PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

UNIVERSITY OF KASDI MERBAH OUARGLA

Faculty of Hydrocarbons, Renewable Energies, Earth & Universe Sciences

Department of Earth & Universe Sciences

MASTER'S THESIS

Submitted in fulfilment of the requirements for the degree of

Master of Science in Applied Geophysics

Advanced Processing, Inversion and Uncertainty Quantification of EM31 HCP/VCP Profiles for Mapping Surface–Subsurface Interaction along the Hamiz River

Presented by:

Larbi Marouf & Tarek Messaoudi

Supervised by:

Nemer Zoubida & Mohamed Djeddi

Jury :

Pr Belekseir Mohamed Salah ……………………………………………………………President

Dr Benkhelifa Randa ……………………………………………………………………Examiner

Dr Nemer Zoubida ………………………………………………………………………Supervisor

Pr Djeddi Mohamed ……………………………………………………………………Co-Supervisor

Academic Year 2025–2026

Contents

List of Figures

List of tables

Abstract

Alluvial aquifers underlying the Mitidja Plain in northern Algeria are vital freshwater sources but remain vulnerable to contamination from Oued El Hamiz, a river carrying substantial industrial, domestic, and agricultural pollutant loads. Because groundwater abstraction has intensified river infiltration along discrete, spatially heterogeneous pathways, conventional point-based monitoring (piezometers, sampling wells) cannot adequately delineate where contaminated river water enters the aquifer. This study presents an integrated geophysical–machine learning framework to map river infiltration vulnerability along a 3 km reach of the Oued El Hamiz corridor. A total of 476 EM31 electromagnetic induction stations were surveyed across three transects parallel to the river bank, with measurements acquired in both horizontal dipole (~3 m depth) and vertical dipole (~6 m depth) configurations. Two-dimensional apparent conductivity images were constructed to visualize subsurface heterogeneity, and a rule-based vulnerability index was developed by integrating hydraulic gradient, sediment texture, and proximity to the river. Unsupervised (K-Means) and supervised (Random Forest classification and regression) machine learning models were then applied to delineate hydrogeological zones, predict infiltration risk, and quantify controlling variables through feature importance analysis. Conductivity imaging revealed discrete low-conductivity "windows" corresponding to permeable sand and gravel deposits that act as preferential infiltration pathways, contrasted with higher-conductivity, clay-dominated zones that impede recharge. K-Means clustering identified four distinct hydrogeological zones, while the Random Forest classifier achieved 97.9% accuracy in predicting vulnerability categories, with the VD/HD conductivity ratio emerging as the dominant predictive feature—indicating that vertical stratification, rather than shallow conductivity alone, governs infiltration potential. Rule-based and machine learning approaches showed strong agreement (99.4%), and 125 critical infiltration zones were identified for priority field verification, including ten stations recommended for immediate monitoring well installation. These findings demonstrate that coupling EM31 geophysical surveying with machine learning provides an efficient, non-invasive methodology for characterizing river–aquifer interaction, offering actionable guidance for groundwater protection and management in the Mitidja Plain and comparable intensively exploited alluvial systems.

Chapter 1: INTRODUCTION

1.1 Background

Groundwater constitutes a vital freshwater resource for agricultural, industrial, and domestic supply in arid and semi-arid regions worldwide. Alluvial aquifers, in particular, are highly productive due to their permeable sediments and hydraulic connection with surface water bodies. However, this same connectivity renders them vulnerable to contamination from polluted rivers, especially in areas where groundwater abstraction intensifies natural infiltration processes (Hami Amira et al., 2026). Understanding where and how river water enters the aquifer is therefore essential for protecting groundwater quality and managing water resources sustainably.

The Mitidja Plain, located in northern Algeria, is a fertile alluvial basin of approximately 1,400 km² that supports intensive agriculture, industrial activities, and domestic supply for over 1.5 million inhabitants, including the eastern suburbs of Algiers. The plain is drained by several perennial rivers, among which Oued El Hamiz represents a major surface water feature. This river originates in the southern Blidean Atlas mountains and flows northward through the plain before discharging into the Mediterranean Sea (Nemer., 2024). However, Oued El Hamiz carries significant contaminant loads from upstream industrial zones, untreated domestic sewage, and agricultural runoff, including elevated concentrations of nitrates, heavy metals, and pathogenic bacteria (Dalale Khous et al., 2026).

The aquifer underlying the Mitidja Plain is predominantly unconfined and receives recharge from river infiltration, direct precipitation, and lateral inflow from adjacent mountainous areas. Groundwater abstraction has intensified over recent decades, creating drawdown cones that may enhance river infiltration. This pumping-enhanced recharge, while potentially increasing available water resources, also creates a pathway for contaminated river water to enter the aquifer, posing serious threats to groundwater quality and public health (Fatima Kastali et al., 2024).

1.2 Problem Statement

The fundamental scientific question driving this research is: Where is contaminated river water infiltrating into the aquifer? This question is critical because infiltration is not uniform along the river corridor. It occurs preferentially through discrete "windows" of clean sand where streambed permeability is high, hydraulic gradients favor river-to-aquifer flow, and pumping wells enhance drawdown. These windows cannot be identified through traditional point measurements such as piezometers or sampling wells alone.

Conventional methods for assessing river-aquifer interactions typically rely on limited point data, failing to capture the spatial heterogeneity that controls infiltration patterns. Geophysical methods, particularly electromagnetic induction, offer the potential to characterize subsurface conductivity continuously along the river corridor, providing proxies for sediment texture, pore-water salinity, and infiltration zones. However, interpreting these geophysical signals for infiltration vulnerability requires integrating multiple factors including hydraulic gradient, sediment permeability, vertical stratification, and anthropogenic influences such as pumping wells.

1.3 Objectives

The primary objective of this study is to assess river infiltration vulnerability along the Oued El Hamiz corridor using an integrated approach combining EM31 geophysical surveying, two-dimensional conductivity imaging, and machine learning. The specific objectives are:

1. To acquire high-resolution EM31 conductivity data at two investigation depths (3 m and 6 m) along three transects parallel to the river bank;
2. To construct two-dimensional conductivity images for each transect to visualize subsurface conductivity patterns and identify potential infiltration zones;
3. To develop a rule-based vulnerability index integrating hydraulic, textural, and spatial factors;
4. To apply unsupervised machine learning (K-Means clustering) for hydrogeological zonation;
5. To apply supervised machine learning (Random Forest classification and regression) for risk prediction and feature importance analysis;
6. To identify critical infiltration zones and anomalous stations requiring field verification; and
7. To provide actionable recommendations for groundwater monitoring and management.

1.4 Significance of the Study

This study makes several important contributions. Methodologically, it demonstrates the value of integrating EM31 geophysical surveying with machine learning for infiltration vulnerability assessment, providing a template applicable to other alluvial settings. Scientifically, it identifies the primary controls on river infiltration, with particular emphasis on vertical stratification as captured by the VD/HD ratio. Practically, it provides specific, actionable information for groundwater managers in the Mitidja Plain, including a prioritized list of critical zones for monitoring well installation and field verification.

1.5 Scope and Limitations

This study focuses on a 3 km segment of Oued El Hamiz within the eastern Mitidja Plain. The analysis is based on EM31 surveys conducted during May of 2018, 2019, and 2020, representing a temporal snapshot during dry season conditions. Piezometric data were obtained from the existing monitoring well network, and pumping well data were compiled from available records. The study does not include direct water quality measurements, relying instead on geophysical proxies for infiltration potential. Temporal variability, seasonal effects, and inter-annual changes are not captured.

1.6 Thesis Organization

This thesis is organized into six chapters.

Chapter 1 provides the introduction, background, problem statement, objectives, significance, scope and limitations, and thesis organization.

Chapter 2 presents the study area description, including the geographic and hydrogeological setting of the Mitidja Plain, the surface water resources of the region, and detailed characterization of Oued El Hamiz.

Chapter 3 presents a review of relevant literature on surface water-groundwater interaction methods, the principles of frequency domain electromagnetic methods and EM31 specifics, and previous geophysical studies in alluvial aquifers.

Chapter 4 describes the data collection procedures, two-dimensional conductivity imaging using Surfer software, feature engineering, rule-based vulnerability index calculation, and the machine learning framework including K-Means clustering, Random Forest classification and regression, and anomaly detection.

Chapter 5 presents the results of the two-dimensional conductivity imaging, clustering analysis, classification and regression models, feature importance ranking, anomaly detection, and critical zone identification, followed by a discussion of the findings in relation to the research objectives and existing literature.

Chapter 6 summarizes the key findings, draws conclusions, and provides recommendations for groundwater management and future research.

Chapter 2: STUDY AREA

2.1 Geographic and Hydrogeological Setting

The Mitidja Plain is an alluvial basin located in northern Algeria, extending approximately 100 km east-west and 10-20 km north-south, covering an area of approximately 1,400 km². The plain is bounded by the Mediterranean Sea to the north, the Blidean Atlas mountains to the south, the Oued Mazafran valley to the west, and the Oued El Hamiz catchment to the east. The aquifer system consists of Quaternary alluvial deposits composed of gravels, sands, silts, and clays, with thickness ranging from 20 to 150 meters. The aquifer is predominantly unconfined and receives recharge from river infiltration, direct precipitation, and lateral inflow from adjacent mountainous areas. Groundwater abstraction has intensified over recent decades to support agricultural irrigation, industrial activities, and domestic supply for approximately 1.5 million inhabitants, including the eastern suburbs of Algiers (Figure 1).

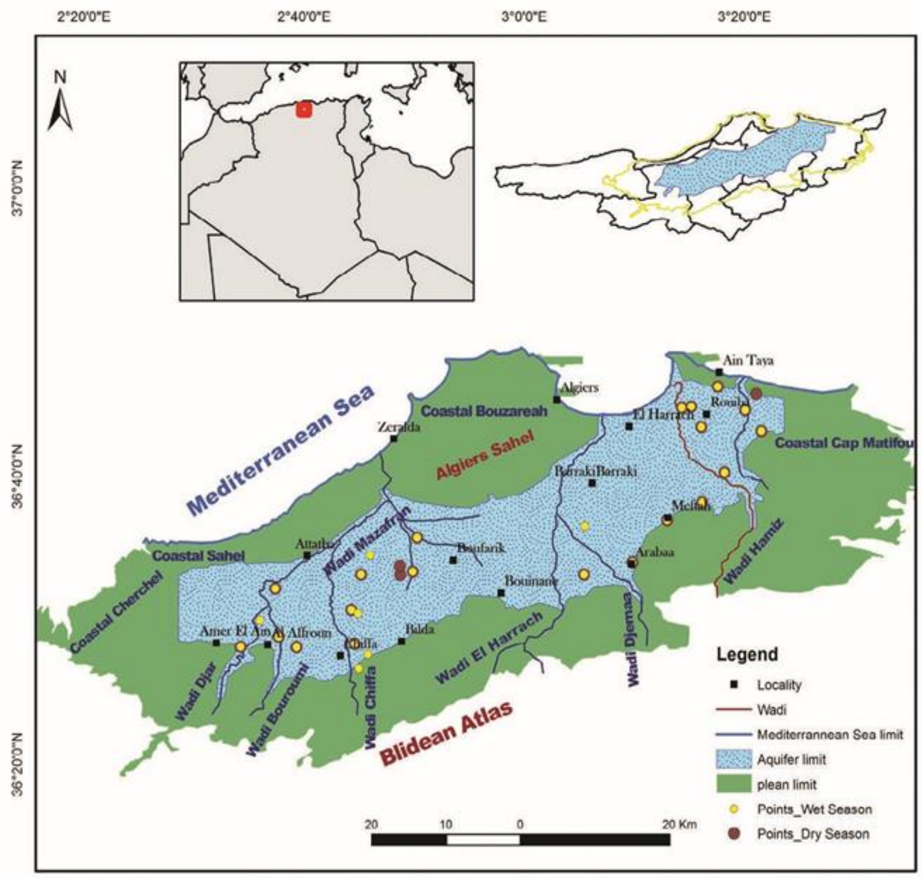

Figure 1:Localization map of Mitidja plain. (Ouahiba Aziez et al., 2025)

2.2 Surface Water Resources

The Mitidja Plain is drained by several perennial and intermittent rivers (oueds) that originate in the southern Blidean Atlas mountains and flow northward toward the Mediterranean Sea. The major surface water features include:

Table 1: Main Rivers (Oueds) of the Mitidja Plain and Their Hydrological Functions

River (Oued)	Length (km)	Catchment (km^2)	Flow regime	Primary function
Oued El Harrach	85	1,500	Perennial	Drainage, irrigation
Oued El Hamiz	55	450	Perennial	Recharge, agriculture
Oued Mazafran	65	1,200	Perennial	Water supply, irrigation
Oued Boudouaou	45	300	Intermittent	Local drainage
Oued Khemis	70	800	Perennial	Irrigation

These rivers historically provided natural recharge to the alluvial aquifer through bank infiltration and streambed percolation. However, urbanization, industrial development, and agricultural intensification have significantly altered both water quantity and quality. Several rivers receive untreated or partially treated wastewater from adjacent cities and industrial zones, raising concerns about groundwater contamination (Fatima Kastali et al., 2024).

