Data-Driven Framework for Predicting Subsurface Contamination Pathways in Complex Remediation Projects

OMOLOLA BADMUS[1], AZEEZ LAMIDI OLAMIDE[2]
[1]*Independent Researcher, USA*
[2]*OKK Global Resources Limited, Nigeria*

Abstract- ***Effective remediation of contaminated sites increasingly depends on advanced predictive capabilities that can accurately characterize and forecast subsurface contamination pathways. Traditional site assessment methods often struggle to capture the spatial heterogeneity, nonlinear contaminant transport dynamics, and multi-source pollution interactions typical of complex remediation projects. This study proposes a comprehensive data-driven framework that integrates geospatial analytics, machine learning models, and hydrogeological simulation to enhance the prediction of contaminant migration in heterogeneous subsurface environments. The framework leverages high-resolution datasets including soil properties, hydrological gradients, geochemical indicators, and historical contaminant concentrations to identify key transport mechanisms and generate predictive contamination plume trajectories. By combining supervised learning algorithms with physics-informed constraints, the model captures both the statistical patterns and mechanistic behaviors governing subsurface pollutant movement. In addition, the framework incorporates uncertainty quantification techniques to evaluate prediction confidence and guide decision-making under data limitations. Case applications demonstrate that the data-driven approach outperforms traditional deterministic models in forecasting plume evolution, delineating risk zones, and identifying potential receptor exposure pathways. Results further show that integrating multi-source datasets significantly improves model robustness, offering actionable insights for remediation design, resource allocation, and long-term monitoring strategies. The study contributes a scalable methodology capable of supporting remediation engineers, environmental regulators, and policymakers in optimizing site-specific and regional contamination management. By bridging advanced analytics with domain knowledge, the proposed framework supports early detection of contamination hotspots, enhances risk assessment, and promotes cost-effective remediation planning. Ultimately, this data-driven predictive architecture represents a transformative tool for managing subsurface contamination under increasing environmental and regulatory pressures, enabling more precise, transparent, and adaptive remediation interventions. Future work will explore real-time data integration, improved interpretability of machine learning models, and incorporation of emerging sensing technologies to further strengthen predictive accuracy and support sustainable environmental restoration.***

Keywords: ***Subsurface Contamination, Data-Driven Modeling, Machine Learning, Hydrogeology, Remediation Projects, Contaminant Transport, Predictive Analytics, Environmental Monitoring, Uncertainty Quantification, Geospatial Analysis.***

I. INTRODUCTION

Subsurface contamination remains one of the most complex and persistent challenges in environmental remediation, largely due to the heterogeneous nature of soil structures, variable hydrogeological conditions, and the dynamic behaviour of contaminant migration. Pollutants originating from industrial spills, leaking storage systems, agricultural runoff, and legacy waste sites often move unpredictably through porous media, creating hidden pathways that threaten groundwater resources, ecological stability, and human health. Traditional predictive approaches primarily deterministic models and manually interpreted hydrogeological assessments frequently struggle to capture these complexities because they rely on limited datasets, oversimplified transport assumptions, and static boundary conditions that fail to reflect real-world variability (Alibakhshi, et al., 2017, Zhang, et akl., 2013). As remediation projects become more intricate, involving multiple contaminant sources, fluctuating hydraulic gradients, and evolving land-use patterns, the shortcomings of conventional tools become increasingly evident. Their inability to integrate high-resolution spatial data, incorporate

temporal changes, or adapt to emerging field information often results in inaccurate plume forecasting, inefficient remediation planning, and elevated project costs (Faseemo, et al., 2009).

The growing availability of multi-source environmental datasets, advancements in sensing technologies, and the rise of scalable analytical methods now provide an opportunity to transform subsurface contamination prediction. Data-driven techniques particularly those leveraging machine learning, geospatial analytics, and hybrid computational models offer the capacity to process complex datasets, identify hidden relationships, and generate more realistic representations of transport mechanisms across diverse soils and aquifers. By integrating physics-informed constraints with data analytics, these approaches address the limitations of purely empirical or purely mechanistic models, enabling more precise delineation of contaminant pathways and more reliable forecasting of plume evolution. The adoption of data-driven frameworks also enhances decision-making by providing probabilistic insights, quantifying uncertainty, and improving the interpretability of subsurface dynamics. In complex remediation environments where uncertainty is high and stakes are significant, such frameworks represent a critical advancement (Manfreda, et al., 2018, Sims & Colloff, 2012). They allow practitioners to optimize resource allocation, accelerate risk assessments, and design interventions that are both targeted and adaptive. In this context, a data-driven framework becomes essential for managing subsurface contamination with greater accuracy, transparency, and operational efficiency.

2.1. Methodology

This study adopts a data-driven, hybrid analytical methodology to predict subsurface contamination pathways in complex remediation projects by integrating heterogeneous environmental data streams, advanced machine learning techniques, and spatial–hydrogeological reasoning. The methodological design is informed by data-centric predictive frameworks applied in environmental monitoring, groundwater contamination mapping, and decision-support systems, emphasizing scalability, uncertainty handling, and real-time adaptability. The approach combines IoT-enabled sensing, remote sensing observations, historical site investigation records, and hydrogeophysical datasets to construct a unified analytical environment capable of learning complex subsurface behavior beyond the limitations of purely deterministic models.

Primary data sources include in-situ sensor networks measuring groundwater quality parameters, soil moisture, hydraulic head, redox potential, and contaminant concentrations, alongside remote sensing products capturing land cover dynamics, surface moisture anomalies, vegetation stress, and terrain attributes. Historical borehole logs, geotechnical profiles, laboratory contaminant analyses, remediation records, and hydrogeological conceptual site models are incorporated to provide contextual grounding. These datasets are ingested through an automated data acquisition pipeline that standardizes formats, timestamps observations, and performs quality control procedures such as noise filtering, missing-value imputation, and outlier detection to ensure analytical robustness.

Feature engineering is conducted to translate raw observations into physically and statistically meaningful predictors of contaminant migration. Derived variables include hydraulic gradients, permeability proxies, lithological continuity indices, contaminant mass flux estimates, and spatio-temporal change metrics extracted from time-series remote sensing imagery. Dimensionality reduction techniques are applied where necessary to manage data redundancy while preserving dominant variance structures relevant to subsurface transport processes. The engineered feature set reflects both intrinsic site vulnerability and dynamic forcing factors such as land-use change, climatic variability, and remediation interventions.

Predictive modeling is implemented using an ensemble machine learning strategy that integrates multiple algorithms, including random forest, gradient boosting, and support vector regression, to capture nonlinear interactions between hydrogeological controls and contaminant behavior. Model training is performed using stratified spatial–temporal sampling to avoid bias and overfitting, with cross-validation employed to evaluate generalization performance.

Ensemble averaging and weighted voting schemes are applied to improve prediction stability and reduce algorithm-specific uncertainty. The models are calibrated to predict contaminant concentration gradients, plume evolution trajectories, and preferential migration pathways across multiple subsurface layers.

Spatial integration is achieved through coupling the trained predictive models with a GIS-based analytical environment. Model outputs are translated into probabilistic contamination pathway maps that visualize likely plume directions, depth-dependent risk zones, and areas of potential exposure. These outputs are dynamically updated as new sensor data become available, enabling near–real-time reassessment of subsurface risk conditions. The framework supports scenario-based analysis by simulating changes in contaminant behavior under alternative remediation strategies, hydrogeological assumptions, or environmental stressors.

Uncertainty quantification is embedded throughout the analytical workflow using ensemble dispersion metrics, sensitivity analysis, and probabilistic output interpretation. This allows decision-makers to distinguish high-confidence predictions from areas requiring additional investigation or monitoring. The final stage of the methodology integrates predictive outputs into a decision-support interface that communicates actionable insights to remediation planners, regulators, and site managers. This interface supports adaptive remediation planning by linking predicted contamination pathways with remediation effectiveness indicators, monitoring priorities, and long-term risk management strategies.

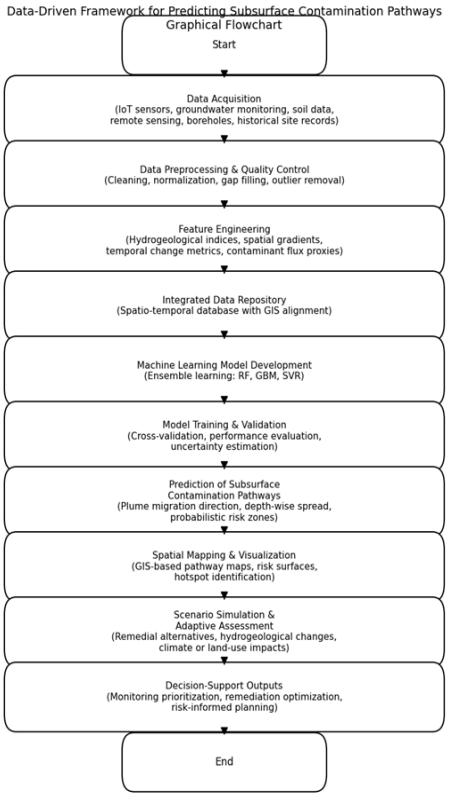

Figure 1: Flowchart of the study methodology

2.2. Problem Context and Environmental Significance

Subsurface contamination represents a persistent environmental challenge that continues to undermine the safety, stability, and sustainability of natural and built ecosystems. The sources of such contamination are diverse and often deeply embedded in industrial, agricultural, and urban activities. Leaking underground storage tanks, refinery operations, pipeline failures, mining residues, landfills, chemical spills, and improper waste disposal practices remain among the most common contributors of hazardous substances into soil and groundwater systems. Agricultural fertilizers, pesticides, and livestock waste also introduce nitrates, phosphates, and microbial contaminants that can migrate far beyond their original application zones. In older industrial regions, legacy contamination from decades of unregulated disposal practices continues to pose significant threats, often compounded by incomplete historical records and limited site characterization (Buma & Livneh, 2017, Zhai, Yue & Zhang, 2016). These diverse sources

introduce contaminants with varying chemical properties, persistence levels, and environmental behaviours, making the identification and prediction of their pathways extraordinarily complex.

