Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Hybrid Data Analytics Framework Integrating Machine Learning and Graph-Based Knowledge Models for Explainable Soil Organic Carbon Assessment

Emmanuel Papadakis
University of Huddersfield
Gulger Mallik
University of Huddersfield
Ahmed Aman Ibrahim
University of Huddersfield
George Baryannis
g.bargiannis@hud.ac.uk
University of Huddersfield
Andreas Altinalmazis-Kondylis
Tierrasphere

Research Article

Keywords: Soil organic carbon, Machine learning, Knowledge graphs, Graph neural networks, Explainable artificial intelligence, Agroforestry

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10069501/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Hybrid Data Analytics Framework Integrating Machine Learning and Graph-Based Knowledge Models for Explainable Soil Organic Carbon Assessment

Emmanuel Papadakis[1], Gulger Mallik[1], Ahmed Aman Ibrahim[1], George Baryannis[1*], Andreas Altinalmazis-Kondylis[2]

[1*]School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, United Kingdom.
[2]Tierrasphere, 50 St Mary's Road, Hemel Hempstead, HP2 5HL, United Kingdom.

*Corresponding author(s). E-mail(s): g.bargiannis@hud.ac.uk;
Contributing authors: e.papadakis@hud.ac.uk; g.mallik3@hud.ac.uk;
ahmedamanibrahim@gmail.com; andreas.altinalmazis@gmail.com;

Abstract

Reliable assessment of soil organic carbon is essential for carbon removal, agroforestry planning, and credible Measurement, Reporting and Verification (MRV) systems. However, existing approaches face persistent challenges related to heterogeneous soil data, inconsistent sampling depths, limited spatial coverage, and the low interpretability of data-driven prediction models. This paper proposes a hybrid data analytics framework that integrates supervised machine learning with graph-based knowledge models to support explainable soil organic carbon assessment in agroforestry systems. The approach combines a data harmonisation pipeline, Random Forest models for estimating organic carbon concentration and carbon stocks, a graph neural network embedding model for identifying spatially coherent soil archetypes, and a Bayesian-network-based explanatory graph for interpreting environmental profiles associated with high and low organic carbon conditions. The framework is evaluated through a Mexico-focused case study using the WoSIS soil profile database enriched with climate, land-cover, ecoregion, and expert-recommended soil covariates. Results show that Random Forest models provide moderate predictive performance, with $\boldsymbol{R^2}$ values ranging from 0.54 to 0.78, while the graph-based component reveals geographically meaningful soil clusters and supports interpretable profiling of organic carbon conditions. The findings indicate that hybrid frameworks combining learning and knowledge can support soil carbon assessment by combining quantitative prediction, spatial contextualisation, and probabilistic explanation. The paper also discusses limitations related to data availability, geographic generalisability, and the need for further validation before operational MRV deployment.

Keywords: Soil organic carbon, Machine learning, Knowledge graphs, Graph neural networks, Explainable artificial intelligence, Agroforestry

1 Introduction

Climate change is strongly linked to anthropogenic greenhouse gas (GHG) emissions, particularly from fossil fuel combustion, land-use change, and industrial activity. In 2024, global GHG emissions reached a new record high of approximately 53.2 Gt CO_2eq. (Crippa et al. 2025). Rising atmospheric GHG concentrations are associated with increasingly frequent and severe climate-related hazards, including floods, droughts, and wildfires, which impose substantial social, economic, and ecological costs (Ferreira 2024). In response to this climate urgency, pathways consistent with limiting global warming to 1.5 °C require rapid emissions reductions together with carbon dioxide removal (CDR) to counterbalance residual emissions and support the transition towards net zero (Rogelj et al. 2018).

A key requirement for credible CDR is robust Measurement, Reporting and Verification (MRV). MRV systems are expected to provide evidence that carbon removal is additional, quantifiable, and sufficiently durable. Where carbon projects are linked to carbon-credit issuance, the credibility of the underlying MRV process directly affects trust in the resulting credits. However, confidence in carbon markets has been weakened by cases in which poor baselines, weak verification practices, and over-crediting have led to inflated climate benefits, particularly in avoidance-based offset projects (Romm et al. 2025; West et al. 2024). These concerns highlight the need for MRV approaches that are not only cost-effective and scalable, but also transparent, auditable, and scientifically grounded (Zakkour and Cook 2024; Mercer et al. 2024).

Soil organic carbon (SOC) assessment is especially challenging in this context. SOC varies across soil type, depth, climate, vegetation, management practice, and landscape position, while available soil datasets are often sparse, heterogeneous, and collected using inconsistent sampling protocols (Ding et al. 2025; Heuvelink et al. 2021). Machine learning offers a promising route for estimating SOC and carbon stocks from multivariate soil and environmental covariates, but purely data-driven models can be difficult to interpret and may treat spatially related observations as independent. This limits their usefulness in MRV settings, where users require not only a predicted carbon value, but also an explanation of the environmental conditions and assumptions associated with that value.

Graph-based knowledge models offer a complementary capability. By representing soil profiles, environmental covariates, spatial relationships, and causal assumptions as connected structures, they can support contextual reasoning and provide interpretable representations of regional soil variability. In particular, graph neural networks can encode spatial similarity between soil profiles, while explanatory graphs can help explain the conditions associated with high or low organic carbon levels. Integrating these graph-based approaches with machine learning therefore offers a route towards hybrid systems that combine predictive performance with contextualisation and explainability.

This paper proposes a hybrid analytics framework for explainable SOC assessment in agroforestry systems. The framework integrates a data harmonisation pipeline, supervised machine learning models for carbon prediction, a graph neural network embedding model for spatial soil representation, and a Bayesian-network-based explanatory graph for interpreting environmental profiles associated with different organic carbon conditions. The approach is demonstrated through a Mexico-focused case study using the World Soil Information Service (WoSIS) soil profile database enriched with climate, land-cover, ecoregion, and expert-recommended soil covariates.

The novelty of the proposed framework lies in the integration of three complementary analytics components within a single soil carbon assessment workflow: supervised prediction, graph-neural-network-based spatial representation, and probabilistic graph-based explanation. The main contributions of this paper are as follows:

1.
2. We propose a hybrid data analytics framework that combines machine learning, graph-based spatial representation, and probabilistic explanation for explainable SOC assessment.
3. We develop a data harmonisation workflow that integrates heterogeneous soil and environmental datasets and standardises soil measurements to a common 0-30 cm depth interval.

4. We evaluate supervised predictive models for estimating organic carbon concentration and carbon stocks in a Mexico-focused case study.
5. We construct a graph neural network embedding model that identifies spatially coherent soil archetypes independently of organic carbon values.
6. We demonstrate how probabilistic graph-based reasoning can support interpretation of environmental profiles associated with high and low organic carbon conditions.

The remainder of the paper is structured as follows. Section 2 reviews related work on machine learning for carbon sequestration prediction, knowledge graphs in environmental domains, and hybrid analytics systems. Section 3 presents the proposed methodology, including the framework architecture, dataset, predictive modelling component, graph-based component, and integration strategy. Section 4 describes the Mexico case study and prototype implementation. Section 5 presents the results and evaluation. Section 6 discusses the implications, strengths, and limitations of the proposed approach. Section 7 concludes the paper and outlines future work.

2 Background and Related Work

2.1 Machine learning for soil organic carbon and carbon sequestration prediction

Machine learning has increasingly been used to estimate, monitor, and predict carbon sequestration across environmental and land-based systems. These methods are attractive because carbon dynamics are influenced by multiple interacting variables, including soil properties, climate, vegetation, topography, and land management. Machine learning models can capture non-linear relationships among these variables and can therefore complement process-based or purely empirical approaches, particularly when heterogeneous observational datasets are available.

A substantial body of work has applied machine learning to carbon-related prediction tasks in different environmental contexts. For example, Chen et al. (2024) used machine learning to predict microalgae biomass production for carbon sequestration, identifying variables such as NaCl concentration, temperature, pH, and light intensity as important predictors. Uniyal et al. (2022) used satellite-derived features and vegetation indices to estimate above-ground biomass as a proxy for carbon sequestration in urban forests. In soil and agroecosystem settings, Sarkar and Somenahally (2023) used regression models to predict soil microbial biomass carbon under different soil health management practices, including cover crops and crop residues. Other studies have considered carbon sequestration in engineered or geological settings, including accelerated carbonation of municipal solid waste incineration fly ash (Chen et al. 2025) and CO_2 foam behaviour for geological carbon storage (Iskandarov et al. 2024).

Although these studies demonstrate the breadth of machine learning applications in carbon sequestration research, soil organic carbon (SOC) assessment presents specific challenges that require dedicated attention. SOC varies strongly with sampling depth, soil texture, climate, vegetation, land use, and management history. Moreover, available soil datasets are often sparse, irregularly sampled, and reported across inconsistent depth intervals. Recent work on SOC prediction has therefore emphasised the need for improved data harmonisation, more robust covariate integration, and careful evaluation across spatially diverse regions (Ding et al. 2025; Grunwald et al. 2025). In practical MRV settings, the challenge is not only to produce accurate estimates, but also to provide interpretable evidence about the environmental and soil conditions associated with those estimates.

