Journal homepage: http://gwse.iheg.org.cn

Research Article

SDPHC: A software tool for detecting petroleum hydrocarbons contamination in active industrial sites using non-invasive survey (NIS) dataset

Yong-jian Gu[1], Qian-kun Luo[1*], Min Zhang[2,3*], Zhuo Ning[2,3], Lin Sun[2,3], Lei Ma[1], Hai-chun Ma[1]

[1] *School of Resources and Environmental Engineering, Hefei University of Technology, Hefei* 230009, *China*.
[2] *Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences, Shijiazhuang* 050061, *China*.
[3] *Key Laboratory of Groundwater Remediation of Hebei Province & China Geological Survey, Shijiazhuang* 050061, *China*.

Abstract: Conventional drilling-based methods for investigating petroleum hydrocarbons (PHCs) contamination in industrial parks are time-consuming, labor-intensive, and disruptive, making them often unsuitable for active industrial sites. Non-invasive survey (NIS) technology has emerged as a promising alternative owing to its cost-effectiveness and minimal environmental disturbance. To enhance the efficiency of NIS-based contamination surveys in active industrial sites and facilitate widespread adoption, this study developed a software tool for detecting petroleum hydrocarbons contamination (SDPHC). SDPHC integrates Python's scientific computing ecosystem with the PySide6 desktop graphical user interface (GUI) framework, achieving a scalable architecture for rapid development. The software provides three complementary analytical methods: empirical threshold analysis (ETA), background level analysis (BLA), and principal component analysis (PCA). Each method is tailored to distinct data scenarios: ETA leverages field-validated thresholds for sites with comprehensive NIS datasets; BLA quantifies site-specific natural baselines for individual indicators to distinguish anthropogenic contamination; and PCA identifies multivariate spatial patterns from correlated soil gas variables (e.g., CO_2, O_2, CH_4), enabling robust contamination zoning even when radon or functional gene data are absent. This modular design allows users to select or combine methods based on data availability and site characteristics. Additionally, SDPHC automates report generation to enhance survey efficiency. Two case studies conducted at active petrochemical parks demonstrate the software's applicability and reliability. SDPHC is anticipated to function as a reliable and powerful tool for conducting NIS-based contamination assessments in industrial parks.

Keywords: Contamination survey; Natural source zone depletion (NSZD); VOCs; Soil gases; Radon; Functional genes

Received: 19 Mar 2025/ Accepted: 26 Oct 2025/ Published: 05 Dec 2025

Introduction

Petroleum hydrocarbons (PHCs) are vital energy resources and industrial raw materials globally (EI 2024). The leakage of PHCs has led to widespread contamination by light non-aqueous phase liquids (LNAPLs), which are currently among the predominant forms of soil and groundwater contamination (Tomlinson et al. 2017; Sookhak Lari et al. 2019b). PHCs, such as gasoline and diesel, have densities lower than water and exhibit low solubility, enabling them to persist as contamination sources in the vadose zone. These contaminants present long-term risks to ecosystems and human health via processes such as diffusion, migration, and volatilization (García-Rincón et al. 2024). Given that LNAPLs often form complex, heterogeneous plumes below the water table, the initial and most crucial step in any remediation effort is to precisely locate and characterize the

*Corresponding author: Qian-kun Luo, Min Zhang, *E-mail address*: qkluo@hfut.edu.cn, minzhang205@live.cn

DOI: 10.26599/JGSE.2026.9280066

Gu YJ, Luo QK, Zhang M, et al. 2025. SDPHC: A software tool for detecting petroleum hydrocarbons contamination in active industrial sites using non-invasive survey (NIS) dataset. Journal of Groundwater Science and Engineering, 13(0): 1-14.

extent of the contaminant body, enabling targeted intervention and preventing further migration.

In current groundwater contamination characterization studies, data acquisition primarily relies on intrusive methods, with direct observations from groundwater samples collected via boreholes serving as the main data source. Corresponding identification approaches are predominantly based on simulation-optimization (S/O) frameworks (Anshuman and Eldho, 2022; Zhu et al., 2024), data assimilation techniques (Zhang et al., 2016; Jiang et al., 2018), and inverse modeling integrated with machine learning algorithms (Pan et al., 2023; Wang et al., 2024). While drilling provides precise measurements, this approach is labor-intensive, costly, and highly disruptive to normal operations. In high-risk industrial parks, drilling activities may potentially cause equipment damage or explosions (Tsai et al. 2020). These limitations limit the effectiveness of conventional drilling-based methods in addressing complex contamination scenarios at active industrial parks.

To address these limitations, non-invasive survey (NIS) technology has emerged as an alternative approach for investigating PHCs contamination, particularly in operational industrial settings where minimizing site disruption is critical. Research utilizing non-invasive data for contamination characterization has been largely dominated by geophysical methods, including electrical resistivity tomography (ERT) (Shao et al., 2019; Xia et al., 2021), ground-penetrating radar (GPR) (Cassidy, 2008; Bertolla et al., 2014) and various other individual or multi-method integrated approaches (Arato et al., 2014; Tsai et al., 2020; Bonnière et al., 2024; Baessa et al., 2024). In contrast, this study employs a distinct NIS framework that integrates geochemical (e.g., CO_2, CH_4, O_2), radiological (^{222}Rn flux), and molecular biological (e.g., *AlkB*, *prmA*) indicators to detect and delineate PHCs contamination. Data for this approach are acquired within the top 0.6 m of the soil profile, rendering the technique minimally invasive and suitable for repeated monitoring under ongoing industrial activities.

The principle of this multi-parameter NIS methodology is closely linked to natural attenuation mechanisms, specifically natural source zone depletion (NSZD) (Verginelli et al. 2024). During the natural attenuation of PHCs leaks, a series of biogeochemical reactions occur. Under aerobic conditions, PHCs are progressively degraded into CO_2 and H_2O via microbial metabolism, such as alkane monooxygenase-catalyzed reactions. Under anaerobic conditions, CH_4 is generated through methanogenic pathways. These degradation products exhibit distinct geochemical gradient characteristics. For instance, CO_2 concentration gradients indicate the intensity of microbial activity, CH_4 accumulation and migration pathways reflect spatial distributions of anaerobic degradation zones, and O_2 consumption rates characterize variations in redox conditions (Garg et al. 2017; Sookhak Lari et al. 2019a; Kulkarni et al. 2022; Smith et al. 2022). Additionally, volatile intermediate degradation products, such as low-molecular-weight organic acids, benzene, toluene, ethylbenzene, and xylenes (BTEX) compounds, enter the soil gas phase via diffusion, thereby forming detectable volatile organic compounds (VOCs) anomaly zones. Radon, naturally emitted from subsurface strata, is absorbed by LNAPL, resulting in differences in background values between contaminated and non-contaminated zones (Davis et al. 2005; Schubert et al. 2007; Schubert 2015; De Simone et al. 2017; Cecconi et al. 2022). Quantitative polymerase chain reaction (qPCR) analysis of functional genes, such as *AlkB*, enables estimation of contaminant degradation rates, facilitating the reconstruction of active contamination zones and revealing spatial relationships between source areas and plumes (Táncsics et al. 2012; Han et al. 2014; Shahi et al. 2016; Fernández-Baca et al. 2018). The NIS technology employs a multi-parameter synergistic monitoring system based on the NSZD mechanism to precisely identify the PHCs contaminant sources and plume ranges in industrial parks without requiring classical drilling. It offers a high-resolution and cost-effective alternative to traditional drilling-based methods for contaminant assessment in active industrial parks.

