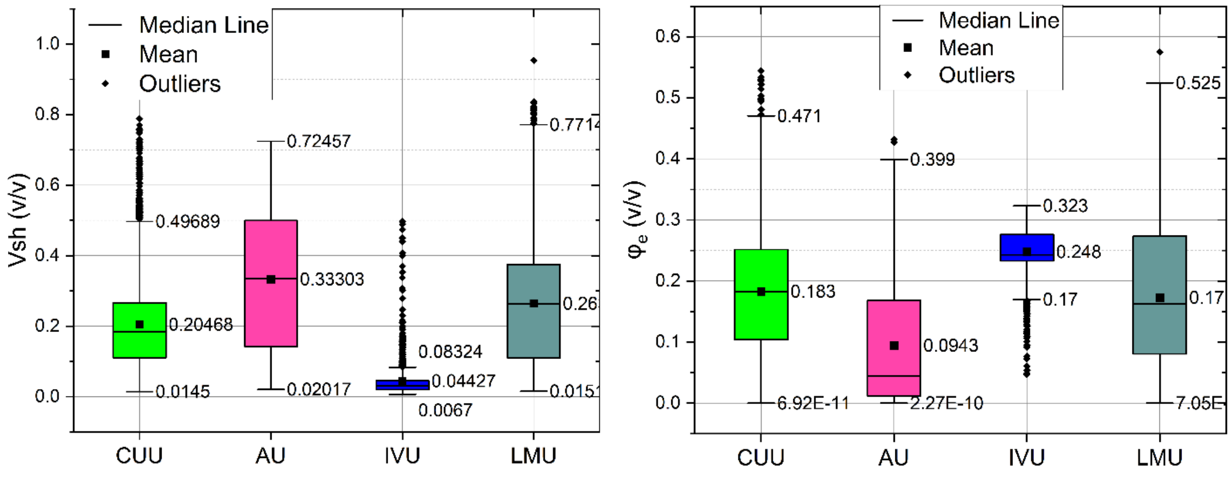


(a)



(b)

Fig. 22 The 2D distribution of the (a) shale volume and (b) effective porosity along the W-E profile shown in Fig. 4.



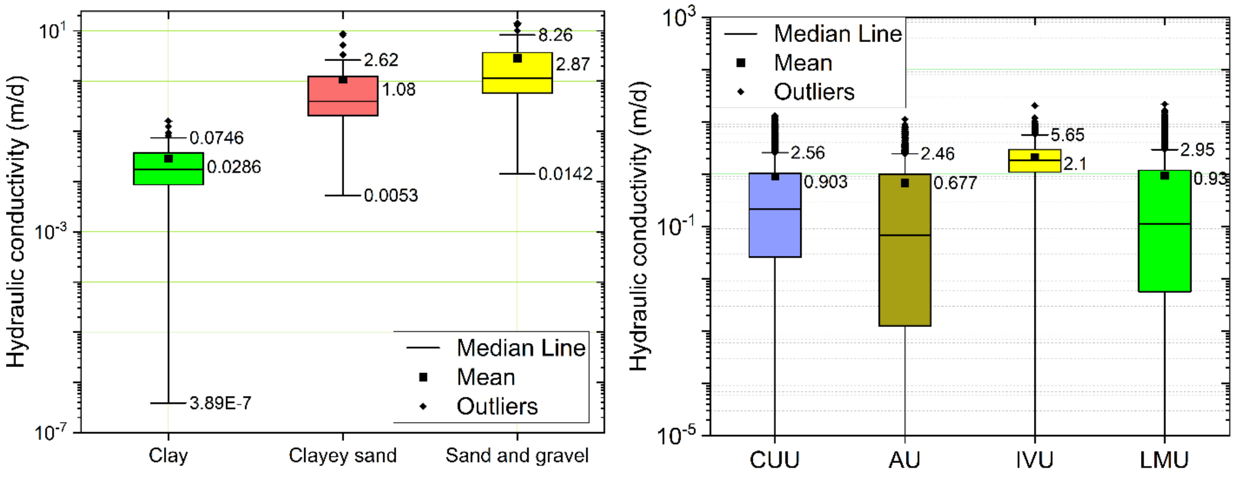
(a) (b)

Fig. 23 Box plot showing the statistical summary of the estimated (a) shale volume and (b) effective porosity for the main hydrostratigraphical unit along the W-E profile presented in Fig. 22. The minimum, maximum, and mean values are shown right to the plot.

3.1.3.2 Hydraulic conductivity estimation by Csókás method

The hydraulic conductivity varies based not only on the rock's characteristics but also on the physical and chemical properties of groundwater (Szabó 2015). The Csókás method effectively accounts for these variations by incorporating physical and petrophysical parameters derived from well-logging data. These parameters influence the ease with which water can flow through the aquifer material. However, to fully characterize hydraulic conductivity, it is also essential to consider the properties of the formation water. In this thesis, those properties were determined through chemical analysis of groundwater samples. The hydraulic conductivity is obtained in 1D, 2D, and 3D, and an example of the 1D estimation along the borehole B1 is illustrated in Fig. 21. The distribution of hydraulic conductivity aligned with identified lithological clusters in which the higher conductivity is indicated in sand and gravel lithology (Cluster 3) and the lower in the clayey layers (Cluster 1). The statistical summary of the hydraulic conductivity for different lithologies is illustrated in Fig. 24a. For the clayey layers, hydraulic conductivity varied between 3.9×10^{-7} to 0.07 m/d, with an average value of 0.03 m/d. Clayey sand layers exhibited hydraulic conductivity between 0.005 to 2.62 m/d, with an average of 1 m/d. In sand and gravel layers, it ranged from 0.01 to 8.26 m/d, with an average value of 2.87 m/d. These values exclude outlying measurements within each unit, as these anomalies do not accurately represent the actual hydrogeological conditions. Accordingly, the estimated hydraulic conductivity for the clusters further confirms the robustness of the MFV and SOM-based cluster formation.

The results of the 2D interpretation along the hydrostratigraphical units are shown in Fig. 25. The hydraulic conductivity estimated along the profile showed high compatibility with 2D clustering and shale volume and effective porosity estimation in which the hydraulic conductivity ranged from relatively high in coarser materials to lower in finer-grained sediments. Figure 24b presents a statistical summary of the hydraulic conductivity for the hydrostratigraphic units along the W-E profile. The CUU showed a wide variation in hydraulic conductivity that ranged between 10^{-7} and 2.65 m/d. The minimum value is indicated in D4 borehole (Fig. 4) in the clayey layers, while the maximum is reported in D5. The intermitted clayey sand layers within the CUU showed an average hydraulic conductivity of 0.9 m/d. The AU, however, showed low hydraulic conductivity compared to that of the CUU, as it ranged between 10^{-8} to 2.46 m/d. The IVU is associated with hydraulic conductivity that ranges between 10^{-7} to 5.65 m/d. The LMU showed hydraulic conductivity fluctuating between 10^{-9} to 2.59 m/d. The lower values in these units are predominantly associated with clayey layers, whereas the higher values correspond to sand and gravel layers, reflecting the influence of lithological composition on hydraulic conductivity distribution.



(a) (b)

Fig. 24 The descriptive statistics of the Csókás method-based hydraulic conductivity for (a) the identified lithology and (b) the main hydrostratigraphical units along the W-E profile

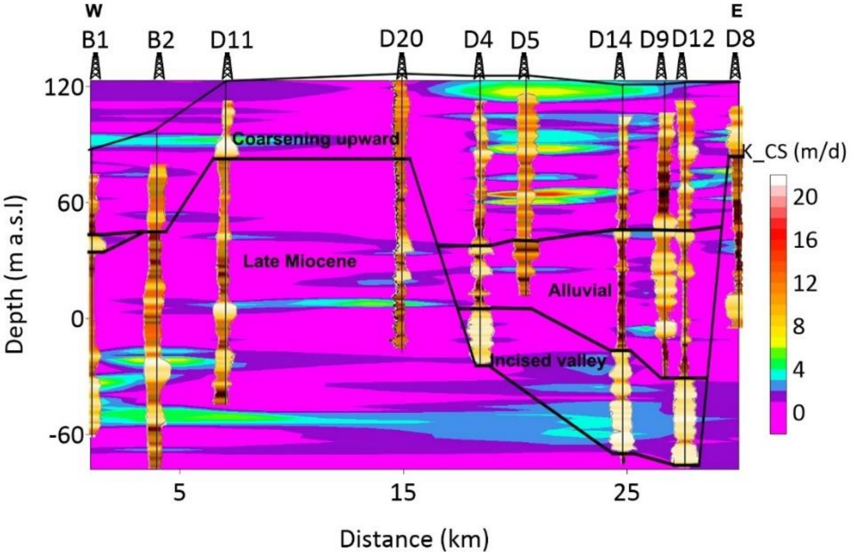


Fig. 25 The result of 2D interpolation of the hydraulic conductivity along the W-E profile.

The estimated hydraulic conductivity is validated by Logan's (1964) method. Logan, method gives only an approximation of the hydraulic conductivity along the saturated, isotropic, and extensive sandy layers in equilibrium conditions. The hydraulic conductivity estimated by the Logan method is highly influenced by screen interval and discharge rate since the higher rate results in higher hydraulic conductivity and vice versa, and more realistic values can be obtained when the pumping rate is optimized. In the study area, the discharge rate highly varied between 90 and 979 m^3/d, resulting in overestimated hydraulic conductivity (Logan 1964). For the CUU, a hydraulic conductivity of 3.2 m/d is indicated. In comparison, the Csókás method provided a mean value of 0.9 m/d. In the AU, the Logan method showed a hydraulic conductivity of 4.1 m/d (Carpio 2024). A higher discharge rate likely influences these high values of hydraulic conductivity. However, in the IVU, a closer agreement is obtained. Logan method demonstrated a comparable value of 4.5 to 2.1 to the Csókás method. The average hydraulic conductivity of the LMU obtained by the Logan and Csókás method is 4.6 and 0.93 m/d, respectively.

The estimated hydraulic conductivity was then interpolated in 3D using the ordinary kriging method to indicate its vertical and horizontal variations throughout the system. The spatial distribution of hydraulic conductivity is depicted in Fig. 26. According to Csókás (1995), aquifers are associated with hydraulic conductivity higher than 0.09 m/d while the aquitards have hydraulic conductivity less than 0.0026 m/d (Szabó et al. 2015a). This indicates that good aquifers are mainly hosted by clayey sand and sand and gravel layers (Püspöki et al. 2013). The productivity of these aquifers varies from being moderate to highly productive (Krásný 1993). Despite their relatively low thickness, the sandy layers within the CUU and AU can still function as good aquifers. The IVU is classified as having a high hydraulic conductivity potential (Marton and Szanyi 1997). The LMU displays a relatively wide range of hydraulic conductivity values. While it includes some lower values, it also contains layers with moderate permeability. Thus, it can be classified as having a low to moderate hydraulic potential (Krásný 1993; Mohammed et al. 2024b). The heterogeneous distribution of hydraulic conductivity by orders of magnitudes within the hydrostratigraphical units confirms their varied lithological compositions. The valley incision unit demonstrated a more uniform distribution of hydraulic parameters, confirming the consistent lithological characteristics (Carpio 2024). The homogeneous lithology as indicated by the clustering and uniformity of petrophysical and hydrogeological parameters in this unit suggests that this unit is characterized by more consistent grain size distribution contributing to a more uniform hydraulic behavior. Consequently, the highly productive IVU can be considered the main groundwater aquifer in the study area (Püspöki et al. 2013).

The Csókás method is a reliable approach for continuously characterizing hydrogeological properties in highly heterogeneous aquifer systems. A key advantage of this method is its ability to estimate hydraulic conductivity solely from well-log data, eliminating the need for costly rock property analyses required by the Kozeny-Carman approach. On the other hand, while pumping tests are valuable and widely used for estimating aquifer properties under real field conditions, they often rely on simplified assumptions that may not capture the full heterogeneity of the subsurface. In contrast, Csókás method provides detailed estimates that more accurately reflect the spatial variability of actual subsurface conditions, complementing the information from pumping tests. These results serve as critical inputs for groundwater modeling, enhancing the accuracy of groundwater flow and contaminant transport simulations, and ultimately improving water resource management.

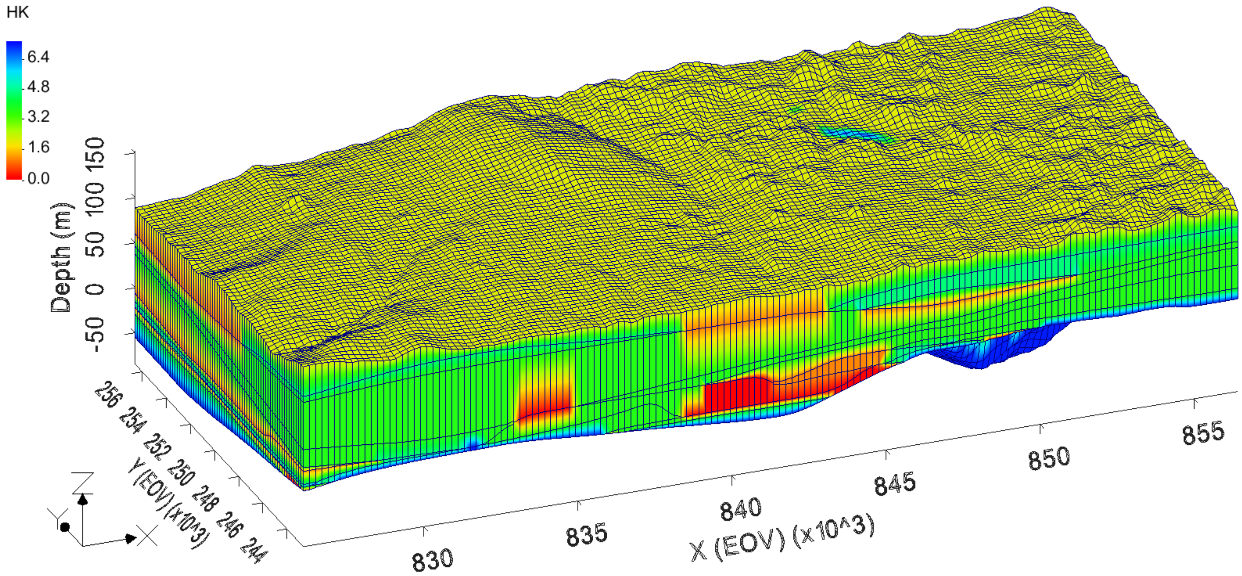


Fig. 26 The 3D distribution of the hydraulic conductivity estimated by Csókás method. The depth is above the sea level and the EOVX and EOVY coordinates are based on the National Projection system (Egységes Országos Vetületi).