Tree-based ensemble methods, such as Random Forests and gradient boosting models, are commonly used in this domain because they are relatively robust to heterogeneous tabular data, can model non-linear interactions, and can operate effectively with moderate-sized datasets. However, such models still provide limited insight into spatial dependence unless spatial features are explicitly engineered. Deep learning models may offer additional representational power, but their performance is often constrained by the size, quality, and spatial representativeness of the available training data. These limitations motivate hybrid approaches that combine the predictive strengths of machine learning with additional mechanisms for contextualisation and explanation.

2.2 Knowledge graphs and graph-based models in environmental decision support

Knowledge graphs and graph-based models provide a complementary approach for representing environmental systems. Rather than treating observations as isolated records, graph-based representations encode entities, attributes, and relationships in a connected structure. This is particularly relevant for environmental applications, where spatial, temporal, ecological, and causal relationships are central to interpretation.

Knowledge graphs have been widely used to integrate heterogeneous domain-specific data sources and support structured information retrieval across complex knowledge-intensive domains (Papadakis et al. 2023; Wu et al. 2022; Hamed et al. 2025). In environmental and sustainability domains, Hamed et al. (2025) developed an ontology-based knowledge graph for wildlife data generated by sensors, enabling structured representation and query-based access to diverse ecological observations. Wu et al. (2022) proposed LinkClimate, a knowledge graph integrating climate observations with sources such as OpenStreetMap and Wikidata to support multi-source climate analysis. These examples illustrate the value of graph-based representations for organising distributed environmental data and making relationships among entities explicit.

Graph-based approaches are also increasingly connected with machine learning. When environmental observations are represented as graphs, graph neural networks (GNNs) can learn embeddings that incorporate both node attributes and relational structure. This is particularly useful for spatial environmental data, where nearby observations may share similar characteristics due to underlying geographic, climatic, or pedological processes. Seri et al. (2025), for example, used spatio-temporal GNNs for greenhouse microclimate modelling, capturing relationships between environmental variables and their evolution over time.

For SOC assessment, graph-based models are relevant for two reasons. First, they can encode spatial relationships between soil profiles, allowing the model to represent similarity across nearby locations rather than relying solely on independent tabular observations. Second, they can support interpretability by grouping soil profiles into spatially coherent archetypes or by representing causal assumptions about how soil and environmental variables influence organic carbon. However, graph-based models alone do not necessarily provide accurate point predictions of SOC or carbon stocks. Their value is therefore strongest when combined with supervised predictive models as part of a hybrid framework.

2.3 Hybrid analytics and explainability gaps addressed by this work

Hybrid analytics frameworks combine data-driven learning with structured knowledge representation or reasoning. In environmental decision-support contexts, this combination is valuable because predictive accuracy alone is insufficient for many operational uses. Users also need to understand why a prediction has been made, what contextual factors influence it, and how reliable it may be under different environmental conditions. These requirements are particularly important for carbon MRV, where estimates may inform project design, monitoring strategies, investment decisions, and eventually carbon-credit claims.

Existing machine learning approaches to carbon prediction often focus on improving numerical performance, while giving less attention to how predictions can be contextualised within spatial, ecological, or causal knowledge. Conversely, knowledge graphs and graph-based environmental models often provide strong representational and reasoning capabilities, but are not always integrated with supervised predictive models for carbon estimation. This creates a methodological gap between accurate but weakly explainable prediction pipelines and interpretable but less directly predictive knowledge-based systems. Similar motivations have been explored in geospatial decision support, where hybrid approaches have combined knowledge-based spatial patterns with data-driven topic modelling to improve the discovery and interpretation of functional regions (Papadakis et al. 2019).

The present work addresses this gap by proposing a hybrid framework that integrates supervised machine learning with graph-based knowledge models for explainable SOC assessment. The machine learning component provides quantitative estimates of organic carbon concentration and carbon stock. The graph neural network component provides spatially grounded soil embeddings that support the identification of soil archetypes across Mexico. The explanatory graph supports interpretation by identifying environmental profiles associated with high and low organic carbon conditions. This combination is designed to support future soil carbon assessment workflows by linking prediction, spatial contextualisation, and explanation within a single data analytics framework.

3 Methodology

3.1 Framework overview

The proposed framework follows a hybrid architecture that combines data-driven machine learning with graph-based knowledge models to support explainable soil organic carbon assessment. The architecture is designed as a modular decision-support pipeline rather than as a standalone carbon-credit verification system. Its purpose is to produce quantitative estimates of soil organic carbon and carbon stocks, while also providing spatial and probabilistic context that can support interpretation in future MRV workflows.

As shown in Figure 1, the framework consists of four main components: (i) a data processing and harmonisation pipeline, (ii) a supervised machine learning component for carbon prediction, (iii) a graph-based knowledge component for spatial representation and probabilistic explanation, and (iv) an integration layer that combines predictive outputs with contextual information.

The data processing pipeline ingests, cleans, and harmonises heterogeneous data from multiple sources, including the WoSIS soil profile database, climate data, land-cover data, and regional ecological datasets. A central task in this pipeline is the standardisation of soil measurements to a common depth interval, enabling profiles collected using different sampling schemes to be compared and used in the same modelling workflow.

The machine learning component acts as the main predictive engine. It uses supervised regression models to estimate organic carbon concentration and carbon stock from harmonised soil and environmental covariates. The graph-based knowledge component complements this predictive layer by representing spatial relationships between soil profiles and by supporting probabilistic interpretation of environmental profiles associated with different organic carbon conditions. The integration layer brings these outputs together to support explainable soil carbon assessment, where a numerical carbon estimate can be interpreted alongside spatially grounded soil archetypes and probabilistic environmental descriptors.

3.2 Dataset and data harmonisation

The primary dataset used in this study is the 2023 snapshot of the WoSIS Soil Profile Database (Batjes et al. 2024). WoSIS provides standardised soil profile observations, including profile identifiers, geographic coordinates, sampling dates, depth intervals, and measured soil properties. For the Mexico-focused case study, records were filtered using the country field and then processed to create an analysis-ready dataset.

The initial extraction included key soil properties such as clay, sand, silt, organic carbon, pH, and electrical conductivity. Following expert consultation, the feature set was expanded to include variables relevant to carbon stock estimation and soil interpretation, including coarse fragment volume, bedrock depth, and soil type classification. The main variables used in the modelling workflow included:

- **Profile identifiers and depth information:** `profile_id`, `upper_depth`, and `lower_depth`. Depth intervals were standardised using `ea_spline` to produce a new feature `depth_category`;
- **Soil properties:** `clay`, `sand`, `silt`, `orgc`, `phaq`, and `elcosp`;
- **Additional soil descriptors:** `coarse_frag`, `bedrock_type` and `soil_type`;
- **Environmental covariates:** temperature, precipitation, ecoregion, and land cover.

A major preprocessing challenge was the heterogeneous structure of the raw soil profile data. Measurements for the same property were often