However, the multi-source datasets generated by NIS technology display substantial heterogeneity. For instance, soil gas data typically comprise discrete concentration values, biogenetic data are presented as abundance matrices, and radon measurements yield radioactive signal intensities. These differences in data formats and response mechanisms prevent the direct integration of existing analytical tools, thereby limiting the precise localization of contamination source zones and plume boundaries. Currently, only a limited number of studies have systematically addressed the standardized processing and synergistic analysis of multi-source NIS datasets, hindering the full utilization of non-invasive technologies and restricting the rapid and widespread adoption of NIS-based investigations for PHCs contamination assessments in active petrochemical parks.

To address these challenges, this study developed SDPHC, a dedicated NIS-based tool for PHCs contamination investigations. The software integrates multi-dimensional indicators derived from NIS datasets, including VOCs, soil gases, radon, and functional genes. By leveraging Python's scientific computing ecosystem and the PySide6 graphical user interface (GUI) framework, SDPHC combines empirical threshold analysis (ETA), background level analysis (BLA), and principal component analysis (PCA) to simultaneously localize source zones and delineate plume boundaries. Additionally, the software incorporates a feature for automated, standardized report generation, which significantly enhances investigation efficiency for contamination assessments in active industrial sites and facilitates large-scale deployment.

The following sections are organized as follows: Section 1 elaborates on the design of the SDPHC software, including implementation framework, architecture, and operational workflows. Section 2 demonstrates the software's applicability and reliability through two real-world PHCs contamination case studies in active industrial parks. Section 3 provides concluding remarks, discusses limitations, and outlines future research directions.

1 Software Design and Workflow

1.1 Implementation framework

1.1.1 Python

As one of the most renowned open-source programming languages in scientific research (IEEE spectrum, 2024), Python has become a foundational technology platform for computational science. Its active developer community and comprehensive scientific computing ecosystem, such as NumPy and SciPy, empower researchers to efficiently leverage well-established modules for building complex models. This significantly enhances development efficiency for scientific and application-oriented software compared to traditional static languages (Crudu 2024). Python's combination of user-friendliness and functional completeness provides distinct advantages in interdisciplinary research. For instance, standardized technical frameworks based on Python have been developed in areas such as geospatial analysis (Chen et al. 2022), ecological modeling, and bioinformatics (Brendel et al. 2020; Chen et al. 2024; López-Fernández et al. 2024). Notably, Python has become the predominant technical platform for data-driven research in environmental sciences (Kadiyala and Kumar 2017).

To facilitate rapid development and deployment, SDPHC utilizes the Python programming language combined with its extensive scientific ecosystem. The incorporation of numerous open-source Python libraries greatly streamlines development by allowing researchers to focus on core algorithm design, while peripheral functionalities are directly implemented through existing library modules. As shown in Table 1, SDPHC systematically integrates Python's comprehensive open-source toolchain to implement core software functionalities.

1.1.2 PySide6

PySide6 (Qt for Python, https://doc.qt.io/qtforpython-6/index.html), the official Python binding for the Qt 6 framework, inherits Qt's extensive widget library and native interface style, thereby providing a robust foundation for the development of high-performance desktop applications. As illustrated in Fig. 1, PySide6's core advantage lies in the integration of the computational efficiency of Qt C++ kernel with Python's syntactic simplicity. This integration supports complex data interaction and interface rendering requirements while substantially enhancing code maintainability. Furthermore, PySide6's deep compatibility with Python's scientific computing ecosystem, exemplified by Matplotlib's customized *NavigationTool-*

Table 1 A list of Python packages during the SDPHC implementation

Category	Library name	Description of the functions	Reference
Data processing	Geopandas	Geospatial data processing and analysis	(Jordahl et al. 2020)
	Shapely	Geometric object manipulation	(Shapely, 2025)
	Scikit-learn	Data standardization, principal component analysis	(Pedregosa et al., 2011)
	Scipy	Data analysis and interpolation	(Virtanen et al., 2020)
Visualization	Matplotlib	Graphics drawing and saving, provision of GUI widgets	(Hunter, 2007)
	Seaborn	Graphic drawing and beautification	(Waskom, 2021)
Report	Python-docx	Word document generation	(python-docx, 2025)
	Pillow	Image format conversion/cropping/filter operations	(Clark & Others, 2015)

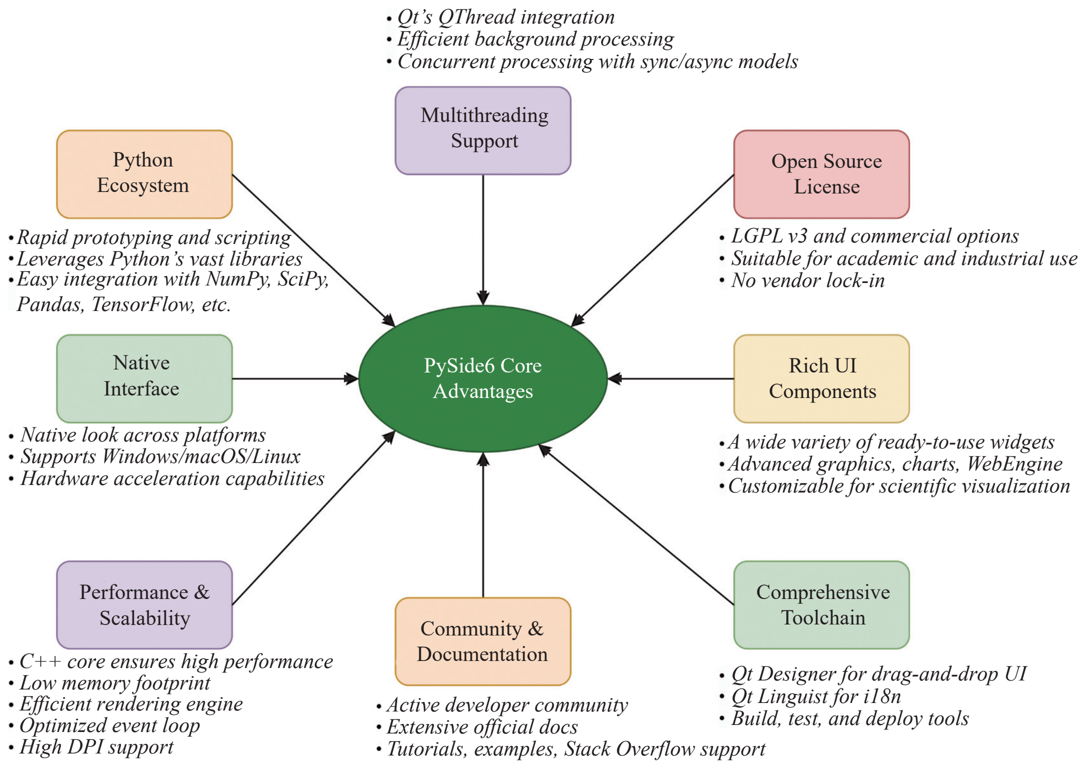

Fig. 1 Core advantages of PySide6 framework

bar2QT and *FigureCanvasQTAgg* modules, enables the seamless integration of data preprocessing, model analysis, and visualization results within a unified interface, thereby substantially reducing user operational barriers. Additionally, PySide6's comprehensive toolchains, such as Qt Linguist for multilingual support, enhance development efficiency through its modular architecture design, thereby offering flexible technical support for software iteration. SDPHC utilizes PySide6 to build a GUI system that combines high-performance rendering with efficient data processing, achieving end-to-end interactive workflows from data preprocessing to visualization while providing an extensible framework for future functional upgrades.

1.2 Software architecture

SDPHC adopts a modular architecture to balance development complexity and extensibility. As illustrated in Fig. 2, the software is structured into six core modules: *Entry file*, *GUIModule*, *CoreModule*, *UtilsModule*, *AssetsModule* and *ReportModule*, each designed for specific functional roles while maintaining intermodular independence.