3.1.3.3 Hydraulic conductivity by DL models

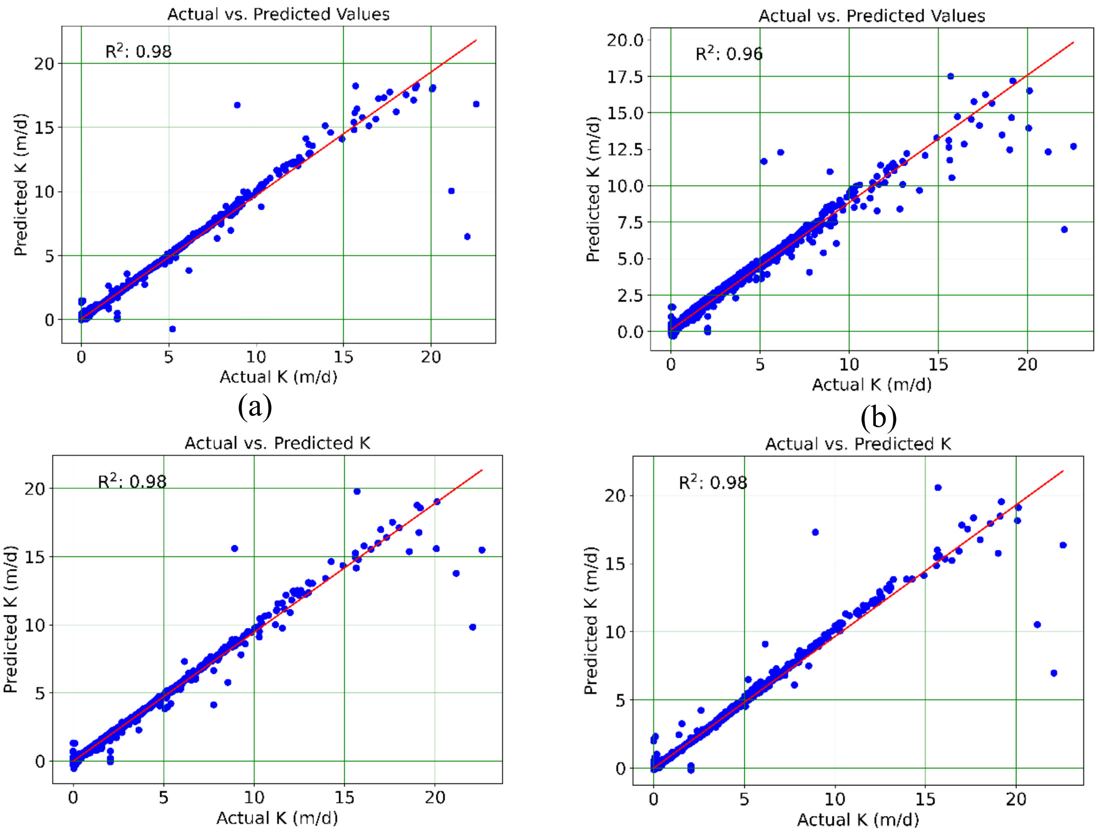
The Csókás method has proven to be highly effective in providing continuous, high-resolution estimations of hydraulic conductivity, allowing for a detailed understanding of its vertical distribution even within a single geological layer. However, the application of the Csókás method requires detailed information about both the formation and water properties (Szabó et al. 2015b), which can vary significantly across different locations. As a result, estimating hydraulic conductivity becomes a borehole-by-borehole process, making it time-consuming (days), particularly when dealing with large numbers of boreholes (Mohammed et al. 2024h). To address this challenge, I implemented DL models including MLPNN, CNN, RNN, and LSTM to enable rapid and efficient prediction of hydraulic conductivity, allowing estimation across the entire dataset in a single processing step.

The DL regression models predicted the hydraulic conductivity using the raw well-logs of SP, NGR, RS, and RD as input and Csókás method-based hydraulic conductivity as output. The architecture of the DL models was fine-tuned through experimentation, with hyperparameters selected from a range of values to ensure optimal performance across different models. For the MLPNN, a configuration of 120 and 60 neurons for the hidden layers was chosen. The dropout rate was set to 0.2 to regularize the model and improve generalization and a learning rate of 0.001 was selected. In CNN, which excels at identifying spatial patterns in data, 64 filters with a kernel size of 5 were used to capture more granular features in the well-log sequences. The fully connected layers were configured with 128 and 64 neurons to allow sufficient capacity for high-level feature extraction while maintaining computational efficiency. A dropout rate of 0.3 and a learning rate of 0.005 were selected. For the RNN, two hidden layers with 150 and 75 units were chosen. The dropout rate was set at 0.15 to prevent overfitting, and the learning rate was set to 0.003. For LSTM, the architecture included two hidden layers with 128 and 56 units. A dropout rate of 0.25 was applied and a learning rate of 0.001 was chosen for fine-tuning updates.

After successful training, the models are validated against the hydraulic conductivity from Csókás method. The MLP-NN demonstrated high performance (Table 3). The RMSE and MAE stood at 0.0001 and 0.06, respectively. Additionally, the coefficient of determination (R^2) exhibited a high value of 0.98, indicating a strong correlation between predicted and actual hydraulic conductivity (Fig. 27a). The CNN model exhibited slightly higher RMSE and MAE values at 0.0002 and 0.1083, respectively, while maintaining a good R^2 score of 0.96 (Fig. 27b). On the other hand, the RNN model produced RMSE, MAE, and R^2 values of 0.0001, 0.0812, and 0.98 (Fig. 27c), respectively, depicting slightly higher errors compared to MLPNN and CNN but still indicating a strong correlation between predicted and actual values. The LSTM model showcased remarkable performance with RMSE, MAE, and R^2 metrics of 0.0001, 0.0679, and 0.98 (Fig. 27d), respectively. The results indicate that MLPNN outperformed the other models in predicting hydraulic conductivity, demonstrating the lowest error and highest accuracy. Its fully connected architecture effectively captures complex relationships within the dataset, making it well-suited for this task. While the models demonstrated promising performance, their limitation in predicting outlying hydraulic conductivity values is observed (Fig. 27). The limitation in predicting outlying high hydraulic conductivity values likely arises due to the imbalanced distribution of training data, where extreme values are underrepresented. Despite this limitation, the comparison between the hydraulic conductivity obtained from the Csókás method and the DL models showed high compatibility (Fig. 28).

Table 3 The performance metrics during validation of the DL regression models against Csókás method-based hydraulic conductivity

Model/metric	MLPNN	CNN	RNN	LSTM
RMSE	0.0001	0.0002	0.0001	0.0001
MAE	0.0606	0.1083	0.0812	0.0679
R^2	0.98	0.96	0.98	0.98



(a) (b) (c) (d)

Fig. 27 Regression between the actual (obtained using Csókás method) and predicted hydraulic conductivity using (a) MLPNN, (b) CNN, (c) RNN, and (d) LSTM neural networks.

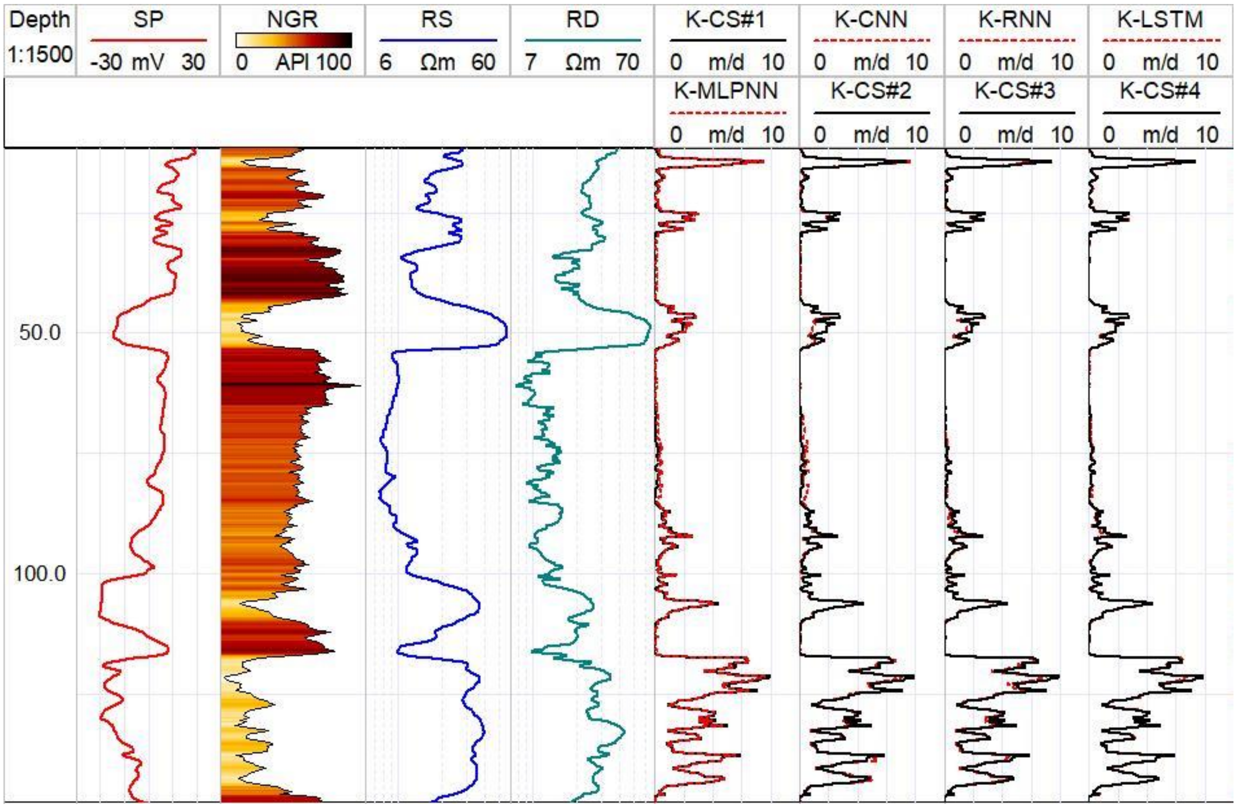


Fig. 28 1D comparison between the hydraulic conductivity logs calculated from Csókás method and DL methods along the D18 borehole.

There is a significant degree of nonlinearity between well-log properties and hydraulic conductivity, primarily because hydraulic conductivity is influenced by a variety of factors beyond just the physical properties of the rock. These complexities make it difficult to establish a direct, linear relationship between well-log data and hydraulic conductivity using traditional methods. However, the developed DL models were able to successfully capture and reveal this inherent nonlinearity. A key benefit of the developed model is its ability to analyze all boreholes in a single processing step, completing the task in just two minutes following hyperparameter optimization. This rapid and automated analysis of comprehensive well-log datasets enables efficient and consistent interpretation across the entire study area. As a result, it significantly facilitates the construction of detailed 2D and 3D hydrogeological models, enhancing the characterization of the groundwater system. These developed models can be independently applied to predict hydraulic conductivity, particularly in regions with similar hydrogeological conditions. However, a key challenge remains in tuning the hyperparameters, which often require multiple iterations to identify the optimal configuration. Nevertheless, even with the need for multiple iterations, the hyperparameter tuning process still requires significantly less time compared to estimating hydraulic conductivity manually for each borehole individually. This makes the approach more efficient and scalable for large-scale hydrogeological assessments.

3.1.3.4 AE-NN-based estimation of shale volume and hydraulic conductivity

The deep AE-NN is utilized to independently estimate shale volume and hydraulic conductivity by exploiting the inherent information from various well logs (SP, NGR, RS, and RD) that are sensitive to these parameters. The motivation for using this model for shale volume and hydraulic conductivity estimation stems from the significant influence of these parameters on all well-log responses (Szabó 2011). These key properties can explain the majority of variance within the dataset, which can be effectively extracted and represented using AE-NN. Various AE-NN architectures were tested, including different numbers of layers and neurons, to determine the optimal structure. The most effective configuration featured an encoder-decoder architecture, where the encoder consisted of hidden layers with 128 and 46 neurons, and the decoder mirrored this arrangement with 46 and 128 neurons, respectively. Due to the limited number of available well logs, only one latent space representation was extracted to ensure an interpretable relationship between the well log inputs and the petrophysical and hydrogeological parameters.