Despite the obvious hydraulic connection between surface water and groundwater in this alluvial setting (Imene Senadi et al., 2026), no comprehensive study has quantitatively evaluated the relationship between these surface water resources and the Mitidja aquifer. Existing investigations have focused either on surface water quality monitoring or on groundwater abstraction statistics, without integrating both systems to assess recharge dynamics or contamination pathways. This knowledge gap is particularly critical given the increasing pressure on water resources from climate variability (reduced precipitation, prolonged droughts) and anthropogenic activities (intensive pumping, pollution loading).

2.3 Research Gap and Study Focus

Understanding how surface water contributes to groundwater recharge, and conversely, how groundwater abstraction influences river infiltration, is essential for sustainable water management. However, the absence of integrated studies has left several fundamental questions unanswered:

- Which rivers actively recharge the aquifer versus those that lose water to the aquifer?
- Where are the preferential infiltration zones along each river corridor?
- How does intensive groundwater pumping modify natural river-aquifer exchange?
- What is the risk of contaminated river water entering production wells?

To address these gaps, this thesis aims to evaluate river-aquifer interactions using an integrated approach combining geophysical surveys, spatial data analysis, and machine learning techniques. The study specifically focuses on one of the major surface water features in the eastern Mitidja Plain: Oued El Hamiz.

2.4 Oued El Hamiz: Detailed Characterization

Oued El Hamiz is a perennial river originating in the southern Blidean Atlas mountains at approximately 400 m above sea level. The river flows northward for approximately 55 km through the Mitidja Plain before discharging into the Mediterranean Sea east of Algiers (Figure 2). The catchment area is approximately 450 km², encompassing diverse land uses including forested headwaters, agricultural plains, and urban-industrial zones near the coast (Fatima Kastali et al., 2024).

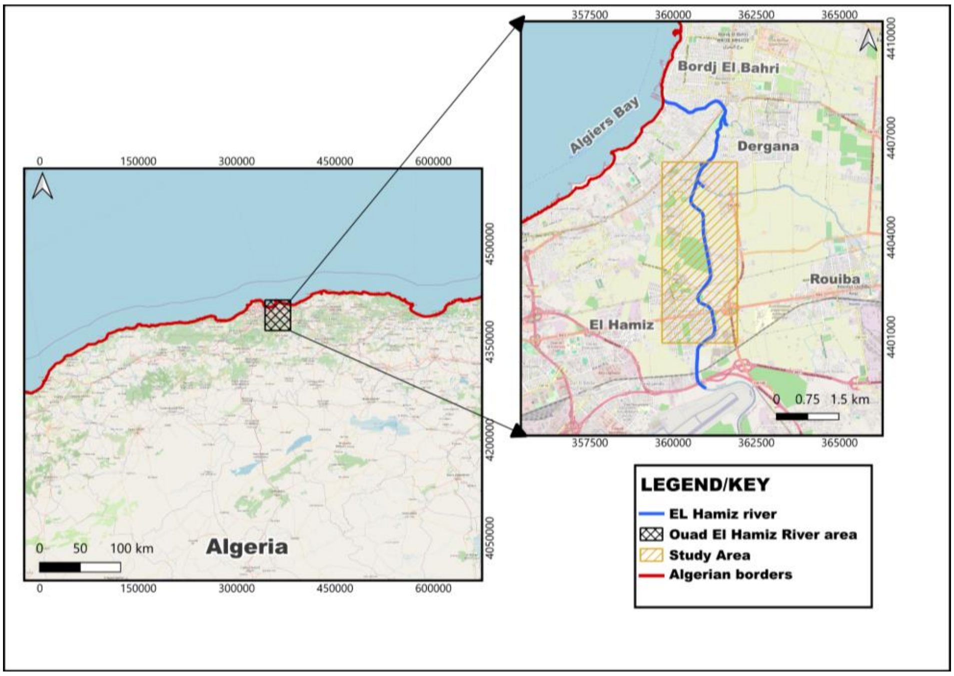

Figure 2: Localization Map of Ouad EL Hamiz

- ### 2.4.1 Hydrological Regime

The flow regime of Oued El Hamiz is characterized by:

Table 2: Hydrological Flow Characteristics of Oued El Hamiz

Parameter	Value
Mean annual flow	1.2 m³/s
Maximum flow (wet season)	15-25 m³/s (November-April)
Minimum flow (dry season)	0.1-0.3 m³/s (June-September)
Baseflow contribution	Approximately 40% from spring discharge
Peak flood events	Associated with Mediterranean rainfall (autumn/spring)

The river exhibits a flashy response to precipitation events, with rapid hydrograph rise following intense rainfall in the mountainous headwaters. During dry periods, baseflow is sustained by springs and shallow groundwater exfiltration, maintaining perennial conditions along most of its course.

- ### 2.4.2 Water Quality Status

Oued El Hamiz receives multiple contamination sources along its trajectory:

Table 3: Main Sources of Water Contamination in Oued El Hamiz

Source type	Location	Primary contaminants
Industrial zone	Reghaïa (mid-reach)	Heavy metals (Cr, Pb, Zn), hydrocarbons
Domestic sewage	Multiple points	Nitrates, phosphates, pathogens (E. coli, coliforms)
Agricultural runoff	Lower plain	Nitrates, pesticides, herbicides
Urban stormwater	Coastal section	Suspended solids, hydrocarbons, trace metals

Previous water quality assessments have documented elevated concentrations of nitrates (>50 mg/L), chromium (>0.05 mg/L), and fecal coliforms (>1,000 CFU/100 mL) in the mid-to-lower reaches, exceeding national and international water quality guidelines for drinking water and irrigation (Khouli et al., 2021).

- ### 2.4.3 Geological Setting

The study focuses on a 3 km segment of Oued El Hamiz located within the eastern Mitidja Plain at coordinates 36.73°N, 3.24°E. This reach is characterized by a channel geometry with a width of 15–25 m, a depth of 1–3 m, and low to moderate sinuosity. The streambed composition consists of alternating gravel bars, sand lenses, and clay-rich sections (Z. Nemer et al., 2023), while the floodplain deposits comprise Quaternary alluvium, including sands, gravels, silts, and clays, reaching thicknesses of up to 50 m (Fatima Kastali et al., 2024). The bank structure features natural levees with localized erosion and undercutting (George N. Zaimes et al., 2019), and the adjacent land use is dominated by intensive agriculture, including market gardening and cereals, alongside scattered rural settlements (Khouli et al., 2021).

- ### 2.4.4 Relevance to River-Aquifer Interaction

Oued El Hamiz was selected for this study due to its strong relevance to river-aquifer interaction, which is demonstrated through several key factors. First, the alluvial aquifer is directly connected to the river, as no continuous confining layer separates surface and groundwater, establishing a significant hydraulic connection. Second, the area faces considerable abstraction pressure, with numerous irrigation wells located within 500 m of the river that potentially induce river infiltration. Third, there is a notable contamination risk, since elevated pollutant concentrations in the river pose a direct threat to groundwater quality in the absence of natural attenuation. Fourth, adequate data availability supports the research, as existing piezometric and pumping data provide essential baseline information for model development. Finally, the site holds substantial management relevance because the river contributes to recharge for wells supplying both agricultural and domestic users, making the understanding of infiltration zones critical for effective wellhead protection (Fatima Kastali et al., 2024).

Chapter 3: LITERATURE REVIEW

3.1 Surface water-groundwater interaction methods

Surface water and groundwater have long been considered separate entities and have traditionally been investigated individually. The chemical, biological, and physical properties of surface water and groundwater are indeed different. However, in the transition zone between them, a variety of processes occur, leading to transport, degradation, transformation, precipitation, or sorption of chemical substances. Water exchange between groundwater and surface water may have a significant impact on the water quality of either hydrological zone. Consequently, the transition zone plays a critical role in mediating interactions between groundwater and surface water (Kalbus et al., 2006).

3.1.1 Classification of Stream-Aquifer Interactions

Under pristine conditions, interactions between streamflow and groundwater can be divided into three categories (Figure 3) (Bierkens and Wada, 2019):

- Gaining streams: Groundwater discharge contributes significantly to streamflow (Figure 3a).
- Losing streams: Groundwater is replenished by infiltrating water from the stream, which adds to recharge from precipitation (Figure 3b).
- Disconnected losing streams (Figure 3c): The groundwater level is not connected to the streambed, and infiltration loss from the stream depends mainly on river level. This infiltration quickly becomes constant at greater groundwater depths.

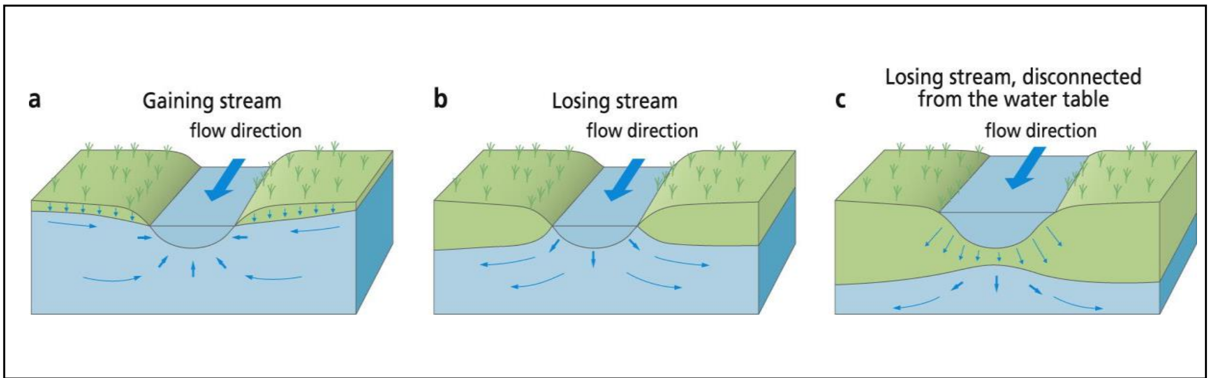

Figure 3: Groundwater-streamflow interaction: (a) gaining stream; (b) losing stream; (c) losing stream disconnected from the water table; modified from (Winter et al 1998) (USGS Bierkens and Wada, 2019).

3.1.2 Hydraulic Controls on Exchange Direction

For groundwater to discharge into a stream, the groundwater level near the stream must be higher than the water level at the stream surface. Conversely, for surface water to infiltrate into groundwater, the groundwater level near the stream must be lower than the elevation of the stream surface. The movement of water between groundwater and surface water represents an important pathway for the transfer of chemical substances and contaminants (Nemer, 2024).

- ### 3.1.3 Influence of Groundwater Pumping

Furthermore, groundwater pumping near streams can significantly influence these exchanges. The distance between a surface water body and a pumping location is a key factor controlling the timing and magnitude of pumping impacts on surface water infiltration (Johnson et al., 2012). When pumping occurs close to a hydraulically connected surface water body, the effects are nearly immediate and proportional to the pumping rate. At greater distances, the effects are distributed over longer periods and may be shared among multiple hydraulically connected surface water bodies (Johnson et al., 2012).

- ### 3.1.4 Aquifer Properties and Pumping Impacts

The physical characteristics of the aquifer, such as stratification, transmissivity, and storage properties, also influence the timing and magnitude of stream depletion caused by pumping. Wells screened in deeper layers may have a more delayed and diffuse impact on surface water bodies compared to wells screened in shallower layers. Highly transmissive aquifers with limited storage capacity transmit impacts more rapidly than less permeable aquifers or those with higher storage capacity (Johnson et al., 2002).

- ### 3.1.5 Common Misconceptions

A common misconception is that the impacts of groundwater pumping follow the estimated groundwater flow paths within an aquifer (Miller et al., 2003). In reality, these impacts propagate radially in all directions, regardless of the groundwater flow direction. This means that pumping effects can be observed upstream, laterally across the hydraulic gradient, and downstream (Miller et al., 2003).