1.2.1 Entry file (Main.py)

Serves as the software's entry point, initializing system settings such as localization, screen scaling, and user interface loading. It manages interactions between the GUI and analytical modules.

1.2.2 GUIModule

Implements the PySide6-based graphical user interface, with independent GUI components for each analytical function (e.g., *empirical_threshold_analysis.py*). Shared GUI elements and custom controls are centralized in *custom_controls.py* to ensure consistency and reduce redundancy across the interface.

1.2.3 CoreModule

This module constitutes the computational backbone of SDPHC, encompassing functionalities for preprocessing, analysis, result storage, and selective visualization of NIS datasets. Each analytical method is implemented in standalone files (e.g., *empirical_threshold_functions.py*) to enhance modularity, maintainability and future expandability.

1.2.4 UtilsModule

A general-purpose utility module providing a collection of shared functions and data structures used across SDPHC. It includes tools for data manipulation, file handling, and configuration management. These utilities help support modularity and code reuse, ensuring consistency and efficiency in development and maintenance.

1.2.5 AssetsModule

Manages static resources including vector icons (.ico) and multilingual translation files (.ts), which are compiled into Qt resource archives (.qrc) via

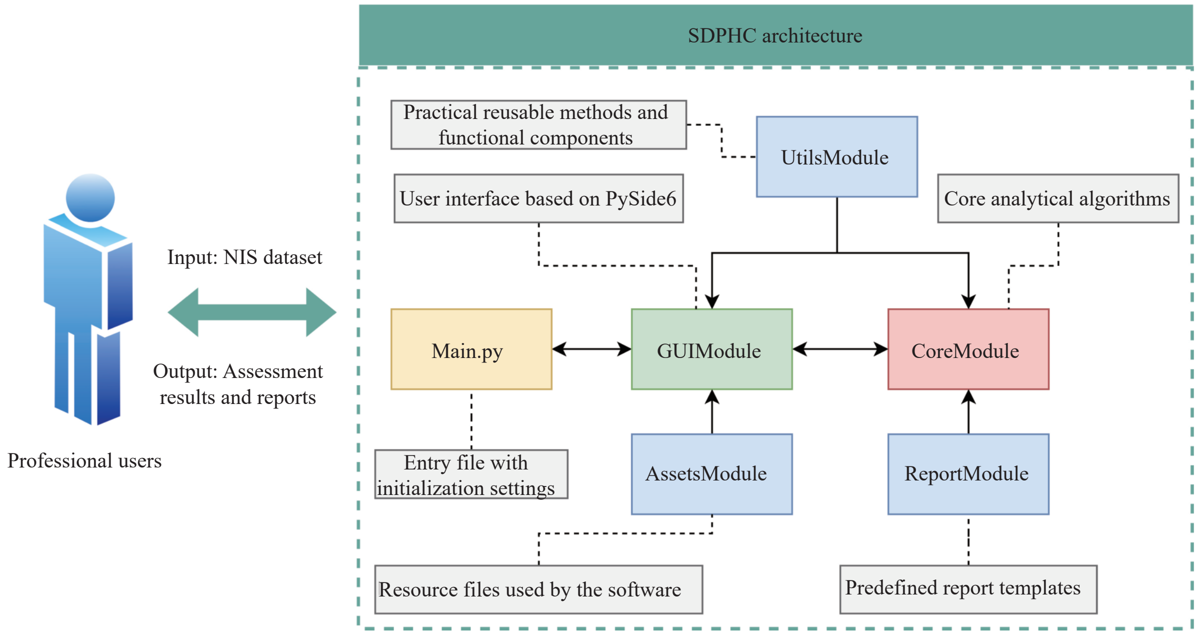

Fig. 2 Diagram of software architecture

PySide6's resource compiler (*pyside6-rcc*).

1.2.6 ReportModule

Enables automated report generation by integrating predefined templates with dynamic analysis outputs. Users can customize templates while leveraging built-in functions to insert formatted data, figures, and annotations (e.g., controlling image size, captions).

1.3 Operational Workflow

The functionalities of SDPHC are systematically introduced through a step-by-step workflow, with corresponding user interfaces illustrated in Fig. 3. As depicted in Fig. 4, the operational process consists of four sequential stages that guide users from data import to final report generation.

Step1. Prepare the NIS dataset and select analysis methods

NIS monitoring data must be preprocessed into standard geospatial formats, such as GeoPackage and Shapefile, using GIS platforms to ensure compatibility with SDPHC. The attribute table must include comprehensive fields for NIS monitoring indicators, which are currently supported by

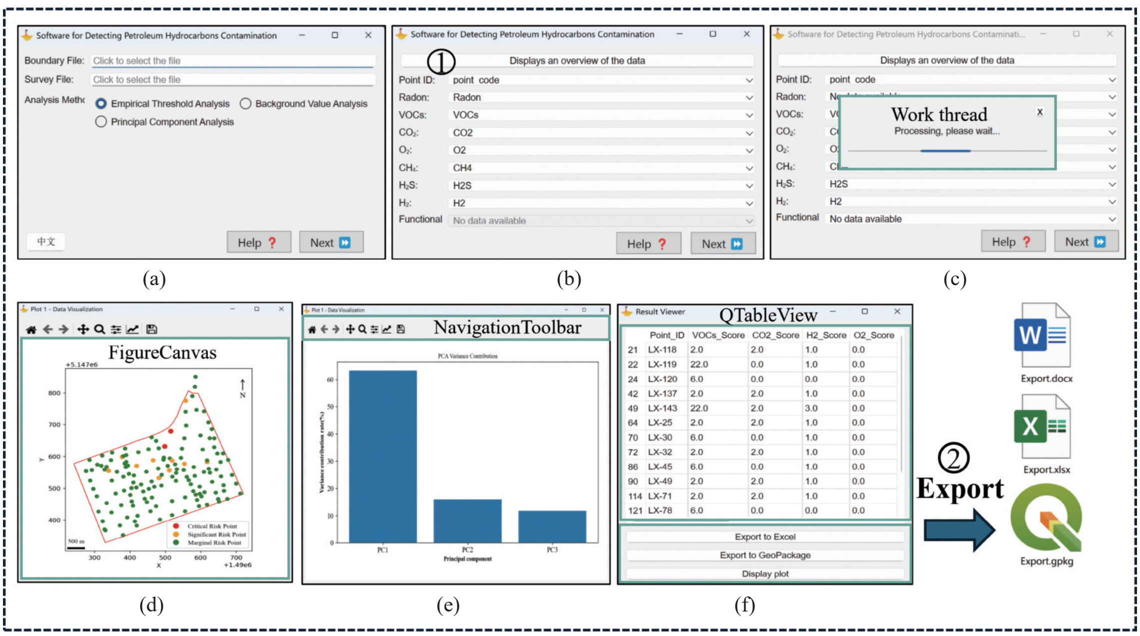

Fig. 3 Software graphical user interface (GUI)