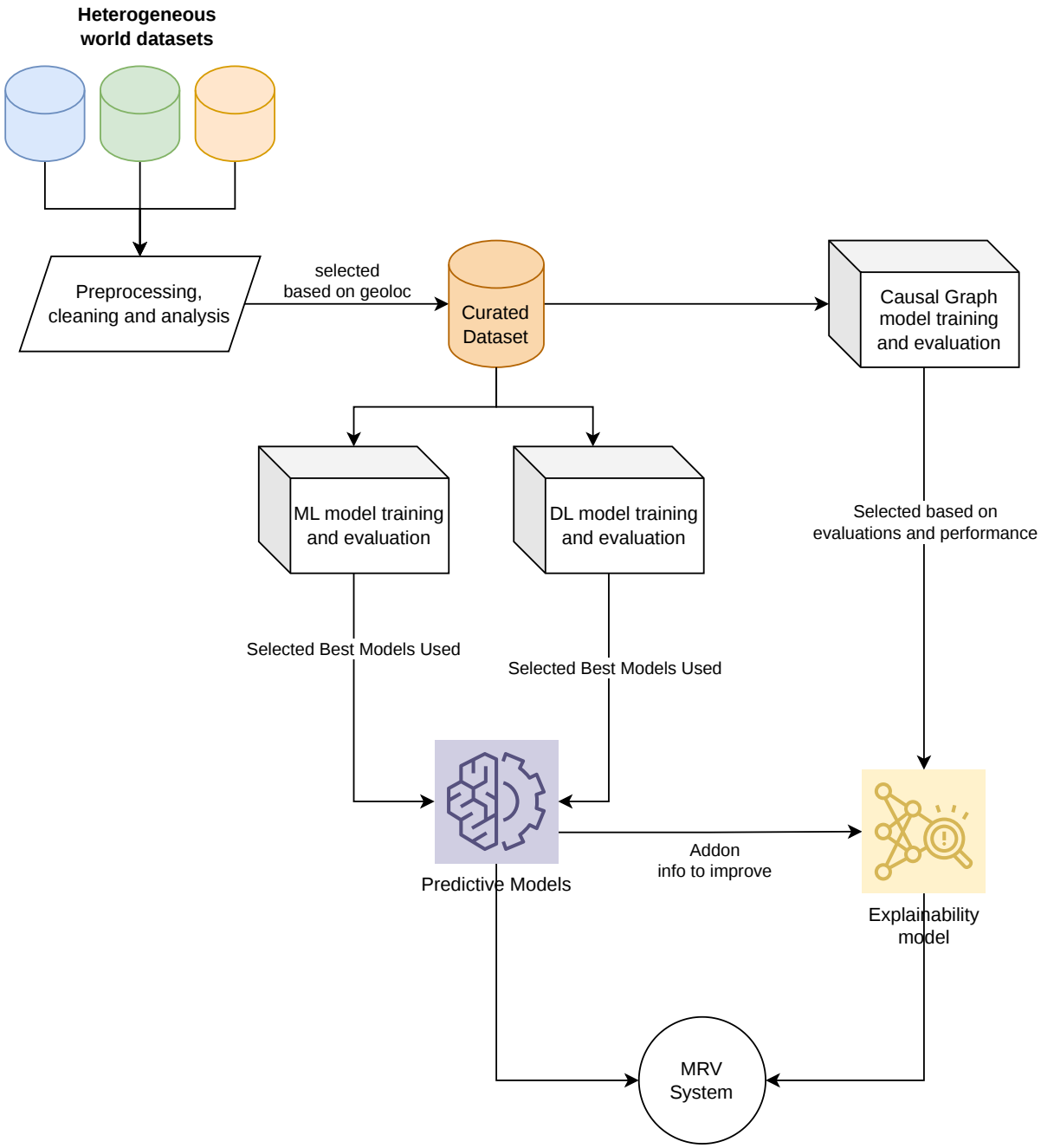


Fig. 1: Conceptual architecture of the proposed hybrid framework

reported over different depth intervals across different profiles, making direct comparison and integration unsuitable. To address this, soil properties were harmonised to a standard 0-30 cm interval using the Equal Area Quadratic Spline method, originally proposed for modelling soil attribute depth functions (Bishop et al. 1999; Malone et al. 2009), implemented through the `ea_spline` function in the `ithir` R package (Smart Digital Agriculture 2024). This method was selected because it preserves the total mass of a soil property when interpolating across depth intervals, which is

important for carbon accounting and profile-level comparison (Hengl et al. 2017).

The harmonisation process was applied profile by profile. Records with missing values in the relevant depth variables were removed, duplicate measurements were aggregated where necessary, and the spline model was used to estimate standardised values for the 0-30 cm layer. This produced a consistent feature representation across profiles and allowed the subsequent machine learning and graph-based components to operate on a common depth basis.

Climate covariates were obtained from the World Bank Climate Change Knowledge Portal (CCKP), which provides gridded climate information derived from ERA5 reanalysis data (World Bank Climate Change Knowledge Portal 2024). For each soil profile location, monthly mean precipitation and near-surface air temperature were extracted according to the available geographic coordinates. Land-cover information was derived from the MODIS MCD12Q1 Land Cover Type dataset (Friedl et al. 2015), using the IGBP classification scheme. Ecoregion information for Mexico was incorporated from the Comisi'on Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) biotic regionalisation dataset (Comisión Nacional para el Conocimiento y Uso de la Biodiversidad 2024). These external covariates were spatially joined to the soil profile data to provide environmental context for modelling and interpretation.

3.3 Machine learning component

The machine learning component was developed to estimate soil carbon outcomes from the harmonised soil and environmental covariates. Two related prediction tasks were considered: estimating organic carbon concentration in grams per kilogram (g/kg), and estimating carbon stock in tonnes per hectare (t/ha). The first task focuses on concentration, while the second is more directly relevant to carbon accounting because it incorporates information required to estimate carbon stored over a defined soil depth.

Before model training, preprocessing was applied to improve data quality and reduce the risk of leakage. Columns with more than 60% missing values were removed. Missing values in retained variables were imputed, categorical variables were encoded using one-hot or label encoding as appropriate, and numerical variables were scaled where required by the model type. All data transformations were fitted only on the training data and then applied to the test data to preserve independence between model development and evaluation.

The data were split using an 80/20 train-test strategy. To avoid leakage from repeated measurements within the same soil profile, the split preserved profile-level separation between training and test sets. This ensured that observations from the same profile did not appear in both subsets.

Several supervised regression models were evaluated during development, including ensemble-based models and an exploratory neural network model. Random Forest regression using default configuration was selected as the main predictive model because it provided the most consistent performance across the carbon prediction tasks and was well suited to the moderate-sized, heterogeneous tabular dataset. For the stock prediction tasks, additional derived variables, including bulk density, organic matter, depth fraction, and coarse fragment correction, were used for target calculation where required. Variables that were directly involved in target construction were excluded from the model inputs to avoid leakage.

The exploratory deep learning model was trained only for organic carbon concentration prediction in g/kg. It was not extended to the stock prediction tasks because the initial concentration model did not outperform the Random Forest models and because stock prediction involves derived targets requiring further validation. The main comparative evaluation therefore focuses on the Random Forest models.

Model performance was assessed using standard regression metrics: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics provide complementary information about prediction error magnitude and the proportion of variance explained by the model.

3.4 Graph-based knowledge component

The graph-based knowledge component was designed to provide spatial contextualisation and interpretable structure around the machine learning predictions. It consists of two elements: a spatial graph used to train a graph neural network embedding model, and a Bayesian-network-based explanatory graph used to interpret environmental profiles associated with high and low organic carbon conditions.

3.4.1 Spatial graph construction

To represent spatial relationships between soil profiles, the harmonised soil dataset was transformed into a graph. Each soil profile was represented as a node, with node features corresponding to its harmonised soil and environmental attributes. Organic carbon and inorganic carbon variables were excluded from the input features used to construct the embedding model so that the resulting soil archetypes were not directly determined by the carbon variables themselves.

Edges were constructed using geographic proximity. The Haversine distance was calculated between the latitude and longitude of each pair of soil profiles. An undirected edge was created between two nodes when the distance between them was less than or equal to 10 km. This threshold was used to encode local spatial relationships between nearby soil profiles, reflecting the assumption associated with Tobler's First Law of Geography (Tobler 1970) that geographically proximate soil observations are more likely to share related environmental and pedological characteristics.

This graph representation allows the model to move beyond independent tabular observations by encoding both node attributes and neighbourhood structure. It is therefore suitable for graph neural network modelling, where each node representation can be updated using information from connected neighbours.

3.4.2 Soil embedding model

A graph neural network was trained to learn low-dimensional embeddings for the soil profile nodes. The purpose of this model was not to predict organic carbon directly, but to identify spatially coherent soil representations that could support clustering, regional interpretation, and future data imputation.

The model used a two-layer Graph Convolutional Network (GCN) architecture (Kipf and Welling 2016). The first `GCNConv` layer transformed the input features into a 128-dimensional hidden representation, followed by a Rectified Linear Unit activation function. The second `GCNConv` layer projected this representation into a 64-dimensional embedding space. The resulting embeddings encode both the intrinsic attributes of each soil profile and the attributes of its spatial neighbours.

The GCN was trained using a self-supervised triplet margin loss, a metric-learning objective that encourages similar instances to be embedded closer together than dissimilar instances (Schroff et al. 2015). For each training step, triplets of nodes were constructed as follows:

- **Anchor:** a selected soil profile node;
- **Positive:** a neighbouring node connected to the anchor within the 10 km spatial graph;
- **Negative:** a node not connected to the anchor.

The objective was to learn embeddings in which the anchor is closer to its positive neighbour than to the negative node by at least a specified margin. The model was trained for 50 epochs using the Adam optimiser with a learning rate of 0.01 and a triplet margin of 1.0.

After training, the 64-dimensional node embeddings were clustered using K-Means. Eight clusters were specified to identify broad soil archetypes across the Mexico case-study region. The resulting clusters were analysed using the original soil and environmental variables to assess whether the learned embeddings corresponded to interpretable regional patterns.

3.4.3 Bayesian-network-based explanatory graph

In addition to the spatial embedding model, an explanatory graph was used to support explanation of organic carbon conditions. The graph represents assumed relationships between organic carbon and relevant soil and environmental variables, including pH, depth, slope, rock fragments, land cover, ecoregion, bedrock, and soil type.

The purpose of this component is explanatory rather than predictive. While the Random

Forest models provide numerical estimates of carbon outcomes, the explanatory graph supports interpretation by identifying which environmental profiles are most strongly associated with high or low organic carbon conditions. Variables were discretised into interpretable states, such as low and high pH, shallow and deeper soil layers, and low and high rock-fragment content. The graph was then queried by conditioning on high or low `orgc` states and examining the most probable associated conditions.

The resulting profiles should be interpreted as data-dependent explanatory summaries rather than as definitive causal proof. Their role in the proposed framework is to make outputs more interpretable by linking carbon conditions to plausible environmental descriptors.

3.5 Integration strategy

The integration strategy combines the outputs of the predictive and graph-based components into a unified decision-support workflow. The Random Forest models provide the primary quantitative estimates of organic carbon concentration and carbon stock. The graph neural network provides spatially grounded soil embeddings and cluster assignments that contextualise each profile within broader soil archetypes. The explanatory graph provides interpretable environmental profiles associated with high and low organic carbon conditions.