The encoding process resulted in the extraction of a 1D latent space (LS) log from the input logs, effectively capturing the underlying essential features of the data while reducing noise and redundancy and preserving key information. Accordingly, the well logs are decoded (Fig. 29) to examine the capability of the LS to reproduce essential features within the data. The correlation analysis between the original and decoded well logs indicated a high R^2 coefficient (0.96) and a low RMSE value (0.04) indicating that the reconstruction of the original logs is performed with minimal information loss. The extracted LS is correlated to the original well logs using Spearman's (1961) correlation matrix to indicate the contribution of each log to the final LS (Fig. 30). The LS exhibited a high positive correlation with NGR (0.95), high negative correlation with resistivity logs, and moderate correlation with SP log.

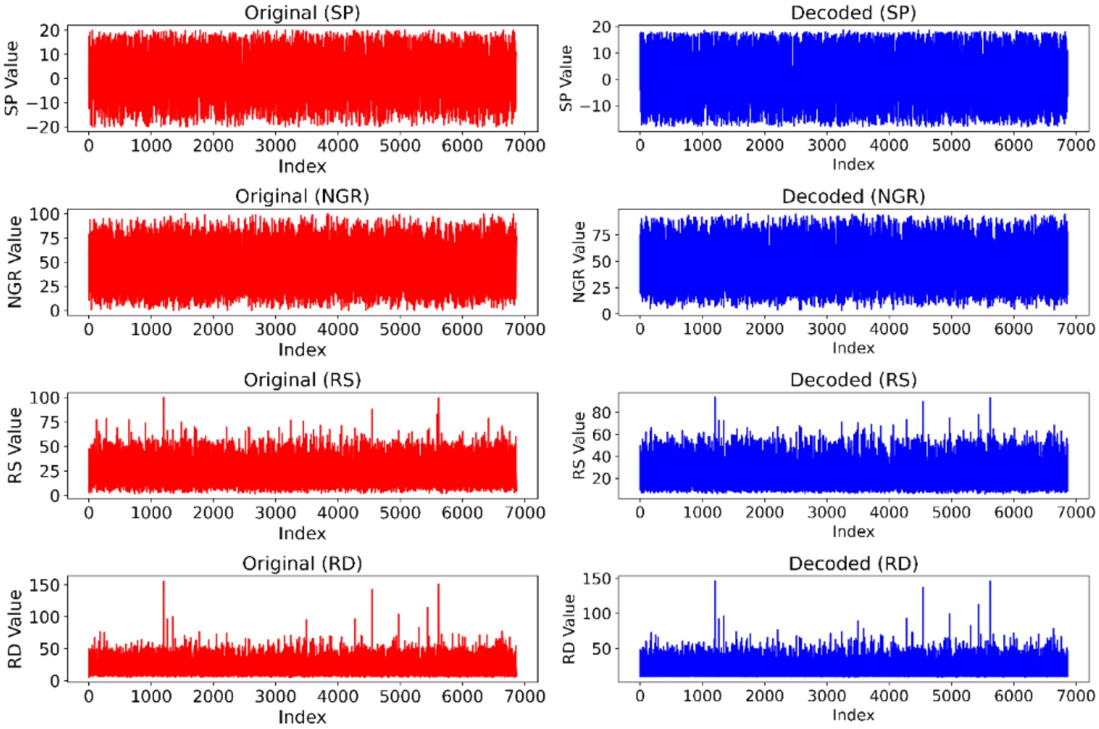
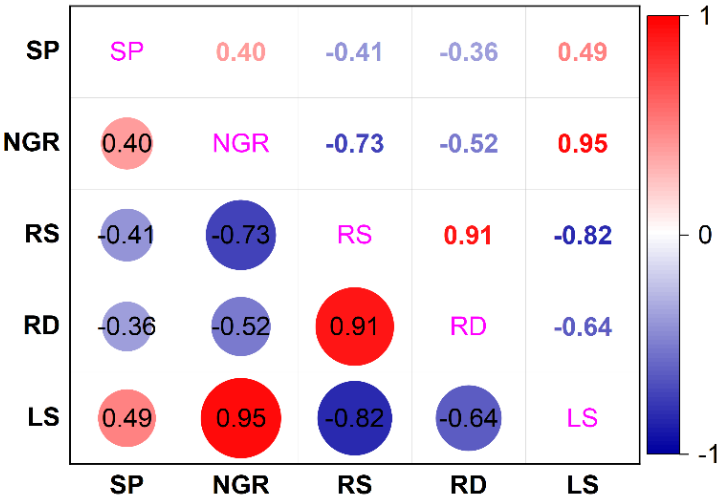


Fig. 29 Representation of the original and decoded well logs based on the extracted 1D LS log.



	SP	NGR	RS	RD	LS
SP	SP	0.40	-0.41	-0.36	0.49
NGR	0.40	NGR	-0.73	-0.52	0.95
RS	-0.41	-0.73	RS	0.91	-0.82
RD	-0.36	-0.52	0.91	RD	-0.64
LS	0.49	0.95	-0.82	-0.64	LS

Fig. 30 Spearman’s rank correlation between well logs and latent space (LS) log.

The LS log extracted by the AE-NN is used for the estimation of the shale volume within the aquifer system. The shale content exerts a significant influence on nearly all well logs as its presence affects the electrical and radioactive properties of the formations (Szabó 2011), leading to distinctive signatures in well log data which is likely to be captured by the extracted latent space. The use of AE-NN for shale volume estimation is also inspired by the strong nonlinear correlation between the extracted LS log and the NGR log as the NGR log is particularly sensitive to the shale content. To establish a quantitative relationship between the normalized LS and shale volume from Larionov method, a regression analysis was conducted (Fig. 31a) that took a Gaussian relationship and expressed as

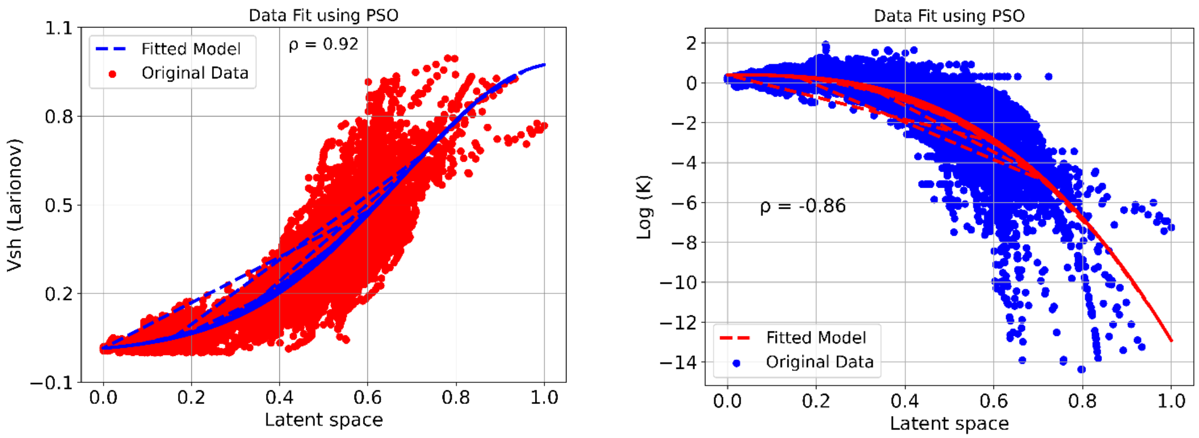
$$V_{sh} = ae^{-\left(\frac{LS-b}{c}\right)^2}, \tag{41}$$

where a, b, and c represent regression parameters that shape the Gaussian curve. Parameter a is the scaling parameter or the maximum value of V_{sh} and determines the height of the curve, b is the location parameter or the center of the curve along the LS axis and represents the value of LS where the peak of the curve occurs, and c is the dispersion parameter or width of the curve and controls how quickly the curve decreases moving away from its center.

Generally, shale volume is inversely proportional to hydraulic conductivity in the clastic primary porosity formations (Shevnin et al. 2006; Szabó 2015). Based on this inverse relationship, I hypothesized that the LS could serve as an indicator of hydraulic conductivity. Furthermore, the strong correlation between the extracted LS and the resistivity logs suggests its effectiveness in capturing information relevant to hydraulic conductivity. This aligns with the principle that the resistivity logs are sensitive to pore-fluid properties and rock matrix composition, which significantly influence hydraulic conductivity. The quantitative relationship between the LS representation and decimal logarithm of hydraulic conductivity calculated by the Csókás method is established by regression analysis (Fig. 31b) and expressed as (Szabó et al. 2022; Mohammed et al. 2025d)

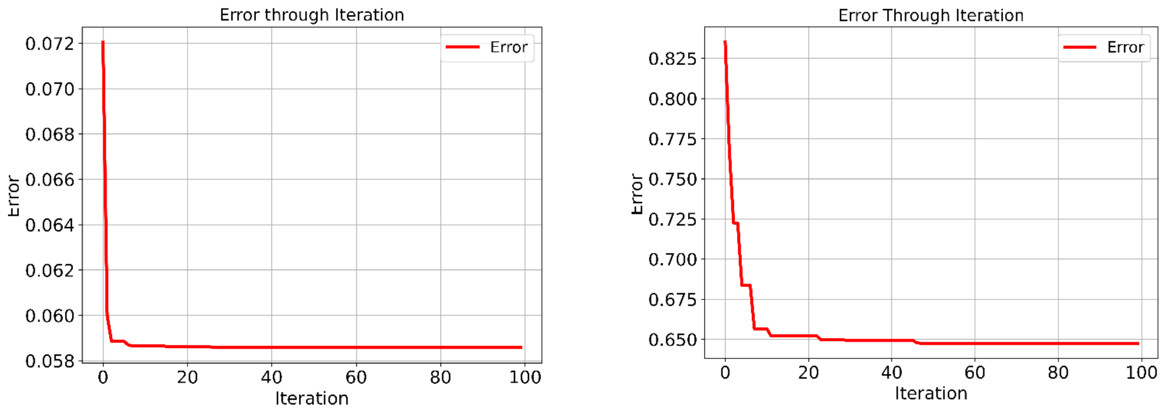
$$\log(K) = a\left(1 - LS^{b}\right)^{c} + d. \tag{42}$$

The regression parameters for shale volume and hydraulic conductivity were estimated using PSO. For shale volume, the initial particle swarm is generated each with random positions within specified bounds, and the best-known positions and errors for each particle are initialized based on their current positions. This process continued for 100 iteration steps (Fig. 32a) optimizing the L_1 objective function. The optimization process achieved errors of 0.059. Accordingly, the values obtained for the regression parameters were a=0.98, b=1.03, and c=0.5 with 95% confidence intervals. The estimated shale volume derived from the AE-PSO model exhibited a strong correlation (0.9) with shale volume estimations obtained through the Larionov method (Fig. 33a). For the hydraulic conductivity, the configuration of the PSO is set similarly to that of the shale volume estimation. PSO required 100 iterations (Fig. 32b) to reach an error of 0.64. As a result, the regression parameters obtained using were a=16.2, b=3.13, c=0.99, and d=−15.85 with 95% confidence interval. Subsequent correlation analysis between the actual hydraulic conductivity obtained from the Csókás method and the predicted values using AE-PSO showed a close agreement, with a correlation coefficient of 0.84 (Fig. 33b).



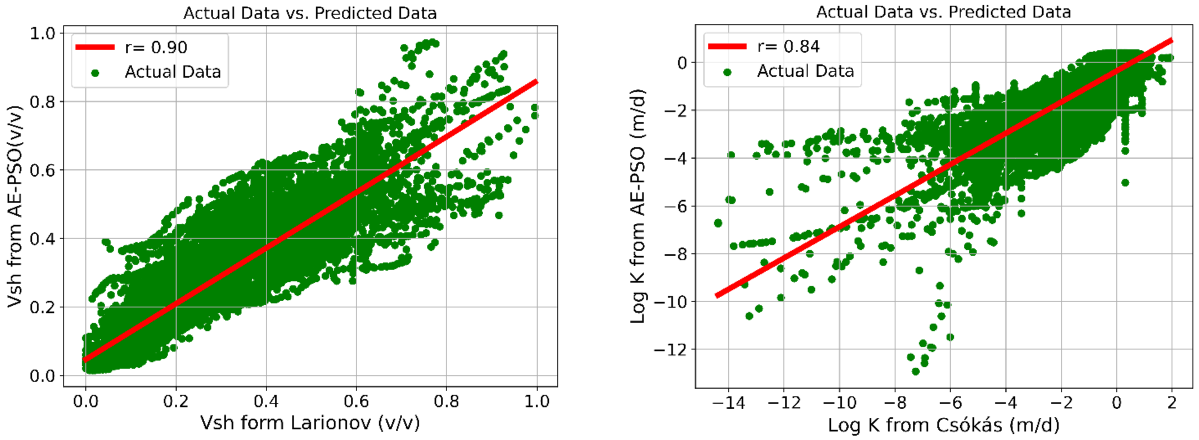
(a) (b)

Fig. 31 Regression relationship between the extracted latent variable and (a) shale volume from Larionov method, and (b) hydraulic conductivity from Csókás method, showing the refinement of the regression parameters across different iterations (blue and red dotted lines).



(a) (b)

Fig. 32 Convergence of the PSO models to optimize the regression parameters for the estimation of (a) shale volume and (b) hydraulic conductivity. The error function is the MAE.



(a) (b)

Fig. 33 (a)The correlation relation between the shale volume estimated by Larionov and AE-PSO (b) The correlation connection between the Csókás and AE-PSO-based hydraulic conductivity.