Complicating the scenario further are the geological and hydrogeological complexities inherent in subsurface environments. Soil and rock layers exhibit heterogeneity in texture, porosity, permeability, and mineral composition, creating preferential flow channels and unpredictable retention zones. Aquifer systems are often dynamic, influenced by seasonal recharge patterns, groundwater extraction, surface infiltration, and climatic variations. Fractured bedrock terrains, karst landscapes, and heterogeneous alluvial deposits amplify the challenge, as contaminants may bypass monitoring points, migrate through unexpected conduits, or become trapped in low-permeability zones before remobilizing under altered hydraulic conditions. Even small spatial variations in soil structure or saturation can significantly alter contaminant transport rates and directions (Schultz & Engman, 2012, Sorooshian, et al., 2014). Because these subsurface conditions are rarely uniform or static, deterministic models that rely on simplified assumptions often fail to capture the nuanced interactions driving contaminant migration.

The uncertainties associated with contaminant transport are further exacerbated by the physicochemical characteristics of the pollutants themselves. Organic solvents such as chlorinated hydrocarbons can form dense non-aqueous phase liquids (DNAPLs) that sink deeply into aquifers, while light non-aqueous phase liquids (LNAPLs) float and spread along the water table. Metals may adsorb to soil particles and later desorb under changes in pH or redox conditions, leading to delayed or secondary contamination plumes. Reactive species may degrade into equally or more harmful by-products, complicating predictions even further. Microbial activity, temperature variations, groundwater velocity changes, and chemical interactions between multiple contaminants add additional layers of uncertainty (Thakur, Singh & Ekanthalu, 2017). As a result, traditional prediction tools often underestimate the spatial extent of contamination, misidentify the direction of plume travel, or overlook contaminant persistence in subsurface reservoirs.

These uncertainties have profound implications for public health, as groundwater remains a primary source of drinking water for millions of people worldwide. Contaminants such as benzene, trichloroethylene, arsenic, nitrates, and heavy metals pose significant health risks, including carcinogenic effects, neurological damage, reproductive issues, and developmental problems in children. When contaminants migrate undetected or unpredictably, they can contaminate wells, irrigation systems, and surface water bodies, exposing communities to long-term health hazards. The latency period of many contamination-related illnesses means harmful exposures may go unnoticed for years, further reinforcing the need for accurate, predictive subsurface contamination models (Andres, et al., 2018, Turczynowicz, Pisaniello & Williamson, 2012). Poorly understood or misrepresented contamination pathways can delay risk communication, hinder effective public-health interventions, and erode community trust in environmental management institutions. Figure 2 shows DPSIR framework and subsurface environmental problems presented by Jago-on, et al., 2009.

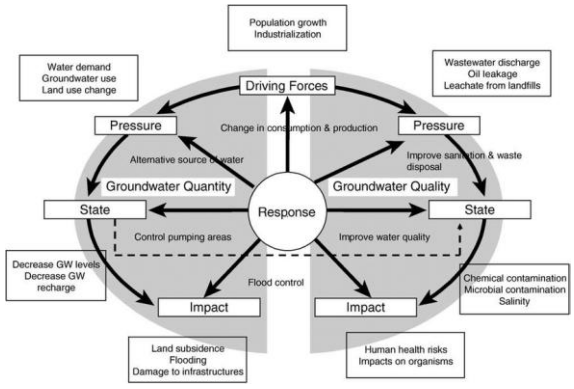

Figure 2: DPSIR framework and subsurface environmental problems (Jago-on, et al., 2009).

Regulatory frameworks governing contaminated sites demand precise assessments and accurate predictions of contamination pathways, making the limitations of conventional approaches a major compliance concern. Environmental Protection Agencies, water authorities, and international regulatory organisations increasingly require detailed hydrogeological characterisation, plume delineation, and predictive modelling as part of site investigations and remediation planning. Inaccurate predictions can result in non-compliance

with cleanup standards, penalties, project delays, or costly redesigns of remediation strategies. Furthermore, underestimating the extent of contamination may lead to insufficient remediation measures, leaving hazardous substances in place that continue to pose long-term risks (McAlary, Provoost & Dawson, 2010, Provoost, et al., 2013). Conversely, overestimating contamination zones may result in unnecessary expenditures, reduced economic viability of redevelopment projects, and inefficient allocation of public or private resources. As environmental regulations become more stringent and climate-related impacts intensify hydrological uncertainties, the need for robust predictive tools becomes even more urgent.

Remediation outcomes are directly tied to the accuracy of subsurface contamination predictions. Successful remediation requires precise identification of contamination hotspots, accurate estimation of plume boundaries, and reliable predictions of how contaminants will move under both current and future conditions. Traditional site investigation methods boring logs, monitoring wells, pump tests, and laboratory analyses provide valuable but often sparse datasets that may inadequately represent the full spatial complexity of the subsurface. As a result, remediation strategies such as pump-and-treat systems, bioremediation, soil vapour extraction, or permeable reactive barriers may be poorly designed, improperly positioned, or insufficiently scaled. Remediation failures not only prolong contamination but also increase operational costs, reduce stakeholder confidence, and complicate long-term site management (Roghani, 2018, Wang, Unger & Parker, 2014).

The integration of data-driven frameworks into contamination pathway prediction directly responds to these limitations by offering more advanced, adaptive, and comprehensive analytical capabilities. Data-driven approaches can fuse multiple datasets geophysical surveys, remote sensing, sensor-based monitoring, historical records, geochemical profiles, and hydraulic measurements to create a more holistic understanding of subsurface conditions. Machine learning algorithms can reveal hidden patterns in contaminant behaviour, identify key transport drivers, and detect anomalies that traditional models might miss (Derycke, et al., 2018, Kulawiak & Lubniewski, 2014). Geospatial analytics enable high-resolution mapping and probabilistic plume delineation, improving both the precision and reliability of predictions. By incorporating temporal datasets, data-driven models can also simulate contaminant migration under varying conditions, capturing the dynamic nature of subsurface environments. Figure 3 shows Conceptual site model of Pb contamination in Klity Creek showing Pb sources, five exposure pathways (fish ingestion, drinking water, soil ingestion and dermal contact, edible plant ingestion, and inhalation), fate and transport mechanisms, and human as well as ecological receptors presented by Phenrat, et al., 2016.

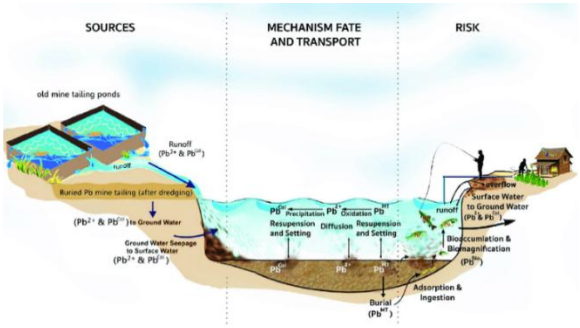

Figure 3: Conceptual site model of Pb contamination in Klity Creek showing Pb sources, five exposure pathways (fish ingestion, drinking water, soil ingestion and dermal contact, edible plant ingestion, and inhalation), fate and transport mechanisms, and human as well as ecological receptors (Phenrat, et al., 2016).

Moreover, data-driven frameworks significantly improve uncertainty quantification, enabling decision-makers to understand and plan for different contamination scenarios. Probabilistic forecasts help identify worst-case pathways, support remediation contingency planning, and strengthen regulatory submissions. Enhanced interpretability and transparency of data-driven outputs also facilitate better communication with stakeholders, regulators, and affected communities, promoting trust and supporting informed decision-making (Hoek, Beelen & Brunekreef, 2011, Levy, 2013).

The environmental significance of adopting a data-driven predictive framework is therefore substantial. By reducing uncertainty, enhancing accuracy, and

enabling more proactive interventions, these models contribute to the protection of water resources, the safeguarding of ecosystems, and the prevention of human exposure to dangerous contaminants. They support more efficient remediation investments, reduce long-term environmental liabilities, and foster sustainable land management practices. In an era marked by increasing environmental pressure, expanding industrial activities, and heightened regulatory expectations, data-driven frameworks offer a transformative pathway toward more effective subsurface contamination management and long-term environmental resilience (Bowen & Wittneben, 2011, Schaltegger & Csutora, 2012).

2.3. Data Acquisition and Characterization

Data acquisition and characterization form the foundational pillar of any data-driven framework designed to predict subsurface contamination pathways, as the quality, diversity, and representativeness of the data directly influence modelling accuracy and remediation outcomes. Subsurface environments are inherently complex, governed by interactions among geological structures, hydrological gradients, chemical processes, and anthropogenic disturbances. Capturing this complexity requires a multifaceted approach to data gathering that integrates hydrogeological datasets, soil geochemical profiles, geospatial mapping, historical monitoring records, real-time sensor inputs, and laboratory test results (Maas, Schaltegger & Crutzen, 2016, Tang & Luo, 2014). Each category of data offers unique insights into contaminant behaviour, transport dynamics, and environmental conditions, and together they create a comprehensive information ecosystem capable of supporting sophisticated predictive analytics.

Hydrogeological datasets are among the most essential components of subsurface contamination analysis, as groundwater movement often dictates the direction, velocity, and dispersion of contaminants. These datasets typically include measurements of hydraulic conductivity, aquifer thickness, groundwater elevation, flow direction, and recharge rates. Pump tests, slug tests, and hydraulic head measurements provide quantitative indicators of how water and consequently contaminants move through porous media. Spatial variations in hydraulic gradients reveal preferential flow paths that may accelerate contaminant migration or redirect plumes unexpectedly. Understanding these parameters is crucial because even minor shifts in groundwater flow can significantly impact plume evolution (Ascui, 2014, Hartmann, Perego & Young, 2013). Furthermore, hydrogeological data enable the calibration of predictive models by defining boundary conditions and governing flow equations, reducing uncertainty in model simulations.

Complementing hydrogeological data is soil geochemistry, which provides detailed information on the chemical composition, mineralogy, and physical properties of subsurface materials. Soil pH, organic content, clay fraction, cation-exchange capacity, and mineral constituents influence contaminant adsorption, desorption, retention, and degradation. For instance, hydrophobic organic contaminants tend to sorb strongly to soils rich in organic matter, while metals may form complexes with mineral surfaces under specific geochemical conditions. These processes determine whether contaminants remain immobile or become mobilized under fluctuating environmental conditions (Ascui & Lovell, 2012, Steininger, et al., 2016). Soil geochemical characterization is especially important for predicting the behaviour of complex contaminants such as DNAPLs, LNAPLs, or reactive species whose fate can change dramatically with shifts in redox conditions, moisture content, or microbial activity. By integrating these geochemical attributes into data-driven models, predictions become more representative of real-world interactions, improving the reliability of pathway forecasting.