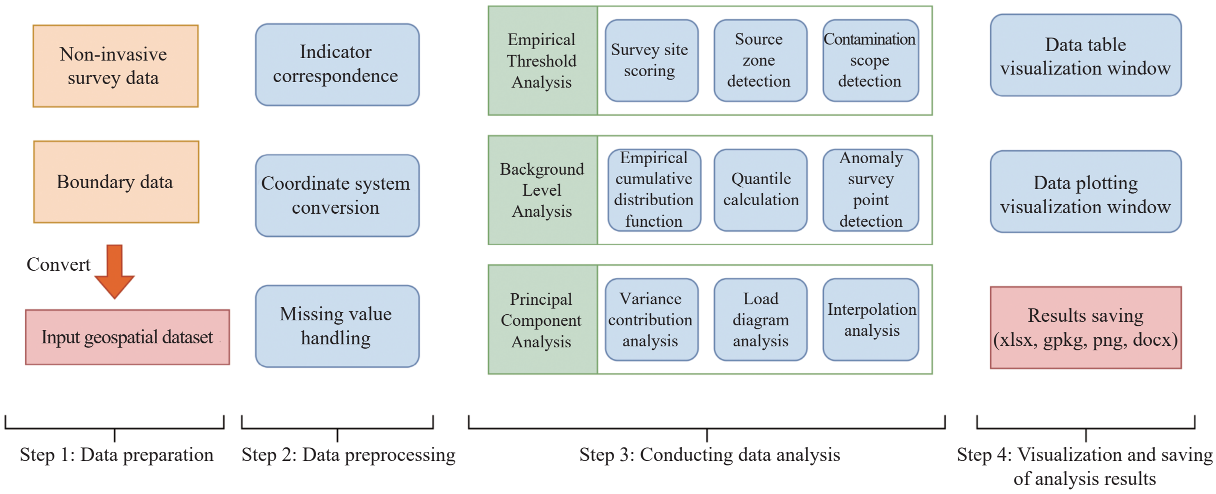

Fig. 4 Operational flow diagram of SDPHC: a four-stage pipeline for NIS data analysis

the software listed in Table 2. These indicators encompass VOCs, soil gases concentrations (CO_2, O_2, CH_4), radon flux, and functional gene abundances (e.g., *AlkB*, *C12O*), each serving as a critical proxy for PHCs contamination dynamics. Detailed information regarding the analytical methods used for these measurements is provided in Appendix A.

Among the three analytical methods currently implemented in SDPHC, ETA distinguishes risk levels of sampling points based on historical NIS survey experience, enabling the delineation of pollution source zones and inferring plume boundaries. This method is particularly suitable for sites with comprehensive datasets. For example, when combined with soil gas and radon surveys, it provides accurate contamination mapping in PHCs-impacted areas. BLA quantifies natural baseline fluctuations and anthropogenic contaminant thresholds, thereby providing a foundation for precise contamination identification and supporting the development of environmental management strategies (Preziosi et al. 2014; Molinari et al. 2019; Azzellino et al. 2019). Given the limited research on background values for PHCs-contaminated sites, SDPHC integrates statistical methods, such as empirical cumulative distribution function (ECDF) and quantiles, with k-means clustering to mitigate misclassification risks caused by skewed data distributions, such as outliers or insufficient sampling density. PCA requires high dataset completeness. For sites where only key soil gas indicators were collected (excluding radon or functional gene data), PCA still provides robust spatial analysis capabilities (Lucas and Jauzein 2008). Users can select individual or combined methods according to site-specific characteristics and data quality.

As illustrated in Fig. 3a, data input is facilitated through interactive input fields that trigger the system file explorer for dataset uploading. Subsequently, users select analytical methods via predefined radio buttons, enabling seamless integration of user preferences with the software's computational workflows.

Step2. Validate and preprocess loaded data

Users can click the “Display an overview of the datasets” button to generate point-boundary over-

Table 2 NIS indicators supported by SDPHC and their relationship to PHCs contamination

Indicator Name	Common Unit	Scientific Significance in PHCs Assessment
VOCs	ppb	Indicators of petroleum leakage; elevated levels suggest direct contamination sources.
CO_2	ppm	Byproduct of aerobic microbial degradation; high concentrations reflect active biodegradation.
O_2	%	Depletion indicates anaerobic conditions linked to hydrocarbon degradation processes.
CH_4	%	Produced under anaerobic conditions; signals methanogenic activity in contaminated zones.
H_2S	mg/L	Anaerobic metabolic byproduct; associated with sulfate reduction in LNAPL-affected areas.
H_2	mg/L	Intermediate in hydrogen fermentation; reflects anaerobic metabolic activity.
Radon	Bq/m³	Adsorbed by LNAPLs; reduced concentrations highlight contaminated subsurface regions.
Functional Genes	copies/g soil	Microbial genetic markers reflecting hydrocarbon degradation pathways and contamination dynamics.

lay maps, thereby verifying spatial topological integrity and confirming accurate data recognition (Fig. 3 marker 1). For NIS datasets, users must match predefined indicators with input dataset attributes, labeling missing fields as "No data available". Coordinate system conversion is performed to ensure visualization accuracy.

As illustrated in Fig. 3b, the supported NIS indicators for each analytical method are dynamically listed in the left panel, while users can select corresponding indicators from a context-aware combo box on the right side based on input dataset attributes. This interactive design ensures method-specific parameter alignment, preventing mismatches between analytical workflows and available data fields.

Step3. Execute analysis workflows and generate plots

Due to computational demands associated with large datasets and complex calculations, this stage requires a substantial amount of processing time. To ensure the stability of the GUI, analytical operations run independently from the GUI thread via Qt's concurrency framework (Fig. 3c). Although specific workflows differ depending on the method, PCA serves as a representative example of the process. As shown in Fig. 4 step3, it starts with handling missing values and standardizing data across all input datasets, followed by PCA computation for extracting eigenvalues and eigenvectors. Variance contribution plots and loading plots are subsequently generated to visualize the relationships among dimensions, while spatial interpolation of principal components enables the analysis of contaminant patterns. SDPHC integrates four interpolation methods (nearest neighbor, cubic spline, inverse distance weighting [IDW] and kriging) to accommodate diverse data distributions. Kriging quantifies spatial autocorrelation via semivariograms for heterogeneous datasets, while IDW performs distance-weighted interpolation for regions with localized variability. All analytical outputs, including data tables and plot objects, are encapsulated in Python dictionary structures before being transmitted to the GUI thread via signal-slot mechanisms for visualization and export. This approach ensures the seamless integration of computational results with interactive display functionalities while preserving system responsiveness during intensive processing tasks.

Step4. Visualize and save results

SDPHC currently supports two visualization components: plotting windows and table views. Plotting windows utilize *FigureCanvas* and *NavigationToolbar* from Matplotlib for interactive visualization of analysis outputs (Fig. 3d-e). Table views are implemented using PySide6's *QTableView*, enabling visualization of analysis results (e.g., risk scores) with bidirectional scrolling and selection capabilities to accommodate large-scale datasets effectively (Fig. 3f). Data export integrates with operating system file explorers, enabling tables to be saved as table files (.xlsx) or geospatial vector files (.gpkg), while graphical outputs are locally stored via toolbar-based operations (Fig. 3 marker 2). For established analytical methods, the software offers customizable report generation based on predefined templates. This feature combines fixed modules (e.g., project information, methodology descriptions) with data-driven components that automatically insert statistical tables and contour maps derived from survey data, thereby significantly reducing repetitive tasks for environmental engineers while enhancing operational efficiency.

1.4 Software Availability and System Requirements

The SDPHC software is implemented in Python 3.13 and currently supports Windows 10 or later operating systems. A complete list of dependencies can be found in the project's *requirements.yml* file. Pre-compiled binary versions are provided for users who prefer to use the software without setting up a development environment.

To promote transparency, reproducibility, and community collaboration, the full source code, detailed installation instructions, and deployment documentation are publicly available on GitHub under the Apache License 2.0 at https://github.com/ArtCoff/SDPHC. Users can clone the repository, create a dedicated Python environment via conda or pip, and run the application by executing main.py in the app directory. Alternatively, precompiled binary versions are available for direct download and execution.

2 Case studies

In this section, the applicability and reliability of SDPHC for contamination detection in active petrochemical parks are demonstrated through two real-world case studies of PHCs contamination investigations.

