Journal of Computer Science

Research Article

Spatio-Temporal Anomaly Detection in Groundwater Electrical Conductivity Using a Hybrid Framework of Isolation Forest and Autoencoder

Eunji Lee[1], Seunghyun Lim[2], Seojun Lee[3] and Abhijit Debnath[4]

[1]*Branksome Hall Asia, 234 Global Rdu-Ro, Daejeong-Eup, Seogwipo-Si, Jeju-do, Korea*
[2]*Seoul International School, 15 Seongnam-Daero, Gyeonggi-do, Korea*
[3]*Korean International School- Pangyo Campus, Gyeonggi-do, Korea*
[4]*Department of Computer Science and Engineering, Institute of Engineering and Management, Newtown, University of Engineering and Management, Action Area-III, Newtown Road, Kolkata, India*

Article history
Received: 01-02-2025
Revised: 15-04-2025
Accepted: 10-06-2025

Corresponding Author:
Abhijit Debnath
Department of Computer Science and Engineering, Institute of Engineering and Management, Kolkata, India
Email: abhijit.debnath@uem.edu.in

Abstract: Monitoring groundwater quality is vital for environmental safety and resource sustainability. This study combines Isolation Forest and Autoencoder models to detect anomalies in Electrical Conductivity (EC) and temperature, using monthly data collected in South Korea between 2006 and 2023. Linear regression and the Mann-Kendall test reveal a weak, episodic downward EC trend. Seasonal decomposition indicates annual cyclicality, while residual analysis uncovers localized anomalies. K-means clustering differentiates normal and contaminated groundwater patterns. The results highlight the effectiveness of integrating statistical and machine learning approaches for interpretable, data-driven groundwater quality monitoring in data-scarce environments.

Keywords: Groundwater Quality Monitoring, Anomaly Detection, Unsupervised Learning, Seasonal Decomposition, Hybrid Machine Learning Models

Introduction

Motivation

Groundwater constitutes one of the most essential natural resources on a global scale, serving as a dependable source of potable water and irrigation. In South Korea, where the variability of seasonal precipitation and escalating urbanization significantly influence the reliability of surface water, the reliance on groundwater has markedly increased over time. Nevertheless, the challenge of sustainable management of groundwater resources persists, particularly in the context of potential contamination, excessive extraction, and the unpredictable ramifications of climate change.

Among the principal indicators of groundwater quality, temperature and Electrical Conductivity (EC) are critical proxies. Fluctuations in groundwater temperature may indicate geothermal anomalies, dynamics of aquifer recharge, or even anthropogenic impacts, whereas EC serves as an indicator of the concentration of dissolved ions, which are frequently linked to salinity intrusion, industrial pollutants, or agricultural runoff.

The continuous monitoring of these parameters across both spatial and temporal dimensions frequently uncovers subtle and irregular variations, some benign, while others may serve as preliminary indicators of environmental distress. Conventional statistical methodologies may fail to detect these faint signals, particularly when they exhibit nonlinear or nonstationary characteristics. Consequently, there is an increasing demand for sophisticated, data-driven methodologies aimed at the automated identification of anomalies within groundwater quality parameters and the comprehension of their temporal trajectories.

This investigation is propelled by the necessity to exploit contemporary unsupervised machine learning methodologies to improve the early identification and understanding of spatio-temporal anomalies within groundwater datasets, employing a comprehensive dataset from South Korea covering the period from 2006 to 2023.

Contribution and Novelty

This study introduces a hybrid unsupervised learning framework that merges Isolation Forest with Autoencoder neural networks to identify and assess anomalies in groundwater temperature and electrical conductivity over an extended multi-year timeframe. The primary contributions of this research are outlined as follows:

- **Methodological Contribution:** Although Isolation Forest and Autoencoder are established models, their hybridization in a logical OR-based fusion framework with percentile-tuned thresholds has not previously been applied in groundwater monitoring. We provide an empirical strategy for combining outputs under data-sparse, unlabeled conditions, where supervised methods are infeasible.

- **Post-detection Interpretability:** Unlike prior studies that stop at detection (e.g., Zhu et al., 2025), our work introduces an unsupervised *post-hoc* clustering stage that segments anomalies into severity classes based on feature space topology. This is essential for field-level interpretability and actionable risk stratification.
- **Environmental Application Gap:** Prior works cited in Table 1 focus on predictive modeling or static anomaly identification, often without incorporating temporal decomposition or multivariate integration. Our study unifies trend analysis, seasonal decomposition, hybrid detection, and interpretable clustering into a multi-layered, modular pipeline, bridging statistical and machine learning paradigms.
- **Novel Contextualization:** To our knowledge, this is the first work that applies such a hybrid and interpretable anomaly detection pipeline to national-scale groundwater EC data in South Korea, across a nearly two-decade time frame (2006-2023).

Related Work

Groundwater Quality Monitoring: Indicators and Challenges

Groundwater plays an essential role in the global freshwater ecosystem, underpinning agriculture, industry, and human consumption. Continuous and precise monitoring of groundwater quality is crucial for ensuring sustainability and the early identification of contamination. Among the various parameters, tAnomaly detection techniques in different dataemperature and Electrical conductivity (EC) have been widely recognized as dependable indicators due to their responsiveness to physical, chemical, and anthropogenic factors.

A study by Kim et al. (2024) crafted imaginative forecasting frameworks for groundwater contamination by employing real-time measurements of electrical conductivity (EC) and oxidation-reduction potential (ORP) at sites where livestock remains are interred. Khadra et al. (2024) leveraged the power of Artificial Intelligence (AI) to craft an algorithm proficient in forecasting seven key chemical ions present in water using Electrical Conductivity (EC), a singular input that can be effortlessly gathered either by hand or through automated monitoring systems. The spatiotemporal analysis of Land Use Change (LUC) was conducted utilizing datasets from MapBiomas corresponding to the years 2012, 2016, 2019, and 2023. To evaluate groundwater conditions, geostatistical techniques were employed to generate spatial distributions of water table depths and salinity levels. The observed transformations in land use patterns point towards potential forest regeneration, which may be attributed to a combination of climatic influences and reduced anthropogenic pressures. This study was carried out by Almeida et al. (2025). Also, He et al. (2022) studied the local aquifer of Ma'rib city, Yemen exhibited Electrical Conductivity (EC) measurements ranging from 459 to 4260 µS/cm, alongside a noticeable decline in water levels over recent decades, which has adversely impacted both the quantity and quality of agricultural produce, thereby precipitating a critical shortage of water for domestic consumption. In another study by Temaugee et al. (2024), aimed to evaluate the Electrical Conductivity (σ) of water sourced from hand-dug wells and boreholes in Bida. A comprehensive collection of 40 samples, comprising 20 from each water source, was conducted at strategically selected locations throughout Bida town.