The predictive capabilities of the regression equation are validated by applying it to an independent dataset. The shale volume and hydraulic conductivity were estimated using AE-PSO models along different boreholes. An example of 1D analysis is illustrated in Fig. 34 where the shale volume and hydraulic conductivity estimations using AE-PSO showed close agreement with the Larionov and Csókás methods, respectively. A similar approach to the deep AE-NN method was previously conducted using factor analysis, where the first extracted factor, which explains the highest percentage of variance, was correlated with shale volume and permeability (Szabó 2011; Szabó et al. 2022; Mohammed et al. 2024a). While the linear factor analysis effectively identifies relationships between well-log data and hydrogeological parameters, the AE-NN does not require strong assumptions about data distribution, making them more flexible and suitable for capturing the complex nonlinear relationships present in the well-logging data (Mohammed et al. 2024f). This efficiency proves the potential of AE-NN as a powerful tool in petrophysical and hydrogeological analysis. The proposed models provide reliable estimates and validate existing hydraulic conductivity estimates obtained from other methods and can be applied directly to areas with similar geological and hydrogeological conditions.

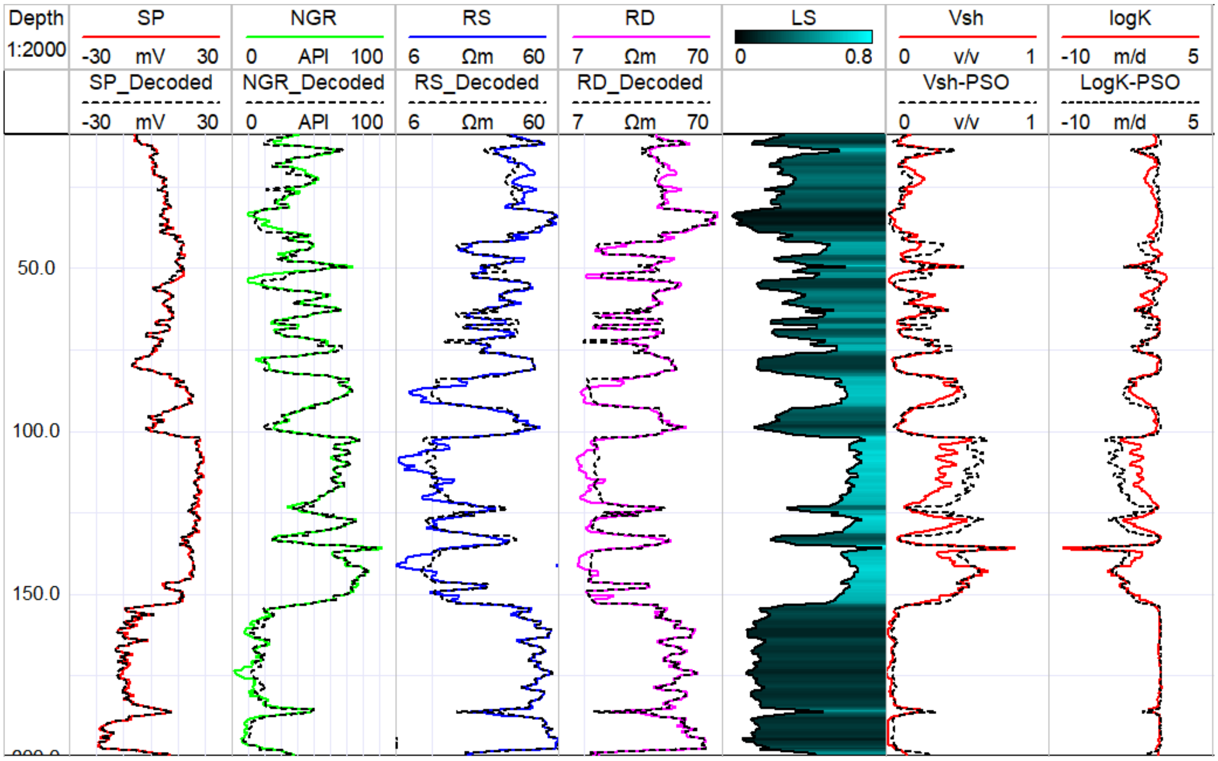


Fig. 34 The 1D estimation of shale volume and hydraulic conductivity using the AE-PSO method. The input and the decoded logs are presented in tracks 1-4, the LS is illustrated in track 5, while the estimated shale volume obtained from Larionov formula (Vsh) and AE-PSO method (Vsh-PSO) as well as hydraulic conductivity obtained from Csókás method (log K) and AE-PSO method (LogK-PSO) are indicated in tracks 6 and 7, respectively.

3.2 Groundwater flow modeling

MODFLOW-USG was employed to simulate steady-state groundwater flow within the Quaternary aquifer system, serving as a means to evaluate the reliability of the conceptual model developed through well-logging data analysis. Materials-based layer property entry method is employed with convertible layers set up. The materials including clay, clayey sand, and sand and gravel are obtained from the results of MFV-CA (Fig. 18). The average values obtained from the Csókás method are used for the horizontal hydraulic conductivity (K_h) (Fig. 26). The anisotropy ratio (horizontal to vertical hydraulic conductivity ratio (K_h/K_v)) is indicated as 1, 7, and 10 for the clay, clayey sand, and sand and gravel layers, respectively (Carpio 2024). Clay has the lowest ratio due to its minimal hydraulic conductivity contrast, while sand and gravel exhibit a higher ratio, reflecting its greater horizontal conductivity (Mohammed et al. 2024e). The translation of the conceptual model into a numerical model is achieved using a 3D CVFD unstructured grid, comprising 10,721 active cells distributed across 151 columns and 71 rows with 13 layers (Fig. 18). Specified head and specified flow boundary conditions are used to constrain the solution of the flow equation (Fig. 35). The specified head boundary is simulated using time-variant specified head package (CHB). The specified head boundary conditions consist of 21 nodes and 19 arcs were derived from observation points obtained from a regional groundwater model (Carpio 2024). This represents a downscaling process in which larger regional models provided boundary conditions for more detailed local models (Mehl and Hill 2006). The specified head boundary conditions were assigned to the northeastern and southwestern parts of the model domain to represent known hydraulic gradients.

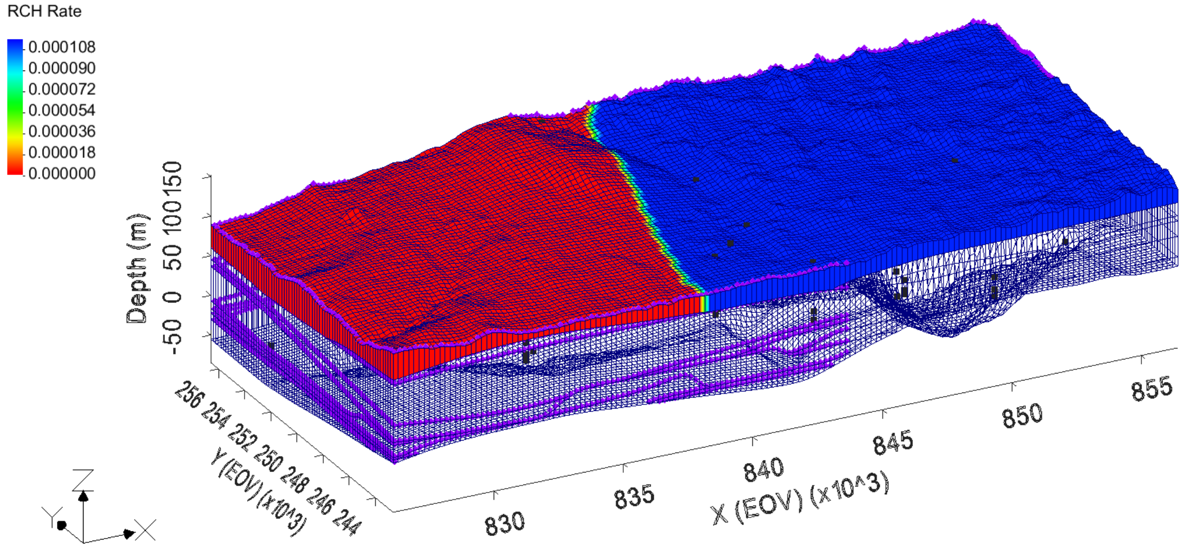
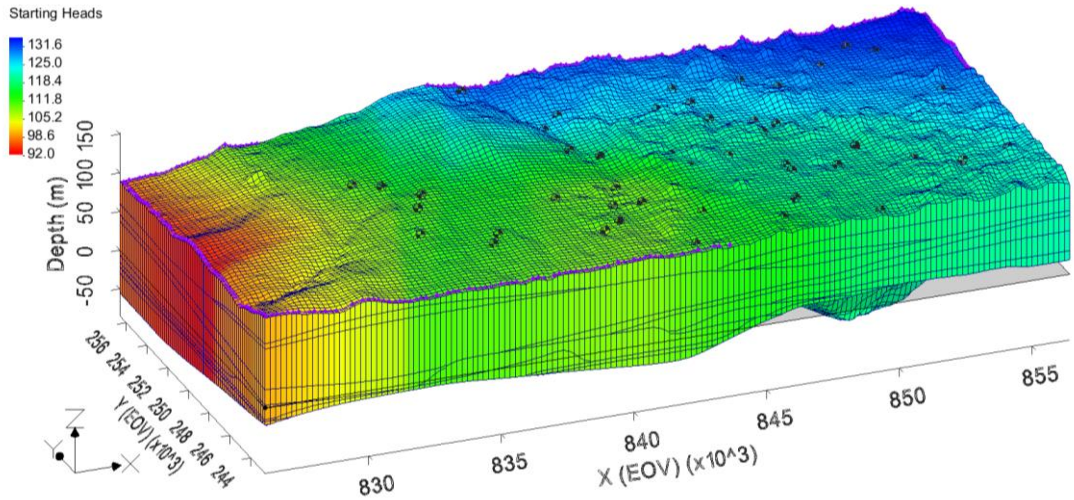


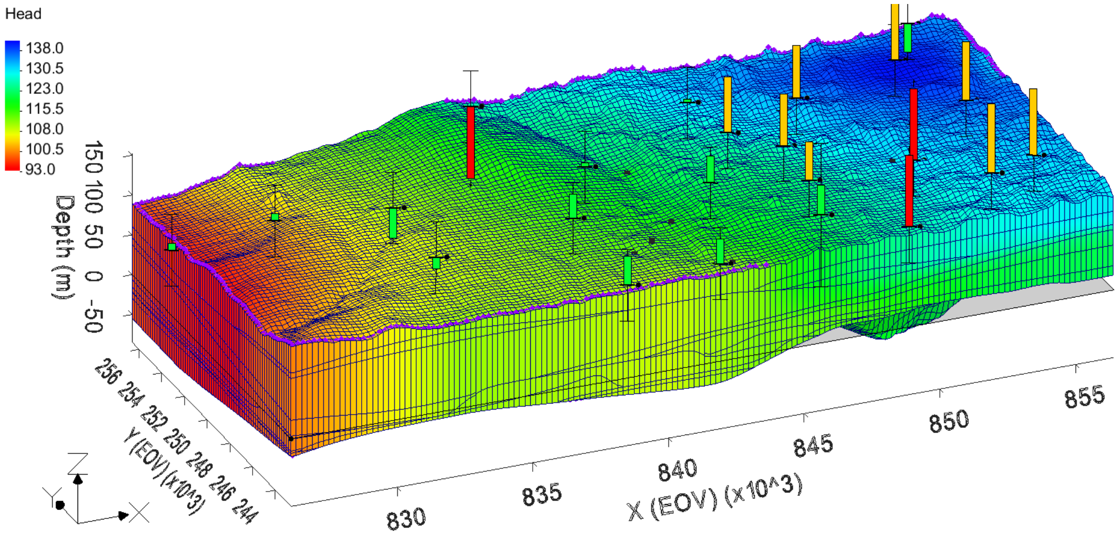
Fig. 35 3D Visualization for the discretization of the model domain with spatially variable recharge rates and specified head boundary conditions (purple arcs).

The specified flow boundaries incorporated recharge and well pumping are simulated using recharge (RCH) and well (WEL) packages. Recharge data is derived from the regional effective infiltration rate map prepared by the Climate of the Carpathian Region- Hungary (CARPATCLIM-HU). Accordingly, the recharge varied from 1.1×10^{-4} m/d to 10^{-6} m/d in the model domain. The recharge is applied to the highest active cells of the model representing the rate at which water enters the groundwater system from precipitation. The recharge distribution shows a spatial pattern with three distinct zones (Fig. 35): a low recharge area in the western parts of the model domain, a moderate recharge area in the eastern parts, and a high recharge area in the northeastern portion with values reaching approximately 0.00011 m/d. The well pumping rates were incorporated into the model using the Well package (WEL), with pumping rates varying between 15 and 2618 m³/d.

The starting head varied between 92.3 to 133.4 m (Fig. 36a) and is uniformly mapped to all layers of the model. This is because the vertical pressure gradient within the 200-300 m of the Quaternary aquifers in the GHP is almost equal (Simon et al. 2023). Consequently, the first run of the model is performed, and the hydraulic head values are calculated, showing agreement with the observed head. The calculated hydraulic head ranges between 93.5 to 139.5 m (Fig. 36b), in which the error compared to the observation wells is represented by box plots. In the box plot, the green box represents errors within the selected variance of 10 m, indicating a good fit. The yellow box corresponds to a variance of 20 m, signifying moderate fit, while the red box, with errors exceeding 20 m, denotes poor fit. The correlation between the observed and calculated head and the resulting residuals is presented in Fig. 37a and b respectively. A moderate correlation is indicated with the coefficient of determination (R^2) of 0.59. The residuals calculated as the difference between the observed and calculated head varied between -10.4 to 10.8 m with overestimation indicated in most of the observation wells.