3.2 Principles of Frequency Domain Electromagnetic (FDEM) Methods and EM31 Specifics

- ### 3.2.1 Basic Principles

Frequency domain electromagnetic (FDEM) methods are used to detect variations in subsurface electrical conductivity (or its inverse, resistivity) using electromagnetic induction (EMI) principles. FDEM devices continuously transmit an electromagnetic (EM) signal generated through the EMI process (Figure 4).

A power source supplies an alternating current (AC) at a specific frequency to a closed, multi-turn coil of wire. Current flow within the transmitter coil generates a magnetic field that surrounds it and penetrates the subsurface. This time-varying primary magnetic field induces an electromotive force in nearby materials due to their EM properties, including electrical conductivity, magnetic susceptibility, and dielectric permittivity (EPA, 2021).

The electromotive force generates electrical eddy currents within subsurface conductors, which in turn produce a secondary magnetic field similar to the primary field. Relative to the primary field, the secondary magnetic field exhibits an amplitude and phase shift due to the EM properties of the subsurface materials. An AC is induced within a surface receiver coil as it is energized by both the primary and secondary magnetic fields and is monitored by a sensor that records the voltage across the coil (EPA, 2021).

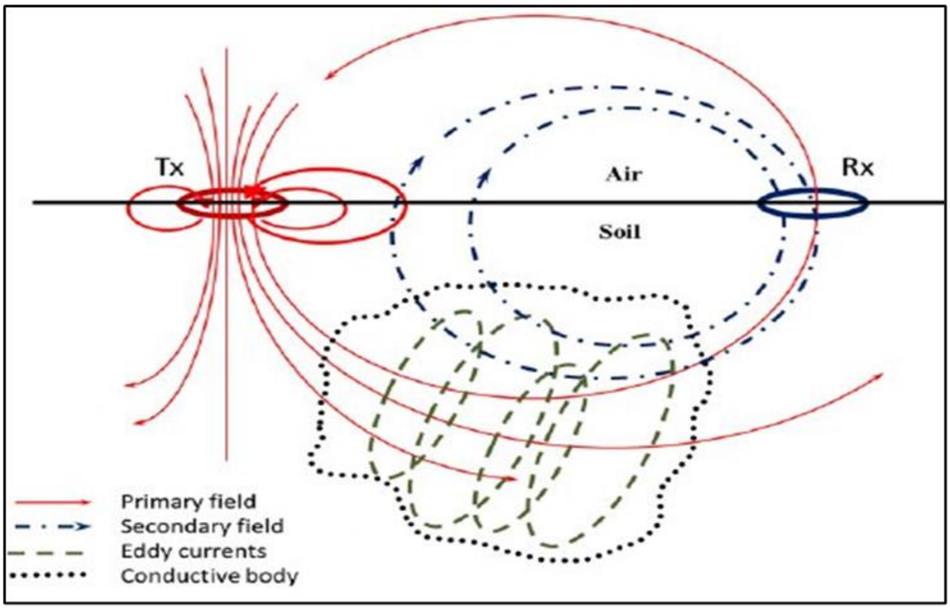

Figure 4: The principle of EM induction. The primary magnetic field (Hp) induces eddy currents in the subsurface, generating a secondary magnetic field (Hs) measured by the receiver coil (Manstein et al., 2015).

- 3.2.2 The EM31 Instrument

The EM31 (Geonics, Mississauga, ON, Canada) is an electromagnetic induction instrument with a transmitter coil and a receiver coil maintained at a constant separation distance of 3.66 m. The instrument is typically operated approximately 1 m above the ground, with an operating frequency of 9.8 kHz. The EM31 measures the terrain's apparent electrical conductivity (ECa) in millisiemens per meter (mS/m) (Figure 5).

The transmitter coil, energized with an alternating current, produces an electromagnetic field (primary field), which induces very small electrical currents in the terrain. These currents, in turn, generate a secondary magnetic field. The receiver coil senses both the primary and secondary fields. Under the operating conditions of the EM31 in a homogeneous half-space, the ratio of the secondary to the primary magnetic field components is linearly proportional to the electrical conductivity of the terrain. The instrument measures this ratio directly. In a layered terrain, the contribution of each layer to the measured apparent conductivity depends on its depth, thickness, and conductivity (Alte-da-Veiga et al., 2022).

Figure 5: Geonics EM31 Ground Conductivity Meter. The instrument showing transmitter and receiver coils at fixed separation (Source: Pyramid Environmental).

Two operating modes are possible with the EM31:

Table 4: Main Characteristics of EM31 Dipole Configurations

Mode	Coil orientation	Investigation depth	Application
Vertical Dipole	Coils horizontal (normal position)	~6 m	Deeper investigation, sensitive to deeper conductivity contrasts
Horizontal Dipole	Coils vertical (instrument rotated 90°)	~3 m	Shallower investigation, sensitive to near-surface conditions

In vertical dipole mode, the penetration depth is approximately 6 meters, while in horizontal dipole mode, the penetration depth is approximately 3 meters (Alte-da-Veiga et al., 2022). This dual-depth capability makes the EM31 particularly suitable for characterizing vertical variations in subsurface conductivity associated with river infiltration.

3.3 Previous Studies in Alluvial Aquifers

Numerous investigations have effectively delineated zones of interaction between surface water and groundwater. A selection of relevant studies is reviewed below.

- ### 3.3.1 Geochemical Approaches

Hassan Mahamat et al. (2025) aimed to identify, better understand, and quantify water exchanges between an alluvial aquifer, the river, and other hydrographic and hydrogeological entities bordering their study site. They employed a combination of classical geochemical tools, including major ions (Ca^{2+}, Mg^{2+}, K^+, Na^+, Cl^-, HCO_3^-, CO_3^{2-}, NO_3^-), isotopic tracers (δ^2H and $\delta^{18}O$ of water), and piezometric data. This multi-method approach allowed them to trace the origin and movement of water masses and identify zones of active exchange.

- ### 3.3.2 EM31 and EM38 for Groundwater Flow Characterization

Andrew Binley et al. (2013) used EM31 and EM38 electromagnetic instruments deployed on the water surface, in combination with piezometric and chemical data, to investigate the spatial variability of groundwater flow. Their primary finding demonstrated that high electrical conductivity, associated with solute-rich groundwater upwelling, was effectively correlated with zones of groundwater discharge to the stream. Conversely, low electrical conductivity corresponded to regions dominated by horizontal hydraulic gradients, where stream water was likely infiltrating into the aquifer.

- ### 3.3.3 Boating EM31 Surveys for Streambed Characterization

Vogelgesang et al. (2024), from the Iowa Geological Survey, conducted a boating EM31 survey along a river to characterize streambed conductivity and groundwater recharge potential in the alluvial system. The survey results showed varying levels of electrical conductivity, ranging from 17 to 205 mS/m. These variations reflected the differences between coarse, sandy sediments (low ECa) and finer materials, such as silts and clays (high ECa), which dominate the streambed and influence infiltration rates and groundwater-surface water interactions.

Chapter 4: DATA AND METHODOLOGY

This chapter presents the data collection and methodological framework implemented to assess river infiltration vulnerability along the Oued El Hamiz corridor. The approach integrates electromagnetic geophysical surveying, spatial data compilation, development of a rule-based vulnerability index, unsupervised machine learning for hydrogeological zonation, supervised machine learning for risk classification and regression, and explainable artificial intelligence for feature interpretation. The methodology is organized into three main sections: data acquisition and preprocessing, feature engineering, and machine learning implementation (Figure 6).

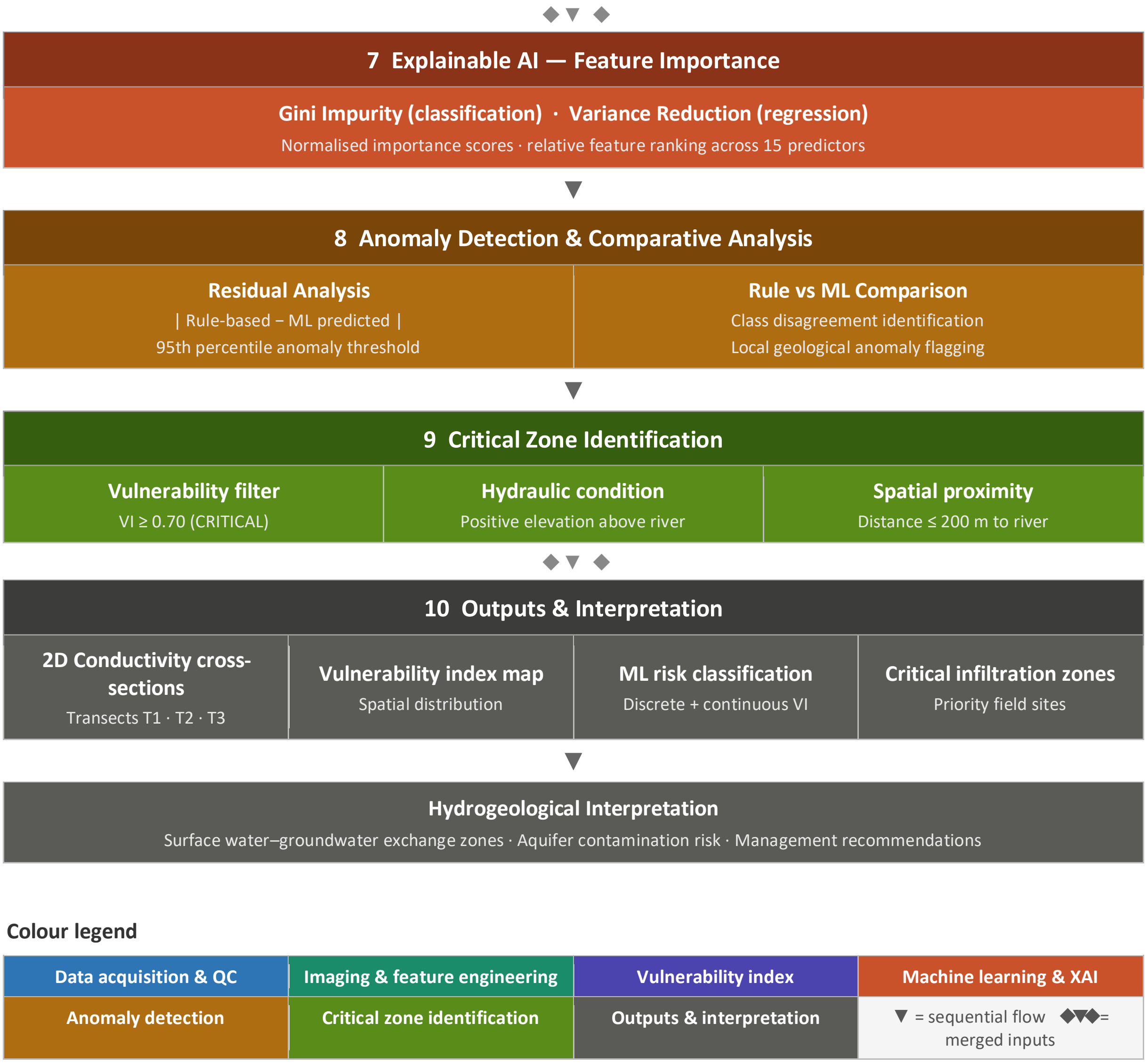

Figure 6: Research Methodology Framework

4.1 Data Acquisition

- ### 4.1.1 EM31 Geophysical Survey

A comprehensive EM31 survey was conducted along three transects parallel to the Oued El Hamiz river. The survey comprised a total of 476 stations distributed as follows: Transect T1 included 168 stations, Transect T2 comprised 160 stations and Transect T3 included 148 stations. Station spacing was maintained at approximately five meters along each transect.

At each station, measurements were recorded in both horizontal dipole mode, providing an effective investigation depth of approximately three meters and sensitive to near-surface conditions including soil moisture, sediment texture, and pore-water salinity, and vertical dipole mode, providing an effective investigation depth of approximately six meters and sensitive to deeper conductivity contrasts and vertical stratification. Additional field data included station identification, date and time of measurement, geographic coordinates in UTM, surface condition codes, and quality control flags

- ### 4.1.2 Data Quality Control

Each measurement was assigned a quality control flag during field data entry based on instrument stability and external interference. Flags included OK for reliable readings, RECHECK for potentially affected readings, and REJECT for invalid readings. Stations flagged as REJECT were excluded from subsequent analysis. The remaining stations were subjected to automated quality screening based on ECa range validation within five to five hundred millisiemens per meter, VD to HD ratio plausibility between 0.3 and 5.0, and consistency with adjacent stations through spatial outlier detection.

- ### 4.1.3 Spatial Data Compilation

Piezometric data were obtained from the existing monitoring well network, comprising 80 wells distributed across the Mitidja Plain with measurements taken during May of 2018, 2019, and 2020. For each monitoring well, the following parameters were recorded: unique well identifier, UTM easting and northing coordinates in Zone 31N, ground elevation in meters above sea level, depth to water table in meters below ground surface, and calculated water table elevation.