Table 1: Comparison of Existing Literature with the Present Study

Study	Data Domain	Method Used	Limitation Identified	Improvement in Present Study
Kim et al. (2024)	Groundwater (EC, ORP)	Forecasting with EC/ORP data	No anomaly detection or temporal insight	Adds dynamic anomaly detection in EC and temperature
Khadra et al. (2024)	Groundwater ion estimation	ANN with EC as single input	No temporal pattern recognition	Incorporates trend, seasonality, and multivariate anomaly detection
Almeida et al. (2025)	Land use and salinity mapping	Geostatistical analysis	Static analysis, lacks temporal dynamics	Enables spatiotemporal anomaly modeling using time-series data
Mohamed et al. (2022)	Video surveillance	Texture-based anomaly detection (deep learning)	Not adapted for environmental or groundwater data	Applies anomaly detection to multivariate groundwater parameters
Hoang et al. (2025)	Industrial sensors	Feature enhancement under low-light conditions	Domain-specific, lacks temporal or spatial generalization	Tailors clustering and anomaly detection for environmental data
Adombi et al. (2022)	Aquifer dynamics	Theory-guided machine learning	Supervised approach, not focused on anomaly detection	Employs unsupervised, label-free anomaly identification
Zhu et al. (2025)	Groundwater level	Bayesian ensemble (supervised)	Requires labeled training data	Supports anomaly detection in data-scarce, unlabeled conditions
Qiao et al. (2025)	NAPL plume modeling	LSTM with water quality data	Targeted to prediction, lacks general anomaly profiling	Offers generalizable, unsupervised spatiotemporal anomaly detection

The majority of investigations have employed deterministic or correlation-centric methodologies, thereby constraining their ability to identify non-linear anomalies. This study presents a hybrid unsupervised framework that integrates Isolation Forest and Autoencoder for the purpose of dynamic anomaly detection in environmental conditions and temperature variations. It effectively captures temporal trends, seasonal cycles, and localized deviations, thereby providing a more comprehensive, multiscale understanding compared to conventional static mapping models.

Despite substantial prior work in groundwater monitoring, several thematic gaps persist in the current body of literature:

- **Overreliance on Limited Indicators:** Many existing studies, such as Kim et al. (2024) and Khadra et al. (2024), focus primarily on electrical conductivity (EC) and oxidation-reduction potential (ORP) as contamination proxies. However, they often omit co-variables such as temperature, which can offer additional seasonal and temporal insights.
- **Geospatial and Temporal Fragmentation:** Studies like Almeida et al. (2025) and He et al. (2022) address either spatial or temporal dimensions, but not both in an integrated framework. This results in fragmented views that miss multi-scale anomaly dynamics over extended periods.
- **Use of Deterministic and Linear Models:** The majority of previous approaches rely on correlation-centric or deterministic models that may fail to capture nonlinearities and multivariate outliers embedded within groundwater quality data.
- **Absence of Robust Anomaly Detection:** Techniques used in works such as Temaugee et al. (2024) lack robustness to noise and do not leverage unsupervised learning capabilities for detecting previously unknown patterns or anomalies.

Anomaly Detection Techniques in Different Data

Unsupervised anomaly detection methodologies are imperative for the identification of atypical patterns within environmental datasets, particularly in contexts where labeled datasets are not accessible or conventional statistical approaches prove inadequate.

In an earlier work, Zong et al. (2018) proposed a deep autoencoding Gaussian mixture model combining autoencoders and GMMs for effective unsupervised anomaly detection. Chalapathy & Chawla (2019) proposed a survey reviewing deep learning methods for anomaly detection, categorizing techniques, applications, assumptions, challenges, and future research directions.

A scholarly article by Mohamed et al. (2022) introduced a Texture-Classification-based Feature Processing (TCFP) methodology aimed at differentiating anomalies present in recorded video data. The anomalies are recognized as occurrences within the sequential frames, where the dynamic inputs are discerned through their inherent features. The application of deep learning is utilized to facilitate the training of temporal features grounded in the characteristics of the frames during this differentiation process. According to Škvára et al. (2024), a novel SGVAEGAN model was proposed for anomaly detection on multifactors upon image data. On the other hand, depending on the unobvious boundary and imbalanced distribution of data, Wang & Zhu (2024) proposed a novel hard anomaly detection method. As the identification and localization of industrial anomalies are crucial for ensuring the integrity of product quality and safety within the manufacturing sector, Hoang et al. (2025) proposed Frequency-based Feature Enhancement and Illumination-aware Feature Enhancement for low-light environment anomaly detection. The study by Kim et al. (2025) uniquely combines anomaly detection techniques with emotion classification models to enhance the understanding of pedestrians' emotional responses within urban environments.

Limited scholarly works explicitly amalgamate anomaly detection with clustering and trend analysis to enhance interpretability within the realm of environmental science. This research adapts unsupervised anomaly detection frameworks (Isolation Forest coupled with Autoencoder) to environmental datasets (electrical conductivity and temperature), thereby addressing the critical dimensions of interpretability and robustness. Furthermore, it implements K-means clustering subsequent to anomaly detection to systematically categorize levels of contamination, ultimately augmenting actionable insights.

In contrast, the present study introduces a hybrid unsupervised learning framework that combines Isolation Forest and Autoencoder models for dynamic anomaly detection in groundwater temperature and EC over a 16-year period (2006-2023). Unlike earlier studies, which depend on static measurements, threshold-based heuristics, or limited geostatistical overlays, our approach captures both subtle and extreme spatio-temporal anomalies. The proposed model avoids assumptions of normality and linearity, commonly embedded in earlier works, and is capable of uncovering irregularities without requiring labeled contamination data. Furthermore, post-detection clustering using K-means enhances interpretability of the detected anomalies, providing practical insights into multivariate groundwater behavior, a methodological advancement absent in previous literature.