Geospatial mapping plays a critical role in visualizing and contextualizing subsurface data, enabling the interpretation of spatial relationships that influence contamination patterns. Geographic Information Systems (GIS) allow the integration of diverse datasets into layered spatial models that highlight geological formations, land use, topographical gradients, drainage networks, and historical industrial activity. Remote sensing data, including satellite imagery and aerial surveys, can reveal surface disturbances, vegetation stress patterns, and potential contamination sources that may not be immediately evident from

ground-based observations (Burritt, Schaltegger & Zvezdov, 2011, Gibassier & Schaltegger, 2015). Geophysical mapping techniques such as electrical resistivity tomography, ground-penetrating radar, and electromagnetic surveys provide non-invasive insights into subsurface structures, identifying fractures, voids, stratigraphic boundaries, and anomalies that could act as conduits or barriers to contaminant transport. These geospatial datasets enhance model inputs by improving spatial resolution and reducing the uncertainty associated with interpolation between sampling points. Figure 4 shows figure of two conceptual plumes caused by DNAPL entry into fractured rock presented by Parker, Cherry & Chapman, 2012.

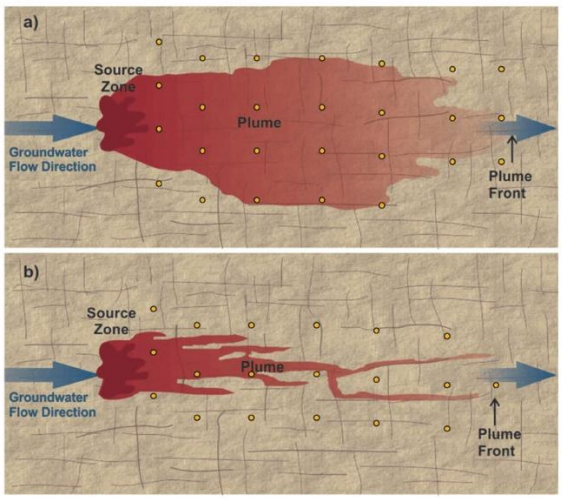

Figure 4: Two conceptual plumes caused by DNAPL entry into fractured rock (Parker, Cherry & Chapman, 2012).

Historical monitoring records are indispensable in understanding long-term contamination trends and validating predictive models. Many contaminated sites have decades of data from groundwater sampling, soil tests, well logs, and site investigations conducted during previous remediation efforts. These records document contaminant concentrations, plume shapes, seasonal fluctuations, and responses to remediation activities. By reconstructing historical plume dynamics, analysts can identify recurring migration patterns, detect plume stability or expansion, and determine whether contaminants are degrading, dissipating, or redistributing (Barzegar, et al., 2018, Karandish, Darzi-Naftchali & Asgari, 2017). Historical datasets also help differentiate between active contamination sources and residual impacts from past events. Incorporating these temporal datasets into data-driven models enhances their ability to simulate future scenarios based on observed trends, increasing predictive confidence and supporting regulatory reporting requirements.

Modern remediation projects increasingly rely on sensor data to provide continuous, real-time insights into subsurface conditions. Advanced sensor technologies measure groundwater levels, moisture content, temperature, electrical conductivity, and in some cases, specific contaminant concentrations. These sensors, often installed in monitoring wells or embedded in soil matrices, transmit data through automated networks, enabling remote monitoring and dynamic model updates. Real-time data capture transient events such as rainfall-driven infiltration, pumping-induced hydraulic shifts, or sudden contaminant releases that would otherwise remain undetected with periodic sampling (Park, et al., 2016, Ransom, et al., 2017). The integration of sensor data strengthens the responsiveness of predictive models, allowing them to adjust forecasts based on current conditions and enhancing the capacity for early warning detection of plume acceleration or deviation.

Laboratory test results provide high-precision measurements that anchor predictive modelling in scientific accuracy. Laboratory analyses quantify concentrations of contaminants, evaluate chemical speciation, and measure reaction rates under controlled conditions. Tests such as grain-size distribution, permeability measurements, sorption isotherms, and batch degradation studies help characterize the fundamental physical and chemical interactions that govern contaminant behaviour. Laboratory microcosms, for example, can be used to simulate biodegradation potential under varying environmental conditions, enabling analysts to assess whether natural attenuation is likely to contribute significantly to remediation (Naghibi, Pourghasemi & Dixon, 2016, Rodriguez-Galiano, et al., 2014). Furthermore, laboratory-derived parameters are essential inputs for both physics-based and machine learning models, ensuring that predictions reflect empirically validated relationships rather than abstract approximations.

The integration of these diverse datasets is not without challenges. Subsurface data are often incomplete, spatially irregular, or collected using different methodologies, resulting in inconsistencies that complicate model calibration. Data gaps may occur in regions where drilling is impractical, where historical records are unavailable, or where sensors fail to capture transient changes. Additionally, variability in sampling frequency and measurement precision can introduce bias if not properly accounted for. Thus, preprocessing methods such as interpolation, normalization, uncertainty quantification, and data fusion become essential steps in preparing the dataset for predictive modelling (Liakos, et al., 2018, Singh, Gupta & Mohan, 2014). Machine learning techniques can help mitigate some of these challenges by identifying patterns within incomplete datasets, estimating missing values, and detecting anomalies indicative of measurement errors or unexpected environmental events.

Despite these challenges, comprehensive data acquisition and characterization significantly enhance the predictive capability of subsurface contamination models. The richness of multi-source datasets allows data-driven frameworks to capture the heterogeneity, nonlinearity, and dynamic behaviour of contamination processes that traditional models often miss. By leveraging high-resolution hydrogeological data, detailed geochemical profiles, robust geospatial mappings, long-term monitoring records, real-time sensor streams, and laboratory analyses, predictive models become more adaptive, accurate, and relevant to real-world remediation demands. Environmental managers can make more informed decisions, regulators gain access to more defensible assessments, and communities benefit from improved protection against subsurface contamination risks (Ahmed, 2017, Karpatne, et al., 2018).

Ultimately, the strength of any data-driven framework lies in the quality and depth of its foundational data. By embracing comprehensive and rigorous data acquisition strategies, remediation practitioners can overcome many of the uncertainties that have historically hindered contamination prediction. This holistic approach not only improves scientific understanding of subsurface systems but also enables more effective and proactive remediation planning an essential step toward achieving long-term environmental sustainability and resilience.

2.4. Analytical and Machine Learning Techniques

Analytical and machine learning techniques form the computational core of a data-driven framework for predicting subsurface contamination pathways, enabling the integration of complex datasets and the modelling of contaminant behaviour in ways that exceed the capabilities of traditional approaches. Subsurface systems are governed by nonlinear interactions among geological structures, hydrogeological dynamics, chemical reactions, and anthropogenic influences (Liakos, et al., 2018, Singh, Gupta & Mohan, 2014). These factors introduce significant uncertainty and spatial variability, making contaminant transport difficult to predict using deterministic or empirical models alone. Machine learning and advanced computational analytics allow researchers and remediation practitioners to uncover hidden patterns, quantify uncertainty, and simulate contaminant migration with greater precision. Several classes of computational techniques including supervised learning, unsupervised clustering, physics-informed models, geospatial analytics, and time-series prediction play distinct but complementary roles in capturing the multidimensional behaviour of subsurface contamination.

Supervised learning algorithms are fundamental tools for predicting contaminant concentrations, plume extents, and transport pathways based on labelled datasets. In these models, historical or experimentally measured contaminant behaviours are used to train algorithms to recognize relationships between input variables and output responses. Techniques such as random forests, support vector machines, gradient boosting, and artificial neural networks can incorporate large numbers of predictors, including soil properties, groundwater velocity, hydraulic gradients, and geochemical indicators (Ahmed, 2017, Karpatne, et al., 2018). These models are particularly effective when complex nonlinearities govern contaminant movement. For instance, neural networks can approximate intricate functional relationships that describe adsorption-desorption cycles or multiphase fluid interactions. Supervised learning is especially

powerful for site-specific forecasting where historical monitoring data provide a strong basis for model calibration. Once trained, these models can generate rapid predictions of plume trajectories or hotspot locations, supporting timely decision-making in remediation design and risk mitigation. However, supervised learning relies on the availability of labelled datasets; when such data are limited or unevenly distributed, alternative or complementary methods become necessary.

Unsupervised clustering offers a valuable approach for identifying patterns and structural behaviours within unlabelled subsurface datasets. Techniques such as k-means clustering, hierarchical clustering, and density-based spatial clustering of applications with noise (DBSCAN) can group areas of similar hydrogeological or geochemical properties, revealing zones that may favour specific contaminant behaviours. Clustering methods also help detect anomalies, such as unexpected contaminant concentration spikes or irregular flow patterns, which may indicate new sources, preferential pathways, or sampling errors (Lemming, 2010, Wang, et al., 2017). In the context of plume delineation, unsupervised algorithms can automatically segment spatial datasets into regions representing high, moderate, or low contaminant levels, improving the efficiency of mapping and monitoring efforts. These unsupervised insights enhance the robustness of supervised models by enabling better feature engineering, reducing dimensionality, and improving data quality through anomaly detection.

Physics-informed models represent another vital advancement in computational techniques for subsurface contamination prediction. These methods integrate machine learning with physical laws governing groundwater flow and contaminant transport, such as Darcy's law, advection-dispersion equations, and mass balance constraints. Physics-informed neural networks (PINNs) enforce these governing principles within the learning architecture, ensuring that model outputs are not only data-driven but also consistent with hydrological realities. This hybrid modelling approach addresses the limitations of purely statistical learning, which may produce accurate predictions within the training domain but generate unrealistic behaviours under extrapolation (An, et al., 2016, Mgbeahuruike, 2018). By embedding physics into the computational framework, researchers can simulate contamination pathways even in areas with sparse data, reducing uncertainty and improving model generalization. PINNs are especially useful in scenarios involving multiphase flow, reactive transport, or heterogeneous media where empirical data alone cannot fully characterize contaminant dynamics.

Geospatial analytics provides the spatial intelligence required to understand how contaminants migrate across complex terrains. Geographic Information Systems (GIS), spatial interpolation methods, and spatial machine learning algorithms allow the integration of multilayer datasets including geological maps, soil classifications, hydrological networks, and historical contamination footprints into high-resolution spatial models. Techniques such as kriging, inverse distance weighting, and neural-network-based spatial prediction help estimate contaminant concentrations at unsampled locations, improving spatial continuity and reducing uncertainty. Spatial autocorrelation measures like Moran's I and Geary's C reveal patterns of plume clustering or dispersion, offering insights into how contaminants respond to subsurface structures (Hardie & McKinley, 2014, Williamson, 2011). More advanced geospatial approaches, including spatial decision-support models and geostatistical simulations, enable practitioners to evaluate multiple contamination scenarios and select optimal remediation strategies. Geospatial analytics is also crucial for integrating remote-sensing data, enabling the detection of surface indicators of subsurface contamination such as vegetation stress or soil anomalies.