The integrated framework therefore separates three complementary functions. First, supervised machine learning estimates the carbon outcome for a profile or region based on harmonised soil and environmental variables. Second, the GNN embedding model identifies the spatial soil archetype to which the profile belongs, supporting regional comparison and stratified interpretation. Third, the explanatory graph provides an explanatory profile that helps users understand which environmental conditions are associated with high or low organic carbon values.

This integration is designed to support transparency in soil carbon assessment. Rather than presenting a carbon estimate as an isolated numerical output, the framework can present the estimate together with information about the profile's spatial context, cluster membership, and likely environmental drivers. In MRV-related applications, this can help users identify where predictions are consistent with known regional patterns, where additional sampling may be needed, and which environmental factors should be considered when interpreting carbon estimates.

4 Case Study

4.1 Study region and application context

The proposed framework was evaluated through a Mexico-focused case study. Mexico was selected because it provides a diverse environmental setting for soil organic carbon assessment, with substantial variation in climate, soil types, elevation, vegetation, and ecological regions. This diversity makes it a relevant test case for examining whether a hybrid approach can support both prediction and contextual interpretation across heterogeneous agroecological conditions.

The case study was also motivated by the wider application context of agroforestry-based carbon assessment. Agroforestry systems are of interest for carbon removal and climate mitigation because they can support biomass accumulation, soil restoration, biodiversity enhancement, and improved land productivity (Lorenz and Lal 2014; Nair et al. 2010). However, assessing their contribution to soil carbon storage requires methods that can operate under data scarcity and spatial variability. In this setting, a decision-support workflow must not only estimate carbon outcomes, but also help users interpret how soil and environmental conditions influence those estimates.

The Mexico case study therefore serves two purposes. First, it provides a geographically and environmentally varied setting in which to evaluate the predictive component of the framework. Second, it provides a realistic context for assessing whether graph-based representations and probabilistic explanations can support more interpretable soil carbon assessment. The focus is on demonstrating the feasibility and analytical value of the hybrid framework, rather than claiming operational readiness for carbon-credit verification.

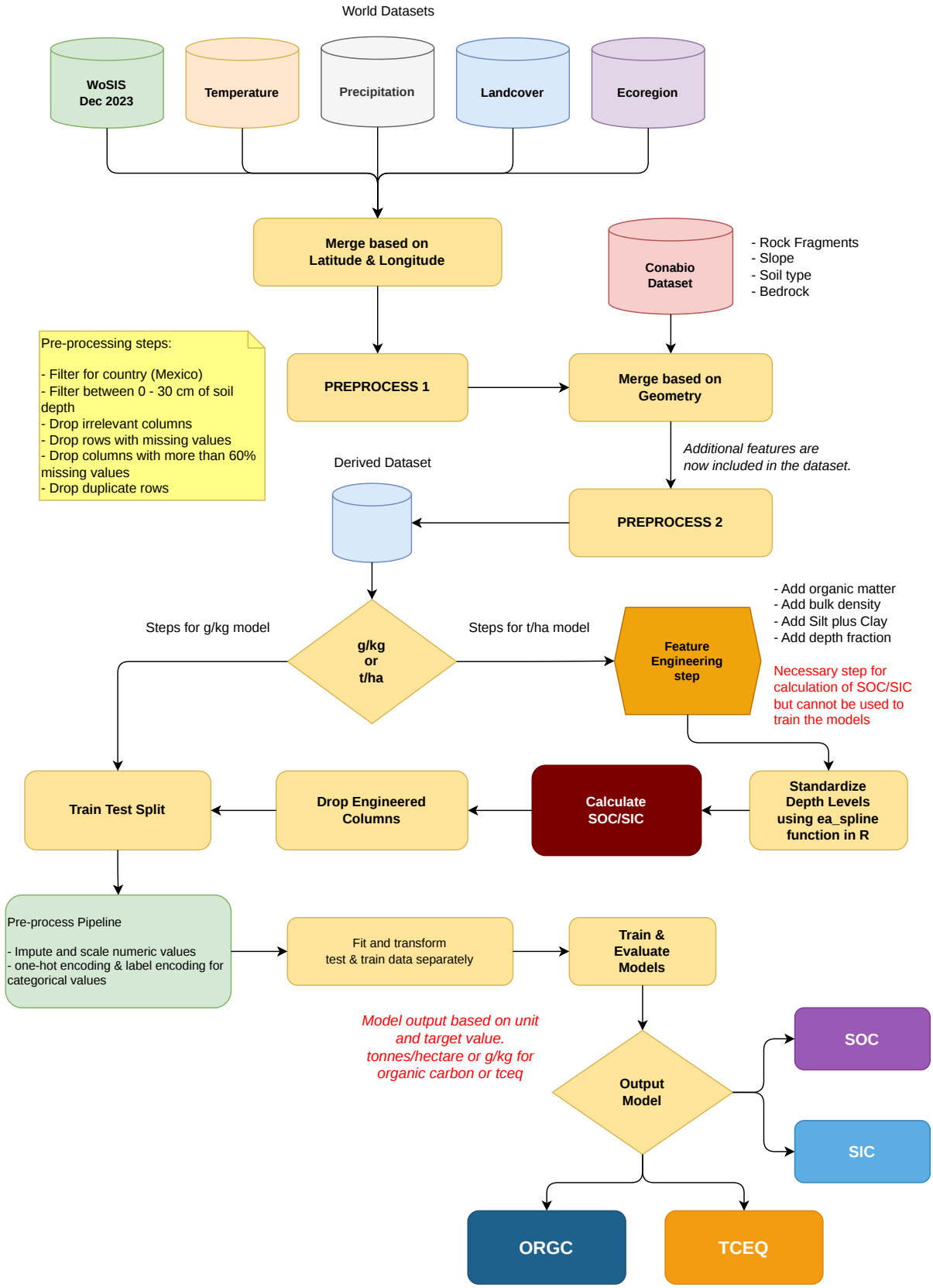


Fig. 2: Data analysis and modelling workflow for the supervised machine learning component.

4.2 Prototype implementation

A prototype analytics dashboard was implemented to demonstrate how the proposed framework could be made accessible to researchers and practitioners. The prototype, referred to as the Tierrasphere internal tool, was developed using Streamlit, a Python-based framework for building interactive data applications. The use of Streamlit enabled rapid development of a dashboard interface while allowing the modelling pipeline, geospatial processing, and visual analytics components to be integrated within a single environment.

The backend workflow integrates harmonised soil profile data with environmental covariates and model outputs. Soil properties were sourced from WoSIS, while additional covariates included climate variables, land-cover classifications, ecological regions, and topographic or soil descriptors where available. The preprocessing and harmonisation steps were performed before deployment so that the dashboard operated on cleaned and standardised inputs. This design reduced computational overhead in the user interface and ensured that displayed outputs were based on a consistent modelling dataset.

The supervised machine learning models were integrated into the prototype to provide estimates of organic carbon concentration and carbon stocks. Model outputs were linked to profile locations and regional descriptors, allowing users to inspect predictions geographically. The graph-based component was incorporated through cluster assignments derived from the soil embedding model, enabling the interface to display spatially coherent soil archetypes alongside the predictive outputs. This allows users to interpret a given prediction not only as an isolated value, but also in relation to similar soil profiles and regional environmental patterns.

The prototype was hosted using Microsoft Azure App Services. Caching mechanisms were used to improve responsiveness when interacting with geospatial layers and model outputs. The implementation was designed as a research prototype, with emphasis on demonstrating analytical functionality and interpretability rather than providing a production-level MRV platform.

4.3 Interface and dashboard functionality

The dashboard interface includes a map-based visualisation panel as its central component, through which users can inspect predicted carbon values and associated environmental variables across the study region. Users can compare locations and relate model outputs to land cover, ecoregion, and graph-derived soil cluster membership.

The interface also includes summary metrics and visual analytics designed to support interpretation. These include distributions of soil properties, model performance indicators, and profile-level information for selected locations. The combination of map-based exploration and tabular or graphical summaries supports movement between regional overview and local inspection. This is important for soil carbon assessment because users may need to identify broad spatial patterns while also examining individual profiles or candidate intervention areas.

A key design goal was to support explainability rather than only prediction display. For this reason, the dashboard was structured to present carbon estimates together with contextual descriptors from the graph-based component. In practical terms, this means that a user can inspect the predicted carbon outcome for a location, its associated soil archetype, and relevant environmental variables. This supports more transparent interpretation of predictions and can help identify cases where further sampling or expert review may be required.

The current dashboard should be understood as a proof-of-concept implementation. It demonstrates how hybrid analytics outputs can be integrated into a usable decision-support interface, but it does not yet provide all functionality required for operational deployment in carbon MRV.