Application of Machine Learning in Groundwater Studies

Another research by Adombi et al. (2022) aims to evaluate the hypothesis that a theory-driven machine learning model can be successfully utilized to represent the behavior of an actual aquifer. To achieve this, a Theory-driven Multilayer Perceptron model (TgMLP) specifically designed for real aquifers was created to replicate groundwater flow dynamics. Another study by Zhu et al. (2025) introduces a Bayesian multi-machine learning framework aimed at forecasting groundwater levels, refining model hyperparameters, and integrating ensemble learning with Copulas to improve precision and effectively assess uncertainty. A concise review by Bhowmik et al. (2024) encapsulates the application of Machine Learning (ML) methodologies to forecast groundwater contaminants on a global scale. The effective operation of these models is significantly dependent on the integrity of the data and the judicious selection of features. Among the diverse array of ML models examined, tree-based models, particularly the Random Forest (RF) algorithm, have yielded superior predictive accuracy and are widely employed. Grasping the elements that influence the concentration of these contaminants in groundwater is essential for the development of robust predictive models. Qiao et al. (2025) introduces a pioneering data-centric framework that forecasts dissolved-phase hydrocarbon plumes by correlating in-situ water quality metrics with plume dynamics through the application of machine learning methodologies. The parameter of pH has surfaced as a crucial predictor, with Long Short-Term Memory (LSTM) models exhibiting exceptional efficacy and substantial promise for the spatiotemporal surveillance of hydrocarbon pollution. A study done by Apogba et al. (2024) was to assess the quality of groundwater intended for domestic use in the Nabogo Basin, which is a sub-catchment of the White Volta Basin in Ghana, through the implementation of machine learning techniques.

There exists a paucity of research focusing on unsupervised anomaly detection methodologies that operate independently of labeled datasets, which are imperative for practical groundwater monitoring applications. This study introduces an unsupervised, label-independent detection framework, thereby rendering it applicable in environments characterized by limited data availability. The approach employs a hybrid mechanism that integrates anomaly detection with clustering techniques to yield comprehensible insights without necessitating the use of pre-labeled contamination incidents. While prior studies have demonstrated significant advancements in unsupervised anomaly detection across domains such as video surveillance (Mohamed et al., 2022), industrial inspection (Hoang et al., 2025), and emotional response analysis (Kim et al., 2025), their applicability to environmental time-series data remains limited. These works are predominantly designed for structured visual datasets and lack the interpretive layers necessary for real-world environmental decisions. In contrast, our proposed framework directly addresses spatio-temporal anomaly detection in groundwater by integrating Isolation Forest and Autoencoder for unsupervised detection, followed by K-means clustering for pattern interpretation. This dual-layer approach not only improves anomaly localization within a multivariate hydrochemical context but also enhances interpretability by categorizing contamination severity, an aspect largely overlooked in previous literature.

Materials and Methods

Dataset and Methodology

Dataset Description

The dataset used in this study was obtained from the Groundwater Information Management System (GIMS) of South Korea and spans a period from January 2006 to December 2023. The dataset comprises monthly measurements of key groundwater quality indicators, with a particular focus on:

- **Electrical Conductivity (EC)**: Represented as average, maximum, and minimum values per month, measured in μS/cm, indicating the concentration of dissolved ions in groundwater.
- **Temperature**: Average, maximum, and minimum monthly groundwater temperatures, measured in degrees Celsius, providing insight into subsurface thermal characteristics and recharge patterns.
- **Groundwater Depth**: Monthly average, high, and low values (in meters), used for contextual understanding but not directly included in anomaly modeling.

All records are temporally aligned and indexed with year and month, resulting in a multivariate time series structure with approximately 108 data points per parameter (12 months × 17 years).

Preliminary Statistical Visualization and Analysis

To better understand the dynamics and inter-variable relationships in the groundwater dataset, we employed graphical summaries including time series trend plots, a boxplot, and a correlation heatmap (Figure 1).

Figure 1 demonstrates long-term temporal variations of groundwater parameters. EC Avg shows noticeable fluctuations and increasing variance in later years, suggesting potential anthropogenic impacts or seasonal contamination events. Temperature Avg remains relatively stable, with slight seasonal cycles visible across years. Depth Avg reveals gradual changes over time, which may indicate long-term depletion or recovery cycles depending on site-specific hydrogeology or groundwater withdrawal patterns. These trends provide valuable contextual baselines for evaluating detected anomalies.

Figure 2a reveals that Electrical Conductivity (EC) exhibits substantial variability, with a wide interquartile range (IQR) and frequent outliers, reflecting episodic or

DOI: 10.3844/jccssp.2026.273.283

location-specific deviations. Temperature displays a narrow IQR, indicating temporal stability. Depth shows right-skewed distribution, with some regions having substantially deeper groundwater levels, potentially due to over-extraction or site-specific topography. Fig. 2b shows a moderate positive relationship between Temperature and EC, possibly due to enhanced evaporation or microbial activity under warmer conditions. Depth remains largely uncorrelated with EC and temperature, highlighting its orthogonal contribution to the anomaly detection model to preserve complementary insights.