2.1 An active petrochemical park

The study area is an active petrochemical park in northeastern China, covering approximately 120,000 m^2. Its primary products include phenol and acetone. The facility layout is illustrated in Fig. 5a: the northern section houses critical production units (e.g., distillation and oxidation workshops), while the southern area accommodates storage facilities, electrical systems, and fire safety infrastructure. To systematically assess potential PHCs contamination, a total of 146 NIS monitoring points were established based on a comprehensive strategy integrating the regional groundwater flow direction (from southwest to northeast), the functional characteristics of production facilities, and preliminary contamination indications. A systematic grid-based sampling approach was adopted, with a density of at least one point per 1,600 m^2, supplemented by targeted densification around initial high-concentration spots of VOCs and CO_2. The collected dataset comprises 144 soil gas samples, 29 functional gene survey points, and 15 radon measurement points, all spatially referenced in Fig. 5a. This comprehensive NIS dataset was subsequently analyzed using the SDPHC software to evaluate the spatial distribution and extent of PHCs contamination within the site.

The park's comprehensive dataset and representative contamination profile warranted the application of ETA. As shown in Fig. 5b, contamination source zones and potential source areas were manually delineated based on anomaly levels at survey points. Key leakage zones were identified at the loading platform and hazardous waste room (red-delineated areas), while orange-delineated zones were designated as priority monitoring areas. Fig. 5c presents the contamination extent evaluated through anomaly scoring across all survey points (red markers indicate anomalies). Results demonstrated widespread contamination throughout the entire facility, with extrapolated risks of off-site migration detected beyond the southwest boundary. The output of ETA was integrated into the park's contamination management practices, while the automated report generation module produced detailed documentation (see Appendix B).

To further investigate the contamination status and guide remediation management, BLA was applied to determine site-specific background values for the contaminated industrial park. By integrating ECDF curves with percentile statistics of NIS indicators (Fig. 6), this approach identified distinct contamination patterns. For VOCs (> 2600

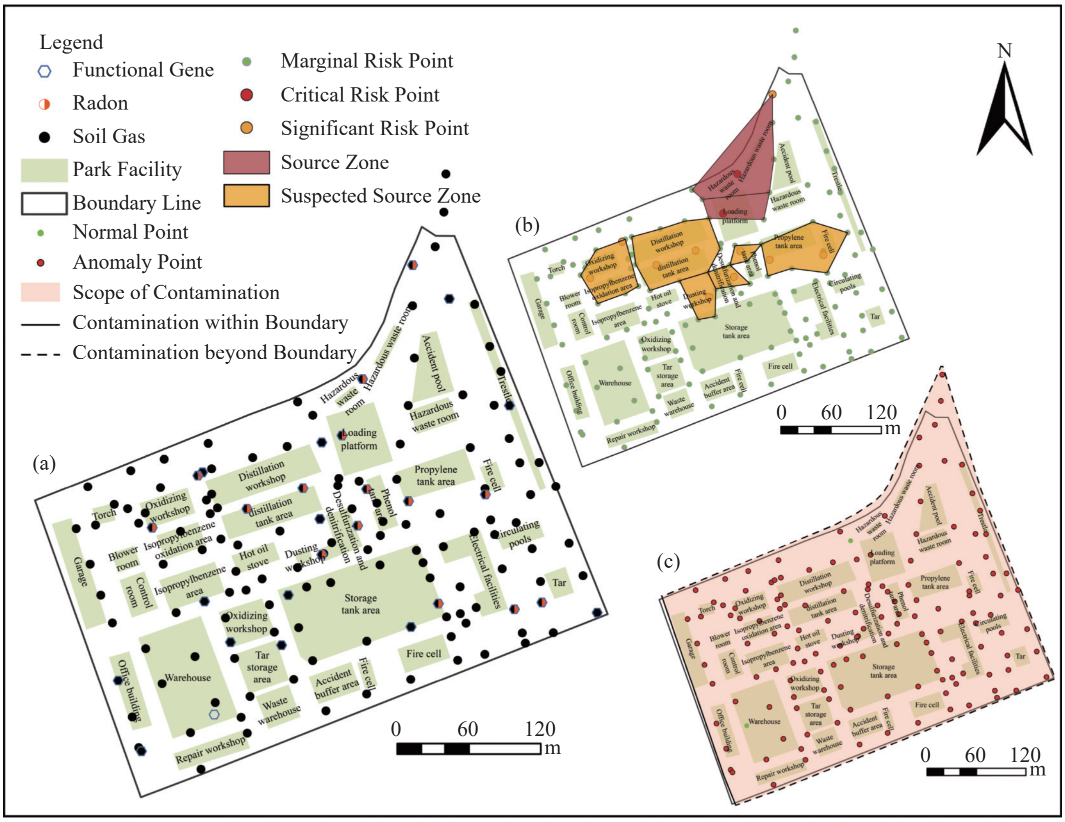

Fig. 5 Overview of the NIS dataset and ETA results in the contaminated petrochemical park (a) Distribution of NIS survey points and data types; (b) Delineation of contaminated source zones; (c) Delineation of contaminated areas; Contamination assessment results are derived from SDPHC analysis

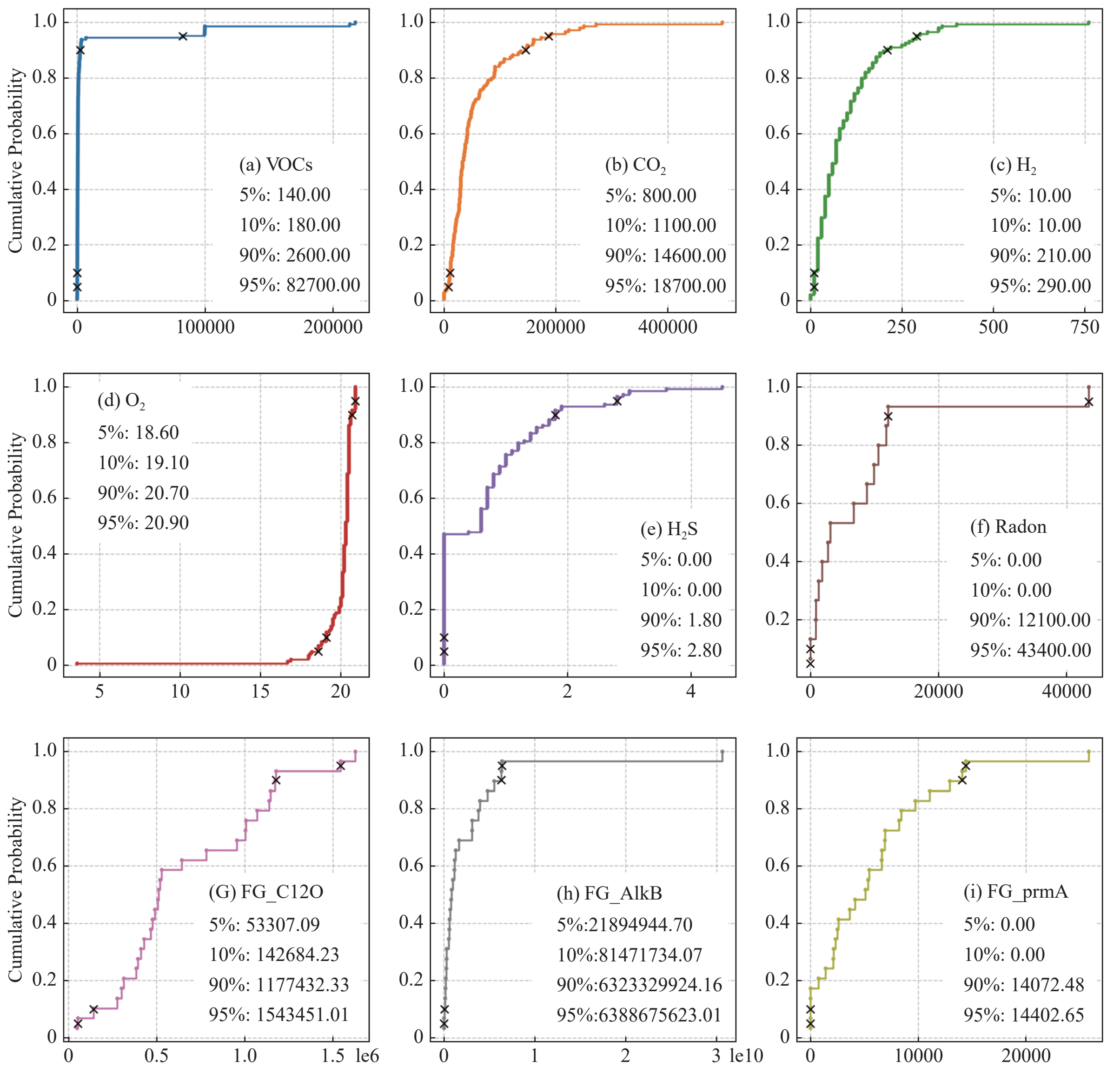

Fig. 6 BLA result of NIS dataset (The curves in the figures are plotted based on empirical cumulative distribution function; the black × symbols represent percentiles, with their specific values annotated in the legend)