Time-series prediction techniques add a temporal dimension to contaminant pathway modelling, enabling the forecasting of plume evolution under changing environmental conditions. Historical monitoring records such as groundwater levels, seasonal recharge rates, and contaminant concentration trends form the basis for predictive models that anticipate future plume behaviour. Machine learning algorithms such as long short-term memory (LSTM) networks, autoregressive integrated moving average (ARIMA) models, and temporal convolutional networks (TCNs) can analyse sequential

data to predict fluctuations in contamination levels or identify early warning signals of plume migration acceleration. Time-series models are particularly valuable for capturing the effects of climatic events, land-use changes, pumping activities, or remediation interventions that alter hydraulic conditions (Cappuyns & Kessen, 2014, Williamson, et al., 2011). They also allow continuous model updating as new data become available, making predictions more adaptive and reflective of real-time environmental dynamics.

The integration of these computational techniques enhances predictive accuracy and reduces the uncertainties that traditionally plague subsurface contamination modelling. By combining supervised models for targeted prediction, unsupervised methods for pattern recognition, physics-informed models for constraint-based learning, geospatial analytics for spatial interpretation, and time-series forecasting for temporal insights, a comprehensive framework emerges that can accommodate the inherent complexity of subsurface environments. This integrated computational approach enables multi-source data fusion, allowing diverse datasets such as hydrogeological measurements, soil geochemistry, geophysical surveys, laboratory analyses, and real-time sensor outputs to inform a unified predictive model (Mitchell, 2012, Sweeney & Kabouris, 2017). Machine learning enhances the ability to detect nonlinear interactions, while physics-based constraints ensure realism and scientific validity. Geospatial and temporal techniques add holistic dimensions that reflect the true variability of environmental systems.

Moreover, advanced analytics and machine learning facilitate uncertainty quantification, an essential component of credible contamination pathway prediction. Techniques such as Monte Carlo simulation, Bayesian inference, and ensemble modelling enable analysts to evaluate model reliability, identify high-risk zones, and support decision-making under uncertainty. These methods provide probabilistic predictions that reflect not only the most likely plume behaviour but also the range of possible outcomes, improving transparency in regulatory reporting and remediation planning (Cheng, et al., 2011, Herat & Agamuthu, 2012).

As computational capacity and data availability continue to expand, machine learning models can be continuously refined, incorporating emerging data streams from remote sensing platforms, autonomous monitoring systems, and advanced laboratory techniques. This adaptability is crucial, as subsurface contamination is rarely static; pathways evolve with shifts in groundwater dynamics, climatic patterns, and human activities. Data-driven computational methods thus enable dynamic modelling frameworks that remain relevant throughout the lifecycle of a remediation project, from initial site assessment to long-term monitoring (Boriana, 2017, Hou & Al-Tabbaa, 2014).

Ultimately, the use of analytical and machine learning techniques in subsurface contamination pathway prediction represents a transformative shift in environmental modelling. These techniques empower practitioners to overcome the limitations of traditional deterministic models, enabling deeper insights into complex environmental behaviours, improving predictive accuracy, and enhancing the effectiveness of remediation strategies. By integrating scientific principles with advanced computation, data-driven frameworks provide a powerful foundation for protecting groundwater resources, ensuring regulatory compliance, and promoting sustainable environmental management in increasingly complex remediation landscapes.

2.5. Framework Architecture for Predicting Contamination Pathways

The architecture of a data-driven framework for predicting subsurface contamination pathways in complex remediation projects must be designed to integrate diverse datasets, sophisticated modelling techniques, and scientifically grounded simulation tools into a unified system capable of producing accurate and actionable predictions. This framework begins with a robust data preprocessing pipeline, progresses into advanced model development and calibration, incorporates seamless coupling with hydrogeological simulations, and culminates in contamination pathway estimation and risk-zone delineation. Each component plays a crucial role in capturing the multifaceted dynamics of contaminant migration, addressing uncertainties, and supporting

effective remediation decision-making (Ferdinand & Yu, 2016, Koop & van Leeuwen, 2017).

Data preprocessing forms the foundation of the framework, ensuring that raw datasets from various sources are cleaned, standardized, and structured for analysis. Subsurface datasets such as soil geochemistry, hydraulic measurements, geospatial maps, laboratory test results, and sensor data often vary widely in temporal frequency, spatial resolution, and measurement units. Many datasets contain missing values, noise, or inconsistencies arising from sampling errors, instrument limitations, or incomplete historical records. Preprocessing therefore includes data cleaning techniques such as noise filtering, interpolation of missing values, normalization of variable scales, and transformation of categorical inputs into quantitative features. Spatial datasets are georeferenced and harmonized to ensure alignment across mapping layers, while temporal datasets are synchronized to capture the sequence of hydrogeological events that influence contaminant transport (Jayasooriya, 2016, Sayles, 2017). Feature engineering may also be applied to extract meaningful indicators such as hydraulic gradients, soil moisture indices, or contaminant decay coefficients, enabling machine learning models to capture complex interactions. The objective of preprocessing is to build a coherent dataset that accurately reflects subsurface conditions and supports reliable model development.

Once data preprocessing is complete, the framework advances to model development, where analytical and machine learning techniques are deployed to learn contaminant behaviour from the input data. Model development typically involves selecting appropriate algorithms such as neural networks, random forests, gradient boosting machines, or physics-informed models based on the characteristics of the contamination problem. For example, neural networks may be suited to capturing nonlinear interactions among geochemical and hydrogeological variables, while physics-informed neural networks integrate governing fluid-flow equations directly into the learning process (Kato, 2010, Meerow & Newell, 2017). In developing these models, training datasets are used to calibrate algorithm parameters, while validation datasets ensure that predictions generalize to unseen conditions. Hyperparameter tuning methods, including grid search, Bayesian optimization, or evolutionary algorithms, refine model performance by identifying optimal configurations for depth, learning rates, regularization strategies, or decision splits. Throughout this process, cross-validation techniques help reduce overfitting and improve the robustness of model outputs. The model development stage is iterative, involving repeated cycles of refinement, performance evaluation, and error analysis to align the predictive model with observed contaminant patterns.

A key strength of the proposed framework lies in its integration of data-driven models with hydrogeological simulations. Traditional hydrogeological models, based on Darcy's law and advection–dispersion principles, offer established scientific grounding and can simulate groundwater flow and contaminant movement under controlled assumptions. However, they often struggle to accommodate the full complexity of real-world subsurface conditions. Coupling these simulations with machine learning models enables the framework to benefit from both empirical and physics-based perspectives (Furniss, 2011, Handmer, et al., 2012). In this integrated architecture, hydrogeological simulations provide boundary conditions, hydraulic heads, aquifer properties, and flow velocities that serve as inputs or constraints for data-driven models. Conversely, machine learning outputs including predicted contaminant concentrations, mobility indicators, or pathway probabilities can refine or adjust simulation parameters, improving the realism of physics-based modelling. This bidirectional coupling allows the system to generate hybrid predictions that are grounded in scientific principles while remaining sensitive to empirical patterns not captured by conventional models.

Pathway estimation represents one of the most critical outcomes of the integrated framework. Once the models have been developed and coupled with hydrogeological simulations, the system predicts contamination pathways by analysing the interactions between contaminant properties, soil characteristics, flow dynamics, and environmental conditions. These predictions often take the form of spatially explicit plume migration maps that trace the expected direction, speed, and extent of contaminant travel. Machine learning models may generate probability

surfaces indicating the likelihood of contaminant presence across the subsurface, enabling environmental managers to identify potential preferential flow paths, plume divergence zones, or stagnation points. Time-series prediction components capture temporal variations, projecting how pathways may shift under seasonal recharge, pumping activities, or remediation interventions (Hubbard, et al., 2018, Singh, van Werkhoven & Wagener, 2014). Incorporating uncertainty quantification adds depth to pathway estimation, allowing analysts to visualize confidence intervals or worst-case scenarios and plan accordingly. This probabilistic perspective is particularly useful in complex remediation environments where geological heterogeneity and incomplete data introduce significant uncertainties.

Risk-zone delineation is the culminating step of the framework, translating pathway predictions into actionable insights for remediation planning, regulatory compliance, and risk communication. Risk zones represent spatial areas categorized according to contamination likelihood, potential exposure severity, or environmental vulnerability. These zones often classified as high-risk, moderate-risk, or low-risk are delineated by combining predicted contaminant concentrations, plume trajectories, groundwater usage patterns, proximity to receptors, and regulatory thresholds. Geospatial analytics play a central role in converting model outputs into risk-zone maps that align with real-world coordinates and site boundaries (Field, 2012, McMillan, et al., 2016). Environmental risk indices may also be calculated to quantify hazards to drinking water sources, ecosystems, infrastructure, or human populations. Such delineation supports strategic decision-making by identifying priority remediation areas, informing the placement of monitoring wells, guiding control measures, and optimizing the allocation of financial and technical resources. Risk-zone mapping further enhances transparency by providing stakeholders such as regulators, community groups, and remediation contractors with clear visualizations of contamination risks (Edwards, et al., 2012, Green, 2016).

The integrated framework architecture also incorporates a feedback mechanism that enables continuous improvement. As new monitoring data become available from field measurements, sensors, or laboratory analyses they are fed back into the preprocessing pipeline, allowing the models to be retrained and refined. This dynamic updating ensures that predictions remain accurate and reflect evolving subsurface conditions. The iterative feedback loop is particularly important for long-term remediation projects, where environmental conditions may change significantly over time due to climatic events, land-use changes, or engineered interventions. The adaptability of the framework allows it to remain relevant and reliable throughout the lifecycle of the remediation process (Viviroli, et al., 2011, Watts, et al., 2015).

Another essential component of the framework is model interpretability and transparency. While machine learning models can produce highly accurate predictions, their complexity can sometimes hinder understanding of the underlying processes. To address this, interpretability tools such as feature importance scores, SHAP values, or sensitivity analyses can be incorporated to reveal which variables most influence contaminant pathways. This not only aids scientific understanding but also increases confidence among regulators and practitioners who rely on the model's outputs for decision-making. Transparent reporting of model assumptions, uncertainties, and limitations further enhances the credibility of the framework (Nelitz, Boardley & Smith, 2013, Perra, et al., 2018).