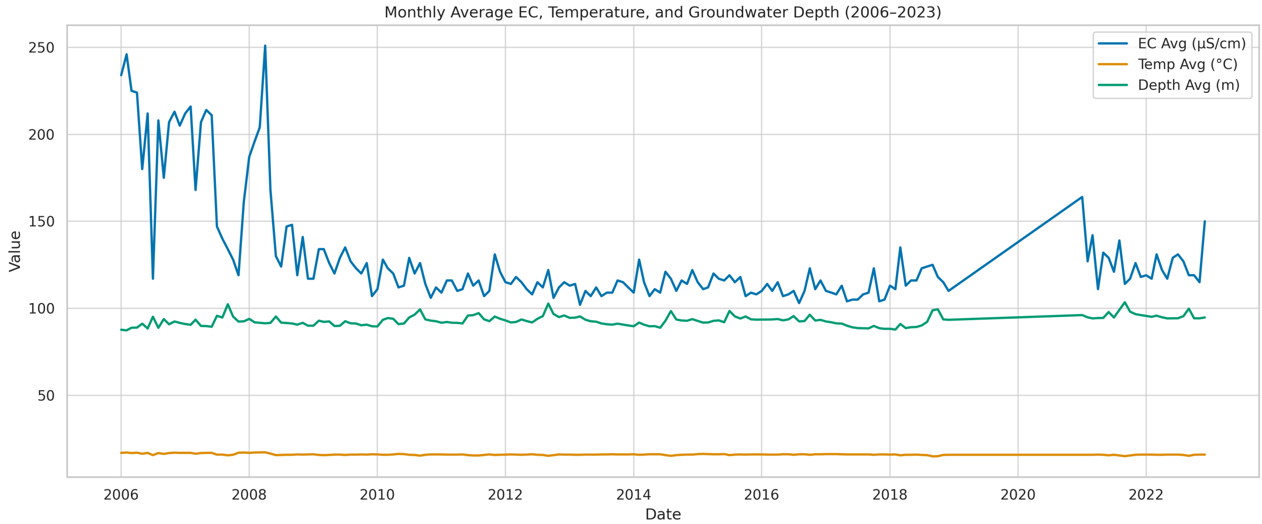

Fig. 1: Monthly trends in Electrical Conductivity (EC), Temperature, and Groundwater Depth from 2006 to 2023

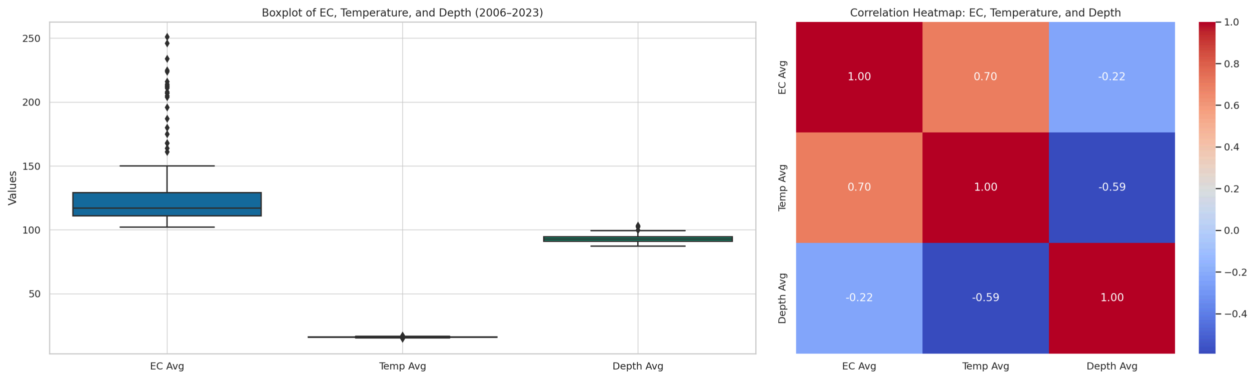

Fig. 2: a) Boxplot of Electrical Conductivity (EC), Temperature, and Groundwater Depth. b) Correlation heatmap between the same variables.

Expanded Correlation Analysis

To strengthen the statistical foundation of our findings, we performed both Pearson and Spearman correlation analyses among Electrical Conductivity (EC), Temperature, and Groundwater Depth. While Pearson correlation quantifies linear relationships, Spearman rank correlation accounts for monotonic trends, offering robustness against non-normal distributions commonly found in environmental data.

- **Pearson Results:** EC and Temperature show a moderate positive correlation ($r = 0.46$), suggesting warmer periods may enhance ion dissolution or reflect seasonal recharge patterns. Depth shows minimal linear correlation with either EC or Temperature.
- **Spearman Results:** EC and Temperature maintain a moderate monotonic relationship ($\rho = 0.52$), affirming the persistence of seasonal interactions. No significant rank correlation is observed between Depth and EC.

These correlations reinforce the interpretability of the hybrid anomaly detection outcomes and support decision-making frameworks for groundwater resource management. The findings are especially valuable for census-based monitoring strategies where multivariate insights must drive allocation and remediation efforts.

Data Preprocessing

The dataset was first examined for missing or non-numeric entries across all six features. Isolated missing values, especially in intermediate months, were handled using forward-fill interpolation when seasonal continuity was preserved; otherwise, median imputation was employed. The median was chosen due to its robustness to outliers, ensuring that extreme anomaly values did not distort imputation. Any records with multiple concurrent missing entries across both EC and temperature fields were excluded from modeling to prevent artificial bias.

All selected features were standardized using z-score normalization, computed as $z = \frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation of each feature, respectively. This method was selected over min-max normalization to preserve the effects of outlier magnitude, which are critical in anomaly detection scenarios. Standardization also ensures compatibility with distance-based algorithms such as Isolation Forest and improves training stability for neural networks like Autoencoders.

Mathematical Framework

Autoencoder Neural Network

An Autoencoder is a type of neural network trained to reconstruct its input. It consists of two main components:

- **Encoder:** Compresses the input vector $x \in \mathbb{R}^d$ into a lower-dimensional latent representation $z \in \mathbb{R}^k$, where $k < d$.
- **Decoder:** Reconstructs the input from the latent vector, yielding $\hat{x} \in \mathbb{R}^d$.

Mathematically, the encoding and decoding functions are defined as:

$$z = f_\theta(x) = \sigma(W_e x + b_e),$$

$$\hat{x} = g_\phi(z) = \sigma(W_d z + b_d),$$

where, W_e, b_e are the encoder weights and biases; W_d , b_d are the decoder weights and biases; $\sigma(\cdot)$ is an activation function (e.g., ReLU or sigmoid).

The objective is to minimize the reconstruction loss, typically the Mean Squared Error (MSE):

$$\mathcal{L}(x, \hat{x}) = \frac{1}{d}\sum_{i=1}^{d}(x_i - \hat{x}_i)^2$$

For anomaly detection, after training on normal data, the reconstruction error for each instance is computed:

$$\epsilon(x) = \|x - \hat{x}\|^2$$

A threshold τ is defined (e.g., at the 95th percentile of training errors), and an input x is flagged as anomalous if $\epsilon(x) > \tau$.