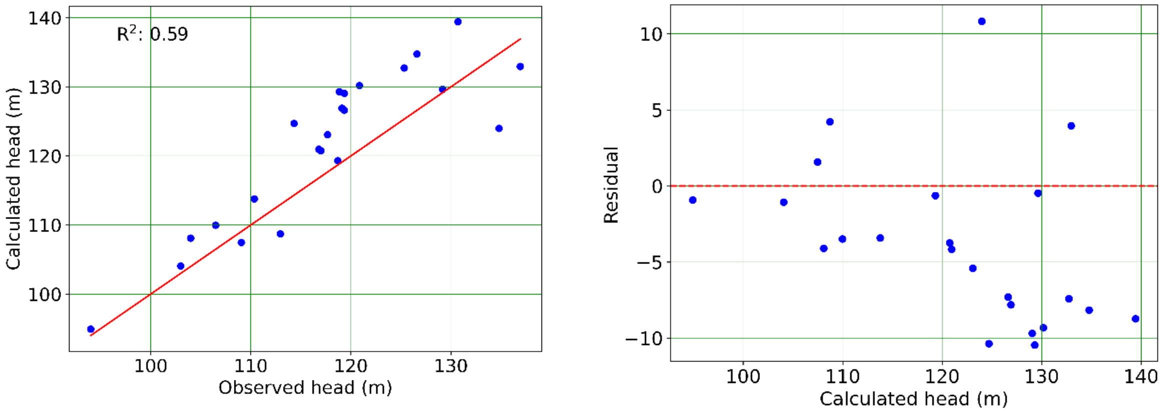


(a)



(b)

Fig. 36 (a) The spatial variation of the initial head obtained from observation wells (black dots). (b) The stationary solution for head from the steady state simulation. The residuals between the simulated and observed heads are represented by box plots in which the green box indicates error within the chosen variance (10 m), while yellow and red indicate moderate and poor fitness, respectively.



(a) (b)

Fig. 37 (a) The correlation between the observed (obtained from observation wells) and calibrated hydraulic head (obtained from numerical calculation). (b) The residual plot shows the error of the calibrated hydraulic head.

A multi-step calibration approach is followed to ensure an accurate representation of the groundwater system. The calibration is conducted using the Head Observation (HOB) package

(Hill et al. 2000) to compare the observed and calculated hydraulic head. A total number of 23 observations well screened in different units is used to calibrate the model. During the initial run of the model, manual calibration was conducted which involved optimizing the horizontal hydraulic conductivity to facilitate the subsequent refinement through the automated techniques. The results of the manual calibration of hydraulic conductivity are illustrated in Table 4. The automatic calibration is conducted within the PEST framework. The primary hydraulic conductivity value obtained from manual calibration served as the starting point for automated calibration. This initial estimate helps guide the optimization process, ensuring that the automated calibration begins with a realistic parameter range. Moreover, constraints including upper and lower bounds are applied to reduce the ambiguity of the numerical solution (Table 4).

Table 4 Solution of the numerical model obtained from manual and automatic calibration with their bounds for given clusters.

Material	Lower bound	Manually calibrated	Upper bound	Automatically calibrated
Clay	0.000001	0.00001	0.01	0.0001
Clayey sand	0.005	1.9	3.5	3
Sand and gravel	0.01	7	10	6.8

During the automated calibration, the PEST code adjusted the hydraulic conductivity of the different materials. The automatically calibrated hydraulic conductivity varied by orders of magnitude from the manually calibrated one, especially for clayey and clayey sand layers. The hydraulic conductivity of the clay experienced a tenfold increase, rising from 10^{-5} to 10^{-4} m/d. Similarly, the hydraulic conductivity of clayey sand layers demonstrated significant enhancement, with values from 1.9 to 3 m/d. Conversely, in sand and gravel layers, a slight decrease in hydraulic conductivity was observed, declining marginally from 7 m/d to 6.8 m/d. Accordingly, the PEST code-based calibrated hydraulic head is obtained that varied between 93.4 and 137.6 m (Fig. 38). A higher correlation between the calibrated and observed head is indicated with an R^2 coefficient of 0.77 (Fig. 39a). The residual for most observations is improved and varied between -7.3 to 10.8 m (Fig. 39b). This error may be attributed to oversimplification of the groundwater system and data quality such as inaccurate measurements of discharge rates that can propagate into the model and lead to errors in the head calculations. Nonetheless, the error in the calibrated hydraulic head fell within the tolerable range (10 m) and successfully represented the main features of the groundwater system. The model exhibited strong numerical stability, with water budget discrepancies maintained below 0.02% during the steady-state stress period. The results of the water budget calculation from the groundwater flow simulation illustrate a well-balanced aquifer system. Most of the inflow is contributed by recharge, with additional input coming from specified head boundaries. Outflow from the system occurs primarily through specified head boundaries and groundwater extraction via wells. The close agreement between total inflow and outflow reflects the accuracy and consistency of the model in representing the overall water balance.

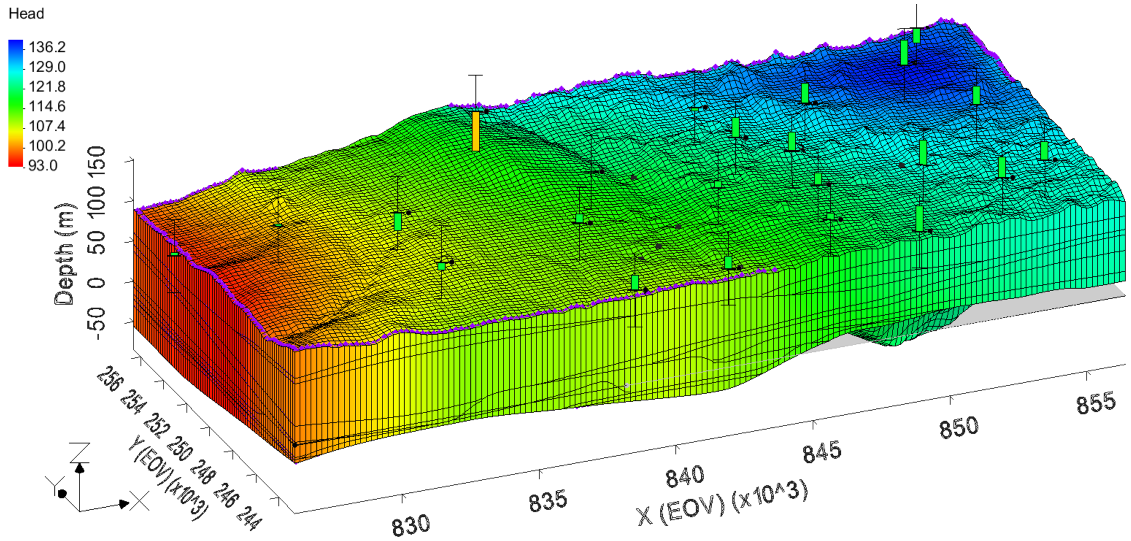
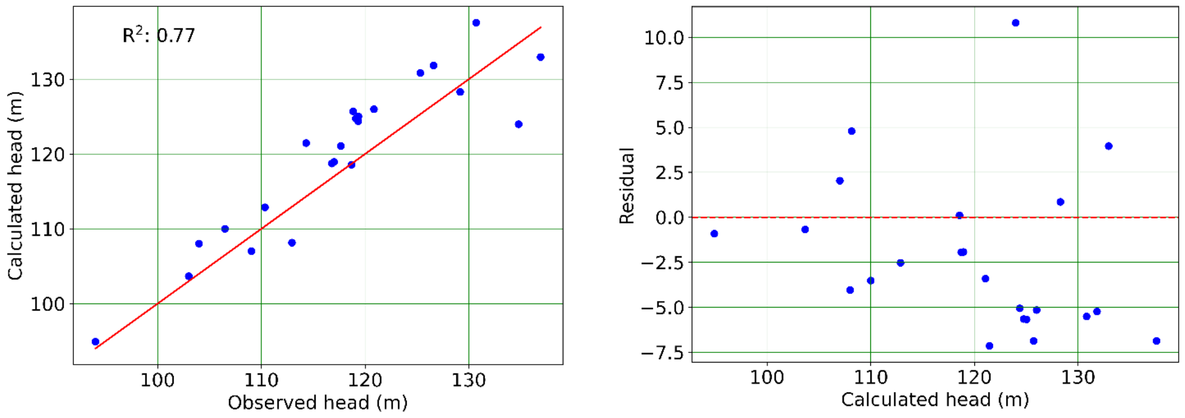


Fig. 38 The distribution of the automatically calibrated hydraulic head using using PEST code. The residuals are represented by box plots in which the green box indicates error within the chosen variance (10 m), while yellow indicates moderate fitness.



(a) (b)

Fig. 39 (a) The correlation between the observed (obtained from observation wells) and calibrated hydraulic head. (b) The residual plot shows the error of the calibrated hydraulic head.

The distribution of calculated hydraulic heads revealed a consistent spatial pattern, with the highest values observed in the northeast and the lowest values in the western parts. Hydraulic head values exhibited a consistent decrease with increasing aquifer depth beneath regions characterized by topographic elevations above 100 to 110 m. In areas where elevations range between 100 and 110 meters, hydraulic head values remain relatively constant (Tóth and Almási 2001). This hydrogeological pattern underscores the significance of topography in controlling the recharge and discharge dynamics in the shallower part of the Great Plain Aquifer (Mádl-Szőnyi and Tóth 2009). In conclusion, the success of the model is largely due to the high-resolution inputs obtained from the analysis of well-logging data using MFV-CA and Csókás method. Given the strong correlations observed between these methods and other techniques explored in this PhD thesis, such as SOM, and supervised and semi-supervised DL models, the results of each method can be reliably used as an input for groundwater modeling. This provides flexibility in conceptual model development, allowing to selection of the most suitable method based on data availability and project requirements.

DOI: 10.14750/ME.2025.025

4. NEW SCIENTIFIC RESULTS

In my PhD research, I developed a hydrogeophysical approach using well-logging data to characterize the Quaternary aquifer system in the Debrecen area. I aimed to enhance the understanding of its structural and dynamic properties. I applied a combination of deterministic methods and machine learning techniques to achieve both qualitative and quantitative characterization. Specifically, I delineated the aquifer geometry, estimated key petrophysical and hydrogeological parameters, and used these geophysics-based inputs to simulate groundwater flow within the system. The following subsection outlines the main scientific contributions of my work.

4.1 Thesis 1. Preprocessing of well logging data: Filling in the missing deep resistivity log

I developed a hybrid deep learning model that integrates a gated recurrent unit (GRU) neural network with the adaptive moment estimation (Adam) optimizer to predict deep resistivity (RD) logs from existing spontaneous potential (SP), natural gamma ray (NGR), and shallow resistivity (RS) logs.

I trained the GRU model on a borehole with a complete log suite located in the central part of the study area to make the GRU model learn from a representative dataset. The model architecture was determined through a trial-and-error approach, testing various configurations of GRU layers, unit sizes, learning rates, and dropout rates. After evaluating several options, the optimal architecture consisted of two GRU layers with 100 and 50 units, respectively. A dropout rate of 0.1 was applied to mitigate overfitting, and a learning rate of 0.0002 ensured stable convergence. The model was trained and validated over 100 epochs to achieve effective learning while preventing overfitting. Accordingly, the RD logs generated with GRU showed a strong correlation with actual RD measurements in validation wells, achieving an average determination coefficient (R^2) of 0.93 and a mean absolute error of 1.07 Ωm. This model provided a cost-effective and efficient solution to generate RD logs, enhancing subsurface characterization without additional fieldwork or digitization.

4.2 Thesis 2. 3D lithological mapping using unsupervised machine learning techniques

I developed a 3D lithological model of the Quaternary aquifer system through the analysis of well-logging data using unsupervised machine learning techniques, including most frequent value-assisted cluster analysis (MFV-CA) and self-organizing maps (SOMs).

The primary statistical analysis, guided by geological knowledge, identified three distinct clusters in the well-logging data, suggesting that the groundwater system is best represented by three lithological facies: clay, clayey sand, and sand and gravel. The automated classifications showed strong consistency with lithologies derived from drilling data, validating the effectiveness of the MFV-CA and SOM methods. Accordingly, I characterized the main hydrostratigraphic units within the Quaternary aquifer system including coarsening upward, alluvial, incised valley, and Late Miocene units. The coarsening upward unit consists of clayey and sandy layers transitioning to coarser sediments, forming the top aquifer. The Alluvial unit is composed of alternating clay and sand layers, with clay predominating, and varies in thickness. The incised valley unit is a continuous sand and gravel deposit, serving as the primary aquifer. The Late Miocene unit consists of fine sand, silt, and clay, marking the base of the Quaternary sediments. These findings led to the development of a 3D lithological model of the Quaternary aquifer system, effectively capturing lithological heterogeneity that traditional drilling and coring methods cannot fully resolve.

4.3 Thesis 3. 3D distribution of hydraulic conductivity using Csókás method

I introduced a novel application of the Csókás method to derive the 3D distribution of hydraulic conductivity from geophysical well logs, enhancing the spatial resolution and extending the reliability of pumping test-based estimations within the Quaternary aquifer system in Debrecen area.