Overall, the proposed integrated framework architecture unites data preprocessing, advanced model development, hydrogeological simulation coupling, pathway estimation, and risk-zone delineation into a cohesive system capable of addressing the complexity of subsurface contamination. By combining machine learning with physics-based modelling and geospatial analysis, the framework overcomes the limitations of traditional methods and provides deeper insights into contaminant behaviour. Its dynamic, data-driven nature ensures adaptability and continuous improvement, while its focus on risk communication and decision support enhances practical applicability (Leibowitz, et al., 2014, Ribeiro Neto, et al., 2014). As contaminated sites grow in number and complexity, such comprehensive frameworks are essential for protecting groundwater resources, ensuring public health, and supporting sustainable environmental remediation practices.

2.6. Uncertainty Quantification and Model Validation

Uncertainty quantification and model validation are essential components of a data-driven framework for predicting subsurface contamination pathways, as they ensure that model outputs are scientifically credible, operationally reliable, and suitable for supporting high-stakes remediation decisions. Subsurface environments are inherently complex, characterized by heterogeneous geological formations, variable hydrogeological conditions, evolving contaminant properties, and incomplete or noisy datasets. These complexities introduce different types of uncertainty ranging from measurement errors and spatial gaps to model structure limitations and parameter variability that can significantly influence the accuracy of contamination pathway predictions (Hanson, et al., 2012, Wagesho, 2014). A robust framework must therefore incorporate systematic approaches for assessing model robustness, evaluating prediction confidence, performing sensitivity analyses, and validating results against real-world observations and historical plume behaviour. By doing so, it strengthens the trustworthiness of the predictive system and provides decision-makers with the information necessary to manage risk effectively.

Model robustness assessment begins with identifying sources of uncertainty within the dataset and the modelling process. In subsurface contamination modelling, uncertainties may arise from sparse monitoring well distributions, irregular sampling intervals, sensor inaccuracies, incomplete geological maps, laboratory measurement variability, or simplifying assumptions in hydrogeological simulations. Data-driven models themselves may introduce uncertainties through algorithm selection, hyperparameter tuning, and the sensitivity of predictions to training data variability. To address these issues, ensemble modelling techniques are often deployed. Ensemble approaches, such as bagging, boosting, or random forests, generate multiple models using variations of the dataset or algorithmic parameters and aggregate the results to produce a more stable and robust prediction (Langat, Kumar & Koech, 2017, Nashwan, et al., 2018). This reduces the influence of outliers and compensates for weaknesses in individual models. Variational methods and dropout sampling in neural networks also provide probabilistic interpretations, allowing the model to quantify uncertainty in its predictions. Such robustness assessments help determine whether the model can generalize beyond training conditions and perform reliably under different contamination scenarios.

Evaluating prediction confidence is another critical dimension of uncertainty quantification. Confidence assessment involves estimating the likelihood that predicted contamination pathways or plume extents fall within acceptable error margins given the available data and model assumptions. Probabilistic modelling techniques, including Bayesian inference, Gaussian processes, and Monte Carlo simulations, are frequently employed to estimate prediction distributions rather than single deterministic outputs. These techniques generate confidence intervals or probability density maps that show the range of possible contamination pathways and highlight areas with high or low prediction certainty (Gober & Kirkwood, 2010, Mark, et al., 2010). For example, Bayesian models treat parameters as probability distributions rather than fixed values, updating their estimates as new data become available. By doing so, they capture the evolving nature of subsurface conditions and provide decision-makers with transparent information about prediction reliability. Monte Carlo simulations allow the model to run thousands of iterations using random variations of key parameters such as hydraulic conductivity, porosity, source concentration, or degradation rates, thereby revealing how parameter variability influences the predicted plume. These confidence measures help remediation engineers plan for best-case and worst-case scenarios, improving contingency planning and risk communication.

Sensitivity analysis plays a vital role in understanding how different input variables influence model outputs. Because subsurface contamination pathways depend on a complex interplay of geological, hydrological, and chemical factors, identifying the most influential parameters helps refine both modelling accuracy and field investigation priorities. Sensitivity analysis methods such as local sensitivity (one-factor-at-a-time) and global sensitivity approaches (Sobol indices, Morris screening) quantify how changes in specific inputs affect the variation in predicted outcomes. For

instance, a sensitivity analysis may reveal that hydraulic conductivity variations significantly alter plume migration directions, while soil organic content plays a lesser role in certain contexts. Such insights guide field teams in prioritizing the collection of high-impact data, optimizing the placement of monitoring wells, and allocating resources toward reducing critical uncertainties (Essaid, Bekins & Cozzarelli, 2015, Kobus, Barczewski & Koschitzky, 2012). Sensitivity analysis also evaluates model stability, determining whether slight changes in environmental conditions or dataset characteristics lead to disproportionately large shifts in predictions. Stable models yield consistent results even when faced with minor variations, while unstable models require further refinement or restructuring. By systematically identifying the drivers of prediction variability, sensitivity analysis enhances both model development and interpretability.

Model validation is arguably the most important step in ensuring that data-driven contamination predictions reflect real-world conditions. Validation involves comparing model outputs with observed field data, historical plume behaviour, or independent datasets that were not used during model training. This step confirms whether the framework can accurately reproduce known contaminant migration patterns and reliably forecast future conditions. Field-based validation may include groundwater sampling, borehole logs, soil core analyses, geophysical surveys, or sensor-based measurements (Kuppusamy, et al., 2016, Majone, et al., 2015). Model predictions of plume boundaries, contaminant concentrations, or migration directions are compared with observed measurements using statistical metrics such as root mean square error (RMSE), Nash–Sutcliffe efficiency, R-squared values, or spatial similarity indices. Discrepancies between predictions and observations highlight areas where the model may require recalibration, additional data inputs, or structural refinements.

Historical plume behaviour offers another important validation avenue. Many contaminated sites have long-term monitoring records documenting how contamination spread over years or decades. By simulating past contamination dynamics and comparing outputs with recorded plume shapes, researchers can assess whether the model captures the temporal evolution of contamination. If the model successfully reconstructs historical patterns, confidence in its predictive power for future scenarios increases. Conversely, if the model fails to align with historical behaviours, this signals gaps in the data inputs, missing physical processes, or inadequacies in the modelling architecture. Time–lag validation, in which models are tested on data from subsequent monitoring periods, further strengthens validation by demonstrating whether predictions remain accurate as conditions evolve (Yaron, Dror & Berkowitz, 2012, Zeidan, 2017).

Cross-validation techniques are essential for model generalization. K-fold cross-validation, leave-one-out validation, and spatial cross-validation allow the model to be tested on multiple partitions of the dataset to ensure it performs consistently across different spatial and temporal subsets. Spatial cross-validation is particularly important for subsurface modelling because spatial autocorrelation can artificially inflate performance metrics if training and validation datasets are too similar. By separating data into geographically distinct areas, spatial cross-validation ensures that the model can predict contamination dynamics in unmonitored regions, a critical requirement for practical application (Binley, et al., 2015, Francisca, et al., 2012).

The final aspect of uncertainty quantification and validation involves integrating results into decision-making tools. Visualization of uncertainty is essential for communicating results to regulators, stakeholders, and remediation engineers. Probabilistic plume maps, uncertainty heatmaps, and confidence-band time-series plots help illustrate where model predictions are strong and where caution is warranted. Decision-support systems may incorporate uncertainty thresholds to trigger alerts, guide monitoring efforts, or prioritize remediation actions. For example, a high-uncertainty zone may prompt additional field sampling, while a high-confidence plume prediction may justify immediate intervention measures such as barrier installation or groundwater extraction (Filippini, 2015, Mallants, et al., 2010).

Ultimately, uncertainty quantification and model validation strengthen the scientific, regulatory, and

operational integrity of the contamination prediction framework. By identifying uncertainties, evaluating prediction confidence, understanding sensitivity to key variables, and validating results against real-world behaviour, the framework becomes more transparent, reliable, and adaptable. These processes not only enhance technical performance but also foster trust among stakeholders, ensuring that remediation decisions are grounded in robust evidence. As subsurface contamination challenges grow more complex and environmental expectations intensify, rigorous uncertainty quantification and validation remain essential pillars of responsible and effective environmental modelling.

2.7. Case Studies and Practical Applications

Case studies and practical applications of data-driven frameworks for predicting subsurface contamination pathways provide essential insights into how these advanced systems perform under real-world conditions and demonstrate their superiority over conventional deterministic models. By analyzing actual and simulated remediation projects, it becomes clear that data-driven approaches not only enhance predictive accuracy but also support more informed decision-making, resource optimization, and long-term environmental monitoring. These case examples illustrate how integrating machine learning, geospatial analytics, physics-informed modelling, and multi-source environmental data can reveal patterns and pathways previously obscured by the limitations of traditional modelling tools (Hipsey, et al., 2015, Scheidt, Li & Caers, 2018).

One compelling example involves an industrial site contaminated with chlorinated solvents leaked from historical degreasing operations. Conventional plume modelling based on deterministic hydrogeological simulations had struggled to accurately represent observed plume behaviour due to the site's heterogeneous subsurface conditions, including fractured bedrock and variable soil permeability. A data-driven framework was implemented to improve pathway predictions by integrating decades of monitoring well records, soil geochemistry, geophysical surveys, and real-time groundwater level sensors. Machine learning algorithms were trained to identify relationships between geological features and contaminant concentrations, while physics-informed neural networks enforced continuity with groundwater flow principles. The resulting predictions showed a significantly improved match with field observations compared to deterministic models, particularly in identifying secondary migration pathways that had previously gone undetected (Deschaine, 2014, Kresic & Mikszewski, 2012). This enhanced predictive capability allowed remediation engineers to redesign extraction well placement, prioritize high-risk zones, and avoid unnecessary drilling in areas that posed minimal contamination risk. Ultimately, the updated remediation strategy reduced operational costs and shortened cleanup timelines, demonstrating the practical value of the data-driven framework.

A second example can be drawn from a simulated petroleum spill scenario designed to test the robustness of different modelling approaches under varying hydrogeological conditions. The simulation incorporated synthetic datasets representing sandy aquifers, clay lenses, and fractured rock systems to mimic real-world complexity. Deterministic advection–dispersion models produced plume predictions that were highly sensitive to small variations in soil permeability and groundwater gradients, resulting in large uncertainties in predicted plume length and direction. In contrast, the data-driven framework utilized unsupervised clustering to classify subsurface regions with similar geophysical characteristics, improving the representation of preferential flow zones (Bello-Dambatta & Javadi, 2010, Felisa, et al., 2015). Supervised learning models, trained on synthetic tracer test results, provided more reliable estimates of contaminant velocity and attenuation rates across different subsurface materials. When predictions were compared to the "true" simulated plume, the data-driven approach demonstrated substantially higher spatial accuracy and reduced prediction error. This case not only validated the framework's predictive strength but also highlighted its resilience to parameter uncertainty an essential advantage when working with incomplete or variable field data.