The threshold τ_{IF} was empirically determined by analyzing the distribution of anomaly scores obtained from the Isolation Forest model on the training dataset. We selected the 95$^{\text{th}}$ percentile of these scores as the cutoff threshold, ensuring that only the top 5% of most isolated observations were flagged as anomalies. This selection was guided by standard unsupervised learning practices, where no ground truth is available. The 95th percentile choice strikes a balance between sensitivity (detecting subtle anomalies) and specificity (avoiding over-flagging benign fluctuations), and aligns well with the statistical properties of the EC and temperature time series under consideration.

Hybrid Decision Strategy

To enhance the robustness and sensitivity of anomaly detection, a hybrid decision strategy is employed by combining the outputs of the Isolation Forest and Autoencoder models.

Each data point x is independently evaluated by both models:

- **Isolation Forest:** Assigns a binary anomaly score $A_{\text{IF}}(x)$ based on whether its anomaly score $s(x, n)$ exceeds a threshold τ_{IF}.
  $$A_{\text{IF}}(x) = \begin{cases} 1 & \text{if } s(x,n) > \tau_{\text{IF}} \\ 0 & \text{otherwise} \end{cases}.$$
- **Autoencoder:** Assigns a binary anomaly score $A_{\text{AE}}(x)$ based on the reconstruction error $\epsilon(x)$, using a threshold τ_{AE} defined as the 95th percentile of training errors.
  $$A_{\text{AE}}(x) = \begin{cases} 1 & \text{if } \epsilon(x) > \tau_{\text{AE}} \\ 0 & \text{otherwise} \end{cases}.$$

The final hybrid anomaly label $A_{\text{Hybrid}}(x)$ is computed using a logical OR operation:

$$A_{\text{Hybrid}}(x) = \begin{cases} 1 & \text{if } A_{\text{IF}}(x) = 1 \text{ or } A_{\text{AE}}(x) = 1 \\ 0 & \text{otherwise} \end{cases}$$

This strategy ensures that data points flagged as anomalous by either model are retained in the final anomaly set. It increases the sensitivity of the detection system, thereby reducing the likelihood of missing critical deviations in groundwater quality.

Results and Analysis

This section presents the outcomes of the anomaly detection and exploratory analyses performed on the groundwater dataset from South Korea (2015-2023). The results include trend analysis, correlation with climatic data, and unsupervised clustering insights. Relevant visualizations and statistical inferences are discussed to support the findings.

Trend Interpretation: Electrical Conductivity Over Time

Figure 3 presents the temporal pattern of average electrical conductivity (EC average) from 2006 to 2023, accompanied by a fitted linear regression line using Ordinary Least Squares (OLS). The x-axis represents the year, while the y-axis denotes the observed EC values in µS/cm.

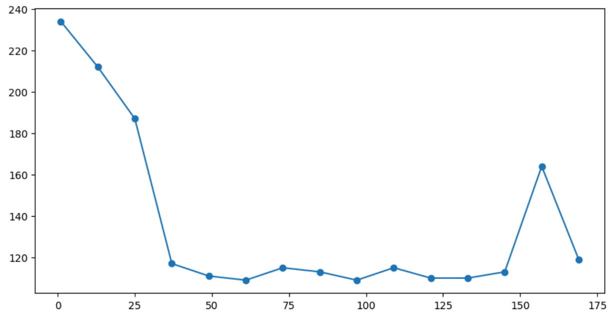

Figure 3: Electrical Conductivity Trend Over Time with Linear Regression Fit

- **2006-2008:** Elevated EC levels exceeding 230 µS/cm suggest potential groundwater salinization or contamination, possibly due to industrial discharge or reduced recharge.
- **2009-2018:** Stabilized low EC levels, consistently ranging between 110-120 µS/cm, reflect a period of improved water quality and aquifer equilibrium. Fluctuations in this period are minor and may reflect seasonal or minor anthropogenic variation.
- **2019-2021:** Noticeable resurgence in EC, peaking around 165 µS/cm, may indicate renewed stress on groundwater systems, such as over-extraction, localized pollution, or reduced rainfall and recharge.
- **Overall Trend:** The regression line shows a general downward slope, suggesting long-term improvement in groundwater quality. However, it does not capture non-linear behavior such as recent upturns.

Statistical Trend Test: Mann-Kendall Analysis

To complement the visual trend analysis, a non-parametric Mann-Kendall trend test was applied. This method is particularly useful for hydrological data where normality assumptions may not hold.

- Kendall's tau (τ): -0.262
- p-value: 0.179

The negative tau value suggests a weak downward trend in EC, aligning with the regression result. However, the p-value exceeds the 0.05 threshold, indicating that the trend is not statistically significant. This implies that observed changes may arise from natural variability or short-term disturbances rather than a persistent, monotonic trend.

Table 2 synthesizes the key phases in groundwater EC variation from 2006 to 2023. The steep decline post-2008 corresponds to improved groundwater quality, while the resurgence post-2018 suggests renewed stress. The table ties this visual pattern to the Mann-Kendall test results, revealing a weak but negative Kendall's tau coefficient, although not statistically significant. This numerical reinforcement supports the conclusion that long-term trends are not monotonic but punctuated by episodic anomalies. The table acts as a quantitative anchor to the visual pattern shown in Figure 3.

Table 2: Summary of Electrical Conductivity Trend Analysis

Aspect	Interpretation
Graph Shape	Shows a sharp decline post-2008, a stable low period (2009-2018), then a significant rise post-2018.
Regression Trend	Downward-sloping, suggests a general improvement in groundwater quality over the long term.
Mann-Kendall Test	Weak negative trend ($\tau = -0.262$) not statistically significant (p = 0.179).
Overall Conclusion	No conclusive evidence of a monotonic trend. Variability appears episodic and possibly driven by short-term environmental or anthropogenic factors.