The Csókás method is developed by Prof. Dr. János Csókás, the former Head of the Geophysical Department at the University of Miskolc. It is a modified version of the Kozeny-Carman equation, uniquely capable of deriving all necessary parameters solely from well log data. These parameters include porosity and resistivity of pore water and resistivity of the saturated formation. Hydraulic conductivity was initially estimated in 1D and validated against independent estimates obtained using the Logan method. After successful validation, the 1D values were interpolated into a three-dimensional (3D) model to capture the spatial variations across the main hydrostratigraphic units of the Quaternary aquifer system (coarsening upward, alluvial, incised valley, and Late Miocene units). Hydraulic conductivity in the coarsening upward Unit ranged from 0.0001 to 11.5 m/d, with an average of 0.9 m/d. In the Alluvial unit, it ranged from 0.00003 to 6.6 m/d, while the incised valley unit exhibited a more uniform distribution, ranging from 0.1 to 8 m/d. The Late Miocene deposits showed hydraulic conductivity values between 0.0053 and 15.3 m/d. The 3D distribution of hydraulic conductivity offers continuous estimates within the heterogeneous system, providing a level of detail unattainable through sparse pumping tests or laboratory analysis. This enhanced resolution facilitates the development of a more reliable conceptual model for groundwater flow simulation.

4.4 Thesis 4. Prediction of hydraulic conductivity using deep autoencoder neural network

I introduced the first application of the semi-supervised deep autoencoder neural network (AE-NN) assisted with particle swarm optimization (PSO) on well logging data for estimating shale volume and hydraulic conductivity.

Initially, I extracted the latent space (LS) log that captures the nonlinear relationships between spontaneous potential (SP), natural gamma ray (NGR), shallow resistivity (RS), and deep resistivity (RD) logs. Using this extracted LS log, I established a regression model for shale volume, estimated using a nonlinear estimation method, and for hydraulic conductivity, calculated using the Kozeny-Carman-based approach. The regression analysis revealed a Gaussian relationship between the LS log and shale volume, as well as a nonlinear relationship between the LS log and hydraulic conductivity. To further improve the predictive accuracy, I specified the regression parameters using the meta-heuristic PSO method. Validation of the models with an independent dataset demonstrated strong correlations between observed and predicted values, with Pearson’s correlation coefficients of 0.9 for shale volume and 0.84 for hydraulic conductivity. Based on these results, I proposed an independent model for estimating these parameters, which provides accurate predictions and validates shale volume and hydraulic conductivity estimates obtained from other methods. This approach can be extended to regions with similar geological and hydrogeological conditions.

4.5 Thesis 5. Lithological and hydrogeological characterization using deep learning models

I designed and implemented hybrid deep learning (DL) models for classification and regression tasks to improve the efficiency of the lithological and hydrogeological characterization of the Quaternary aquifer system based on well-logging data.

The deep learning (DL) models developed in this thesis include multilayer perceptron neural networks (MLPNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM), all hybridized with Adam optimizer. To optimize model performance, I experimentally tuned key hyperparameters, such as the number of layers, units, learning rate, and dropout rate, by systematically exploring a range of values. I trained the models using raw well-logging data as inputs, with outputs derived from the most frequent value-assisted cluster analysis (MFV-CA) for lithology and the Csókás method for hydraulic conductivity. For classification tasks (Lithology), model performance was evaluated using accuracy, recall, and precision metrics. In regression tasks (Hydraulic conductivity), performance was assessed using mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2). The DL models effectively predicted lithology and hydraulic conductivity directly from raw well logs with over 95% accuracy, while significantly reducing processing time to less than 3 minutes for lithology and 2 minutes for hydraulic conductivity after hyperparameters tuning. These models can be independently and reliably applied to aquifers with similar hydrogeological characteristics, particularly within the Great Hungarian Plain.

4.6 Thesis 6. 3D numerical modeling by incorporating geophysical inputs

I derived reliable input parameters including lithology and hydraulic conductivity from geophysical well logs and integrated them into constructing a 3D conceptual model of the Quaternary aquifer system in Debrecen area.

I integrated the 3D lithological model obtained from the most frequent value-assisted cluster analysis (MFV-CA), the 3D hydraulic conductivity derived from the Csókás method, and water level data to construct a comprehensive conceptual model of the Quaternary aquifer system. Then, I translated the conceptual model into a numerical model within the MODFLOW-USG framework. I applied a combination of manual and automated calibration (inverse modeling) to optimize the hydraulic conductivity values of various lithological layers within defined bounds, enhancing the agreement between observed and simulated head data and reducing ambiguity in the numerical solution. After several iterations, the residuals for most observations improved significantly, ranging from -7.3 to 10.8 m. This demonstrates the effectiveness of geophysical data as a reliable input for groundwater flow simulation in heterogeneous systems, providing a cost-effective and time-efficient approach for local and regional-scale modeling.

DOI: 10.14750/ME.2025.025

5. SUMMARY

In my PhD thesis, I developed a hydrogeophysical approach to characterize the Quaternary aquifer system in the Debrecen area, Eastern Hungary. This aquifer system, shaped by complex geological structures, faces increasing pressure from agricultural activities and domestic water use. Accurate characterization of its lithological and hydrogeological properties is crucial for assessing groundwater capacity and evaluating long-term sustainability. However, the high heterogeneity of aquifer systems poses significant challenges, as drilling provides only point-based data that may not adequately capture the full extent of subsurface variability. Likewise, hydrogeological and petrophysical parameters obtained through field tests and laboratory analyses are constrained by limited sampling density and high costs. These limitations underscore the need for more efficient techniques that can deliver continuous, high-resolution data over large areas. To address this, I utilized well-logging data to gain a detailed understanding of the geological and hydrogeological features of this heterogeneous groundwater system.

The well-logging data used in this thesis was initially incomplete, with the deep resistivity (RD) log missing in most boreholes. I began the investigation by preprocessing the dataset and imputing the missing RD logs, as this log is essential for capturing the intrinsic properties of subsurface materials beyond the invaded zone. The RD log was imputed using a gated recurrent unit (GRU) neural network, trained on the available spontaneous potential (SP), natural gamma ray (NGR), and shallow resistivity (RS) logs. The GRU model was highly effective due to its ability to capture complex patterns among the well logs, with a strong correlation between the predicted and actual RD measurements in the validation wells ($R^2 = 0.93$, RMSE = 0.06 Ωm, MAE = 1.07 Ωm). This preprocessing step significantly enhanced the quality and consistency of the input data, ensuring more robust subsequent analyses. These improved data were then used to characterize the main hydrostratigraphic units within the Quaternary aquifer system.

I analyzed the preprocessed well-logging data using the most frequent value-assisted cluster analysis (MFV-CA) and self-organizing maps (SOMs) to perform lithological characterization of the primary hydrostratigraphic units within the Quaternary system. To ensure reliable clustering, I selected the optimal number of clusters using the elbow method and geological context. The elbow method identified three clusters as optimal, aligning with the geological description of the Quaternary sediments. Based on log responses, I assigned these clusters to clay, clayey sand, and sand and gravel. The MFV-CA and SOM methods effectively produced geologically realistic results that closely matched lithological logs from drilling. The clustering-based lithology was first analyzed in 1D, then interpolated into 2D, and ultimately developed into a 3D model to fully characterize the hydrostratigraphic units. The upper part of the system comprises the coarsening upward (CUU) and alluvial (AU) units, consisting of alternating layers of sand, clayey sand, and clay, while the lower incised valley unit (IVU) contains continuous coarse-grained sediment. The significance of this approach lies in its ability to enhance hydrogeological characterization by providing high-resolution, data-driven lithological models that can be employed for improved conceptual site models.

I used the Csókás method to determine hydraulic conductivity distributions across the aquifer system. This method offered continuous estimations that are superior to point-specific measurements. I estimated the hydraulic conductivity in 1D, 2D, and 3D. The estimated conductivity aligned with lithological clusters: sand and gravel exhibited the highest conductivity, while clayey layers showed the lowest. The 2D analysis along the main hydrostratigraphic units revealed varying conductivity values. The CUU demonstrated wide variability, ranging from 10^{-7} and 2.65 m/d, while the AU showed lower conductivity values, between 10^{-8} to 2.46 m/d. The IVU ranged from 10^{-7} to 5.65 m/d, and the LMU fluctuated between 10^{-9} and 2.59 m/d. Based on these findings, I classified the IVU as having the highest potential for groundwater production. This method significantly improves the characterization of aquifer systems by enabling high-resolution spatial mapping of hydraulic conductivity, which is essential for groundwater flow modeling and resource management. Based on the robust performance demonstrated in this work, I strongly recommend hydrogeologists adopt the Csókás method for well-log interpretation and aquifer characterization, especially within the Great Hungarian Plain.

I proposed hybrid deep learning (DL) models to enhance the efficiency of the relatively computationally intensive MFV-CA and Csókás methods. These models include multi-layer perceptron neural networks (MLPNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) networks, all hybridized with the Adam optimizer for classification and regression tasks. The deep learning networks demonstrated the ability to significantly reduce computation time while effectively handling complex nonlinear relationships between the raw well logs and lithology and hydraulic conductivity. They achieved an accuracy of approximately 98% for lithological classification, closely aligning with the results of MFV-CA while decreasing processing time from 10 to less than 3 minutes after hyperparameters tuning. For hydraulic conductivity prediction, the models demonstrated high predictive capability, achieving an R^2 coefficient of 0.98. The developed models can be used independently for the characterization of groundwater systems with similar geological and hydrogeological settings, particularly in the Great Hungarian Plain.

I used a semi-supervised deep autoencoder neural network (AE-NN) to analyze well logging data and estimate shale volume and hydraulic conductivity. The AE-NN encoded the input logs, extracting a 1D latent space (LS) log that effectively captured the key features within the data while reducing noise and redundancy. The model successfully decoded the logs with high accuracy ($R^2 = 0.92$), reconstructing them with minimal loss. I then developed regression models to link the extracted LS log to both shale volume and hydraulic conductivity. These models revealed a Gaussian relationship with shale volume and a nonlinear relationship with hydraulic conductivity. To improve prediction accuracy, I optimized the regression parameters using particle swarm optimization (PSO). The models were validated against independent datasets. The predicted shale volume correlated strongly ($R = 0.9$) with estimates from the Larionov method, while hydraulic conductivity predictions showed good agreement with the Csókás method ($R = 0.84$). The derived universal equations provide a reliable alternative for estimating aquifer parameters in regions with comparable hydrogeological settings, particularly where pumping tests or well-logging data are unavailable.

I introduced an innovative approach to simulate 3D steady-state flow conditions by using geophysical data as the sole input. The results from the MFV-CA and Csókás methods, along with available hydrogeological data, were integrated to construct conceptual models. I then translated the conceptual models into a numerical model using MODFLOW-USG. For model calibration, I implemented both manual and automated approaches, incorporating 23 head observation points distributed throughout the model domain. The automated calibration significantly improved the performance, increasing the coefficient of determination (R^2) from 0.59 to 0.77. The final calibrated model demonstrated residuals ranging from -7.3 to 10.8 m. This demonstrated that geophysical well-logging data analyzed with different deterministic and machine-learning models could serve as a reliable input for groundwater flow modeling, reducing dependence on extensive hydrogeological field surveys.

The scientific findings of this PhD thesis offer practical applications for groundwater management and development. The analysis of well-logging data established a robust framework for accurately mapping lithological variations and estimating key petrophysical and hydrogeological parameters. The integration of geophysical data notably reduces—though does not replace—the need for extensive laboratory and pumping tests, thereby enhancing the efficiency and cost-effectiveness of aquifer characterization. Moreover, due to their broad spatial coverage, geophysical measurements can support the extrapolation of known parameter values to wider areas. To further enhance spatial resolution, the number of clusters used in unsupervised classification can be increased in future studies, potentially capturing subsurface heterogeneities more effectively. The accuracy of hydraulic conductivity estimation can also be significantly improved by optimizing the Csókás constants and refining the regression coefficients that link the latent features from AE-NN to hydraulic conductivity. As additional well data and new measurements become available, these localized empirical relationships can be recalibrated, resulting in a more precise and adaptable predictive model. Since the efficiency and reliability of machine learning approaches are highly dependent on the amount, completeness, and quality of the input data, assembling a rich and high-quality dataset is essential for producing valid and generalizable outcomes in complex hydrogeological settings. The numerical model is initially calibrated using historical data; therefore, obtaining new water level measurements and recalibrating the model accordingly is essential to maintain its accuracy and relevance over time. Additionally, incorporating transient flow conditions, climate change scenarios, and anthropogenic influences will allow for more accurate predictions of long-term aquifer behavior and better inform sustainable water resource management strategies.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to the Institute of Water Resources and Environmental Management at the University of Miskolc for providing an exceptional research environment and valuable resources during my PhD studies. The support and encouragement from the dedicated staff have been instrumental in the successful completion of this research. I am deeply indebted to my supervisors, Prof. Dr. Péter Szűcs and Prof. Dr. Norbert Péter Szabó, for their guidance, insightful feedback, and constant encouragement throughout my PhD journey. Their expertise, commitment, and mentorship have been valuable, shaping both the scope and quality of this dissertation.

I would like to thank the Supervisory Authority for Regulatory Affairs (SARA) for providing the necessary data to carry out this research. This dissertation represents the culmination of collective efforts, and I am grateful to all those who have played a role in its realization. I would also like to express my appreciation to my colleagues for their collaborative spirit and intellectual contributions.

This research is carried out within the framework of the Széchenyi Plan Plus program with the support of the RRF 2.3.1 21 2022 00008 project. The research was partially conducted within the framework of the Sustainable Development and Technologies National Program of the Hungarian Academy of Sciences (FFT NP FTA)

Publication List of the Candidate

Mohammed MAA, Szabó NP, Eltijani A, Szűcs P., An integrated workflow combining machine learning and wavelet transform for automated characterization of heterogeneous groundwater systems. Scientific Reports 1–20, (2025).