A third case study involved a former agricultural site impacted by nitrate leaching from fertilizer applications. Predicting nitrate migration is particularly challenging because its movement

depends on dynamic interactions among soil moisture, microbial processes, groundwater recharge, and agricultural practices. Deterministic models often oversimplify these interactions, leading to inaccurate forecasts of groundwater contamination hotspots. In this agricultural case, the data-driven framework integrated remote-sensing data on vegetation health, rainfall records, soil moisture sensors, groundwater monitoring data, and laboratory nitrate measurements. Time-series prediction techniques, such as long short-term memory (LSTM) networks, captured seasonal variations in nitrate mobility, while geospatial analytics mapped the spatial distribution of high-risk zones (Liang, 2018, McGrath, Reid & Tran, 2017). The resulting predictions enabled water managers to implement targeted mitigation strategies such as buffer strips, adjusted fertilizer application schedules, and enhanced monitoring in areas projected to have elevated nitrate concentrations. By comparing these predictions with field measurements collected over several years, the framework demonstrated high temporal accuracy and provided valuable insights for regulatory agencies tasked with protecting drinking water sources.

Another practical application can be observed in urban brownfield redevelopment projects, where complex mixtures of contaminants often coexist due to historical industrial usage. One such project involved a former manufacturing site with petroleum hydrocarbons, heavy metals, and polycyclic aromatic hydrocarbons (PAHs). Traditional plume models could not adequately capture interactions between these contaminants or account for the influence of urban infrastructure on subsurface flow patterns. The data-driven framework implemented for this project used multi-contaminant modelling techniques, integrating soil vapor intrusion measurements, groundwater data, building foundation maps, and chemical degradation profiles (Bello-Dambatta, 2010, Leeson, et al, 2013). Machine learning algorithms identified zones where co-contaminant interactions accelerated degradation or mobilization, while geospatial analytics mapped risk zones under existing and future land-use scenarios. This allowed urban planners and developers to make informed decisions regarding excavation requirements, building foundation design, and long-term monitoring plans. The framework also facilitated compliance with regulatory risk assessment requirements by producing transparent and scientifically defensible predictions.

A further example involves a coastal industrial facility where saline intrusion complicates contaminant transport dynamics. Deterministic models struggled to account for density-driven flow processes affecting the migration of heavy metals and industrial solvents. A data-driven approach combined seawater intrusion models, tidal fluctuation data, electrical conductivity measurements, and chemical concentration profiles to better represent the dynamic interface between freshwater and saline water. Physics-informed machine learning models simulated contaminant movement under fluctuating tidal conditions, revealing periodic shifts in plume direction and intensity that were missed by traditional methods (Awe, Akpan & Adekoya, 2017, Osabuohien, 2017). These insights helped engineers design adaptive remediation systems capable of responding to tidal cycles, including adjustable pumping schedules and dynamic barrier controls, ultimately improving remediation efficiency.

Across all these cases, a consistent theme emerges: data-driven frameworks provide superior predictive power by capturing the nonlinearity, heterogeneity, and temporal variability inherent in subsurface environments. Their ability to leverage diverse datasets allows them to uncover subtle patterns and interactions that deterministic models lack the flexibility to represent. Additionally, the probabilistic outputs generated by many data-driven models enhance decision-making by quantifying uncertainty, allowing remediation planners to make risk-informed choices rather than relying on single deterministic estimates (Awe & Akpan, 2017).

Decision-support relevance is another major benefit highlighted in these case studies. The integration of data-driven outputs into geospatial visualization tools enables stakeholders to interact with contamination maps, identify priority zones, and evaluate remediation alternatives. These visuals help regulators, community members, and project engineers understand contamination risks more clearly, fostering transparency and collaborative decision-making. Furthermore, many data-driven frameworks support scenario analysis, allowing users

to simulate the effects of different remediation strategies, climatic events, or operational changes. This level of adaptability is essential in the face of evolving environmental conditions and regulatory expectations (Akpan, et al., 2017, Oni, et al., 2018).

Importantly, the transition from deterministic to data-driven approaches does not eliminate the role of traditional hydrogeological modelling; rather, it enhances it. By coupling data-driven insights with physics-based simulations, the hybrid models produced in many of these case studies capture both empirical patterns and scientifically grounded processes. This integration leads to more reliable forecasts, better alignment with field observations, and improved remediation outcomes (Ike, et al., 2018).

Overall, the case studies demonstrate that data-driven frameworks provide tangible advantages across a wide range of contamination scenarios. They improve predictive accuracy, reduce uncertainty, optimize remediation resource allocation, and strengthen regulatory compliance. Whether applied to industrial sites, agricultural landscapes, urban environments, or coastal zones, these frameworks consistently outperform conventional models and represent a transformative advancement in environmental remediation practice. As environmental challenges intensify and data availability expands, data-driven approaches will continue to play a central role in safeguarding groundwater resources, protecting public health, and enabling effective and sustainable remediation strategies (Awe, 2017, Osabuohien, 2019).

2.8. Conclusion

The development of a data-driven framework for predicting subsurface contamination pathways offers a transformative advancement in environmental remediation, providing a more accurate, adaptive, and holistic understanding of contaminant behaviour beneath the ground. The key findings across the framework components reveal that integrating diverse datasets ranging from hydrogeological measurements and soil geochemistry to geospatial mapping, monitoring records, sensor networks, and laboratory analyses greatly enhances the capacity to represent the complexity of subsurface systems. Machine learning techniques, physics-informed modelling, geospatial analytics, and time-series prediction collectively address the nonlinearities and uncertainties that traditional deterministic models often fail to capture. Through this multidimensional analytical architecture, the framework generates more precise plume forecasts, identifies previously undetected migration pathways, delineates high-risk zones, and improves the scientific and operational foundation for remediation decision-making.

In practical terms, the framework significantly strengthens remediation engineering by improving the accuracy of contamination assessments and enabling more strategic resource allocation. Enhanced pathway prediction supports optimized well placement, targeted soil or groundwater treatment, and more efficient remediation system designs. The integration of uncertainty quantification further enriches decision-making, providing probabilistic insights that guide risk-based planning and regulatory compliance. The case studies demonstrate that this approach not only improves predictive reliability but also reduces long-term remediation costs, shortens project timelines, and enhances stakeholder confidence by offering transparent, data-supported outcomes. The framework's ability to incorporate historical plume behaviour and dynamically update predictions as new information becomes available highlights its value for long-term monitoring and adaptive management.

Despite its strengths, several limitations remain. The accuracy of predictions is still constrained by data availability, spatial coverage, and sensor reliability. Heterogeneous subsurface environments can produce complex interactions that challenge even advanced machine learning models. Computational demands may be significant when integrating large datasets or running hybrid simulations. Moreover, while physics-informed models improve predictive realism, they still depend on accurate representation of physical processes and high-quality boundary conditions. There is also an ongoing need to enhance model interpretability, ensuring that highly technical outputs can be clearly understood by regulators, engineers, and community stakeholders.

Future research should prioritize real-time data integration to enhance responsiveness to evolving subsurface conditions. Incorporating data streams

from advanced sensing technologies such as distributed fiber-optic sensing, autonomous subsurface probes, and novel geophysical imaging methods will increase temporal and spatial resolution, improving both predictive accuracy and early-warning capabilities. Expanding the use of hybrid models that further unify machine learning with multiphase flow and reactive transport simulations will help capture more complex contamination behaviours. Research should also explore cloud-based platforms and edge computing to support scalable, real-time modelling across large or remote sites. Additionally, developing explainable AI tools for environmental applications will help bridge the gap between computational sophistication and practical usability.

In conclusion, the data-driven framework represents a powerful and forward-looking approach to understanding and managing subsurface contamination. By integrating multi-source data with advanced analytical tools, it provides deeper insight, greater predictive confidence, and more effective remediation strategies than traditional methods. Continued innovation in real-time monitoring, advanced sensing, hybrid modelling, and explainability will further strengthen this framework, supporting sustainable environmental protection and more resilient remediation practices in increasingly complex contamination scenarios.

REFERENCES

[1] Ahmed, F. (2017, October). An IoT-big data based machine learning technique for forecasting water requirement in irrigation field. In *International conference on research and practical issues of enterprise information systems* (pp. 67-77). Cham: Springer International Publishing.

[2] Akpan, U. U., Adekoya, K. O., Awe, E. T., Garba, N., Oguncoker, G. D., & Ojo, S. G. (2017). Mini-STRs screening of 12 relatives of Hausa origin in northern Nigeria. Nigerian Journal of Basic and Applied Sciences, 25(1), 48-57.

[3] Alibakhshi, S., Groen, T. A., Rautiainen, M., & Naimi, B. (2017). Remotely-sensed early warning signals of a critical transition in a wetland ecosystem. *Remote Sensing*, *9*(4), 352.

[4] An, C. J., McBean, E., Huang, G. H., Yao, Y., Zhang, P., Chen, X. J., & Li, Y. P. (2016). Multi-soil-layering systems for wastewater treatment in small and remote communities. *J. Environ. Inform*, *27*(2), 131-144.

[5] Andres, L., Boateng, K., Borja-Vega, C., & Thomas, E. (2018). A review of in-situ and remote sensing technologies to monitor water and sanitation interventions. *Water*, *10*(6), 756.

[6] Ascui, F. (2014). A review of carbon accounting in the social and environmental accounting literature: what can it contribute to the debate?. *Social and Environmental Accountability Journal*, *34*(1), 6-28.

[7] Ascui, F., & Lovell, H. (2012). Carbon accounting and the construction of competence. *Journal of Cleaner Production*, *36*, 48-59.

[8] Awe, E. T. (2017). Hybridization of snout mouth deformed and normal mouth African catfish Clarias gariepinus. Animal Research International, 14(3), 2804-2808.

[9] Awe, E. T., & Akpan, U. U. (2017). Cytological study of Allium cepa and Allium sativum.

[10] Awe, E. T., Akpan, U. U., & Adekoya, K. O. (2017). Evaluation of two MiniSTR loci mutation events in five Father-Mother-Child trios of Yoruba origin. Nigerian Journal of Biotechnology, 33, 120-124.