5 Results and Evaluation

5.1 Supervised prediction performance

The supervised machine learning component was evaluated using standard regression metrics, including mean squared error (MSE), root

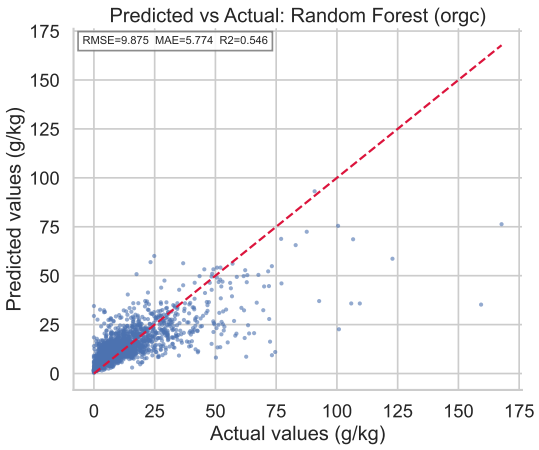
mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The evaluation focused on the Random Forest models, which provided the most consistent performance across the main prediction tasks. The results are summarised in Table 1.

The Random Forest models achieved moderate predictive performance, with R^2 values ranging from 0.54 to 0.78. These values indicate that the models captured meaningful signal in the data, while also leaving a substantial proportion of variance unexplained. This level of performance is consistent with the difficulty of soil carbon prediction from heterogeneous observational data, where carbon outcomes are shaped by interacting soil, climatic, ecological, and management-related factors (Ding et al. 2025; Zhang et al. 2024).

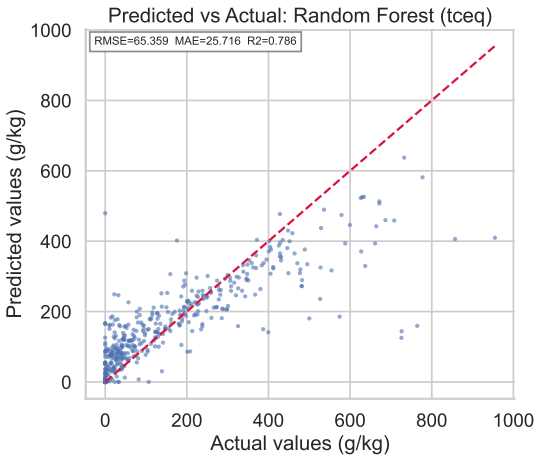
The strongest performance was observed for Total Carbon Equivalent (TCEQ) concentration, with an R^2 of 0.78. This is practically relevant because stock estimates are more directly aligned with carbon accounting than concentration estimates alone. However, the associated RMSE of 65.35 g/kg and MAE of 25.71 g/kg indicate considerable prediction uncertainty in absolute terms. The Organic Carbon (ORGC) concentration model showed lower error magnitudes (RMSE 9.87 g/kg, MAE 5.77 g/kg) but also a lower R^2 of 0.54, suggesting more constrained predictive capacity.

For stock-based predictions, both models achieved an R^2 of 0.54. The ORGC stock model produced an RMSE of 8.89 t/ha and MAE of 6.20 t/ha, while the TCEQ stock model exhibited substantially larger absolute errors (RMSE 128.67 t/ha, MAE 52.99 t/ha) due to the broader dynamic range of stock values. These results suggest that while the models capture broad spatial patterns in carbon stocks, they are less reliable for precise point-level estimation.

The prediction plots in Figures 3 and 4 further illustrate this pattern. The models reproduce general trends in the data, but the dispersion around the 1:1 line highlights residual uncertainty and limited precision for individual soil profiles. This reinforces the need to interpret model outputs alongside contextual information and to employ uncertainty-aware decision processes, particularly in applications related to MRV and carbon accounting.



(a) Predictions vs. actual values for Organic Carbon (g/kg).



(b) Predictions vs. actual values for Total Carbon Equivalent (g/kg).

Fig. 3: Model performance plots for carbon predictions in g/kg.

An exploratory deep learning model was also trained for `orgc` prediction in g/kg. This model was not included in the main comparative table because the available evaluation outputs were not directly comparable across all metrics reported for the Random Forest models. The convergence curves indicate stable training dynamics, with both loss and MAE decreasing for the training and validation sets. However, the final validation MAE stabilised around 7-8 g/kg and the model achieved a relatively low R^2 of 0.26, as shown in Figure 5. This suggests that, although the neural

Table 1: Evaluation metrics for Random Forest predictive models

Prediction Task	Unit	MSE	RMSE	MAE	R^2
ORGC (Figure 3a)	g/kg	97.51	9.87	5.77	0.54
TCEQ (Figure 3b)	g/kg	4271.75	65.35	25.71	0.78
ORGC Stock (Figure 4a)	t/ha	79.20	8.89	6.20	0.54
TCEQ Stock (Figure 4b)	t/ha	16557.87	128.67	52.99	0.54

network learned some patterns in the data, it did not generalise as effectively as the Random Forest models.

Deep learning was therefore not extended to the `orgc` stock prediction task in the present study. Since the stock targets depend on additional derived variables and pedotransfer-based calculations, a defensible deep learning comparison would require further model tuning, validation, and sensitivity analysis. The results reported here therefore focus on the Random Forest models as the most robust and consistently evaluated predictive component of the proposed framework.

5.2 Graph embedding model

The graph embedding model was evaluated by examining whether the learned soil embeddings produced spatially coherent and environmentally interpretable clusters. The model was trained using soil and environmental variables while excluding organic and inorganic carbon from the input features. This design ensured that the resulting clusters were not directly determined by the carbon variables themselves, allowing them to be interpreted as carbon-independent soil archetypes.

Figure 7 shows the geographic distribution of the eight learned clusters. The clusters exhibit clear regional structure, suggesting that the graph embedding model captured spatially meaningful similarities between soil profiles. Tables 2 and 3 provide summary statistics for the numerical and categorical variables associated with each cluster. Note that, for all clusters, *bedrock/mode* is *rocas_sedimentarias*, while *soil_type/mode* is REGOSOL EUTRICO.

The cluster summaries indicate that the learned embeddings distinguish profiles along several interpretable soil and environmental dimensions. Clusters 0 and 2 show the highest mean organic carbon values, despite organic carbon not being used as an input to the embedding model. Both clusters are associated with comparatively acidic pH values and higher silt-clay fractions than several other clusters. Cluster 2 has the lowest mean sand content and the highest mean silt-clay fraction, suggesting a finer-textured soil profile. In contrast, Cluster 3 is characterised by high pH, high mean inorganic carbon, and low slope, indicating a distinct alkaline and arid-region soil profile. Cluster 1 also shows relatively high pH and inorganic carbon, but differs from Cluster 3 in its modal land-cover class and lower inorganic carbon mean.

Clusters 4-7 are less strongly differentiated in several of the aggregate numerical descriptors, although they vary in sand and silt-clay composition and occupy different positions in the learned embedding space. Their shared modal categorical values suggest that they may represent finer spatial or textural subgroups within a broader soil and ecoregion context. This highlights both the interpretive value and the limitation of aggregate cluster statistics: the embedding model can identify spatial groupings, but detailed interpretation requires further examination of the underlying profiles and regional conditions.

Overall, the graph embedding results support the use of the GNN component as a spatial contextualisation layer. The model does not replace the supervised predictive models, but it provides complementary information about regional soil similarity, which can support stratified sampling, interpretation of model outputs, and identification of areas where additional data collection may be useful.

5.3 Environmental profiles for organic carbon

The Bayesian-network-based explanatory graph was used to examine the environmental profiles associated with high and low organic carbon conditions. This analysis provides an explanatory

Table 2: Cluster statistics for numeric soil and environmental variables. Top half: texture, conductivity, slope, and organic carbon. Bottom half: Inorganic carbon, silt-clay, pH, and rock fragments

Cluster	sand		elcosp		slope		orgc	
	mean	std	mean	std	mean	std	mean	std
0	53.36	13.99	0.087	0.138	5.51	4.91	19.50	20.73
1	64.14	14.43	0.377	1.16	1.17	1.35	7.39	7.59
2	45.30	13.70	0.109	0.171	5.56	4.63	21.77	20.29
3	69.24	12.91	0.342	0.730	0.93	1.17	6.40	5.05
4	72.04	13.32	0.093	0.170	2.98	3.19	10.77	10.67
5	65.53	14.19	0.090	0.163	2.50	3.10	11.23	11.00
6	67.97	13.00	0.090	0.163	2.50	3.10	11.23	11.00
7	59.99	13.57	0.090	0.163	2.50	3.10	11.23	11.00

Cluster	inorgc		silt_clay		phaq		rock_fragments	
	mean	std	mean	std	mean	std	mean	std
0	1.12	7.88	46.65	13.99	5.78	0.62	19.28	6.38
1	82.22	132.62	35.90	14.40	7.83	1.20	8.09	7.19
2	4.14	42.55	54.71	13.70	5.52	0.80	6.09	6.01
3	215.87	226.25	30.77	12.89	8.39	0.57	9.55	7.75
4	0.28	3.46	27.96	13.32	5.78	0.77	5.83	6.39
5	0.28	3.46	34.47	14.19	5.78	0.77	5.83	6.39
6	0.28	3.46	32.03	13.00	5.78	0.77	5.83	6.39
7	0.28	3.46	40.01	13.57	5.78	0.77	5.83	6.39

Table 3: Mode values of categorical variables for each cluster

Cluster	landcover/mode	ecoregion/mode
0	9	4
1	17	5
2	9	4
3	7	5
4	8	2
5	8	2
6	8	2
7	8	2

layer by identifying which soil and environmental states become more probable when the graph is conditioned on high or low `orgc`. Table 4 summarises the resulting profiles.