A pumping well inventory was compiled from the well registry and supplemented with field surveys, resulting in 45 wells with recorded parameters including unique identifiers, UTM coordinates, pumping rates in m3/day, operating hours per day, calculated annual pumping volumes, and use type categorized as agriculture, urban, or industrial

- ### 4.1.4 Elevation and Distance Data

Ground elevation relative to river level was recorded at each EM31 station. This parameter serves as a proxy for the hydraulic gradient direction, with positive values indicating the station is above river level where the river can potentially infiltrate into the aquifer, and negative values indicating the station is below river level where the river cannot infiltrate due to opposing hydraulic gradient. The Euclidean distance from each EM31 station to the Oued El Hamiz river channel was calculated using QGIS.

4.2 Two-Dimensional Conductivity Imaging

To visualize the spatial distribution of subsurface electrical conductivity along each transect and to identify zones of active infiltration, two-dimensional conductivity images were constructed using the EM31 measurements. These images provide a continuous representation of conductivity variations along the length of each transect, enabling the identification of lateral heterogeneity and preferential infiltration pathways that may not be apparent from point measurements alone.

- ### 4.2.1 Data Preparation for Imaging

For each transect, the EM31 measurements were organized into datasets containing station coordinates along the transect and the corresponding $□□□_{□□}$ and $□□□_{□□}$ values. Both horizontal dipole mode and vertical dipole mode measurements were integrated into a single dataset for each transect to generate comprehensive images that capture conductivity variations at both shallow and deeper investigation depths simultaneously. Station positions were represented by their distance along the transect from a defined origin, typically the starting point of the survey line.

- ### 4.2.2 Gridding and Interpolation

The combined conductivity data were imported into Surfer software (Golden Software, LLC) for gridding and visualization. Gridding was performed using the kriging interpolation method, which provides optimal interpolation weights based on the spatial correlation structure of the data.. The kriging parameters were configured with a linear variogram model, and the grid dimensions were set to maintain a resolution consistent with the station spacing of approximately five meters. The search radius for interpolation was defined to include a sufficient number of neighboring points while preserving local anomalies.

Importantly, the gridding process incorporated both □□□□□ and □□□□□ measurements together, allowing the generation of a single continuous conductivity image per transect that reflects the combined information from both dipole modes. This approach provides a more comprehensive representation of subsurface conditions by integrating the shallow sensitivity of the horizontal dipole mode with the deeper penetration of the vertical dipole mode

- ### 4.2.3 Image Generation and Interpretation

The gridded combined data were visualized as two-dimensional color contour maps with distance along the transect plotted on the horizontal axis and investigation depth represented by the combined response of both dipole modes. A consistent color scale was applied across all images to facilitate comparison between transects. The color scale was selected to highlight conductivity contrasts, with low ECa values typically represented by cool colors indicating sand-dominated sediments or fresh water, and high ECa values represented by warm colors indicating clay-rich sediments or elevated pore-water salinity.

The resulting images were interpreted to identify zones of anomalously low conductivity that may represent clean sand windows with high infiltration potential, zones of anomalously high conductivity that may represent clay lenses acting as natural barriers or zones of contaminated water infiltration, lateral continuity or discontinuity of conductive layers, and spatial patterns of conductivity that correlate with distance from the river or with known surface conditions (Philip Brunner et al., 2017). The integration of both dipole modes in a single image allows for the identification of vertical conductivity gradients and stratification patterns that might otherwise be missed when viewing each mode separately.

- ### 4.2.4 Integration with Point Data

The two-dimensional conductivity images were used in conjunction with the point measurements to guide the selection of features for the machine learning analysis. Zones identified as having distinct conductivity signatures in the combined HD-VD images were targeted for detailed examination of individual station data, and spatial patterns observed in the images informed the engineering of interaction features that capture lateral heterogeneity. The combined images also facilitated the identification of stations where the relationship between shallow and deep conductivity deviates from the surrounding area, indicating potential infiltration zones or geological boundaries.

4.3 Data Pre-processing and Feature Engineering

- ### 4.3.1 Data Cleaning

The raw EM31 data were processed through a sequential cleaning procedure. Stations with QC Flag equal to REJECT were removed from the dataset. Missing values were handled by removing stations with missing $\square\square\square_{\square\square}$, $\square\square\square_{\square\square}$, latitude, or longitude values. Extreme ECa values < 5 mS/m or > 500 mS/m were flagged and reviewed, though no stations were removed based solely on ECa range unless accompanied by a REJECT flag. The VD to HD ratio was calculated as $\square\square\square_{\square\square}/\square\square\square_{\square\square}$, and ratios below 0.3 or above 5.0 were flagged for review, with all flagged stations passing manual inspection.

- ### 4.3.2 Derived EM31 Features

From the raw EM31 measurements, several derived features were calculated. The VD to HD ratio was computed as $\square\square\square_{\square\square}/\square\square\square_{\square\square}$, providing an indicator of vertical homogeneity where a value near 1 indicates a homogeneous subsurface profile, a value greater than 1.3 indicates increasing conductivity with depth often associated with clay-rich layers, and a value < 0.8 indicates decreasing conductivity with depth often associated with sand-dominated profiles (John Triantafilis et al., 2010). The ECa difference was computed as $\square\square\square_{\square\square} - \square\square\square_{\square\square}$, providing a measure of the depth gradient of conductivity (John Triantafilis et al., 2010). The ECa product was computed as $\square\square\square_{\square\square} \times \square\square\square_{\square\square}$, providing a combined measure of overall conductivity magnitude.

- ### 4.3.3 Spatial Feature Engineering

Additional spatial features were created through mathematical transformations to capture non-linear relationships. The log-transformed distance was computed as $D_{\log} = \ln(d+1)$

where d is the distance to the river (m). The squared elevation was computed as elevation above the river raised to the second power, capturing non-linear gradient effects: $E_{\text{sq}} = E^2$

where E is the elevation above the river (m). Two interaction terms were created to allow the machine learning models to capture situations where the influence of one variable depends on the value of another. The first interaction term combines distance and elevation: $I_{dE} = d \times E$

The second interaction term combines $\square\square\square_{\square\square}$ and distance to the river: $I_{\text{ECa},d} = \text{ECa}_{\text{HD}} \times d$

These engineered features provide the machine learning models with flexible representations of the coupled spatial and geophysical controls on river infiltration vulnerability.

- ### 4.3.4 Pumping-Related Features

For each EM31 station, two pumping-related features were calculated. The distance to the nearest pumping well was computed as the minimum Euclidean distance to any well with a pumping rate greater than zero cubic meters per day:

$$d_{\text{well}} = \min_i \sqrt{(x_s - x_i)^2 + (y_s - y_i)^2}, \quad \forall i \text{ where } Q_i > 0$$

Where $(\square_{\square}, \square_{\square})$ are the station coordinates, $(\square_{\square}, \square_{\square})$ are the coordinates of well i , and $\square_{\square}$ is the pumping rate of well *i* (m³/day). The cumulative pumping within 500 meters was calculated as the sum of pumping rates for all wells located within a 500 meter radius of the station:

$$Q_{\text{cum},500} = \sum_{i=1}^{n} Q_i \cdot \mathbb{I}\left(d_i \leq 500\right)$$

where $\square_{\square}$ is the Euclidean distance from the station to well $\square$ and $\mathbb{I}(\cdot)$ is the indicator function equal to 1 when the condition is satisfied and 0 otherwise. These features quantify the localized abstraction pressure that may induce river infiltration and alter hydraulic gradients near the stream-aquifer interface.

4.3.5 Categorical Variable Encoding

Surface condition codes recorded during field data collection were converted to numerical format using one-hot encoding. The original categories included SAND, GRAVEL, CLAY, SILT, MOIST, DRY, WET, MIXED, VEG, CROP, and URBAN. Each category was transformed into a separate binary indicator variable that takes a value of one when the corresponding surface condition is present and zero otherwise. Similarly, transect assignment was one-hot encoded into three binary indicators for T1, T2, and T3.

4.3.6 Feature Standardization

All numerical features were standardized prior to machine learning model training using the StandardScaler function from the scikit-learn library, which transforms each feature to have a mean of zero and a standard deviation of one. This ensures that features with larger numerical ranges do not dominate those with smaller ranges.

4.4 Rule-Based Vulnerability Index

4.4.1 Component Development

A rule-based vulnerability index was developed to integrate the multiple factors controlling river infiltration into a single continuous score ranging from 0 to 1. Five components were selected based on the conceptual model of river infiltration processes (B. Dixon., 2005).

The hydraulic regime component uses elevation above the river as a proxy for the hydraulic gradient direction. The score for this component was calculated as:

$$S_{\text{hydraulic}} = \min\left(\frac{E}{2}, 1\right)$$

The permeability component uses $\square\square\square_{\square\square}$ as an indicator of sediment permeability. The score was calculated as:

$$S_{\text{perm}} = \max\left(0, \min\left(1 - \frac{\text{ECa} - 20}{80}, 1\right)\right)$$

This formulation assigns a full score of one to stations with ECa less than 20 mS/m, indicating clean sand with high permeability, and a score of zero to stations with ECa greater than 100 mS/m (Tóth, T et al., 2006).

The vertical texture component uses the VD to HD ratio as an indicator of vertical homogeneity. The score was calculated as:

$$S_{\text{vertical}} = \max\left(0, \min\left(1 - \frac{|R-1|}{1.5}, 1\right)\right)$$

This formulation assigns a full score of 1 to stations with a VD to HD ratio exactly equal to 1, indicating perfect vertical homogeneity.

The river proximity component uses distance to the river in meters. The score was calculated as

$$S_{\text{prox}} = \max\left(0, \min\left(1 - \frac{d}{200}, 1\right)\right)$$

This assigns a score of 1 to stations at the river bank and progressively lower scores with increasing distance.

The surface condition component uses the visual surface classification recorded during field surveys. Each surface type was assigned a weight based on its expected permeability, with SAND and GRAVEL receiving a weight of 1, CLAY and SILT receiving a weight of 0.2, and other surfaces receiving intermediate weights

- ### 4.4.2 Composite Index Calculation

The composite vulnerability index was calculated as the weighted sum of the five component scores. The hydraulic regime component received the highest weight of 35%. The permeability component received a weight of 25%. The vertical texture component received a weight of 15%. The river proximity component received a weight of 15%. The surface condition component received a weight of 10%

Risk levels were assigned based on the vulnerability index values. Stations with a vulnerability index of 0.7 or higher were classified as CRITICAL risk. Stations with a vulnerability index between 0.5 and 0.69 were classified as HIGH risk. Stations with a vulnerability index between 0.3 and 0.9 were classified as MODERATE risk. Stations with a vulnerability index below 0.3 were classified as LOW risk

4.5 Machine Learning Framework

- ### 4.5.1 Feature Matrix Construction

The final feature matrix for machine learning analysis comprised 15 features. Raw EM31 features included □□□□□, □□□□□, VD to HD ratio, distance to river, and elevation above river. Derived EM31 features included ECa difference and ECa product. Engineered features included log-transformed distance, squared elevation, distance-elevation interaction, and ECa-distance interaction. Categorical features included one-hot encoded surface condition variables for the DRY surface type and one-hot encoded transect variables for T1, T2, and T3.

- ### 4.5.2 Unsupervised Learning: K-Means Clustering

K-Means clustering was applied to identify natural groupings of stations based on their hydrogeophysical characteristics without using any prior information about risk levels or vulnerability. The algorithm partitions observations into a specified number of clusters by minimizing the within-cluster sum of squares (Abiodun M. Ikotun et al., 2023). The optimal number of clusters was determined using the elbow method, which involves computing the within-cluster sum of squares for values of k ranging from 2 to 8 and identifying the point where the rate of decrease substantially slows (Indra Herdiana et al., 2025). The features were scaled using StandardScaler prior to clustering, and the K-Means algorithm was applied with four clusters, random state of 42, and 10 initializations

- ### 4.5.3 Supervised Learning: Random Forest Classification

The Random Forest classifier was trained to predict categorical risk levels from the EM31 and spatial features. Random Forest is an ensemble learning method that constructs a large number of decision trees during training and outputs the mode of the individual tree predictions for classification tasks (Leo Breiman., 2001). The model was configured with 100 trees, maximum depth of 10 levels, minimum samples required to split an internal node set to 5, minimum samples required at a leaf node set to 2, bootstrap sampling enabled, and random state of 42.