ppb at the 90th percentile) and CO_2 (>14,600 ppm at the 90th percentile), high-value exceedances of ECDF thresholds directly indicate industrial leakage (Fig. 6a-b). Conversely, low-value anomalies in O_2 (<18.6% at the 5th percentile) and radon (<3000 Bq/m^3, determined by empirical thresholds due to limited sampling density and outlier effects) correspond to aerobic degradation zones (Fig. 6d) and extensive LNAPL accumulation in groundwater, respectively. Functional gene analysis further demonstrated that threshold exceedances for *C12O* (>1.18×10^6 copies/g at the 90th percentile) and *AlkB* (>6.32×10^9 copies/g at the 90th percentile) correlate with aromatic hydrocarbon pollution and alkane leakage while revealing microbial degradation potential (Fig. 6g-h). The *prmA* gene showed low values (0 at the 5th percentile) indicating anaerobic inhibition zones and high values (>1.41×10^4 copies/g at the 90th percentile) reflecting methane oxidation activity (Fig. 6i), with distinct concentration ranges corresponding to specific biogeochemical processes.

2.2 An active chemical-pharmaceutical manufacturing park

The study area is an active chemical-pharmaceutical manufacturing park located in eastern China, covering approximately 50,000 m^2. The facility layout and NIS monitoring point distribution are illustrated in Fig. 7a. The primary NIS survey indicators included VOCs, H_2S, CH_4, CO_2, O_2, and H_2 concentrations. Given the absence of radon and functional gene survey data but the high completeness of monitoring point datasets, PCA was

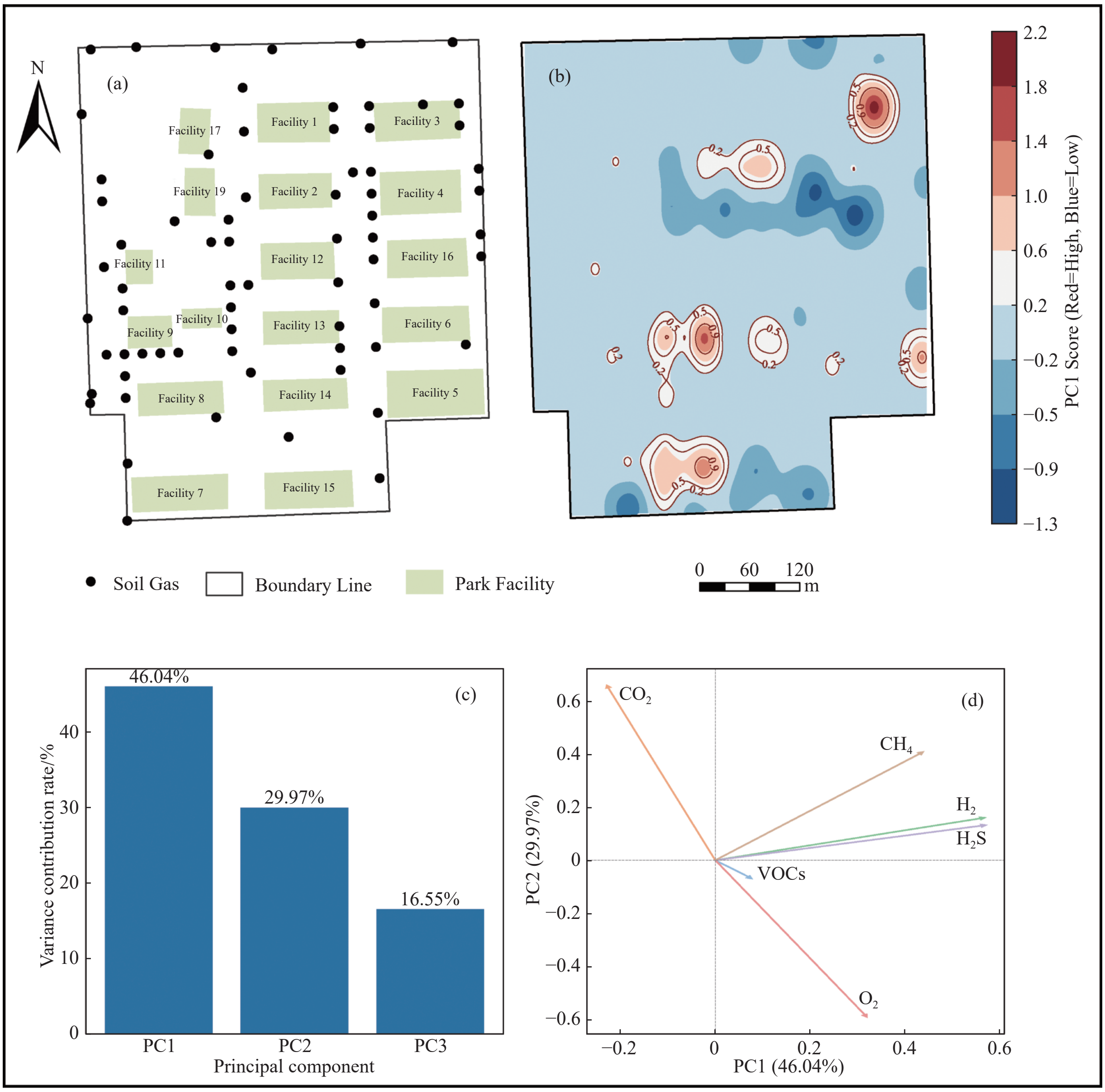

Fig. 7 Overview of the NIS dataset and PCA result in the contaminated chemical-pharmaceutical manufacturing park (a) Distribution of NIS survey points and data types; (b) Contamination assessment results based on PC1; (c)–(d) PCA analysis outcomes of soil-gas data. Panels(b), (c), and (d) are direct outputs from SDPHC

selected to assess the contamination status of the park.

Fig. 7c-d presents the PCA results for the NIS survey dataset from this industrial park. The first three principal components accounted for 92.56% of the total variance, with PC1 contributing 46.04%, PC2 contributing 29.97%, and PC3 contributing 16.55% (Fig. 7c), thereby indicating that these components effectively capture the major patterns of variation in the dataset. The loading plot (Fig. 7d) illustrated variable differentiation patterns along PC1 and PC2: PC1 was dominated by strong positive loadings for H_2 (0.57), H_2S (0.57), and CH_4 (0.44), representing the continuous process of anaerobic metabolic activity (hydrogen fermentation → sulfate reduction → methanogenesis). In PC2, the strong negative loading of O_2 (−0.59) and the strong positive loading of CO_2 (0.66) showed a significant negative correlation ($r = -0.85$), highlighting the distinct contrast between aerobic and anaerobic environments.