[11] Barzegar, R., Moghaddam, A. A., Deo, R., Fijani, E., & Tziritis, E. (2018). Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. *Science of the total environment*, *621*, 697-712.

[12] Bello-Dambatta, A. (2010). *The development of a web-based decision support system for the sustainable management of contaminated land*. University of Exeter (United Kingdom).

[13] Bello-Dambatta, A., & Javadi, A. A. (2010). Contaminated land decision support: a review of concepts, methods and systems. *Modelling of Pollutants in Complex Environmental Systems*, *2*, 43.

[14] Binley, A., Hubbard, S. S., Huisman, J. A., Revil, A., Robinson, D. A., Singha, K., & Slater, L. D. (2015). The emergence of hydrogeophysics for improved understanding

of subsurface processes over multiple scales. *Water resources research*, *51*(6), 3837-3866.

[15] Boriana, V. (2017). *Urban Regeneration of Underused Industrial Sites in Albania* (Doctoral dissertation).

[16] Bowen, F., & Wittneben, B. (2011). Carbon accounting: Negotiating accuracy, consistency and certainty across organisational fields. *Accounting, Auditing & Accountability Journal*, *24*(8), 1022-1036.

[17] Buma, B., & Livneh, B. (2017). Key landscape and biotic indicators of watersheds sensitivity to forest disturbance identified using remote sensing and historical hydrography data. *Environmental Research Letters*, *12*(7), 074028.

[18] Burritt, R. L., Schaltegger, S., & Zvezdov, D. (2011). Carbon management accounting: explaining practice in leading German companies. *Australian accounting review*, *21*(1), 80-98.

[19] Cappuyns, V., & Kessen, B. (2014). Combining life cycle analysis, human health and financial risk assessment for the evaluation of contaminated site remediation. *Journal of Environmental Planning and Management*, *57*(7), 1101-1121.

[20] Cheng, F., Geertman, S., Kuffer, M., & Zhan, Q. (2011). An integrative methodology to improve brownfield redevelopment planning in Chinese cities: A case study of Futian, Shenzhen. *Computers, environment and urban systems*, *35*(5), 388-398.

[21] Derycke, V., Coftier, A., Zornig, C., Leprond, H., Scamps, M., & Gilbert, D. (2018). Environmental assessments on schools located on or near former industrial facilities: Feedback on attenuation factors for the prediction of indoor air quality. *Science of the Total Environment*, *626*, 754-761.

[22] Deschaine, L. M. (2014). *Decision support for complex planning challenges* (Doctoral dissertation, Ph. D. Dissertation, Chalmers University of Technology, Göteborg, Sweden, 233p).

[23] Edwards, F. K., Baker, R., Dunbar, M., & Laizé, C. (2012). A review of the processes and effects of droughts and summer floods in rivers and threats due to climate change on current adaptive strategies.

[24] Essaid, H. I., Bekins, B. A., & Cozzarelli, I. M. (2015). Organic contaminant transport and fate in the subsurface: Evolution of knowledge and understanding. *Water Resources Research*, *51*(7), 4861-4902.

[25] Faseemo, O., Massot, J., Essien, N., Healy, W., & Owah, E. (2009, August). Multidisciplinary Approach to Optimising Hydrocarbon Recovery From Conventional Offshore Nigeria: OML100 Case Study. In SPE Nigeria Annual International Conference and Exhibition (pp. SPE-128889). SPE.

[26] Felisa, G. (2015). Dynamics of coastal aquifers: data-driven forecasting and risk analysis.

[27] Ferdinand, A. V., & Yu, D. (2016). Sustainable urban redevelopment: Assessing the impact of third-party rating systems. *Journal of Urban Planning and Development*, *142*(1), 05014033.

[28] Field, C. B. (Ed.). (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

[29] Filippini, M. (2015). Geological and hydrogeological features affecting migration, multi-phase partitioning and degradation of chlorinated hydrocarbons through unconsolidated porous media.

[30] Francisca, F. M., Carro Perez, M. E., Glatstein, D. A., & Montoro, M. A. (2012). Contaminant transport and fluid flow in soils. *Horizons in Earth Research. Nova Science Publishers, New York*, 179-214.

[31] Furniss, M. J. (2011). *Water, climate change, and forests: watershed stewardship for a changing climate*. DIANE Publishing.

[32] Gibassier, D., & Schaltegger, S. (2015). Carbon management accounting and reporting in practice: a case study on converging emergent approaches. *Sustainability Accounting, Management and Policy Journal*, *6*(3), 340-365.

[33] Gober, P., & Kirkwood, C. W. (2010). Vulnerability assessment of climate-induced water shortage in Phoenix. *Proceedings of the*

National Academy of Sciences, *107*(50), 21295-21299.

[34] Green, T. R. (2016). Linking climate change and groundwater. In *Integrated groundwater management: Concepts, approaches and challenges* (pp. 97-141). Cham: Springer International Publishing.

[35] Handmer, J., Honda, Y., Kundzewicz, Z. W., Arnell, N., Benito, G., Hatfield, J., ... & Yamano, H. (2012). Changes in impacts of climate extremes: human systems and ecosystems. *Managing the risks of extreme events and disasters to advance climate change adaptation special report of the intergovernmental panel on climate change*, 231-290.

[36] Hanson, R. T., Flint, L. E., Flint, A. L., Dettinger, M. D., Faunt, C. C., Cayan, D., & Schmid, W. (2012). A method for physically based model analysis of conjunctive use in response to potential climate changes. *Water Resources Research*, *48*(6).

[37] Hardie, S. M. L., & McKinley, I. G. (2014). Fukushima remediation: status and overview of future plans. *Journal of environmental radioactivity*, *133*, 75-85.

[38] Hartmann, F., Perego, P., & Young, A. (2013). Carbon accounting: Challenges for research in management control and performance measurement. *Abacus*, *49*(4), 539-563.

[39] Herat, S., & Agamuthu, P. (2012). E-waste: a problem or an opportunity? Review of issues, challenges and solutions in Asian countries. *Waste Management & Research*, *30*(11), 1113-1129.

[40] Hipsey, M. R., Hamilton, D. P., Hanson, P. C., Carey, C. C., Coletti, J. Z., Read, J. S., ... & Brookes, J. D. (2015). Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. *Water Resources Research*, *51*(9), 7023-7043.

[41] Hoek, G., Beelen, R., & Brunekreef, B. (2011). Methodological issues and statistical analysis in land use regression modeling. *Epidemiology*, *22*(1), S101.

[42] Hou, D., & Al-Tabbaa, A. (2014). Sustainability: A new imperative in contaminated land remediation. *Environmental Science & Policy*, *39*, 25-34.

[43] Hubbard, S. S., Williams, K. H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., ... & Varadharajan, C. (2018). The East River, Colorado, Watershed: A mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. *Vadose Zone Journal*, *17*(1), 1-25.

[44] Ike, P. N., Aifuwa, S. E., Nnabueze, S. B., Olatunde-Thorpe, J., Ogbuefi, E., Oshoba, T. O., & Akokodaripon, D. (2018). Utilizing Nanomaterials in Healthcare Supply Chain Management for Improved Drug Delivery Systems. medicine (Ding et al., 2020; Furtado et al., 2018), 12, 13.

[45] Jago-on, K. A. B., Kaneko, S., Fujikura, R., Fujiwara, A., Imai, T., Matsumoto, T., ... & Taniguchi, M. (2009). Urbanization and subsurface environmental issues: An attempt at DPSIR model application in Asian cities. Science of the total environment, 407(9), 3089-3104.

[46] Jayasooriya, V. M. (2016). *Optimization of green infrastructure practices for industrial areas* (Doctoral dissertation, Victoria University).

[47] Karandish, F., Darzi-Naftchali, A., & Asgari, A. (2017). Application of machine-learning models for diagnosing health hazard of nitrate toxicity in shallow aquifers. *Paddy and Water environment*, *15*(1), 201-215.

[48] Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544-1554.

[49] Kato, S. (2010). *Greenspace conservation planning framework for urban regions based on a forest bird-habitat relationship study and the resilience thinking*. University of Massachusetts Amherst.

[50] Kobus, H., Barczewski, B., & Koschitzky, H. P. (Eds.). (2012). *Groundwater and subsurface remediation: research strategies for in-situ technologies*. Springer Science & Business Media.

[51] Koop, S. H., & van Leeuwen, C. J. (2017). The challenges of water, waste and climate change in cities. *Environment, development and sustainability*, *19*(2), 385-418.

[52] Kresic, N., & Mikszewski, A. (2012). *Hydrogeological conceptual site models: data analysis and visualization*. CRC press.

[53] Kulawiak, M., & Lubniewski, Z. (2014). SafeCity A GIS-based tool profiled for supporting decision making in urban development and infrastructure protection. *Technological Forecasting and Social Change*, *89*, 174-187.

[54] Kuppusamy, S., Palanisami, T., Megharaj, M., Venkateswarlu, K., & Naidu, R. (2016). In-situ remediation approaches for the management of contaminated sites: a comprehensive overview. *Reviews of Environmental Contamination and Toxicology Volume 236*, 1-115.

[55] Langat, P. K., Kumar, L., & Koech, R. (2017). Temporal variability and trends of rainfall and streamflow in Tana River Basin, Kenya. *Sustainability*, *9*(11), 1963.

[56] Leeson, A., Stroo, H., Crane, C., Deeb, R., Kavanaugh, M., Lebron, C., ... & Simpkin, T. (2013). SERDP and ESTCP workshop on long term management of contaminated groundwater sites.

[57] Leibowitz, S. G., Comeleo, R. L., Wigington Jr, P. J., Weaver, C. P., Morefield, P. E., Sproles, E. A., & Ebersole, J. L. (2014). Hydrologic landscape classification evaluates streamflow vulnerability to climate change in Oregon, USA. *Hydrology and Earth System Sciences*, *18*(9), 3367-3392.

[58] Lemming, G. (2010). *Environmental assessment of contaminated site remediation in a life cycle perspective*. Technical University of Denmark.

[59] Levy, L. C. (2013). Chasing fumes: The challenges posed by vapor intrusion. *Nat. Resources & Env't*, *28*, 20.

[60] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674.

[61] Liang, J. (2018). *Development of Physically-Based and Data-Driven Models to Predict Contaminant Loads in Runoff Water From Agricultural Fields*. University of California, Riverside.

[62] Maas, K., Schaltegger, S., & Crutzen, N. (2016). Integrating corporate sustainability assessment, management accounting, control, and reporting. *Journal of cleaner production*, *136*, 237-248.