Mohammed, M. A., Szabó, N. P., & Szűcs, P. High-resolution characterization of complex groundwater systems using wireline logs analyzed with machine learning classifiers and isometric mapping techniques. Modeling Earth Systems and Environment, (2024)

Mohammed, M. A., Szabó, N. P., & Szűcs, P. Robust estimation of hydrogeological parameters from wireline logs using semi-supervised deep neural networks assisted with global optimization-based regression methods. Groundwater for Sustainable Development, 27, 101348. (2024).

Mohammed, M. A., Szabó, N. P., Kilik, R., & Szűcs, P. Examining innovative unsupervised learning techniques for automated characterization of complex groundwater systems. Results in Engineering, 23, 102594. (2024).

Mohammed, M. A., Szabó, N. P., & Szűcs, P. Multi-step modeling of well logging data combining unsupervised and deep learning algorithms for enhanced characterization of the Quaternary aquifer system in Debrecen area, Hungary. Modeling Earth Systems and Environment, 10(3), 3693-3709. (2024).

Mohammed, M. A., Flores, Y. G., Szabó, N. P., & Szűcs, P. Assessing heterogeneous groundwater systems: Geostatistical interpretation of well logging data for estimating essential hydrogeological parameters. Scientific Reports, 14(1), 7314, (2024).

Mohammed MAA, Szabó NP, Flores YG, 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. Groundwater for Sustainable Development 101086. (2024)

Mohammed MAA, Szabó NP, & Szűcs P Joint interpretation and modeling of potential field data for mapping groundwater potential zones around Debrecen. Acta Geodaetica et Geophysica. (2024)

Mohammed, M. A., Szabó, N. P., Alao, J. O., & Szűcs, P. Geophysical characterization of groundwater aquifers in the Western Debrecen area, Hungary: insights from gravity, magnetotelluric, and electrical resistivity tomography. Sustainable Water Resources Management, 10(2), 67. (2024).

Mohammed, M. A., Mohammed, S. H., Szabó, N. P., & Szűcs, P. Geospatial modeling for groundwater potential zoning using a multi-parameter analytical hierarchy process supported by geophysical data. Discover Applied Sciences, 6(3), 121. (2024).

Mohammed MAA, Szabó NP, Mikita V, & Szűcs P. Tracking the spatiotemporal evolution of groundwater chemistry in the Quaternary aquifer system of Debrecen area, Hungary: integration of classical and unsupervised learning methods. Environmental Science and Pollution Research 1–20. (2025).

Mohammed MAA, Szabó N.P., Mikita V., Szűcs, P. Long-term assessment of groundwater contamination and associated health risks: A study of heavy metals in the Quaternary aquifer system in Pannonian Basin. Physics and Chemistry of the Earth, Parts A/B/C 139:103935. https://doi.org/https://doi.org/10.1016/j.pce.2025.103935

Mohammed MAA, Mohamed A, Alarifi S. S., Mahmoud A., Alao J. O., Norbert P Szabó, & Péter Szűcs, Investigation of petrophysical and hydrogeological parameters of the transboundary Nubian Aquifer system using geophysical methods. Frontiers in Earth Science 1–15, (2024).

Mohammed MAA, Mohamed A, Szabó NP, & Szűcs P., Development of machine learning - based models for identifying the sources of nitrate and fluoride in groundwater and predicting their human health risks. International Journal of Energy and Water Resources, (2024).

Mohammed MAA, Kaya F, Mohamed A, Alarifi S., Abdelrady A., Keshavarzi A., Szabó, N. P, & Szűcs, P., Application of GIS-based machine learning algorithms for prediction of irrigational groundwater quality indices. Frontiers in Earth Science 1–19, (2023).

Mohammed MAA, Abdelrahman MMG, Szabó NP, & 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. Sustainable Water Resources Management 9:1–16, (2023).

Mohammed MAA, Szabó NP, & Szűcs P., Assessment of the Nubian aquifer characteristics by combining geoelectrical and pumping test methods in the Omdurman area, Sudan. Modeling Earth Systems and Environment 9:4363–4383, (2023).

Mohammed MAA, Szabó NP, Szűcs P., Delineation of groundwater potential zones in northern Omdurman area using electrical resistivity method. IOP Conference Series: Earth and Environmental Science 1189:12012, (2023).

Mohammed MAA, Szabó NP, & Szűcs P., Characterization of groundwater aquifers using hydrogeophysical and hydrogeochemical methods in the eastern Nile River area, Khartoum State, Sudan. Environmental Earth Sciences (2023).

Mohammed MAA, Eltijani A, Szabó NP, & Szűcs P., Hydro-chemometrics of the Nubian Aquifer in Sudan: an integration of groundwater quality index, multivariate statistics, and human health risk assessment. Discover Water 3:15, (2023).

Mohammed, Musaab, Kovács, B., Szabó, N.P., & Szűcs, P. Steady-state simulation of groundwater flow in Khartoum state, Sudan, Pollack Periodica, 18(3). doi:10.1556/606.2023.00758, (2023).

Mohammed, M. A., Szabó, N. P., & Szűcs, P. Exploring hydrogeological parameters by integration of geophysical and hydrogeological methods in northern Khartoum state, Sudan. Groundwater for Sustainable Development, 20, 100891, (2023).

Mohammed M, Szabó NP, Szucs P, Prediction of protective capacity of the Nubian Aquifer using electrical resistivity method in Bahri City, Sudan. Geosciences and Engineering 11:5–19, (2023).

Mohammed M, Abba SI, Szabó N, & Szűcs P, Management of agricultural groundwater in Sudan: The use of artificial intelligence algorithms in Khartoum State. Geosciences and Engineering 11:, (2023).

Mohammed, M. A., Khleel, N. A., Szabó, N. P., & Szűcs, P. Modeling of groundwater quality index by using artificial intelligence algorithms in northern Khartoum State, Sudan. Modeling Earth Systems and Environment, 1-16, (2022).

Mohammed, M. A., Szabó, N. P., & Szűcs, P. Multivariate statistical and hydrochemical approaches for evaluation of groundwater quality in north Bahri city-Sudan. Heliyon, 8(11), e11308, (2022).

Mohammed, M. A., Elskiekh, A., E., Szabó, N. P., & Szűcs, P. Hydrogeological investigations in basement terrains by using geological, geomorphological and geophysical methods, western Hamissana area, NE Sudan, Geosciences and engineering journal, (2022).

Mohammed M. A. A., The use of Landsat ETM+ in hydrogeological investigation in basement terrain, Hamissana area, NE Sudan, Humanitarian and Natural Sciences Journal, (2020).

Abdelrady M, Moneim MA, Alarifi SS, et al. Geophysical investigations for the identification of subsurface features influencing mineralization zones. Journal of King Saud University - Science 35:102809. (2023)

Eltijani A, Mohammed MAA, Abuobida Y, Yousif IM. Integrating CoDA and PCA for enhanced characterization of fluvial depositional processes: a case study of the Shendi formation, Sudan. Discover Geoscience 2:. (2024)

Mohamed A, Alarifi SS, Al-Kahtany K, Mohammed MAA. Application of gravity and remote sensing data to groundwater storage variation in Wadi Al Dawasir, Saudi Arabia. Journal of King Saud University - Science 36:103172. (2024)

Mohamed A, Alarifi SS, Mohammed MAA. Geophysical monitoring of the groundwater resources in the Southern Arabian Peninsula using satellite gravity data. Alexandria Engineering Journal 86:311–326. (2024b)

Mohamed A, Asmoay A, Alarifi SS, Mohammed MAA. Simulation of Surface and Subsurface Water Quality in Hyper-Arid Environments. Hydrology. (2023)

Omeiza J, Ghassan H, Ayejoto DA, et al. Evaluation of Groundwater contamination and the Health Risk Due to Landfills using integrated geophysical methods and Physiochemical Water Analysis. Case Studies in Chemical and Environmental Engineering 8:100523. (2023)

Alao JO, Ayejoto DA, Fahad A, Mohammed MA, Saqr AM, Joy AO. Environmental burden of waste generation and management in Nigeria. InTechnical Landfills and Waste Management: Volume 2: Municipal Solid Waste Management (pp. 27-56). Cham: Springer Nature Switzerland. (2024)

Abdelmalek D, Azzeddine R, Mohamed A, Faouzi Z, Galal WF, Alarifi SS, Mohammed MA. Groundwater quality assessment using revised classical diagrams and compositional data analysis (CoDa): Case study of Wadi Ranyah, Saudi Arabia. Journal of King Saud University-Science. 1;36(10):103463. (2024)

Dione PM, Faye C, Mohamed A, Alarifi SS, Mohammed MA. Assessment of the impact of climate change on current and future flows of the ungauged Aga-Foua-Djilas watershed: a comparative study of hydrological models CWatM under ISIMIP and HMF-WA. Applied Water Science.14(7):163. (2024)

Abdelrady M, Pham LT, Mohamed A, Alarifi SS, Duong VH, Mohammed MA. Application of the new edge filters of aeromagnetic data to detect the subsurface structural elements controlling the mineralization in the Barramiya area, Eastern Desert of Egypt. Journal of King Saud University-Science. 1;36(11):103539. (2024)

Mohamed A, Othman A, Asmaoy A, Galal WF, Mohammed MA. Assessment of heavy metal pollution of groundwater at the upper stream of Wadi Ranyah, Saudi Arabia, using multivariate statistical approach. Applied Water Science. 15(4):72. (2025)

Mohammed SH, Mohammed MA, Karim HA, Al-Manmi DA, Aziz BQ, Mustafa AI, Szűcs P. Integrating geospatial, hydrogeological, and geophysical data to identify groundwater recharge potential zones in the Sulaymaniyah basin, NE of Iraq. Scientific Reports. 22;15(1):9920. (2025)

Mohieldain AA, Dobróka M, Mohammed MA, Szabó NP. Gravity-based structural and tectonic characterization of the Shendi-Atbara Basin, Central Sudan. Journal of African Earth Sciences. 8:105571. (2025)

REFERENCES

Alger RP (1971) Interpretation of electric logs in fresh water wells in unconsolidated formations. SPE Reprint Series 1:255

Anderson MP, Woessner WW, Hunt RJ (2015) Applied groundwater modeling: simulation of flow and advective transport. Academic press

Archie GE (1942) The electrical resistivity log as an aid in determining some reservoir characteristics. Transactions of the AIME 146:54–62

Asfahani J, Ahmad Z, Ghani BA (2018) Self organizing map neural networks approach for lithologic interpretation of nuclear and electrical well logs in basaltic environment, Southern Syria. Applied Radiation and Isotopes 137:50–55. https://doi.org/10.1016/j.apradiso.2018.03.008

Bear J, Verruijt A (2012) Modeling groundwater flow and pollution. Springer Science \& Business Media

Bendefy L (1968) Debrecen városi belsosége süllyedésének hidrogeológiai vonatkozásai (Hydrogeologic Aspects of Settlement Subsidences Observed in Debrecen). Hidrológiai Közlöny 48:549–559

Buday T, Püspöki Z (2011) Facies Variations Detected by Well Log Correlation in a Geothermal Reservoir (Újfalu Formation) around Debrecen, Hungary. In: 6th Congress of the Balkan Geophysical Society. p cp--262

Buday T, Szűcs P, Kozák M, et al (2015) Sustainability aspects of thermal water production in the region of Hajdúszoboszló-Debrecen, Hungary. Environmental Earth Sciences 74:7511–7521. https://doi.org/10.1007/s12665-014-3983-1

Carman PC (1937) Fluid flow through granular beds. Trans Inst Chem Eng 15:150–166

Carpio YGF (2024) High resolution facies interpretation of Quaternary sedimentary series for hydrogeological modeling of the Nyírség-Hajdúság groundwater body, NE Hungary. PhD Thesis, University of Miskolc

Cho K, Van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:14061078

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:14123555

Csókás J (1995) Determination of yield and water quality of aquifers based on geophysical well logs. Magyar Geofizika 35:176–203

Czauner B, Szijártó M, Sztanó O, et al (2024) Re-interpreting renewable and non-renewable water resources in the over-pressured Pannonian Basin. Scientific Reports 14:24586. https://doi.org/10.1038/s41598-024-76076-8

Dayhoff JE (1990) Neural network architectures: an introduction. Van Nostrand Reinhold Co.

Doll HG t (1949) The SP log: Theoretical analysis and principles of interpretation. Transactions of the AIME 179:146–185

Dramsch JS (2020) 70 Years of Machine Learning in Geoscience in Review. Advances in Geophysics 61:1–55. https://doi.org/10.1016/bs.agph.2020.08.002

Erdélyi M (1976) Outlines of the hydrodynamics and hydrochemistry of the Pannonian Basin. Acta Geologica Hungarica 20:287–309

Géron A (2022) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc."