This ensemble consisted of 100 decision trees. In the original paper on Random Forests, (Leo Breiman., 2001) showed that the error rates approach a lower bound as the size of the ensemble grows, and that random forests with 100 or so trees attain the best achievable accuracy on a dataset of moderate size.

The maximum depth of each decision tree in the ensemble was fixed at 10. Without a depth constraint (max_depth = None), trees will tend to grow until they are fully-grown, such that they fit the training data perfectly but predict poorly for unseen samples: a condition called overfitting (James et al., 2013).

The number of observations required in an internal node to proceed with splitting was set to 5. By default (minimum value of 2), very small samples can be split, which can lead to a risk of overfitting by allowing splits which are essentially driven by noise in the predictor variables; thus we increase the limit to 5 in order to ensure that any split must have a statistical significance, which results in more generalised trees (Hastie et al., 2009).

The minimum samples a leaf node required was 2. This parameter prevents the occurrence of leaf nodes that contain single samples which yield overly localized predictions with no statistical backing, and that overemphasize the impact of outlying samples in the training data (Leo Breiman., 2001).

The random state of 42 is set so that the experiment can be fully reproduced. Since there are a number of steps in the creation of a forest that use randomization (selecting bootstrap samples, selecting random sub-set of features for splitting, creation of tree), without fixing the random state the same forest may not be recreated when the analysis is run more than once, or shared with colleagues. Setting a random state ensures that exactly the same forest will be recreated, and that code can be reproduced by others (Peng, 2011; Gundersen and Kjensmo, 2018).

The dataset was divided into training and testing subsets using a stratified split to preserve the proportion of risk levels in both subsets. 70% of the stations were assigned to the training set and 30% to the testing set. Model performance was evaluated using accuracy, precision, recall, F1-score, and a confusion matrix comparing predicted risk levels to the rule-based risk levels used as training targets.

- #### 4.5.4 Supervised Learning: Random Forest Regression

The Random Forest regressor was trained to predict continuous vulnerability index values from the same feature set used for classification. The model configuration was identical to the classifier, with the output type set to regression. Performance was evaluated using the coefficient of determination R-squared, which indicates the proportion of variance in the vulnerability index explained by the model (Davide Chicco et al., 2021), and the root mean square error, which measures the average magnitude of prediction errors (Xuanhui Wang et al., 2018).

- #### 4.5.5 Explainable Artificial Intelligence: Feature Importance

Feature importance was extracted from the trained Random Forest models to identify which parameters most strongly control infiltration vulnerability. Importance is calculated as the average decrease in node impurity, measured by the Gini index for classification and variance for regression, across all trees in the forest (Leo Breiman., 2001). The values are normalized to sum to one, allowing direct comparison of the relative contribution of each feature.

- #### 4.5.6 ML Predictions and Comparative Analysis

The trained Random Forest models were used to generate predictions for all EM31 stations. ML-predicted vulnerability values were obtained from the regression model, and ML-predicted risk levels were obtained from the classification model. A comparative analysis between rule-based and ML predictions was conducted by calculating the residual as the absolute difference between rule-based vulnerability and ML-predicted vulnerability, and by identifying disagreements between rule-based risk levels and ML-predicted risk levels.

4.6 Anomaly Detection

Stations where the machine learning prediction substantially deviated from the rule-based vulnerability index were identified as anomalies requiring further investigation. The residual for each station was calculated as the absolute difference between the rule-based vulnerability index and the machine learning predicted vulnerability index. The anomaly threshold was set at the 95 percentiles of the residual distribution, meaning that only the 5% of stations with the largest discrepancies were flagged as anomalies. Stations exceeding this threshold may represent potential measurement errors, local geological anomalies not captured by regional patterns, or unique hydrogeological conditions with exceptional infiltration or barrier features (Haoxin Shi et al.,2023).

4.7 Critical Zones Identification

Critical infiltration zones representing the highest priority areas for field verification and potential remediation were identified by applying a set of logical filters to the station data. A station was classified as belonging to a critical zone if it satisfied three conditions simultaneously: the station must have a vulnerability index of 0.7 or higher, indicating CRITICAL risk; the station must have a positive elevation above river, indicating that the river is hydraulically capable of infiltrating at that location; and the station must be located within two 200 meters of the river, reflecting the spatial decay of river influence.

4.8 Software Implementation

All data processing, analysis, and visualization were performed using the Python programming language version 3.12. The pandas library version 2.0 was used for data manipulation and cleaning. The numpy library version 1.24 was used for numerical computing. The scikit-learn library version 1.3 was used for machine learning implementations including Random Forest classification and regression, K-Means clustering, and feature standardization. The scipy library was used for spatial interpolation. The matplotlib library version 3.7 and the seaborn library version 0.12 were used for visualization. Two-dimensional conductivity imaging was performed using Surfer software (Golden Software, LLC) for gridding and contouring. Geographic information system calculations including distance to river were performed using QGIS version 3.28.

Chapter 5: RESULTS AND DISCUSSION

This chapter presents the results of the infiltration vulnerability assessment along the Oued El Hamiz river corridor. The results are organized into five main sections. The first section presents the two-dimensional conductivity images generated for each transect, revealing subsurface conductivity patterns and identifying potential infiltration zones. The second section describes the unsupervised machine learning results, including the K-Means clustering analysis that identified distinct hydrogeological zones. The third section presents the supervised machine learning results, including Random Forest classification and regression performance, feature importance analysis, and comparison between rule-based and ML predictions. The fourth section identifies critical infiltration zones and anomalous stations requiring further investigation. The fifth section synthesizes the findings into a summary of key results.

5.1 Two-Dimensional Conductivity Imaging

- ### 5.1.1 Transect T1

This profile is the most structurally "clean." (Figure 7) The thin conductive surface layer likely represents moist alluvial fine sediments (silts, clays) or a capillary fringe fed by bank moisture (Wesley McCall., 1996). The thick resistive lower unit is consistent with unsaturated or low-salinity sandy/gravelly alluvium with minimal groundwater influence. The eastward thinning of the conductive cap may indicate a transition to drier, coarser facies or a reduction in riparian moisture (Yong Zeng et al., 2020). Surface water–groundwater interaction here appears limited and shallow , the riverbank is effectively buffering the hyporheic zone (F. Boano et al., 2014).

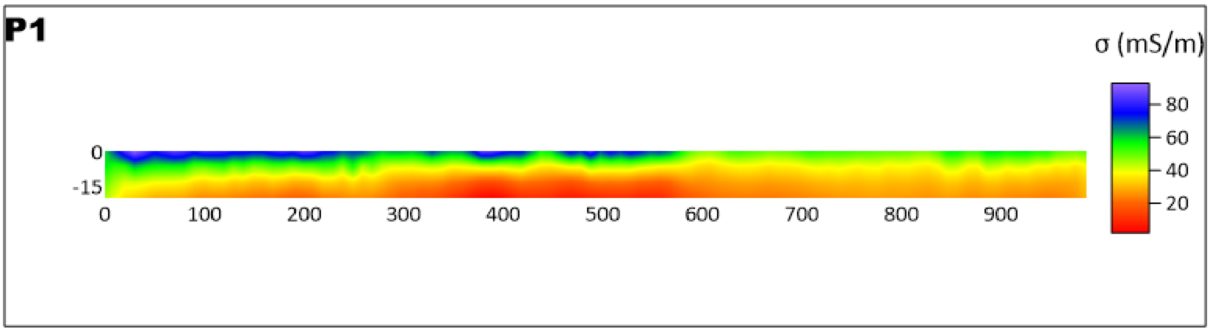

Figure 7: 2D electrical conductivity cross-section of Transect T1

- ### 5.1.2 Transect T2

The lateral patchiness of near-surface conductivity anomalies is geophysically significant: these are likely preferential infiltration windows (Figure 8), zones where river water penetrates the bank through more permeable or fractured sediment corridors (Martin A. Briggs et al., 2018). This heterogeneous pattern is characteristic of active hyporheic exchange, where bank storage operates episodically rather than diffusely (Dylan J. Irvine et al., 2024). The higher $\sigma_{\max}$ (up to 120 mS/m) compared to T1 may also reflect locally more mineralized water (possibly mixing of river water with deeper, longer-residence groundwater) (Suwannee River Water Management District, n.d.).

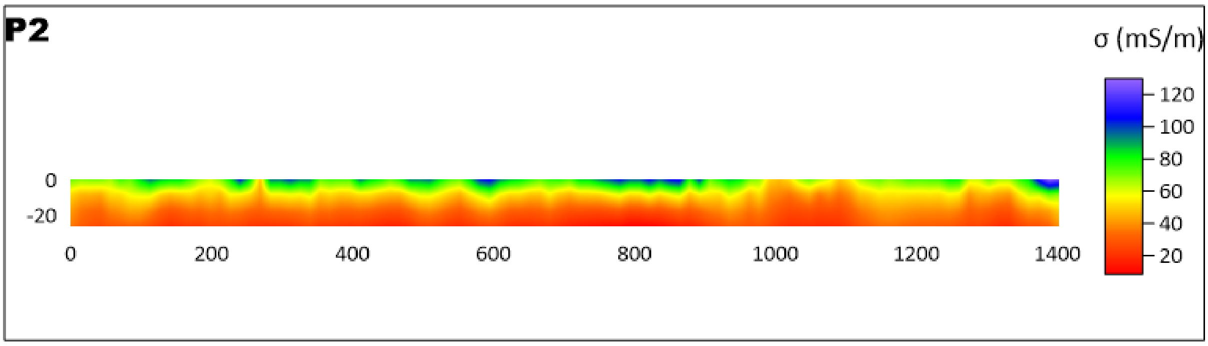

Figure 8: 2D electrical conductivity cross-section of Transect T2

5.1.3 Transect T3

T3 represents a fundamentally different hydrogeological environment (Figure 9). The elevated background σ_a indicates widespread saturation with mineralized pore fluids, this section of the bank is likely closer to or more directly connected with the river water column, or the local lithology has a higher clay matrix contribution (George R. Jiracek). The deep blue anomalies are the most interpretively important features of the entire dataset: they are consistent with either:

- Active bank infiltration plumes: river water (if the river is more mineralized than background groundwater) is infiltrating deeply along preferential pathways, classic bank filtration / bank storage signature (Ian Cartwright et al., 2020).
- Or conversely, groundwater discharge zones where longer-residence, more mineralized groundwater is upwelling toward the surface, potentially feeding or mixing with river water, an effluent seepage scenario (Antoine Biehler et al., 2020).

The two spatially separated blue anomalies suggest two distinct preferential flow corridors at this bank section, possibly controlled by buried palaeochannel structures or fault/fracture zones in the alluvium (Fayçal Rejiba et al., 2018).

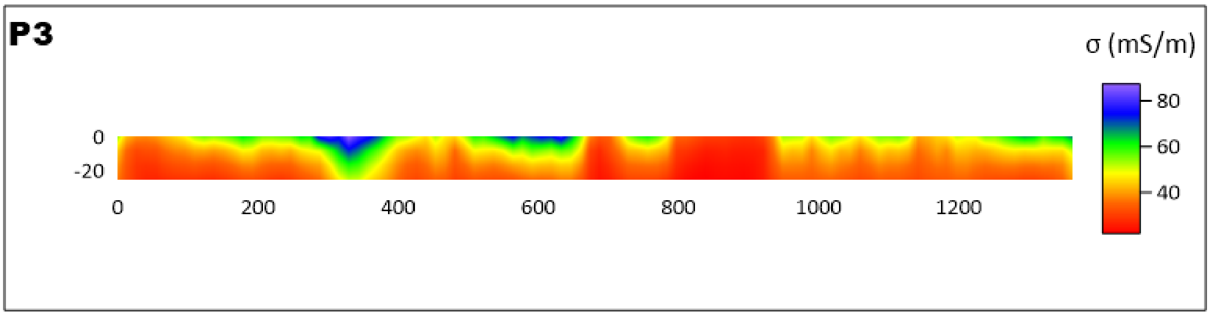

Figure 9: 2D electrical conductivity cross-section of Transect T3

5.1.4 Comparison Across Transects

A systematic comparison of the three transect images reveals distinct conductivity patterns among the transects. Transect T1 exhibits the greatest number of low conductivity anomalies, consistent with the presence of permeable sediments. Transect T2 shows intermediate characteristics, with a mix of low and high conductivity features. Transect T3 is dominated by higher conductivity values.

The discrete low conductivity windows observed in T1 and T2 provide evidence for the existence of preferential pathways that could serve as conduits for contaminated river water to enter the aquifer.

5.2 Unsupervised Learning: K-Means Clustering

- ### 5.2.1 Cluster Identification

K-Means clustering was applied to identify natural groupings of stations based on their hydrogeophysical characteristics without using prior information about risk levels. The elbow method indicated that four clusters provided an optimal partition of the data, balancing the number of clusters with the within-cluster sum of squares.