[63] Majone, M., Verdini, R., Aulenta, F., Rossetti, S., Tandoi, V., Kalogerakis, N., ... & Fava, F. (2015). In situ groundwater and sediment bioremediation: barriers and perspectives at European contaminated sites. *New biotechnology*, *32*(1), 133-146.

[64] Mallants, D., Van Genuchten, M. T., Šimůnek, J., Jacques, D., & Seetharam, S. (2010). Leaching of contaminants to groundwater. In *Dealing with Contaminated Sites: From Theory towards Practical Application* (pp. 787-850). Dordrecht: Springer Netherlands.

[65] Manfreda, S., McCabe, M. F., Miller, P. E., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., ... & Toth, B. (2018). On the use of unmanned aerial systems for environmental monitoring. *Remote sensing*, *10*(4), 641.

[66] Mark, B. G., Bury, J., McKenzie, J. M., French, A., & Baraer, M. (2010). Climate change and tropical Andean glacier recession: Evaluating hydrologic changes and livelihood vulnerability in the Cordillera Blanca, Peru. *Annals of the Association of American geographers*, *100*(4), 794-805.

[67] McAlary, T. A., Provoost, J., & Dawson, H. E. (2010). Vapor intrusion. In *Dealing with Contaminated Sites: From Theory towards Practical Application* (pp. 409-453). Dordrecht: Springer Netherlands.

[68] McGrath, R., Reid, R., & Tran, P. (2017, January). EPRI Report: Review of Geostatistical Approaches to Characterization of Subsurface Contamination-17442. In *43rd Annual Waste Management Conference (WM2017)*.

[69] McMillan, H., Montanari, A., Cudennec, C., Savenije, H., Kreibich, H., Krueger, T., ... & Xia, J. (2016). Panta Rhei 2013–2015: global perspectives on hydrology, society and change. *Hydrological Sciences Journal*, *61*(7), 1174-1191.

[70] Meerow, S., & Newell, J. P. (2017). Spatial planning for multifunctional green infrastructure: Growing resilience in Detroit. *Landscape and urban planning*, *159*, 62-75.

[71] Mgbeahuruike, L. U. (2018). *An investigation into soil pollution and remediation of selected polluted sites around the globe* (Doctoral dissertation, Manchester Metropolitan University).

[72] Mitchell, M. (2012). *Long-Term Monitoring and Maintenance Plan for the Mixed Waste Landfill March 2012* (No. SAND2012-1957P). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

[73] Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environmental monitoring and assessment*, *188*(1), 44.

[74] Nashwan, M. S., Shahid, S., Chung, E. S., Ahmed, K., & Song, Y. H. (2018). Development of climate-based index for hydrologic hazard susceptibility. *Sustainability*, *10*(7), 2182.

[75] Nelitz, M., Boardley, S., & Smith, R. (2013). Tools for climate change vulnerability assessments for watersheds. *Prepared by ESSA Technologies Ltd. for the Canadian Council of Ministers of the Environment*.

[76] Oni, O., Adeshina, Y. T., Iloeje, K. F., & Olatunji, O. O. (2018). Artificial Intelligence Model Fairness Auditor For Loan Systems. Journal ID, 8993, 1162.

[77] Osabuohien, F. O. (2017). Review of the environmental impact of polymer degradation. Communication in Physical Sciences, 2(1).

[78] Park, Y., Ligaray, M., Kim, Y. M., Kim, J. H., Cho, K. H., & Sthiannopkao, S. (2016). Development of enhanced groundwater arsenic prediction model using machine learning approaches in Southeast Asian countries. *Desalination and Water Treatment*, *57*(26), 12227-12236.

[79] Parker, B. L., Cherry, J. A., & Chapman, S. W. (2012). Discrete fracture network approach for studying contamination in fractured rock. AQUA mundi, 3(2), 101-116.

[80] Perra, E., Piras, M., Deidda, R., Paniconi, C., Mascaro, G., Vivoni, E. R., ... & Meyer, S. (2018). Multimodel assessment of climate change-induced hydrologic impacts for a Mediterranean catchment. *Hydrology and Earth System Sciences*, *22*(7), 4125-4143.

[81] Phenrat, T., Otwong, A., Chantharit, A., & Lowry, G. V. (2016). Ten-year monitored natural recovery of lead-contaminated mine tailing in Klity Creek, Kanchanaburi Province, Thailand. Environmental Health Perspectives, 124(10), 1511.

[82] Provoost, J., Tillman, F., Weaver, J., Reijnders, L., Bronders, J., Van Keer, I., & Swartjes, F. (2013). Vapour intrusion into buildings–aliterature review. *Soil contamination and indoor air quality*, *15*.

[83] Ransom, K. M., Nolan, B. T., Traum, J. A., Faunt, C. C., Bell, A. M., Gronberg, J. A. M., ... & Harter, T. (2017). A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA. *Science of the Total Environment*, *601*, 1160-1172.

[84] Ribeiro Neto, A., Scott, C. A., Lima, E. A., Montenegro, S. M. G. L., & Cirilo, J. A. (2014). Infrastructure sufficiency in meeting water demand under climate-induced socio-hydrological transition in the urbanizing Capibaribe River basin–Brazil. *Hydrology and Earth System Sciences*, *18*(9), 3449-3459.

[85] Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Science of the Total Environment*, *476*, 189-206.

[86] Roghani, M. (2018). Investigation of Volatile Organic Compounds (VOCs) Detected at Vapor Intrusion sites.

[87] Sayles, L. R. (2017). *Managing large systems: organizations for the future*. Routledge.

[88] Schaltegger, S., & Csutora, M. (2012). Carbon accounting for sustainability and management.

Status quo and challenges. *Journal of cleaner production*, *36*, 1-16.

[89] Scheidt, C., Li, L., & Caers, J. (Eds.). (2018). *Quantifying uncertainty in subsurface systems*. John Wiley & Sons.

[90] Schultz, G. A., & Engman, E. T. (Eds.). (2012). *Remote sensing in hydrology and water management*. Springer Science & Business Media.

[91] Sims, N. C., & Colloff, M. J. (2012). Remote sensing of vegetation responses to flooding of a semi-arid floodplain: Implications for monitoring ecological effects of environmental flows. *Ecological Indicators*, *18*, 387-391.

[92] Singh, K. P., Gupta, S., & Mohan, D. (2014). Evaluating influences of seasonal variations and anthropogenic activities on alluvial groundwater hydrochemistry using ensemble learning approaches. *Journal of Hydrology*, *511*, 254-266.

[93] Singh, R., van Werkhoven, K., & Wagener, T. (2014). Hydrological impacts of climate change in gauged and ungauged watersheds of the Olifants basin: a trading-space-for-time approach. *Hydrological Sciences Journal*, *59*(1), 29-55.

[94] Sorooshian, S., Nguyen, P., Sellars, S., Braithwaite, D., AghaKouchak, A., & Hsu, K. (2014). Satellite-based remote sensing estimation of precipitation for early warning systems. *Extreme natural hazards, disaster risks and societal implications*, *1*, 99.

[95] Steininger, K. W., Lininger, C., Meyer, L. H., Muñoz, P., & Schinko, T. (2016). Multiple carbon accounting to support just and effective climate policies. *Nature Climate Change*, *6*(1), 35-41.

[96] Sweeney, M. W., & Kabouris, J. C. (2017). Modeling, instrumentation, automation, and optimization of water resource recovery facilities. *Water Environment Research*, *89*(10), 1299-1314.

[97] Tang, Q., & Luo, L. (2014). Carbon management systems and carbon mitigation. *Australian Accounting Review*, *24*(1), 84-98.

[98] Thakur, J. K., Singh, S. K., & Ekanthalu, V. S. (2017). Integrating remote sensing, geographic information systems and global positioning system techniques with hydrological modeling. *Applied Water Science*, *7*(4), 1595-1608.

[99] Turczynowicz, L., Pisaniello, D., & Williamson, T. (2012). Health risk assessment and vapor intrusion: A review and Australian perspective. *Human and Ecological Risk Assessment: An International Journal*, *18*(5), 984-1013.

[100] Viviroli, D., Archer, D. R., Buytaert, W., Fowler, H. J., Greenwood, G. B., Hamlet, A. F., ... & Woods, R. (2011). Climate change and mountain water resources: overview and recommendations for research, management and policy. *Hydrology and Earth System Sciences*, *15*(2), 471-504.

[101] Wagesho, N. (2014). Catchment dynamics and its impact on runoff generation: coupling watershed modelling and statistical analysis to detect catchment responses. *International Journal of Water Resources and Environmental Engineering*, *6*(2), 73-87.

[102] Wang, H., Cai, Y., Tan, Q., & Zeng, Y. (2017). Evaluation of groundwater remediation technologies based on fuzzy multi-criteria decision analysis approaches. *Water*, *9*(6), 443.

[103] Wang, X., Unger, A. J., & Parker, B. L. (2014). Risk-Based Characterization for Vapour Intrusion at a Conceptual Brownfields Site: Part 2. Pricing the Risk Capital. *Journal of Civil Engineering*, *3*(4), 189-208.

[104] Watts, G., Battarbee, R. W., Bloomfield, J. P., Crossman, J., Daccache, A., Durance, I., ... & Wilby, R. L. (2015). Climate change and water in the UK–past changes and future prospects. *Progress in Physical Geography*, *39*(1), 6-28.

[105] Williamson, M. (2011). Advanced simulation capability for environmental management (ASCEM): An overview of.

[106] Williamson, M., Meza, J., Moulton, D., Gorton, I., Freshley, M., Dixon, P., ... & Collazo, Y. T. (2011). Advanced simulation capability for environmental management (ASCEM): an overview of initial results. *Technology & Innovation*, *13*(2), 175-199.

[107] Yaron, B., Dror, I., & Berkowitz, B. (2012). *Soil-subsurface change: chemical*

pollutant impacts. Springer Science & Business Media.

[108] Zeidan, B. A. (2017). Groundwater degradation and remediation in the Nile Delta Aquifer. In *The Nile Delta* (pp. 159-232). Cham: Springer International Publishing.

[109] Zhai, X., Yue, P., & Zhang, M. (2016). A sensor web and web service-based approach for active hydrological disaster monitoring. *ISPRS International Journal of Geo-Information*, *5*(10), 171.

[110] Zhang, Y., Peng, C., Li, W., Fang, X., Zhang, T., Zhu, Q., ... & Zhao, P. (2013). Monitoring and estimating drought-induced impacts on forest structure, growth, function, and ecosystem services using remote-sensing data: recent progress and future challenges. *Environmental Reviews*, *21*(2), 103-115.