When the graph is conditioned on high `orgc`, the inferred profile is associated with acidic conditions, shallow soil depth, temperate subhumid and arid or semi-arid ecoregions, and landcover classes including grasslands, savannahs, and open shrublands. These conditions are consistent with the expectation that organic carbon is concentrated near the surface and influenced by vegetation-derived inputs and environmental conditions affecting decomposition.

When the graph is conditioned on low `orgc`, the inferred profile shifts towards alkaline conditions, deeper soil layers, arid and semi-arid ecoregions, lower slopes, and higher rock-fragment content. This profile is consistent with environments in which organic matter accumulation may be more constrained.

Some background characteristics, particularly sedimentary bedrock and Regosol Eutrico as the most common soil type, remain relatively stable across both scenarios. This suggests that variables such as pH, depth, slope, rock fragments, and ecoregion provide more discriminative explanatory information for distinguishing high and low organic carbon conditions in the present dataset.

Table 4: Comparative summary of Bayesian-network-derived environmental profiles associated with high and low organic carbon (`orgc`) conditions

Variable	High `orgc`	Low `orgc`
pH (`phaq`)	Low / acidic pH ($\leq$ 5.95), 41.6% probability	High / alkaline pH ($\geq$ 8.08), 43.8% probability
Soil depth	Shallow layer (0–5 cm), 40.4% probability	Deeper layer (15–30 cm), 40.7% probability
Dominant ecoregion	Templada subhúmeda, 34.4%, followed by Árida y semiárida, 27.5%	Árida y semiárida, 59.8% probability
Land cover	Grasslands, 24.5%; savannahs, 19.8%; and open shrublands, 18.5%	No single land-cover class was sufficiently dominant to be highlighted in the inferred profile
Slope	Higher slope, 36.3% probability	Lower slope, 37.6% probability
Rock fragments	Low rock-fragment content, 36.1% probability	High rock-fragment content, 40.4% probability
Stable background characteristics	Predominantly sedimentary bedrock and Regosol Eutrico as the most common soil type across both scenarios	

This analysis should be interpreted with caution because land-use data were unavailable, limiting our ability to assess the expected influence of plant-derived inputs on soil organic carbon.

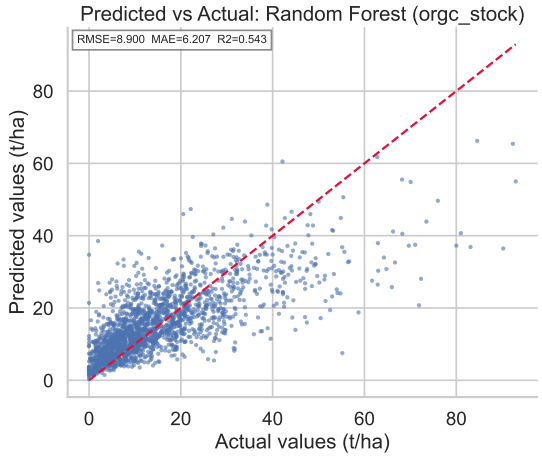
6 Discussion

6.1 Value of the hybrid framework

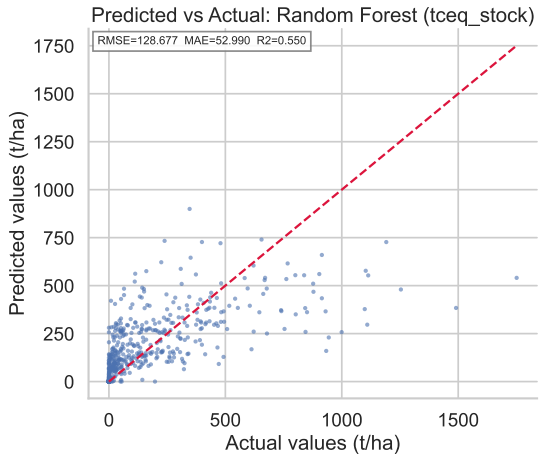
The results demonstrate the value of combining supervised machine learning with graph-based knowledge models for explainable soil organic carbon assessment. The Random Forest models provide the main quantitative prediction capability, while the graph-based components provide spatial contextualisation and probabilistic interpretation. This division of roles is important because soil carbon assessment requires more than a numerical estimate: agroforestry carbon outcomes emerge from interactions among soil properties, climate, vegetation, and management interventions, reflecting the kind of interdependent behaviour associated with complex adaptive systems (Ahmad et al. 2024). For decision-support and future MRV applications, users therefore need to understand how predictions relate to regional soil patterns, environmental conditions, and plausible explanatory factors, reflecting wider concerns around interpretability and trust in AI-supported environmental decision-making (Schiller et al. 2025).

The supervised models achieved moderate predictive performance, with R^2 values between 0.54 and 0.78 across the main prediction tasks. These results indicate that the models capture meaningful signal in the harmonised soil and environmental data, but also that substantial variability remains unexplained. In practical terms, the models are therefore better suited to exploratory assessment, relative comparison, and decision support than to precise point estimation. This is an important distinction for carbon MRV, where outputs may eventually inform financial or regulatory claims and therefore require high levels of confidence, auditability, and validation.

The graph embedding component adds a complementary form of value. As shown in Figure 7, the learned embeddings produce spatially coherent clusters across Mexico. Tables 2 and 3 further show that these clusters correspond to interpretable differences in soil and environmental characteristics, including pH, texture, inorganic



(a) Predictions vs. actual values for Organic Carbon (t/ha).



(b) Predictions vs. actual values for Total Carbon Equivalent (t/ha).

Fig. 4: Model performance plots for carbon predictions in t/ha.

carbon, slope, and soil type. Since organic and inorganic carbon were excluded from the embedding model inputs, the resulting clusters can be interpreted as carbon-independent soil archetypes. This supports regional interpretation by helping users understand whether a given profile belongs to a broader group of environmentally similar soils.

The Bayesian-network-based graph provides a further explanatory layer. By conditioning the graph on high and low `orgc` states, the framework identifies contrasting environmental profiles associated with organic carbon conditions. These profiles are not presented as definitive causal proof, but they help translate model outputs into interpretable environmental descriptors. For example, high organic carbon conditions were associated with shallower soil layers and more acidic pH, while low organic carbon conditions were associated with deeper layers, alkaline conditions, and higher rock-fragment content. This type of explanation is valuable in decision-support contexts because it links predicted or observed carbon conditions to domain-relevant variables that can be examined by soil scientists, project developers, or MRV practitioners.

The hybrid framework therefore contributes to literature by combining three types of evidence: numerical prediction, spatial similarity, and probabilistic explanation. This combination is more useful for explainable soil carbon assessment than any single component alone. The machine learning models estimate carbon outcomes, the graph embeddings contextualise profiles within spatially grounded soil archetypes, and the explanatory graph supports interpretation of the conditions associated with high or low organic carbon. Together, these components provide a foundation for more transparent and interpretable carbon assessment workflows.

6.2 Implications for soil carbon analytics and MRV workflows

The proposed framework has several implications for soil carbon analytics and future MRV workflows. First, it demonstrates how heterogeneous soil and environmental data can be harmonised into a consistent modelling workflow. This is a practical requirement for soil carbon assessment because available datasets often differ in sampling depth, spatial resolution, measurement coverage, and variable definitions. The use of depth harmonisation and covariate integration provides a reproducible basis for constructing analysis-ready datasets from imperfect observational sources.

Second, the framework supports a more transparent form of carbon assessment. In many carbon-related applications, a prediction alone is insufficient. Users need to know where the estimate comes from, which environmental conditions

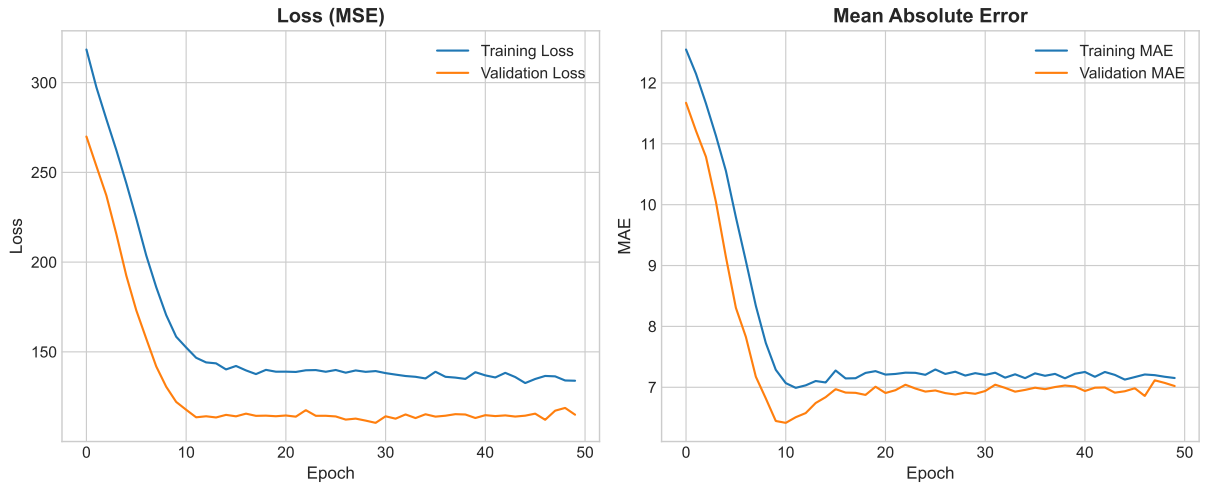


Fig. 5: Training and validation loss and MAE for the exploratory deep learning model for Organic Carbon prediction in g/kg.