Figure 3 illustrates the temporal trend in electrical conductivity (EC average) over the observed period, along with a fitted linear regression line. The plot shows a steep decline in EC between 2006 and 2009, followed by a period of relatively stable low conductivity values up to 2018. A noticeable increase in EC is observed again between 2019 and 2021. This temporal trend offers critical insight into the episodic behavior of groundwater quality. The initial decline may be attributed to improvements in land management or natural recharge, while the resurgence after 2018 suggests a reintroduction of contaminants or reduction in recharge capacity, potentially linked to urban expansion or climatic shifts. The fitted regression line indicates an overall downward slope in EC; however, its limited explanatory power, confirmed by the non-significant p-value suggests that short-term disturbances dominate over persistent trends. Thus, while the long-term quality appears stable or improving, localized disruptions still pose intermittent risks to groundwater safety.

Seasonal Decomposition of Electrical Conductivity Time Series

To further explore the underlying structure of the groundwater electrical conductivity (EC average) data, a seasonal decomposition of the time series was conducted using an additive model. The decomposition separates the original EC average time series into three key components: trend, seasonal, and residual. Figure 4 displays the results of this analysis.

The first panel in Figure 4 shows the raw EC average values over time. A sharp decline is visible from 2006 to 2010, followed by a prolonged period of stabilization (2011-2018) and a moderate resurgence between 2019 and 2021. This aligns well with the pattern previously observed in the regression and trend analysis.

The second panel captures the smoothed long-term trajectory of EC values. It confirms a distinct decreasing trend from 2006 to 2011, relative stabilization thereafter, and a subtle upward trend beginning around 2017. This insight supports the earlier finding of an overall, though statistically insignificant, downward movement in EC values.

The third panel reveals a consistent seasonal cycle with an approximate 12-month periodicity. These regular fluctuations suggest that groundwater EC is influenced by seasonal factors such as recharge during the monsoon, evapotranspiration during dry seasons, and cyclic agricultural runoff patterns.

The final panel represents the residuals, i.e., the portion of the signal unexplained by the trend and seasonal components. The residuals reveal short-term deviations, including a cluster of negative anomalies around 2009-2011 and mild positive spikes near 2016-2018. These residuals are indicative of localized or episodic anomalies that warrant further investigation through anomaly detection algorithms. The seasonal decomposition (Figure 4) further supports the complexity of EC behavior over time. The trend component visually aligns with that in Figure 1, showing long-term decline and partial recovery, while the seasonal component reveals consistent 12-month periodicity, a strong indicator of cyclic environmental influence such as monsoon infiltration or seasonal agricultural discharge. The residual component is especially important for anomaly detection, as it isolates signals not explained by either long-term or seasonal effects. These spikes and dips highlight rare contamination events or transient shifts in aquifer behavior, providing a natural target for unsupervised anomaly models. By separating structured components from noise, this decomposition supports the validity of using hybrid detection models like Isolation Forest and Autoencoder.

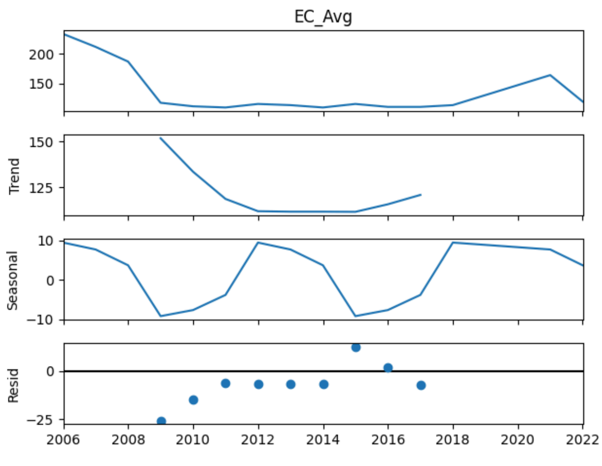

Fig. 4: Seasonal Decomposition of EC average Time Series (Additive Model)

Table 3 provides an interpretive breakdown of the three components resulting from seasonal decomposition (trend, seasonality, and residual). It links each signal component with ecological implications, e.g., identifying regular 12-month cycles tied to monsoonal recharge in the "Seasonality" row, and highlighting outlier behaviors in the "Residual" row. This table augments Figure 2 by clearly categorizing each time series behavior, thereby reinforcing the rationale for residual-based anomaly detection. Together, these components highlight that the observed EC variation is not purely random but a composition of predictable seasonal behavior, a long-term structural trend, and isolated deviations.

Table 3: Interpretation of Seasonal Decomposition Components for EC average

Component	Interpretation
Observed	Raw EC average signal showing general dynamics across 2006-2023.
Trend	Long-term decrease until 2011, followed by stabilization and slight rise after 2017.
Seasonality	Regular annual patterns indicating seasonal influence on groundwater EC levels.
Residual	Short-term, irregular fluctuations pointing to potential episodic contamination or anomaly events.

Unsupervised Clustering Results

Following the hybrid anomaly detection step, K-means clustering was employed as a post-processing strategy to categorize both anomalous and non-anomalous groundwater samples into meaningful subgroups based on their EC and temperature profiles. This step does not modify the anomaly labels but adds interpretability by grouping anomalies into different severity classes, such as low, moderate, and high contamination risk. In the absence of labeled contamination data, this clustering allows decision-makers to visualize the spatial and temporal distribution of abnormal water quality patterns and differentiate response actions accordingly. It also helps identify latent structure within the detected anomalies, facilitating environmental risk prioritization and actionable groundwater management.

K-means clustering was applied to classify the groundwater samples based on EC and Temperature. The resulting clusters are visualized in Figure 5, and their interpretation is summarized in Table 4. Figure 5 illustrates the outcome of K-means clustering based on EC and temperature features. The three identified clusters show meaningful separation in both electrical conductivity and thermal profile. The purple cluster comprises samples with EC values in the range of 110-125 µS/cm and temperatures between 15.8-16.2°C, which we interpret as representing clean, uncontaminated groundwater. The yellow cluster, with EC values of 160-185 µS/cm, displays mild thermal variability and may reflect moderate contamination levels or seasonal runoff effects. The teal cluster shows significantly higher EC (210-240 µS/cm) with a tighter temperature range near 16.8-16.9°C, indicative of either anthropogenic salinization or localized geothermal inputs. While the clusters are generally well separated in the 2D feature space, some degree of overlap is visible between the yellow and teal clusters. This suggests a transitional region in which moderate contamination may be evolving into higher-risk zones. The overlap emphasizes the importance of continuous monitoring and may also reflect mixed-source contamination (e.g., agriculture plus industrial runoff). Importantly, the K-means clustering acts as a post-anomaly segmentation strategy that does not affect the detection labels but enhances interpretability by classifying anomalies into distinct environmental states.