The resulting cluster distribution shows a pronounced dominance of one cluster containing one hundred twenty-five stations, representing the predominant hydrogeological condition across the study area. The remaining clusters contain zero to two stations, indicating that these represent rare or localized conditions that deviate substantially from the dominant pattern (Figure 10).

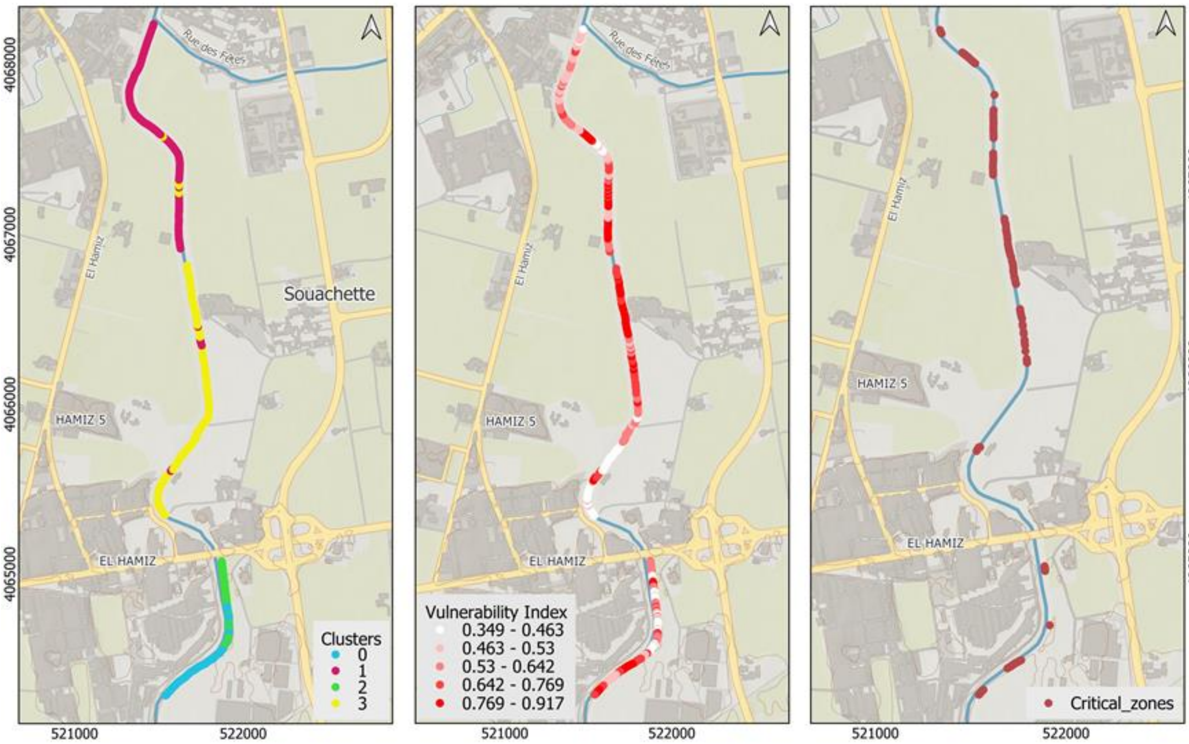

Figure 10: Spatial vulnerability map showing the distribution of clusters and critical zones across the study area. The dominant cluster (Cluster 0) is distributed across all transects, while minor clusters occur at specific locations. Critical zones (highlighted) represent priority areas for field verification.

- ### 5.2.2 Cluster Characterization

Cluster zero, containing the majority of stations, is characterized by moderate vulnerability values and represents the typical alluvial conditions of the Mitidja Plain. Stations in this cluster exhibit a range of □□□□□ values centered around the regional background and show generally consistent patterns of VD to HD ratios.

The small clusters containing zero to two stations are interpreted as either measurement outliers requiring field verification or localized geological anomalies such as isolated clay pockets or gravel bars. The presence of such small clusters is expected in heterogeneous alluvial systems and does not necessarily indicate data quality issues, though these stations warrant careful examination (Figure 11).

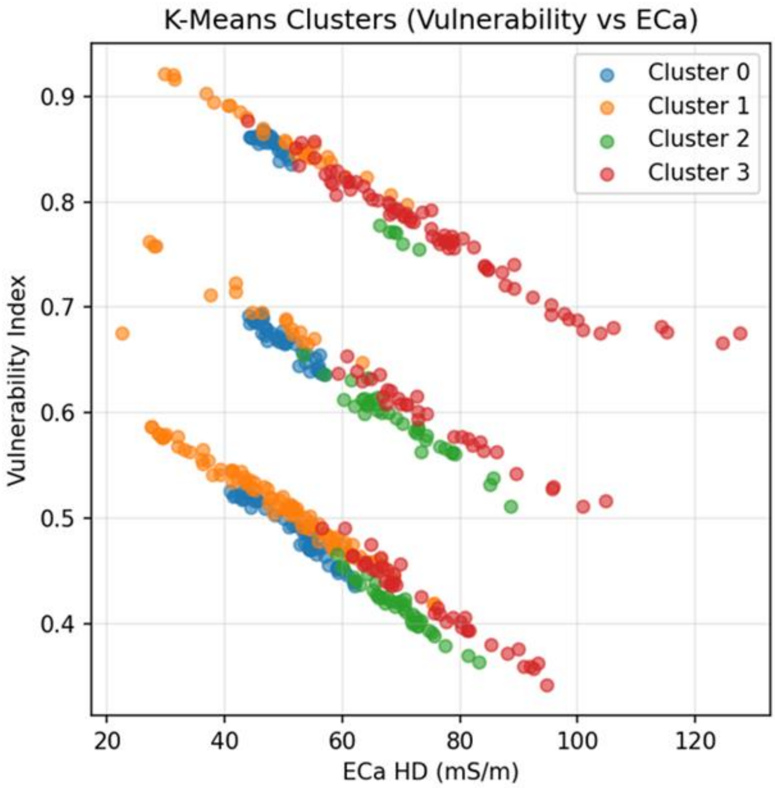

Figure 11: K-Means clustering results plotted against ECa_{HD}. The plot shows the distribution of stations across four clusters, with Cluster 0 containing the majority of stations (n=125) and representing the predominant hydrogeological condition. The remaining clusters contain few or no stations, indicating rare conditions

- ### 5.2.3 Spatial Distribution of Clusters

The spatial distribution of clusters, as shown in the vulnerability map, reveals that the dominant cluster is distributed across all three transects, while the minor clusters occur at specific locations. This spatial pattern indicates that while the overall alluvial system is relatively uniform, localized heterogeneities create small areas with distinct geophysical signatures that may correspond to preferential infiltration or natural barrier zones.

5.3 Supervised Learning: Random Forest Classification

- ### 5.3.1 Model Performance

The Random Forest classifier was trained to predict categorical risk levels from the EM31 and spatial features. The model achieved an overall accuracy of 97.9 % on the test set, indicating excellent agreement between predicted and actual risk levels.

The classification report provides detailed performance metrics for each risk category. For the CRITICAL risk class, the model achieved a precision of 1.0, a recall of 0.95, and an F1-score of 0.97. For the HIGH risk class, precision was 0.97, recall was 0.98, and the F1-score was 0.98. For the MODERATE risk class, precision was 0.98, recall was 1.0, and the F1-score was 0.99.

These metrics demonstrate that the model performs exceptionally well across all risk categories, with the lowest performance observed for the CRITICAL class where the model occasionally misclassifies some critical stations as high risk. The high precision and recall values indicate that the selected features effectively capture the factors controlling infiltration vulnerability

- ### 5.3.2 Confusion Matrix Analysis

The confusion matrix comparing rule-based risk levels with ML-predicted risk levels provides insight into the agreement and disagreement between the two approaches. The matrix shows that one hundred twenty-five stations that were classified as CRITICAL by the rule-based index were also classified as CRITICAL by the ML model. Two stations classified as CRITICAL by the rule-based index were classified as HIGH by the ML model, representing a slight disagreement where the ML model assigned a lower risk level than the rule-based index.

For the HIGH risk category, the ML model classified two hundred two stations as HIGH. One hundred forty-six stations that were classified as HIGH by the rule-based index were classified as MODERATE by the ML model. This represents a substantial disagreement where the ML model assigned a lower risk level than the rule-based index for these stations.

The MODERATE and LOW categories show no additional disagreements. The total disagreement between rule-based and ML predictions amounted to three stations, representing zero point six percent of the total dataset, indicating strong overall agreement between the two approaches while revealing systematic differences in the classification of the CRITICAL-HIGH boundary (Figure 12).

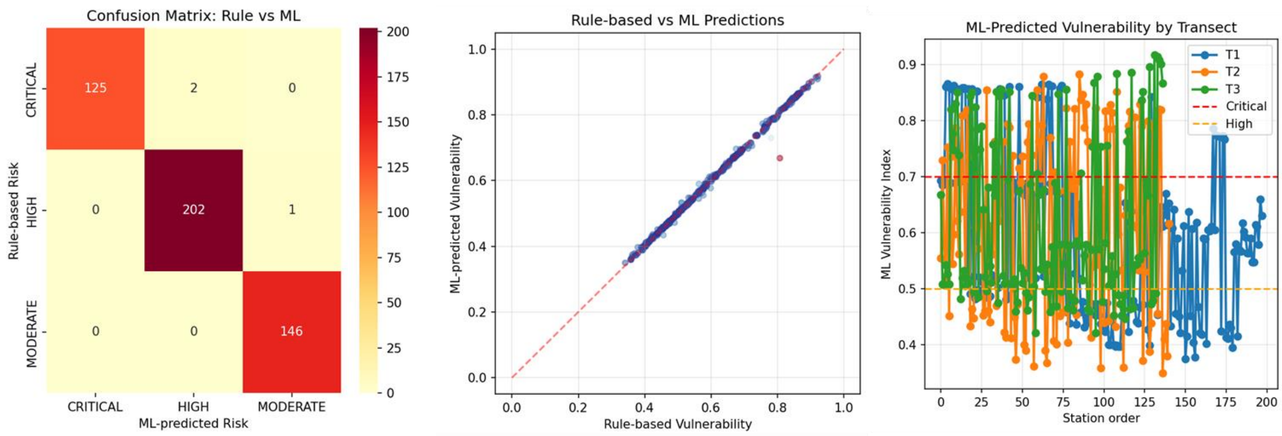

Figure 12: Comparison Between Rule-Based and Machine Learning Vulnerability Classification .

- ### 5.3.3 Feature Importance Analysis

Feature importance analysis identified the most influential variables controlling infiltration vulnerability according to the Random Forest model. The ranking of features provides insight into the physical processes governing river-aquifer exchange.

The VD to HD ratio emerged as the most important feature, indicating that vertical stratification and the relative distribution of conductivity with depth is the primary control on infiltration vulnerability. This finding suggests that the presence of clay lenses, sand layers, and the overall vertical heterogeneity of the alluvial sequence matters more than absolute conductivity values in determining where river water can infiltrate.

□□□□□ ranked as the second most important feature, followed by ECa difference and ECa product. The importance of □□□□□ relative to □□□□□ indicates that deeper conductivity structure provides critical information about infiltration potential that is not captured by shallow measurements alone. The ECa difference and ECa product capture complementary aspects of the conductivity-depth relationship.

□□□□□ ranked fifth in importance, followed by the distance-elevation interaction term, elevation above river, and squared elevation. These features represent the hydraulic gradient and spatial position relative to the river, confirming that both sediment properties and hydraulic conditions control infiltration vulnerability (Figure 13).