Haas J (2012) Geology of Hungary. Springer Science \& Business Media

Hill MC, Banta ER, Harbaugh AW, Anderman ER (2000) MODFLOW-2000, the US Geological Survey modular ground-water model; user guide to the observation, sensitivity, and parameter-estimation processes and three post-processing programs

Hochreiter S (1997) Long Short-term Memory. Neural Computation MIT-Press

Holland JH (1992) Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press

Horrocks T, Holden E-J, Wedge D (2015) Evaluation of automated lithology classification architectures using highly-sampled wireline logs for coal exploration. Computers & Geosciences 83:209–218. https://doi.org/https://doi.org/10.1016/j.cageo.2015.07.013

Hussain W, Luo M, Ali M, et al (2023) Machine learning - a novel approach to predict the porosity curve using geophysical logs data: An example from the Lower Goru sand reservoir in the Southern Indus Basin, Pakistan. Journal of Applied Geophysics 214:105067. https://doi.org/10.1016/j.jappgeo.2023.105067

Imamverdiyev Y, Sukhostat L (2019) Lithological facies classification using deep convolutional neural network. Journal of Petroleum Science and Engineering 174:216–228. https://doi.org/10.1016/j.petrol.2018.11.023

Jazayeri A, Werner AD (2019) Boundary condition nomenclature confusion in groundwater flow modeling. Groundwater 57:664–668

Jorgensen DG (1988) Estimating permeability in water-saturated formations. The Log Analyst 29:

Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. pp 1942–1948

Keys WS (1990) Borehole geophysics applied to ground-water investigations. US Department of the Interior, US Geological Survey

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Klute A, Dirksen C (1986) Hydraulic conductivity and diffusivity: Laboratory methods. Methods of soil analysis: Part 1 physical and mineralogical methods 5:687–734

Kobr M, Mareš S, Paillet F (2005) Geophysical well logging: Borehole geophysics for hydrogeological studies: Principles and applications. Hydrogeophysics 291–331

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological cybernetics 43:59–69

Kozeny J (1927) Uber kapillare leitung der wasser in boden. Royal Academy of Science, Vienna, Proc Class I 136:271–306

Krásný J (1993) Classification of Transmissivity Magnitude and Variation. Groundwater 31:230–236

Larionov V V (1969) Radiometry of boreholes. Nedra, Moscow 127:

LeCun Y, Boser B, Denker JS, et al (1989) Backpropagation applied to handwritten zip code recognition. Neural computation 1:541–551

Lillicrap TP, Santoro A (2019) Backpropagation through time and the brain. Current opinion in neurobiology 55:82–89

Logan J (1964) Estimating transmissibility from routine production tests of water wells. Groundwater 2:35–37

Mádl-Szőnyi J, Tóth J (2009) A hydrogeological type section for the Duna-Tisza Interfluve, Hungary. Hydrogeology Journal 17:961–980. https://doi.org/10.1007/s10040-008-0421-z

Mannor S, Peleg D, Rubinstein R (2005) The cross entropy method for classification. In: Proceedings of the 22nd international conference on Machine learning. pp 561–568

Marton L, Szanyi J (1997) Kelet-magyarországi pleisztocén üledékek geostatisztikai vizsgálata. A rétegek közötti területi átszivárgás meghatározása (Geostatistical investigation of Pleistocene sediments in Eastern Hungary. Determination of local recharge rates). Hidrológiai Közlöny 77:241–248

Mehl SW, Hill MC (2006) MODFLOW-2005, the US Geological Survey modular ground-water model-documentation of shared node local grid refinement (LGR) and the boundary flow and head (BFH) package

Mohammed MAA, Abdelrahman MMG, Szabó NP, Szűcs P (2023) Innovative hydrogeophysical approach for detecting the spatial distribution of hydraulic conductivity in Bahri city , Sudan : A comparative study of Csókás and Heigold methods. Sustainable Water Resources Management 9:1–16. https://doi.org/10.1007/s40899-023-00885-4

Mohammed MAA, Flores YG, Szabó NP, Szűcs P (2024a) Assessing heterogeneous groundwater systems : Geostatistical interpretation of well logging data for estimating essential hydrogeological parameters. Scientific Reports 1–17. https://doi.org/10.1038/s41598-024-57435-x

Mohammed MAA, Mohammed SH, Szabó NP, Szűcs P (2024b) Geospatial modeling for groundwater potential zoning using a multi-parameter analytical hierarchy process supported by geophysical data. Discover Applied Sciences 6:. https://doi.org/10.1007/s42452-024-05769-6

Mohammed MAA, Szabó NP, Alao JO, Szűcs P (2024c) Geophysical characterization of groundwater aquifers in the Western Debrecen area , Hungary : insights from gravity , magnetotelluric , and electrical resistivity tomography. Sustainable Water Resources Management 10:. https://doi.org/10.1007/s40899-024-01062-x

Mohammed MAA, Szabó NP, Eltijani A, Szűcs P (2025a) An integrated workflow combining machine learning and wavelet transform for automated characterization of heterogeneous groundwater systems. Scientific Reports 1–20

Mohammed MAA, Szabó NP, Flores YG, Szűcs P (2024d) Multi-well clustering and inverse modeling-based approaches for exploring geometry, petrophysical, and hydrogeological parameters of the Quaternary aquifer system around Debrecen area, Hungary. Groundwater for Sustainable Development 101086. https://doi.org/https://doi.org/10.1016/j.gsd.2024.101086

Mohammed MAA, Szabó NP, Kilik R, Szűcs P (2024e) Examining innovative unsupervised learning techniques for automated characterization of complex groundwater systems. Results in Engineering 23:1–10. https://doi.org/10.1016/j.rineng.2024.102594

Mohammed MAA, Szabó NP, Mikita V, Szűcs P (2025b) Long-term assessment of groundwater contamination and associated health risks: A study of heavy metals in the Quaternary aquifer system in Pannonian Basin. Physics and Chemistry of the Earth, Parts A/B/C 139:103935. https://doi.org/https://doi.org/10.1016/j.pce.2025.103935

Mohammed MAA, Szabó NP, Mikita V, Szűcs P (2025c) Tracking the spatiotemporal evolution of groundwater chemistry in the Quaternary aquifer system of Debrecen area, Hungary: integration of classical and unsupervised learning methods. Environmental Science and Pollution Research 1–20

Mohammed MAA, Szabó NP, Szűcs P (2024f) Robust estimation of hydrogeological parameters from wireline logs usingsemi-supervised deep neural networks assisted with global optimization-based regression methods. Groundwater for Sustainable Development 27:. https://doi.org/10.1016/j.gsd.2024.101348

Mohammed MAA, Szabó NP, Szűcs P (2024g) Joint interpretation and modeling of potential field data for mapping groundwater potential zones around Debrecen. Acta Geodaetica et Geophysica. https://doi.org/10.1007/s40328-023-00433-8

Mohammed MAA, Szabó NP, Szűcs P (2024h) Multi - step modeling of well logging data combining unsupervised and deep learning algorithms for enhanced characterization of the Quaternary aquifer system in Debrecen area , Hungary. Modeling Earth Systems and Environment. https://doi.org/10.1007/s40808-024-01986-5

Mohammed MAA, Szabó NP, Szűcs P (2025d) High - resolution characterization of complex groundwater systems using wireline logs analyzed with machine learning classifiers and isometric mapping techniques. Modeling Earth Systems and Environment 1:1–17. https://doi.org/10.1007/s40808-024-02263-1

NATéR National Adaptation Geo-information System. https://map.mbfsz.gov.hu/nater/

Nelson PH (1994) Permeability-porosity relationships in sedimentary rocks. The log analyst 35:

Oliver MA, Webster R (1990) Kriging: a method of interpolation for geographical information systems. International Journal of Geographical Information System 4:313–332

Paillet FL, Crowder RE (1996) A generalized approach for the interpretation of geophysical well logs in ground-water studies—theory and application. Groundwater 34:883–898

Panday S, Langevin CD, Niswonger RG, et al (2013) MODFLOW--USG version 1: An unstructured grid version of MODFLOW for simulating groundwater flow and tightly coupled processes using a control volume finite-difference formulation

Popescu M-C, Balas VE, Perescu-Popescu L, Mastorakis N (2009) Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems 8:579–588

Püspöki Z, Demeter G, Tóth-Makk Á, et al (2013) Tectonically controlled Quaternary intracontinental fluvial sequence development in the Nyírség-Pannonian Basin, Hungary. Sedimentary Geology 283:34–56. https://doi.org/10.1016/j.sedgeo.2012.11.003

Püspöki Z, Fogarassy-Pummer T, Thamó-Bozsó E, et al (2021) 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. https://doi.org/10.1111/bor.12471

Püspöki Z, Kovács IJ, Fancsik T, et al (2016) Magnetic susceptibility as a possible correlation tool in Quaternary alluvial stratigraphy. Boreas 45:861–875. https://doi.org/10.1111/bor.12196

Rubin Y, Hubbard SS (2006) Hydrogeophysics. Springer Science \& Business Media

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. nature 323:533–536

Schlumberger (1991) Log interpretation principles/applications. Schlumberger Educational Services

Schlumberger (1984) Schlumberger Log Interpretation Charts. Schlumberger Well Services, Houston 1–21

Serra OE (1983) Fundamentals of well-log interpretation

Shevnin V, Delgado-Rodr\'\iguez O, Mousatov A, Ryjov A (2006) Estimation of hydraulic conductivity on clay content in soil determined from resistivity data. Geof{\'\i}sica internacional 45:195–207

Shi Y, Eberhart R (1998) A modified particle swarm optimizer. In: 1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360). pp 69–73

Simon S, Déri-takács J, Szijártó M, et al (2023) Wetland Management in Recharge Regions of Regional Groundwater Flow Systems with Water Shortage , Nyírség. Water (Switzerland) 15:

Spearman C (1961) The proof and measurement of association between two things.

Steiner F. (1988) Most frequent value procedures (a short monograph). Geophysical Transactions 34:139–260

Szabó NP (2011) Shale volume estimation based on the factor analysis of well-logging data. Acta Geophysica 59:935–953. https://doi.org/10.2478/s11600-011-0034-0

Szabó NP (2015) Hydraulic conductivity explored by factor analysis of borehole geophysical data. Hydrogeology Journal 23:869–882

Szabó NP, Abordán A, Dobróka M (2022) Permeability extraction from multiple well logs using particle swarm optimization based factor analysis. GEM - International Journal on Geomathematics 13:1–27. https://doi.org/10.1007/s13137-022-00200-x

Szabó NP, Braun BA, Abdelrahman MMG, Dobróka M (2021) Improved well logs clustering algorithm for shale gas identification and formation evaluation. Acta Geodaetica et Geophysica 56:711–729. https://doi.org/10.1007/s40328-021-00358-0

Szabó NP, Kiss A, Halmágyi A (2015a) Hydrogeophysical characterization of groundwater formations based on well logs: case study on cenozoic clastic aquifers in East Hungary. Geosciences and Engineering 4:45–71

Szabó NP, Kormos K, Dobróka M (2015b) Evaluation of hydraulic conductivity in shallow groundwater formations: a comparative study of the Csókás’ and Kozeny–Carman model.

Acta Geodaetica et Geophysica 50:461–477. https://doi.org/10.1007/s40328-015-0105-9

Szanyi J (2004) The environmental effects of underground water production on the example of Dél Nyírség. University of Szeged (Hungary)

Székely F, Deák J, Szűcs P, et al (2020) Verification of Radiocarbon Transport Predicted by Numerical Modeling in the Porous Formation of NE Hungary Considering Paleo-Hydrogeology. Radiocarbon 62:219–233. https://doi.org/10.1017/RDC.2019.84

Sztanó O, Magyar I, Csillag G, et al (2023) Felső miocén - pliocén. In: Babinszki E (ed) Magyarország litosztratigráfiai egységeinek leírása II. Szabályozott Tevékenységek Felügyeleti Hatósága, Budapest, pp 117–137

Szűcs P, Civan F, Virág M (2006) Applicability of the most frequent value method in groundwater modeling. Hydrogeology Journal 14:31–43. https://doi.org/10.1007/s10040-004-0426-1

Szűcs P, Szabó NP, Zubair M, Szalai S (2021) Innovative Hydrogeophysical Approaches as Aids to Assess Hungarian Groundwater Bodies. Applied Sciences 11:2099. https://doi.org/10.3390/app11052099

Theim G (1906) Hydrologische methoden. Gebhardt, Leipzig 56:475

Timur A (1968) An investigation of permeability, porosity, and residual water saturation relationships. In: SPWLA Annual Logging Symposium. p SPWLA--1968

Tóth J (2009) Gravitational systems of groundwater flow: theory, evaluation, utilization. Cambridge University Press

Tóth J, Almási I (2001) Interpretation of observed fluid potential patterns in a deep sedimentary basin under tectonic compression: Hungarian Great Plain, Pannonian Basin. Geofluids 1:11–36. https://doi.org/10.1046/j.1468-8123.2001.11004.x

Tselentis G-A (1985) The processing of geophysical well logs by microcomputers as applied to the solution of hydrogeological problems. Journal of Hydrology 80:215–236

Varsányi I (1992) Temporal variability in groundwater chemistry in the great hungarian plain during the period 1975–1989. Hydrological Sciences Journal 37:119–128. https://doi.org/10.1080/02626669209492572

Varsányi I, Palcsu L, Ó Kovács L (2011) Groundwater flow system as an archive of palaeotemperature: Noble gas, radiocarbon, stable isotope and geochemical study in the Pannonian Basin, Hungary. Applied Geochemistry 26:91–104. https://doi.org/https://doi.org/10.1016/j.apgeochem.2010.11.006

Wang W, Huang Y, Wang Y, Wang L (2014) Generalized autoencoder: A neural network framework for dimensionality reduction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp 490–497

Wiener J, Rogers J, Moll B (1995) Predict permeability from wireline logs using neural networks. Petroleum Engineer International 68:

Wiener JM, Rogers JA, Rogers JR, Moll RF (1991) Predicting carbonate permeabilities from wireline logs using a back-propagation neural network. In: SEG Technical Program Expanded Abstracts 1991. Society of Exploration Geophysicists, pp 285–288

Williams J, Paillet F (2023) Geophysical logging for hydrogeology. New York Water Science Center

Yang Y, Aplin AC, Larter SR (2004) Quantitative assessment of mudstone lithology using geophysical wireline logs and artificial neural networks. Petroleum Geoscience 10:141–151

Zhu L, Li H, Yang Z, et al (2018) Intelligent logging lithological interpretation with convolution neural networks. Petrophysics 59:799–810