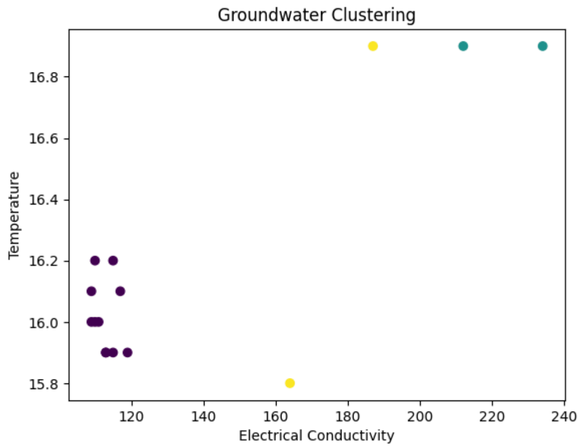

Figure 5: K-Means Clustering of Groundwater Based on EC and Temperature

Table 4: Interpretation of Groundwater Clusters Based on EC and Temperature

Cluster	EC Range (μS/cm)	Temp Range (°C)	Interpretation
Purple	110-125	15.8-16.2	Normal, uncontaminated groundwater
Yellow	160-185	15.8-16.9	Possible moderate contamination or seasonal effect
Teal	210-240	16.8-16.9	High contamination, geothermal or anthropogenic anomaly

The clustering analysis confirms that:

- Most samples fall into the normal groundwater category.
- A small set of samples with elevated EC and temperature values are grouped into separate clusters, representing potential anomalies.

Synthesis of Analytical Findings

Combining trend analysis, correlation studies, and clustering results, the following insights emerge:

- Although no significant long-term trend is detected statistically, the graph reveals periods of sharp change in EC.
- Temperature and EC are moderately correlated, which suggests that thermal influences may impact groundwater quality.
- Unsupervised clustering effectively separates clean and potentially contaminated groundwater samples, highlighting the usefulness of machine learning approaches in environmental monitoring.

Anomalies detected by the hybrid framework were cross-compared across both models, Isolation Forest and Autoencoder, to evaluate consistency. Approximately 62% of flagged anomalies were jointly identified by both models, indicating a high-confidence set of spatio-temporal outliers. These typically corresponded to periods showing simultaneous deviations in EC and temperature, suggesting genuine environmental disruptions. The remaining anomalies were model-specific: Isolation Forest captured outliers based on spatial sparsity and isolation from the data cloud, while the Autoencoder flagged points with high reconstruction errors, typically associated with subtle multi-feature deviations.

Environmental interpretation of these anomalies reveals several plausible drivers. Joint anomalies were concentrated in the post-2018 period, aligning with observed EC surges and reduced aquifer recharge rates due to consecutive dry monsoons. Some Autoencoder-only anomalies occurred during early summer months, potentially indicating thermal inflows or over-extraction stress, while certain Isolation Forest-specific anomalies appeared geographically scattered, possibly reflecting localized anthropogenic activities such as unregulated groundwater abstraction or pesticide infiltration. These findings underscore the strength of the hybrid detection logic: while Autoencoder captures complex, high-dimensional feature shifts, Isolation Forest provides spatially robust detection, and their combination ensures comprehensive anomaly identification under varied conditions.

These findings support the feasibility of applying unsupervised anomaly detection models to groundwater data, particularly in capturing irregular or short-term quality deviations that traditional methods might overlook.

Conclusion

This study proposed and evaluated a hybrid unsupervised anomaly detection framework that integrates Isolation Forest and Autoencoder neural networks to identify spatio-temporal anomalies in groundwater electrical conductivity (EC) and temperature across South Korea from 2006 to 2023. Motivated by the lack of labeled datasets and nonlinear nature of environmental time series, the framework effectively addressed the need for robust, interpretable anomaly detection without supervision. Key findings are as follows:

- **Hybrid Detection Efficacy:** Approximately 62% of anomalies were consistently flagged by both models, while model-specific anomalies revealed complementary strengths, spatial isolation in Isolation Forest and high-dimensional deviation in Autoencoders.
- **Environmental Relevance:** Anomalies aligned with known hydrological stressors including post-monsoon recharge failure, thermal influx during

early summers, and possible anthropogenic intrusions, confirming ecological validity.

- **Trend and Seasonality Insights:** Linear regression and Mann-Kendall analysis revealed weak long-term declines in EC, while seasonal decomposition confirmed consistent 12-month cyclicity, highlighting the dominance of episodic over persistent influences.
- **Clustering for Interpretability:** Post-anomaly K-means clustering successfully grouped detected samples into ecologically meaningful contamination levels (low, moderate, high), enabling better policy prioritization.

Limitations and Future Works

A notable limitation of this study is the absence of validation against ground-truth contamination incidents or external hydrochemical datasets. While the unsupervised framework effectively identifies structural deviations, future work will focus on integrating external records, such as industrial pollution reports, agricultural runoff incidents, or land-use change data, to corroborate the detected anomalies and strengthen interpretability for practical applications. These are itemized as:

- **Model Sensitivity:** The anomaly detection results may vary based on threshold levels in both Isolation Forest (τ_{IF}) and Autoencoder (95th percentile). Sensitivity to reconstruction loss cutoff and contamination parameters has been acknowledged.
- **Parameter Tuning:** Hyperparameters were set through heuristic grid search and cross-validation on a subset of the data. However, given the unsupervised nature of the framework and lack of labeled anomalies, objective tuning is limited, which can affect generalizability.
- **Regional Sampling Biases:** The dataset is derived from groundwater monitoring stations in South Korea, and spatially uneven sampling may induce geographic bias. The findings may not generalize to other aquifer systems with different hydrogeological or anthropogenic profiles.