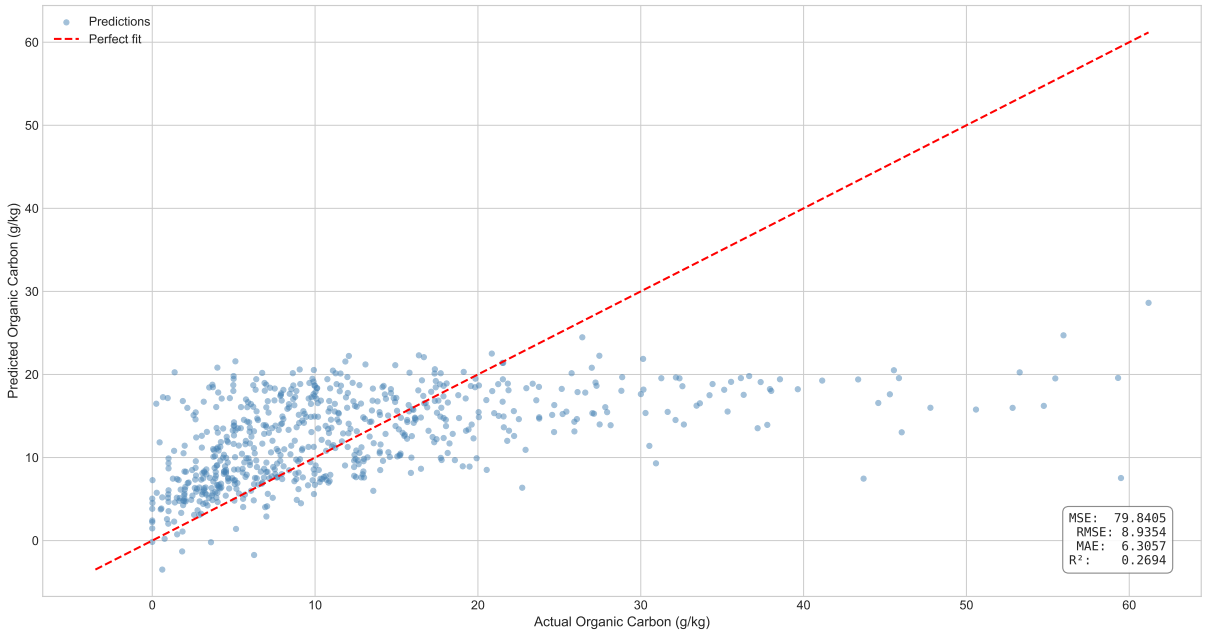


Fig. 6: Prediction performance plot for the exploratory deep learning model for Organic Carbon in g/kg.

are associated with it, and whether it is consistent with broader regional patterns. The proposed framework addresses this by connecting predictive outputs with graph-derived clusters and probabilistic environmental profiles. This does not remove the need for field validation, but it can help prioritise sampling, identify anomalous cases, and support expert review.

Third, the framework architecture is modular. The predictive component, graph embedding component, explanatory graph, and dashboard interface can be updated independently as new data or improved models become available. This is important for climate-tech applications, where systems must adapt to new field measurements, changing project boundaries, updated MRV standards, and

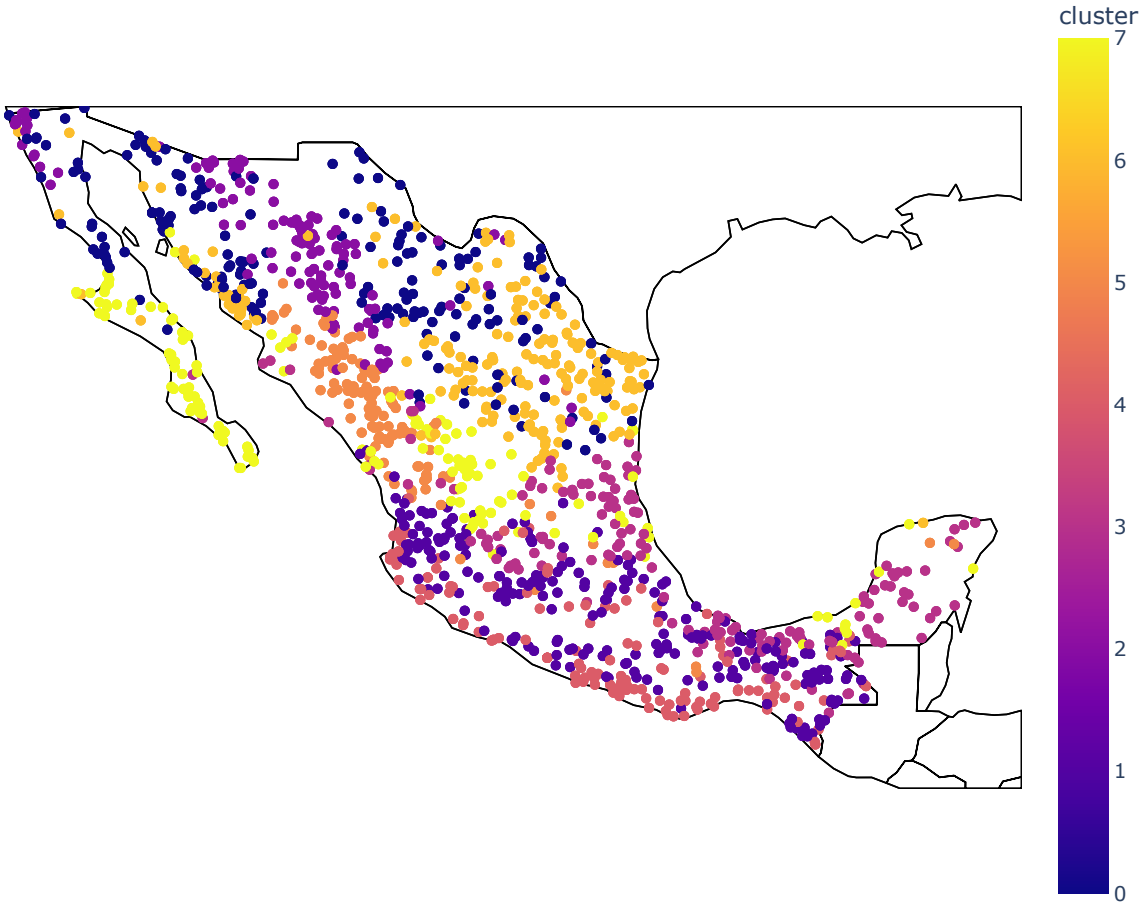


Fig. 7: Geographic distribution of the eight soil embedding clusters across Mexico, generated using the graph neural network embedding model. Each point represents a soil profile, coloured according to its cluster assignment. Organic and inorganic carbon variables were excluded from the embedding model inputs, so the clusters reflect spatial and environmental similarity rather than direct carbon values. Map lines delineate study areas and do not necessarily depict accepted national boundaries.

evolving scientific knowledge. For example, future versions could incorporate improved pedotransfer functions, higher-resolution covariates, field validation data, or uncertainty estimation methods without redesigning the full architecture.

Fourth, the prototype demonstrates how hybrid analytics outputs can be embedded in an interactive decision-support interface. This is important because MRV workflows involve multiple stakeholders, including researchers, land managers, project developers, auditors, and investors. A dashboard that presents predictions alongside spatial and explanatory context can support communication between technical and non-technical users. However, the current implementation should be viewed as a proof of concept rather than an operational MRV platform.

6.3 Limitations

Several limitations must be acknowledged. The first concerns data availability and quality. After

filtering and cleaning the WoSIS data for Mexico, the effective modelling dataset was relatively modest in size. This constrained the ability to train more complex models and contributed to the moderate predictive performance observed in the results. Soil carbon is influenced by many interacting processes, and the available covariates do not fully capture management history, vegetation inputs, disturbance, fertilisation, grazing, irrigation, or other land-use practices that can strongly affect organic carbon dynamics.

A second limitation concerns the distinction between land cover and land use. While land-cover information was available through remote-sensing-derived classifications, detailed land-use and management data were not available. This limits the ability of the model and explanatory graph to assess the expected influence of plant-derived inputs, agroforestry management, and human intervention on soil organic carbon. This is particularly important for agroforestry systems, where management practices can be central to carbon accumulation.

A third limitation relates to geographic generalisability. The case study focuses on Mexico, which provides substantial environmental diversity, but the trained models should not be assumed to generalise to other regions without further validation. Applying the framework to other countries or agroecological zones would require representative soil data, harmonised covariates, and region-specific evaluation (Zhang et al. 2024). Without this, model outputs may be unreliable in areas with soil types, climates, or land-use systems that are underrepresented in the training data.