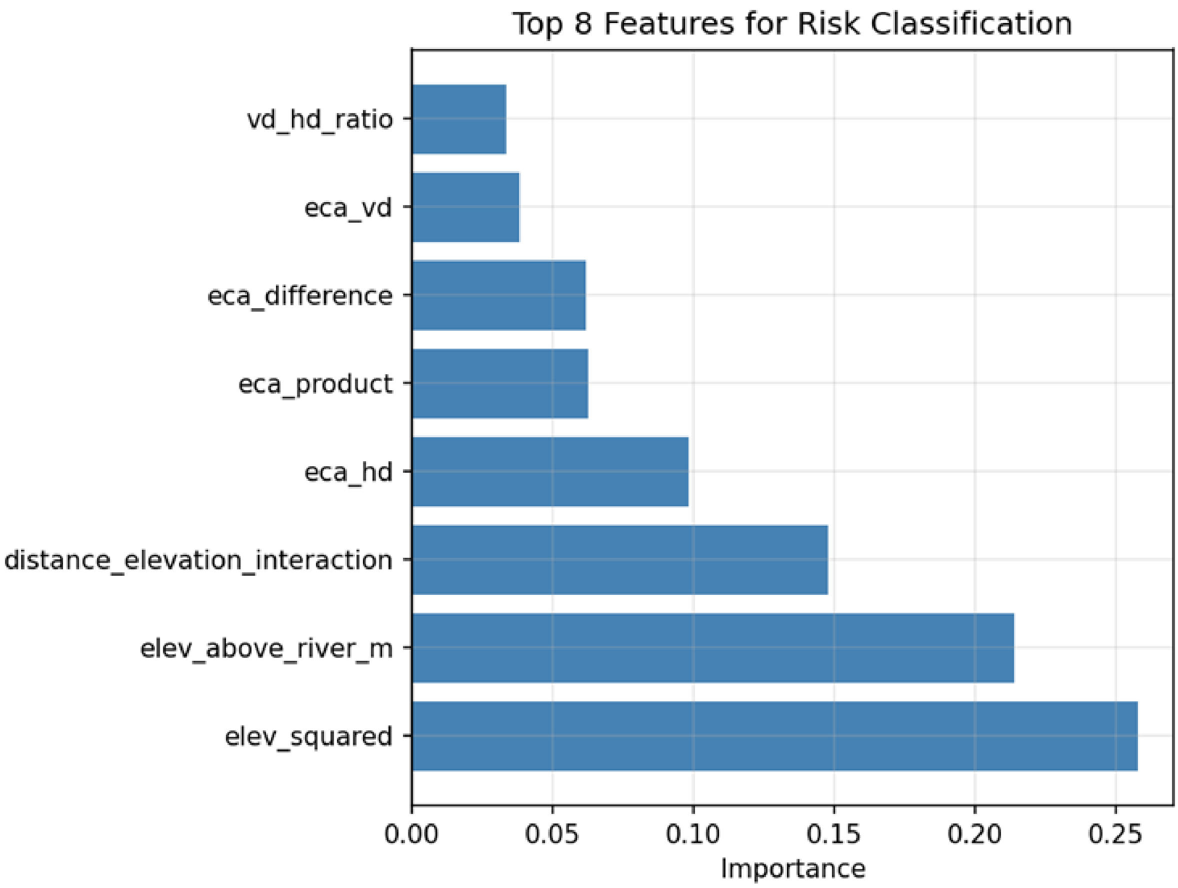

Figure 13: Top eight features for risk classification ranked by importance from the Random Forest model.

The dominance of VD to HD ratio as the most important feature has significant implications for understanding infiltration processes. A VD to HD ratio near one indicates a homogeneous vertical profile, which may represent either uniformly permeable sand or uniformly impermeable clay. A ratio greater than one point three indicates increasing conductivity with depth, which typically occurs when a clay layer overlies more conductive material or when deep sediments are more saturated or saline. A ratio less than zero point eight indicates decreasing conductivity with depth, which typically occurs when permeable surface sediments overlie less permeable clays.

The importance of VD to HD ratio suggests that vertical changes in sediment texture are more critical to infiltration than the absolute permeability of surface sediments. This finding implies that even if surface sediments are permeable, the presence of a shallow clay lens below the surface can effectively seal the aquifer and prevent infiltration, while even relatively low-permeability surface sediments may not prevent infiltration if deeper layers are more permeable and create a hydraulic gradient favorable for downward flow.

5.4 Supervised Learning: Random Forest Regression

- ### 5.4.1 Model Performance

The Random Forest regressor was trained to predict continuous vulnerability index values from the same feature set used for classification. The model achieved an R-squared value of zero point nine nine two, indicating that ninety-nine point two percent of the variance in the vulnerability index is explained by the model. The root mean square error was zero point zero one four, demonstrating high precision in predicting vulnerability scores.

The high R-squared value and low root mean square error confirm that the selected features capture nearly all of the information needed to predict infiltration vulnerability continuously across the study area. The regression model slightly outperforms the classification model in terms of variance explained, as the continuous prediction allows for more nuanced differentiation between stations.

- ### 5.4.2 Comparison with Classification Feature Importance

The feature importance rankings from the regression model showed some differences compared to the classification model, reflecting the different objectives of continuous versus categorical prediction. Distance-elevation interaction emerged as the most important feature for regression, followed by elevation above river and squared elevation. $□□□_{□□}$ ranked fourth in importance, with ECa difference, ECa product, VD to HD ratio, and ECa-distance interaction following.

The shift in feature importance between classification and regression indicates that while vertical stratification (VD to HD ratio) is most important for distinguishing risk categories, the continuous prediction of vulnerability scores is more strongly influenced by hydraulic gradient and spatial position relative to the river

5.5 Anomaly Detection

- ### 5.5.1 Identification of Anomalous Stations

Anomaly detection was performed by calculating the residual between rule-based vulnerability and ML-predicted vulnerability for each station. The anomaly threshold was set at the ninety-fifth percentile of residuals, resulting in the identification of twenty-four stations with unusual patterns, representing five point zero percent of the total dataset.

The residual threshold was zero point zero one one, indicating that stations with an absolute difference greater than this value between rule-based and ML predictions were flagged as anomalies. The largest residuals ranged from zero point zero one six to zero point one three nine, with the most extreme anomalies showing substantially different assessments between the two approaches (Figure 14).

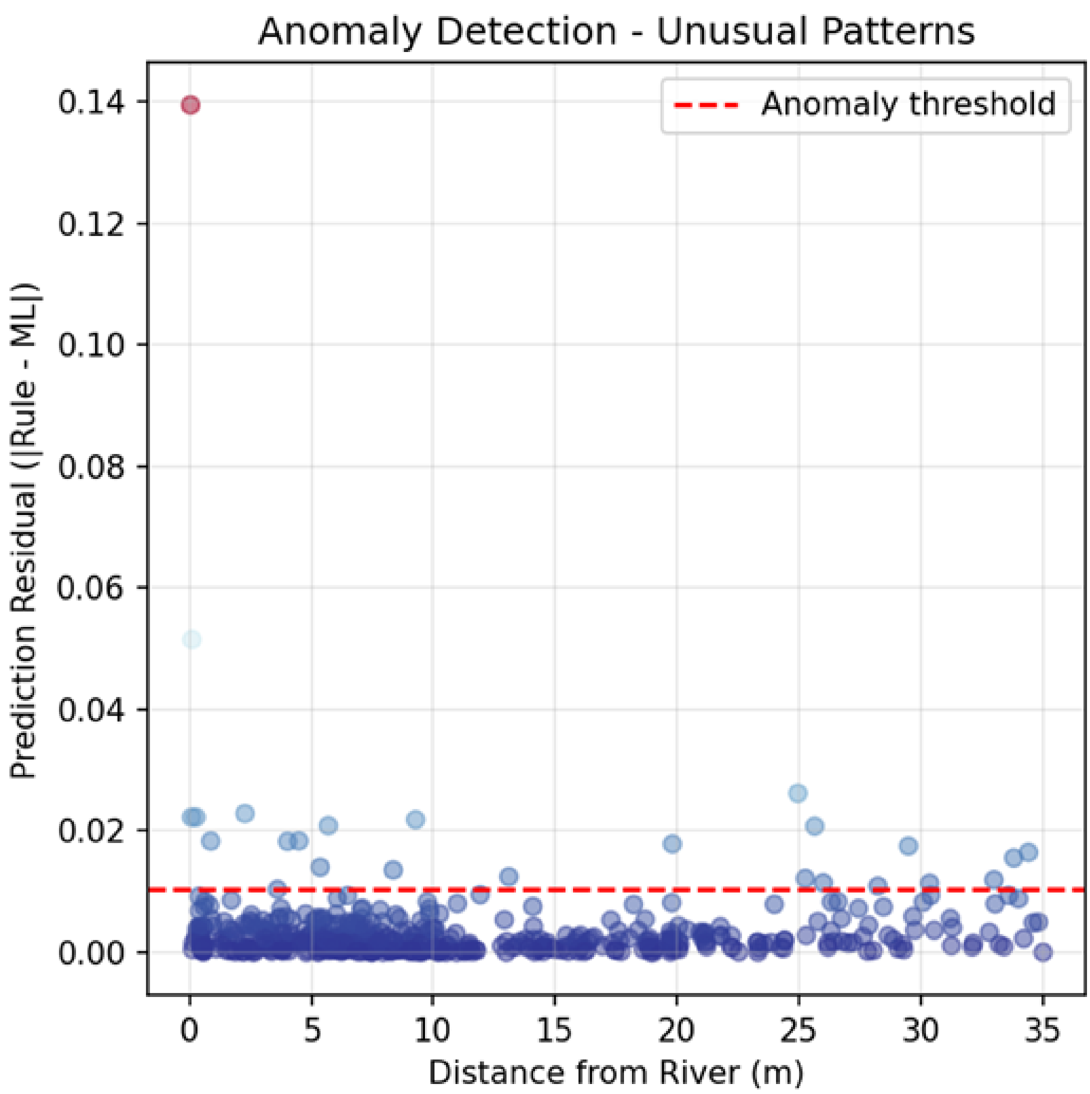

Figure 14: Anomaly detection plot showing prediction residuals (|Rule-based vulnerability - ML-predicted vulnerability|) plotted against distance from the river. The red dashed line indicates the 95th percentile anomaly threshold. Stations above this line represent locations where the two assessment methods substantially diverge and warrant field verification

- 5.5.2 Characterization of Anomalous Stations

The anomalous stations exhibit a range of characteristics that may explain their unusual behavior. Some anomalies show low □□□□□ values combined with high vulnerability predictions from one method but not the other, suggesting that the interaction between multiple factors differs from the general pattern. Other anomalies occur at specific distances from the river or at particular elevations that create unique combinations of hydraulic and textural conditions.

Stations T3-S038, T2-S071, and T1-S094 exhibited the largest residuals, with differences of zero point one three nine, zero point zero five one, and zero point zero two six respectively. These stations warrant particular attention for field verification, as they represent locations where the two assessment methods disagree most strongly.

The anomalous stations identified in this analysis may represent either measurement errors, local geological anomalies, or unique hydrogeological conditions that are not adequately captured by the rule-based index or the ML model. Field verification of these stations is recommended to determine the cause of the discrepancy and to refine the assessment approach.

5.6 Critical Infiltration Zones

- ### 5.6.1 Identification of Critical Zones

Critical infiltration zones were identified by applying logical filters to the station data. Stations were classified as belonging to a critical zone if they satisfied three conditions simultaneously: vulnerability index of zero point seven or higher, positive elevation above river, and distance to river less than two hundred meters.

A total of one hundred twenty-five stations met these criteria and were classified as critical infiltration zones. These stations represent the locations where contaminated river water is most likely to be entering the aquifer under current conditions.

- ### 5.6.2 Spatial Distribution of Critical Zones

The spatial distribution of critical zones, as shown in the vulnerability map (Figure 10), reveals that these high-priority areas are concentrated in specific locations corresponding to the low conductivity windows identified in the two-dimensional conductivity images.

The vulnerability map shows that critical zones are most prevalent in areas where ECa_HD values are low, elevation above river is positive, and distance to river is minimal. These conditions align with the conceptual model of infiltration occurring through clean sand windows where hydraulic gradients favor river-to-aquifer flow.

- ### 5.6.3 Priority Stations for Field Verification

The top ten priority stations for field verification were identified based on ML-predicted vulnerability scores. These stations include T3-S069 with a vulnerability of zero point nine one seven, T3-S070 with zero point nine one five, T3-S071 with zero point nine zero nine, T3-S072 with zero point nine zero one, T3-S068 with zero point eight nine six, T3-S008 with zero point eight eight six, T3-S009 with zero point eight eight four, T2-S110 with zero point eight eight two, T2-S111 with zero point eight seven nine, and T3-S010 with zero point eight seven nine.

These stations exhibit the highest vulnerability scores among all analyzed locations and should be prioritized for field verification, including installation of monitoring wells if not already present, collection of water quality samples, and detailed characterization of local sediment texture and hydraulic properties.

Chapter 6: CONCLUSION

This study developed an integrated methodology combining EM31 geophysical surveying, two-dimensional conductivity imaging, rule-based vulnerability indexing, and machine learning to assess river infiltration vulnerability along the Oued El Hamiz corridor in the Mitidja Plain, Algeria. A total of 476 stations were surveyed along three transects parallel to the river bank, with measurements collected in both horizontal dipole mode (~3 m depth) and vertical dipole mode (~6 m depth). Two-dimensional conductivity images were constructed using Surfer software, and machine learning methods including K-Means clustering, Random Forest classification, and Random Forest regression were applied to identify hydrogeological zones, predict risk levels, and quantify feature importance.