High PC1 values corresponded to elevated $H_2/H_2S/CH_4$ levels coupled with O_2 depletion, aligning with anaerobic metabolic characteristics of LNAPL residual zones. Low PC1 values reflected aerobic degradation-dominated margins of contamination plumes or background areas. Visualizing the spatial gradient of PC1 clearly illustrated the transition from contamination source zones (red core) to background regions (blue transition). The kriging interpolation results for contamination identification in this park are shown in Fig. 7b, with multiple source zones delineated by red contour lines.

3 Conclusions and outlook

This study introduces SDPHC, an integrated analytical software designed for assessing PHCs contamination in active industrial parks using NIS datasets. By integrating the Python programming language, the PySide6 GUI framework, and open-source packages, SDPHC enables efficient data extraction, analysis, and computation while ensuring high development efficiency and scalability. The software incorporates three analytical methods, including ETA, BLA and PCA, to accommodate varying data completeness in NIS datasets. Its applicability has been demonstrated through two case studies.

The first case study involved a comprehensive investigation of radon, VOCs, O_2, CO_2, and functional gene data within an active petrochemical park. The combined application of ETA and BLA facilitated contamination characterization. Automated report templates based on empirical thresholds were developed and implemented (see Appendix B). Background value calculations further quantified natural baseline fluctuations for NIS indicators, providing critical insights for contamination source identification and remediation planning. The second case, a chemical-pharmaceutical manufacturing park, focused on VOCs, O_2, CO_2, and H_2S soil gas surveys. PCA exhibited strong suitability for contamination mapping, with the first three principal components accounting for 92.56% of data variance. The successful application of SDPHC in these active industrial parks highlights its robust capability for multi-dimensional NIS data integration, enabling precise localization of contamination source zones, accurate boundary delineation, and workflow automation through standardized processing pipelines and report generation modules. These features significantly enhance engineering efficiency and are anticipated to promote large-scale adoption of NIS-based environmental risk management frameworks and optimized contaminant control strategies in industrial parks, accelerating the widespread implementation of NIS technologies for sustainable site remediation.

However, several limitations should be acknowledged. The effectiveness of NIS is influenced by site-specific conditions, such as vadose zone characteristics and hydrogeological properties. In environments with low-permeability soils or high moisture content, the migration of gaseous indicators may be impeded, potentially affecting the sensitivity of the survey. Moreover, a current limitation of the method lies in its inability to accurately determine the vertical extent or depth of contamination based solely on surface-level NIS datasets.

Future research will focus on addressing this challenge through a deeper mechanistic understanding of NSZD processes and the development of high-fidelity, three-dimensional PHCs degradation simulation models. These advancements are expected to enable more precise depth estimation and volumetric characterization of contaminant plumes. The modular software architecture enables seamless integration of new functionalities, while existing toolkits (e.g., property window classes, plotting modules) simplify development complexity. The integration of advanced NSZD-based models into SDPHC holds strong promise for significantly enhancing the accuracy and reliability of PHCs contamination assessments in complex industrial settings.

Acknowledgements

This research was financially supported by the National Key Research and Development Program of China (No.2022YFC3702200), the S&T Program of Hebei (No.242S4201Z), the National Natural Science Foundation of China (No. 42372279), and the Natural Science Foundation of Anhui Province (No. JZ2022AKZR0451).

Appendix

Appendix Content related to this article can be found at https://github.com/ArtCoff/SDPHC/tree/main/APPENDIX.

References

Andrew Murray, Hugo van Kemenade, wiredfool, et al. 2025. python-pillow/Pillow: 12.0.0 (12.0.0). Zenodo. Doi：10.5281/zenodo.17362416.

Anshuman A, Eldho TI. 2022. Entity aware sequence to sequence learning using LSTMs for estimation of groundwater contamination release history and transport parameters. Journal of Hydrology, 608: 127662. Doi：10.

1016/j.jhydrol.2022.127662.

Arato A, Wehrer M, Biró B, et al. 2014. Integration of geophysical, geochemical and microbiological data for a comprehensive small-scale characterizationof an aged LNAPL-contaminated site. Environmental Science and Pollution Research, 21(15): 8948−8963. Doi：10.1007/s11356-013-2171-2.

Azzellino A, Colombo L, Lombi S, et al. 2019. Groundwater diffuse pollution in functional urban areas: The need to define anthropogenic diffuse pollution background levels. Science of The Total Environment, 656: 1207−1222. Doi：10.1016/j.scitotenv.2018.11.416.

Baessa MPM, Chang HK, Teramoto EH, et al. 2024. Integrating high-resolution investigation and soil TPH analysis for accurate estimation of LNAPL volume in the subsurface. Environmental Earth Sciences, 83(13): 415. Doi：10.1007/s12665-024-11621-2.

Bertolla L, Porsani JL, Soldovieri F, et al. 2014. GPR-4D monitoring a controlled LNAPL spill in a masonry tank at USP, Brazil. Journal of Applied Geophysics, 103: 237−244. Doi：10.1016/j.jappgeo.2014.02.006.

Bonnière A, Khaska S, Salle CLGL, et al. 2024. Long-term impact of wastewater effluent discharge on groundwater: Identification of contaminant plume by geochemical, isotopic, and organic tracers' approach. Water Research, 257: 121637. Doi：10.1016/j.watres.2024.121637.

Brendel CE, Dymond RL, Aguilar MF. 2020. Integration of quantitative precipitation forecasts with real-time hydrology and hydraulics modeling towards probabilistic forecasting of urban flooding. Environmental Modelling & Software, 134: 104864. Doi：10.1016/j.envsoft.2020.104864.

Cassidy NJ. 2008. GPR attenuation and scattering in a mature hydrocarbon spill: A modeling study. Vadose Zone Journal, 7(1): 140−159. Doi：10.2136/vzj2006.0142.

Cecconi A, Verginelli I, Baciocchi R. 2022. Modeling of soil gas radon as an in situ partitioning tracer for quantifying LNAPL contamination. Science of The Total Environment, 806: 150593. Doi：10.1016/j.scitotenv.2021.150593.

Chen CJ, Judge J, Hulse D. 2022. PyLUSAT: An open-source Python toolkit for GIS-based land use suitability analysis. Environmental Modelling & Software, 151: 105362. Doi：10.1016/j.envsoft.2022.105362.

Chen GD, Zhang K, Wang S, et al. 2024. PHyL v1.0: A parallel, flexible, and advanced software for hydrological and slope stability modeling at a regional scale. Environmental Modelling & Software, 172: 105882. DOI: 10.1016/j.envsoft.2023.105882.

Crudu V. 2024. Evaluating the efficiency of python versus C++ in handling scientific computing projects. https://moldstud.com/articles/p-evaluating-the-efficiency-of-python-versus-c-in-handling-scientific-computing-projects. Accessed 7 May 2025

Davis BM, Istok JD, Semprini L. 2005. Numerical simulations of radon as an in situ partitioning tracer for quantifying NAPL contamination using push–pull tests. Journal of Contaminant Hydrology, 78(1−2): 87−103. DOI: 10.1016/j.jconhyd.2005.03.003.

De Simone G, Lucchetti C, Pompilj F, et al. 2017. Soil radon survey to assess NAPL contamination from an ancient spill. Do kerosene vapors affect radon partition? Journal of Environmental Radioactivity, 171: 138−147. Doi：10.1016/j.jenvrad.2017.02.014.

EI. 2024. Statistical Review of World Energy. Energy Institute

Fernández-Baca CP, Truhlar AM, Omar AH, et al. 2018. Methane and nitrous oxide cycling microbial communities in soils above septic leach fields: Abundances with depth and correlations with net surface emissions. Science of The Total Environment, 640: 429−441. Doi：10.1016/j.scitotenv.2018.05.303.