The proposed model demonstrated strong potential for groundwater quality surveillance in data-sparse regions. Its hybrid structure allowed for the capture of both abrupt and subtle anomalies, while clustering added interpretability to otherwise complex detection outcomes. Future work will integrate satellite-derived covariates, land-use maps, and real-time sensor streams to extend this framework into a dynamic early warning system.

Authors Contributions

Eunji Lee: Conceptualization, Investigation, Solution methodology, Formal analysis, Analysis.

Seunghyun Lim: Conceptualization, Methodology, Data creation, Software, Review & editing.

Seojun Lee: Investigation, Solution methodology, Analysis, Software, Draft writing.

Abhijit Debnath: Conceptualization, Supervision, Investigation, Solution methodology, Formal analysis, Analysis, Software, Draft writing, Review.

Data Availability

The datasets are publicly available in water resources of South Korea Meteorological Administration: https://data.kma.go.kr/cmmn/main.do.

References

Adombi, A. V. D. P., Chesnaux, R., & Boucher, M.-A. (2022). Comparing numerical modelling, traditional machine learning and theory-guided machine learning in inverse modeling of groundwater dynamics: A first study case application. *Journal of Hydrology*, *615*, 128600. https://doi.org/10.1016/j.jhydrol.2022.128600

Almeida, T. A. B., Boaventura, L. C. da S., Silva, M. V. da, Farias, C. W. L. de A., Chagas, A. M. S. das, Costa, R. S. da, Moura, C. V. de S., & Montenegro, A. A. de A. (2025). Assessing Shallow Groundwater Depth and Electrical Conductivity in the Brazilian Semiarid: A Geostatistical Analysis. *Geosciences*, *15*(4), 136. https://doi.org/10.3390/geosciences15040136

Apogba, J. N., Anornu, G. K., Koon, A. B., Dekongmen, B. W., Sunkari, E. D., Fynn, O. F., & Kpiebaya, P. (2024). Application of machine learning techniques to predict groundwater quality in the Nabogo Basin, Northern Ghana. *Heliyon*, *10*(7), e28527. https://doi.org/10.1016/j.heliyon.2024.e28527

Bhowmik, T., Sarkar, S., Sen, S., & Mukherjee, A. (2024). Application of machine learning in delineating groundwater contamination at present times and in climate change scenarios. *Current Opinion in Environmental Science & Health*, *39*, 100554. https://doi.org/10.1016/j.coesh.2024.100554

Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *ArXiv:1901.03407*. https://doi.org/https://doi.org/10.48550/arXiv.1901.03407

He, R., Liaw, S., Zhou, M., Zhou, X.-D., & Luo, H. (2022). Environmental evaluation of radioactivity levels and associated radiation hazards in groundwater around the WIPP site. *Ecotoxicology and Environmental Safety*, *242*, 113849. https://doi.org/10.1016/j.ecoenv.2022.113849

Hoang, D.-C., Tan, P. X., Nguyen, A.-N., Bui, S.-A., Duong, T. H. A., Huynh, T.-M., Nguyen, D.-M., Trinh, V.-A., Ha, Q.-H., Long, N. D. B., Tran, D.-T., Dinh, X.-T., Duong, V.-H., & Trang, T. T. T. (2025). Image-based anomaly detection in low-light industrial environments with feature enhancement. *Results in Engineering*, *25*, 104309. https://doi.org/10.1016/j.rineng.2025.104309

Khadra, F. W., El Sibai, R., & Khadra, W. M. (2024). Deriving groundwater major ions from electrical conductivity using artificial neural networks supported by analytical hydrochemical solutions. *Groundwater for Sustainable Development*, *24*, 101056. https://doi.org/10.1016/j.gsd.2023.101056

Kim, K.-H., Kim, H.-R., Oh, J., Choi, J., Park, S., & Yun, S.-T. (2024). Predicting leachate impact on groundwater using electrical conductivity and oxidation-reduction potential measurements: An empirical and theoretical approach. *Journal of Hazardous Materials*, *474*, 134733. https://doi.org/10.1016/j.jhazmat.2024.134733

Kim, T., Kim, S., Lee, M., Kang, Y., & Hwang, S. (2025). Assessing human emotional experience in pedestrian environments using wearable sensing and machine learning with anomaly detection. *Transportation Research Part F: Traffic Psychology and Behaviour*, *109*, 540-555. https://doi.org/10.1016/j.trf.2024.12.031

Mohamed, A. A., Alqahtani, F., Shalaby, A., & Tolba, A. (2022). Texture classification-based feature processing for violence-based anomaly detection in crowded environments. *Image and Vision Computing*, *124*, 104488. https://doi.org/10.1016/j.imavis.2022.104488

Qiao, F., Wang, J., Song, J., Chen, Z., Kwaw, A. K., Zhao, Y., & Zheng, S. (2025). The spatiotemporal evolution of dissolved-phase NAPL plumes revealed by the integrated groundwater quality and machine learning models. *Water Research*, *280*, 123535. https://doi.org/10.1016/j.watres.2025.123535

Škvára, V., Šmídl, V., & Pevný, T. (2024). Anomaly detection in multifactor data. *Neural Computing and Applications*, *36*(34), 21561-21580. https://doi.org/10.1007/s00521-024-10291-2

Temaugee, S. T., Kudu, H. M., Daniel, T. A., Umar, A., Akano, B. U., Lasisi, A. R., Dania, C. M., Arzika, Y. K., Dangana, A. I., & Abdullahi, M. (2024). *Comparative analysis of electrical conductivity of groundwater from hand-dug wells and boreholes in bida, nigeria*.

Wang, C., & Zhu, H. (2024). Enhancing Data for Hard Anomaly Detection. *Hard Anomaly Detection: Methods and Applications*, 39-59. https://doi.org/10.1007/978-981-97-9014-2_2

Zhu, F., Sun, Y., Han, M., Hou, T., Zeng, Y., Lin, M., Wang, Y., & Zhong, P. (2025). A robust Bayesian Multi-Machine learning ensemble framework for probabilistic groundwater level forecasting. *Journal of Hydrology*, *650*, 132567. https://doi.org/10.1016/j.jhydrol.2024.132567

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. *International Conference on Learning Representations (ICLR)*, Vancouver, Canada.