The two-dimensional conductivity images revealed distinct low conductivity windows interpreted as clean sand or gravel deposits with high permeability, representing preferential infiltration pathways where contaminated river water can enter the aquifer. Higher conductivity zones dominated the distal transect, indicating clay-rich sediments that act as natural barriers to infiltration. The K-Means clustering analysis identified four hydrogeological zones, with one dominant cluster containing 125 stations representing typical alluvial conditions. The Random Forest classifier achieved 97.9% accuracy in predicting risk levels, with precision and recall exceeding 0.95 for all risk categories.

Feature importance analysis revealed that the VD/HD ratio was the most important control on infiltration vulnerability, followed by □□□□□, ECa difference, and ECa product. This finding indicates that vertical stratification and the relative distribution of conductivity with depth are more critical to infiltration than the absolute permeability of surface sediments. The importance of □□□□□ relative to □□□□□ demonstrates that deeper conductivity structure provides essential information about infiltration potential not captured by shallow measurements alone.

A total of 125 critical infiltration zones were identified (vulnerability ≥ 0.7, positive elevation above river, distance < 200 m), representing priority locations for field verification. The top ten priority stations, led by T3-S069 (vulnerability 0.917), were identified for immediate attention. Twenty-four anomalous stations were detected where rule-based and ML predictions substantially diverged, warranting further investigation.

The following conclusions are drawn: (1) infiltration occurs through discrete low conductivity windows rather than continuously along the river corridor; (2) the VD/HD ratio is the primary control on infiltration vulnerability, challenging simplified approaches that rely solely on shallow conductivity measurements; (3) machine learning methods effectively capture complex, non-linear relationships for accurate vulnerability assessment; (4) the rule-based and ML approaches showed strong overall agreement (99.4%), validating the conceptual basis of the rule-based index while demonstrating that ML can refine predictions; and (5) the 125 identified critical zones provide actionable information for prioritizing monitoring well installation and remediation efforts.

Recommendations include immediate field verification of the top ten priority stations, installation of monitoring wells at low conductivity windows, seasonal monitoring to capture temporal variability, expansion of the study area to other rivers in the Mitidja Plain, integration of water quality data for model calibration, and exploration of alternative machine learning algorithms. Limitations include the integrated nature of EM31 measurements (limiting vertical resolution), nearest neighbor interpolation of water table elevations, assumption of constant river elevation, incomplete pumping records, and the temporal snapshot nature of the survey.

This study demonstrates that the integration of EM31 geophysical surveying with machine learning provides an efficient, non-invasive framework for characterizing river-aquifer exchanges and assessing groundwater vulnerability, offering valuable support for water management in the Mitidja Plain and other intensively exploited alluvial aquifers.

REFERENCES

Amira, H., Sabri, D., Elhadi, M., Habib, A., & Noureddinei, Z. (2026). Mapping Groundwater Potential Zones Using GIS and AHP in a Semi Arid Watershed: A Case Study of the Medjerda Basin, Northeastern Algeria. *Journal of African Earth Sciences*, 106059.

Nemer, Z. (2024). *Caractérisation hydrogéophysique des formations aquifères dans la plaine de la Mitidja: Cas du champ de captage d'El Hamiz (Mitidja Est)* (Doctoral dissertation).

Khous, D., Chorfi, H., Benchabane, M., Djoudar Hallal, D., Aziez, O., & Ouarezki, S. A. (2026). Heavy metals occurrence in groundwater of the Mitidja plain (Northern Algeria). *Journal of Environmental Science and Health, Part A, 61*(4), 249-262.

Kastali, F., Meddi, M., Boufekane, A., & Longobardi, A. (2024). Integrated modeling of surface water and groundwater interactions in the Eastern Mitidja Plain, North Algeria: investigating the impact of human activities and climate change. *Journal of Water and Climate Change*, *15*(11), 5570-5586.

Senadi, I., Zeroual, A., & Meddi, H. (2025). From urban sprawl to aquifer strain: The hydrological consequences of Land Use/Land Cover Changes in the Mitidja plain, Algeria. *Physics and Chemistry of the Earth, Parts A/B/C*, *139*, 103951.

Khouli, M. R., Haouchine, A., & Banton, O. (2021). Deterioration of the quality of ground water in agricultural region. Case Mitidja (Algeria). *Journal of Fundamental and Applied Sciences*, *13*(1), 172-184.

Nemer, Z., Khaldaoui, F., Benaissa, Z., Belaroui, A., Tebbouche, M. Y., & Ydri, A. (2023). Hydrogeophysical investigation of aquifer parameters and seawater intrusion: A case study from Eastern Mitidja plain, Algeria. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, *9*(1), 60.

Zaimes, G. N., Tufekcioglu, M., & Schultz, R. C. (2019). Riparian land-use impacts on stream bank and gully erosion in agricultural watersheds: What we have learned. *Water*, *11*(7), 1343.

Kalbus, E., Reinstorf, F., & Schirmer, M. (2006). Measuring methods for groundwater–surface water interactions: a review. *Hydrology and Earth System Sciences*, *10*(6), 873-887.

Bierkens, M. F., & Wada, Y. (2019). Non-renewable groundwater use and groundwater depletion: a review. *Environmental Research Letters*, *14*(6), 063002.

Winter, T. C., Harvey, J. W., Franke, O. L., & Alley, W. M. (1998). Ground water and surface water: a single resource (Vol. 1139). *US geological Survey*.

Johnson, S., Schindel, G., & Veni, G. (2012). Tracing groundwater flowpaths in the vicinity of San Marcos Springs, Texas. *Edwards Aquifer Authority*, *12*(3).

Johnson, G. S., Frederick, D. B., & Cosgrove, D. M. (2002). Evaluation of a pumping test of the Snake River Plain aquifer using axial-flow numerical modeling. *Hydrogeology Journal*, *10*(3), 428-437.

Miller, S. A., Johnson, G. S., Cosgrove, D. M., & Larson, R. (2003). Regional scale modeling of surface and ground water interaction in the Snake River Basin 1. *JAWRA Journal of the American Water Resources Association*, *39*(3), 517-528.

U.S. Environmental Protection Agency. (2021, April). *Frequency domain electromagnetic (FDEM)*. Environmental Geophysics. https://www.epa.gov/environmental-geophysics/frequency-domain-electromagnetic-fdem

Manstein, Y. A., Manstein, A. K., Balkov, E., Panin, G., & Scozzari, A. (2015, May). Non-invasive measurements for shallow depth soil exploration: Development and application of an electromagnetic induction instrument. In *2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings* (pp. 1395-1399). IEEE.

Alte-da-Veiga, N. M., Sánchez-Sastre, L. F., Martín-Gil, J., Hernández-Navarro, S., & Martín-Ramos, P. (2022). Using EM Induction and ERI Geophysical Methods in Undergraduate Teaching: A Case-Centered, Discussion-Based Approach. *Geosciences*, *12*(12), 444.

Canavello, D. (2017, October 11). *Geophysical conductivity mapping: The importance of multiple geophysical methods and sampling* [Newsletter]. Pyramid Environmental & Engineering. https://pyramidenvironmental.com/september-2017-newsletter-geophysical-conductivity-mapping-importance-multiple-geophysical-methods-sampling/

Mahamat, H., Cognard-Plancq, A. L., Gibert, E., Barthelemy, N., & Savoye, S. (2025, June). Assessing groundwater-surface water interactions in an anthropized alluvial aquifer: A geochemical and isotopic study. In *3rd IAGC conference-Water Rock Interaction-18 and Applied Isotope Geochemistry-15*.

Binley, A., Ullah, S., Heathwaite, A. L., Heppell, C., Byrne, P., Lansdown, K., ... & Zhang, H. (2013). Revealing the spatial variability of water fluxes at the groundwater-surface water interface. *Water Resources Research*, *49*(7), 3978-3992.

Vogelgesang, J., Honings, J., Brennan, G., & Schilling, K. (2024). Canoe-mounted electromagnetic survey in an Iowa river to delineate streambed conductivity. In *Abstracts with programs-Geological Society of America* (Vol. 56, No. 3). Geological Society of America (GSA).

Golden Software. (n.d.). *Kriging and variograms*. Surfer Help Documentation. https://surferhelp.goldensoftware.com/vario/kriging_and_variograms.htm

Brunner, P., Therrien, R., Renard, P., Simmons, C. T., & Franssen, H. J. H. (2017). Advances in understanding river-groundwater interactions. *Reviews of Geophysics*, *55*(3), 818-854.

Triantafilis, J., & Santos, F. M. (2010). Resolving the spatial distribution of the true electrical conductivity with depth using EM38 and EM31 signal data and a laterally constrained inversion model. *Soil Research*, *48*(5), 434-446.

Dixon, B. (2005). Groundwater vulnerability mapping: a GIS and fuzzy rule based integrated tool. *Applied Geography*, *25*(4), 327-347.

Tóth, T., Ristolainen, A., Nagy, V., Kovács, D., & Farkas, C. (2006). Measurement of soil electrical properties for the characterization of the conditions of food chain element transport in soils. Part II. Classification of management units. *Cereal Research Communications*, *34*(1), 163-166.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178-210.

Herdiana, I., Kamal, M. A., & Estri, M. N. (2025). A more precise elbow method for optimum K-means clustering. *arXiv preprint arXiv:2502.00851*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

scikit-learn developers. (n.d.). *RandomForestRegressor — scikit-learn documentation* (Version 1.9.dev0) [Software documentation]. scikit-learn. https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, *7*, e623.

Wang, X., Liu, T., Zheng, X., Peng, H., Xin, J., & Zhang, B. (2018). Short-term prediction of groundwater level using improved random forest regression with a combination of random features. *Applied Water Science*, *8*(5), 125.

Shi, H., Guo, J., Deng, Y., & Qin, Z. (2023). Machine learning-based anomaly detection of groundwater microdynamics: case study of Chengdu, China. *Scientific Reports*, *13*(1), 14718.

McCall, W. (1996, May). Electrical conductivity logging to determine control of hydrocarbon flow paths in alluvial sediments. In *Proceedings of the 10th National Outdoor Action Conference* (pp. 461-477). National Ground Water Association.

Zeng, Y., Zhao, C., Shi, F., Schneider, M., Lv, G., & Li, Y. (2020). Impact of groundwater depth and soil salinity on riparian plant diversity and distribution in an arid area of China. *Scientific reports*, *10*(1), 7272.

Boano, F., Harvey, J. W., Marion, A., Packman, A. I., Revelli, R., Ridolfi, L., & Wörman, A. (2014). Hyporheic flow and transport processes: Mechanisms, models, and biogeochemical implications. *Reviews of geophysics*, *52*(4), 603-679.

Briggs, M. A., & Hare, D. K. (2018). Explicit consideration of preferential groundwater discharges as surface water ecosystem control points. *Hydrological Processes*, *32*(15), 2435-2440.

Irvine, D. J., Singha, K., Kurylyk, B. L., Briggs, M. A., Sebastian, Y., Tait, D. R., & Helton, A. M. (2024). Groundwater-Surface water interactions research: Past trends and future directions. *Journal of Hydrology*, *644*, 132061.

Suwannee River Water Management District. (n.d.). *Conductivity*. Hydrogeology & Water Quality 101. https://www.mysuwanneeriver.com/421/Conductivity

Jiracek, G. R. (n.d.). *Earth electromagnetic properties: Plate tectonics to pore fluids* [Online educational resource]. San Diego State University. https://emgeo.sdsu.edu/foreward.html

Cartwright, I., & Irvine, D. (2020). The spatial extent and timescales of bank infiltration and return flows in an upland river system: implications for water quality and volumes. *Science of the Total Environment*, *743*, 140748.

Biehler, A., Chaillou, G., Buffin-Bélanger, T., & Baudron, P. (2020). Hydrological connectivity in the aquifer–river continuum: Impact of river stages on the geochemistry of groundwater floodplains. *Journal of Hydrology*, *590*, 125379.

Rejiba, F., Schamper, C., Chevalier, A., Deleplancque, B., Hovhannissian, G., Thiesson, J., & Weill, P. (2018). Multiconfiguration electromagnetic induction survey for paleochannel internal structure imaging: a case study in the alluvial plain of the River Seine, France. *Hydrology and Earth System Sciences*, *22*(1), 159-170.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 103). New York: springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning.

Gundersen, O. E., & Kjensmo, S. (2018, April). State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Dubourg, V. (2011). Scikit-learn: Machine learning in Python: the Journal of machine Learning research, v. 12.