García-Rincón J, Gatsios E, Lenhard RJ, et al. 2024. Advances in the characterisation and remediation of sites contaminated with petroleum hydrocarbons. Cham: Springer Nature.

Garg S, Newell CJ, Kulkarni PR, et al. 2017. Overview of natural source zone depletion: Processes, controlling factors, and composition change. Groundwater Monitoring & Remediation, 37(3): 62−81. Doi：10.1111/gwmr.12219.

GeoPandas. https://geopandas.org/en/stable/#. Accessed 24 Mar 2025a

Han XM, Liu YR, Zheng YM, et al. 2014. Response of bacterial pdo1, nah, and C12O genes to aged soil PAH pollution in a coke factory area. Environmental Science and Pollution Research, 21(16): 9754−9763. Doi：10.1007/s11356-014-2928-2.

Hunter JD. 2007. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3): 90−95. Doi：10.1109/MCSE.2007.55.

IEEE spectrum Top programming languages. 2024. https://spectrum.ieee.org/top-programming-languages-2024.

Jiang SM, Fan JH, Xia XM, et al. 2018. An effective Kalman filter-based method for groundwater pollution source identification and plume morphology characterization. Water, 10(8): 1063. Doi：10.3390/w10081063.

Jordahl, K., Bossche, J.V. den, Fleischmann, M., et al. 2020. geopandas/geopandas: v0.8.1. Doi：10.5281/zenodo.3946761.

Kadiyala A, Kumar A. 2017. Applications of Python to evaluate environmental data science problems. Environmental Progress & Sustainable Energy, 36(6): 1580−1586. Doi：10.1002/ep.12786.

Kulkarni PR, Walker KL, Newell CJ, et al. 2022. Natural source zone depletion (NSZD) insights from over 15 years of research and measurements: A multi-site study. Water Research, 225: 119170. Doi：10.1016/j.watres.2022.119170.

López-Fernández A, Gómez-Vela FA, Gonzalez-Dominguez J, et al. 2024. bioScience: A new Python science library for high-performance computing bioinformatics analytics. SoftwareX, 26: 101666. Doi：10.1016/j.softx.2024.101666.

Lucas L, Jauzein M. 2008. Use of principal component analysis to profile temporal and spatial variations of chlorinated solvent concentration in groundwater. Environmental Pollution, 151(1): 205−212. Doi：10.1016/j.envpol.2007.01.054.

Molinari A, Guadagnini L, Marcaccio M, et al. 2019. Geostatistical multimodel approach for the assessment of the spatial distribution of natural background concentrations in large-scale groundwater bodies. Water Research, 149: 522−532. Doi：10.1016/j.watres.2018.09.049.

Pan ZD, Lu WX, Wang H, et al. 2023. Groundwater contaminant source identification based on an ensemble learning search framework associated with an auto XGBoost surrogate. Environmental Modelling & Software, 159: 105588. Doi：10.1016/j.envsoft.2022.105588.

Pedregosa F, Varoquaux G, Gramfort A, et al. 2011. Scikit-learn: machine learning in python. Journal of Machine Learning Research, 12: 2825−2830

Preziosi E, Parrone D, Del Bon A, et al. 2014. Natural background level assessment in groundwaters: Probability plot versus preselection method. Journal of Geochemical Exploration, 143: 43−53. Doi：10.1016/j.gexplo.2014.03.015.

python-docx. 2025. python-docx documentation. Available on https://python-docx.readthedocs.io/en/latest/

Schubert M. 2015. Using radon as environmental tracer for the assessment of subsurface Non-Aqueous Phase Liquid (NAPL) contamination–A review. The European Physical Journal Special Topics, 224(4): 717−730. Doi：10.1140/epjst/e2015-02402-3.

Schubert M, Paschke A, Lau S, et al. 2007. Radon as a naturally occurring tracer for the assessment of residual NAPL contamination of aquifers. Environmental Pollution, 145(3): 920−927. Doi：10.1016/j.envpol.2006.04.029.

Shahi A, Aydin S, Ince B, et al. 2016. Evaluation of microbial population and functional genes during the bioremediation of petroleum-contaminated soil as an effective monitoring approach. Ecotoxicology and Environmental Safety, 125: 153−160. Doi：10.1016/j.ecoenv.2015.11.029.

Shao S, Gao C, Guo XJ, et al. 2019. Mapping the contaminant plume of an abandoned hydrocarbon disposal site with geophysical and geochemical methods, Jiangsu, China. Environmental Science and Pollution Research, 26(24): 24645−24657. Doi：10.1007/s11356-019-05780-0.

Shapely. 2025. The Shapely User Manual. Available on https://shapely.readthedocs.io/en/stable/manual.html.

Smith JWN, Davis GB, DeVaull GE, et al. 2022. Natural source zone depletion (NSZD): From process understanding to effective implementation at LNAPL-impacted sites. Quarterly Journal of Engineering Geology and Hydrogeology, 55(4): qjegh2021−qjegh2166. Doi：10.1144/qjegh2021-166.

Sookhak Lari K, Davis GB, Rayner JL, et al. 2019. Natural source zone depletion of LNAPL: A critical review supporting modelling approaches. Water Research, 157: 630−646. Doi：10.1016/j.watres.2019.04.001a.

Sookhak Lari K, Rayner JL, Davis GB. 2019. Toward optimizing LNAPL remediation. Water Resources Research, 55(2): 923−936. Doi：10.1029/2018WR023380.

Táncsics A, Szoboszlay S, Szabó I, et al. 2012. Quantification of subfamily I. 2. C catechol 2, 3-dioxygenase mRNA transcripts in groundwater samples of an oxygen-limited BTEX-contaminated site. Environmental Science & Technology, 46(1): 232−240. DOI: 10.1021/es201842h.

Tomlinson DW, Rivett MO, Wealthall GP, et al. 2017. Understanding complex LNAPL sites: Illustrated handbook of LNAPL transport and fate in the subsurface. Journal of Environmental Management, 204: 748−756. Doi：10.1016/j.jenvman.2017.08.015.

Tsai YJ, Chou YC, Wu YS, et al. 2020. Noninvasive survey technology for LNAPL-contaminated site investigation. Journal of Hydrology, 587: 125002. Doi：10.1016/j.jhydrol.2020.125002.

Verginelli I, Lahvis MA, Jourabchi P, et al. 2024. Soil gas gradient method for estimating natural source zone depletion rates of LNAPL and specific chemicals of concern. Water Research, 267: 122559. Doi：10.1016/j.watres.2024.122559.

Virtanen P, Gommers R, Oliphant TE, et al. 2020. SciPy 1. 0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3): 261−272. Doi：10.1038/s41592-019-0686-2.

Wang CQ, Dou Z, Zhu Y, et al. 2024. Breaking the mold of simulation-optimization: Direct forward machine learning methods for groundwater contaminant source identification. Journal of Hydrology, 642: 131759. Doi：10.1016/j.jhydrol.2024.131759.

Waskom M. 2021. Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60): 3021. Doi：10.21105/joss.03021.

Xia T, Dong YH, Mao DQ, et al. 2021. Delineation of LNAPL contaminant plumes at a former perfumery plant using electrical resistivity tomography. Hydrogeology Journal, 29(3): 1189−1201. Doi：10.1007/s10040-021-02311-5.

Zhang JJ, Li WX, Zeng LZ, et al. 2016. An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. Water Resources Research, 52(8): 5971−5984. Doi：10.1002/2016WR018598.

Zhu LZ, Lu WX, Luo CM, et al. 2024. An ensemble optimizer with a stacking ensemble surrogate model for identification of groundwater contamination source. Journal of Contaminant Hydrology, 267: 104437. Doi：10.1016/j.jconhyd.2024.104437.