A fourth limitation concerns the graph-based component. The graph embedding model produced interpretable spatial clusters, but the present study did not evaluate whether these embeddings improve supervised prediction when used as model inputs. Similarly, the explanatory graph supports explanatory profiling, but its outputs should be understood as probabilistic associations derived from the available data and assumed graph structure, not as definitive causal evidence. A stronger validation would require ablation studies, expert review of the causal structure, and comparison against independent field observations.

A fifth limitation concerns uncertainty. Although the evaluation reports standard regression metrics, the current prototype does not yet provide full uncertainty quantification at the level required for operational MRV. For high-integrity carbon accounting, users would need confidence intervals, error propagation across derived carbon stock calculations, sensitivity analysis, and clear communication of uncertainty in the dashboard. These elements are essential before the framework could support high-stakes verification or crediting decisions.

7 Conclusion and Future Work

This paper presented a hybrid analytics framework for explainable soil organic carbon assessment in agroforestry systems. The proposed approach integrates supervised machine learning, graph neural network embeddings, probabilistic graph-based explanation, and an interactive prototype dashboard. The aim is to support carbon assessment workflows in which numerical predictions are accompanied by spatial context and interpretable environmental profiles.

The Mexico-focused case study demonstrated the feasibility of the framework using harmonised soil profile data enriched with climate, land-cover, ecoregion, and soil covariates. The Random Forest models achieved moderate predictive performance across the main prediction tasks, with R^2 values ranging from 0.54 to 0.78. These results indicate that the models capture meaningful patterns in the available data, while also highlighting the difficulty of precise soil carbon prediction from heterogeneous observational datasets. The graph embedding model identified spatially coherent soil archetypes across Mexico, providing a complementary layer of regional interpretation. The explanatory graph further supported explainability by identifying environmental profiles associated with high and low organic carbon conditions.

The main contribution of the paper is therefore a hybrid analytics framework for decision support that links prediction, spatial contextualisation, and explanation. Rather than positioning the framework as an operational MRV platform, the paper shows how machine learning and graph-based knowledge models can be combined to support more transparent and interpretable soil carbon assessment. This is particularly relevant

for future carbon MRV workflows, where estimates must be supported by contextual evidence, uncertainty awareness, and expert review.

Future work should proceed in several directions. First, the predictive models should be validated using additional field measurements, including independent agroforestry datasets and longitudinal observations. This would enable stronger assessment of model generalisability and support evaluation across different soil types, climates, and management systems. Second, future versions should incorporate more detailed land-use and management data, including vegetation inputs, agroforestry practices, disturbance history, and fertilisation or grazing regimes, as these factors are central to soil organic carbon dynamics. Third, the graph-derived embeddings should be evaluated directly as predictive features through ablation studies comparing models with and without graph-based representations. This would clarify whether the spatial graph component improves predictive performance in addition to supporting interpretation.

Further work is also needed to strengthen the explainability and MRV readiness of the framework. The explanatory graph should be refined through expert elicitation and compared against independent soil science evidence. The prototype dashboard should incorporate uncertainty visualisation, data provenance, audit trails, and reporting functions aligned with specific MRV standards. Finally, the framework should be tested beyond Mexico to assess its transferability to other agroecological regions. These developments would support the transition from a research prototype towards a more robust, transparent, and auditable decision-support workflow for soil carbon assessment.

Acknowledgements. The authors thank Marcela Flores of Tierrasphere for supporting the project and facilitating engagement with the industrial context that motivated this research.

Declarations

7.1 Funding

This work was supported by Innovate UK [project reference 10121892]. The funder had no role in the design of the study, analysis and interpretation of data, writing of the manuscript, or decision to submit the article for publication.

7.2 Competing interests

The authors have no relevant financial or non-financial interests to disclose.

7.3 Data availability

The data used in this study were derived from publicly available sources, including the WoSIS Soil Profile Database, World Bank Climate Change Knowledge Portal climate data, MODIS land-cover data, and CONABIO ecoregion data. The processed datasets generated during the current study are available from the corresponding author on request.

7.4 Code availability

The code used to preprocess the data, train the models, and generate the analyses is available from the corresponding author on request.

References

Ahmad MA, Baryannis G, Hill R (2024) Defining Complex Adaptive Systems: An Algorithmic Approach. Systems 12(2). https://doi.org/10.3390/systems12020045

Batjes NH, Calisto L, de Sousa LM (2024) Providing quality-assessed and standardised soil data to support global mapping and modelling (wosis snapshot 2023). Earth System Science Data 16(10):4735–4765

Bishop T, McBratney A, Laslett G (1999) Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma 91(1):27–45. https://doi.org/10.1016/S0016-7061(99)00003-8

Chen J, Wang R, Chen Z, et al (2025) A new insight for co2 sequestration with heavy metal immobilization by calcium recycling using interpretative machine learning. Chemical Engineering Journal 511:161959

Chen M, Chen Y, Zhang Q (2024) Assessing global carbon sequestration and bioenergy potential

from microalgae cultivation on marginal lands leveraging machine learning. Science of The Total Environment 948:174462

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (2024) Biotic regions of mexico: Gis layer. http://www.conabio.gob.mx/informacion/gis/?vns=gis_root/region/biotic/zecol1mgw, accessed 29 April 2026

Crippa M, Guizzardi D, Pagani F, et al (2025) Ghg emissions of all world countries. Publications Office of the European Union, Luxembourg 10:953322

Ding Z, Liu K, Grunwald S, et al (2025) Advancing soil organic carbon prediction: A comprehensive review of technologies, ai, process-based and hybrid modelling approaches. Advanced Science 12(31):e04152. https://doi.org/10.1002/advs.202504152

Ferreira S (2024) Extreme weather events and climate change: economic impacts and adaptation policies. Annual Review of Resource Economics 16

Friedl MA, Sulla-Menashe D, Tan B, et al (2015) Modis/terra+aqua land cover type yearly l3 global 500m sin grid (mcd12q1). URL https://doi.org/10.5067/MODIS/MCD12Q1.006, accessed: 2025-09-11

Grunwald D, Poeplau C, Koch HJ, et al (2025) Effect of sugar beet (beta vulgaris l.) cultivation on soil organic carbon stocks in germany. Soil Use and Management 41(1):e70022. https://doi.org/10.1111/sum.70022

Hamed N, Rana O, Orozco Ter Wengel P, et al (2025) Foods: Ontology-based knowledge graphs for forest observatories. ACM Journal on Computing and Sustainable Societies 3(1):1–42

Hengl T, Mendes de Jesus J, Heuvelink GB, et al (2017) Soilgrids250m: Global gridded soil information based on machine learning. PLoS one 12(2):e0169748. https://doi.org/10.1371/journal.pone.0169748

Heuvelink GB, Angelini ME, Poggio L, et al (2021) Machine learning in space and time for modelling soil organic carbon change. European Journal of Soil Science 72(4):1607–1623

Iskandarov J, Fanourgakis GS, Ahmed S, et al (2024) Machine learning prediction and optimization of co2 foam performance for enhanced oil recovery and carbon sequestration: Effect of surfactant type and operating conditions. Geoenergy Science and Engineering 240:213064

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907

Lorenz K, Lal R (2014) Soil organic carbon sequestration in agroforestry systems. a review: K. lorenz, r. lal. Agronomy for Sustainable Development 34(2):443–454

Malone B, McBratney A, Minasny B, et al (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154(1):138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

Mercer L, Burke J, Rodway-Dyer S (2024) Towards improved cost estimates for monitoring, reporting and verification of carbon dioxide removal. Grantham Research Institute on Climate Change and the Environment

Nair PR, Nair VD, Kumar BM, et al (2010) Carbon sequestration in agroforestry systems. Advances in agronomy 108:237–307

Papadakis E, Gao S, Baryannis G (2019) Combining design patterns and topic modeling to discover regions that support particular functionality. ISPRS International Journal of Geo-Information 8(9). https://doi.org/10.3390/ijgi8090385

Papadakis E, Baryannis G, Batsakis S, et al (2023) Adhd-kg: a knowledge graph of attention deficit hyperactivity disorder. Health information science and systems 11(1):52

Rogelj J, Shindell D, Jiang K, et al (2018) Mitigation pathways compatible with 1.5 c in the context of sustainable development. In: Global

warming of 1.5 C. Intergovernmental Panel on Climate Change, p 93–174

Romm J, Lezak S, Alshamsi A (2025) Are carbon offsets fixable? Annual Review of Environment and Resources 50(1):649–680

Sarkar R, Somenahally A (2023) Machine learning soil-environmental impacts on agroecosystems for relating microbial biomass to soil carbon sequestration. Smart Agricultural Technology 4:100208

Schiller J, Stiller S, Ryo M (2025) Artificial intelligence in environmental and earth system sciences: explainability and trustworthiness. Artificial Intelligence Review 58(10):316

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823

Seri E, Petitta M, Papaioannou C, et al (2025) Sustainable greenhouse microclimate modeling: A comparative analysis of recurrent and graph neural networks. Building and Environment p 113473

Smart Digital Agriculture (2024) ithir R Package. https://smartdigiag.com/software/#s-1, accessed 29 April 2026

Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Economic geography 46(sup1):234–240

Uniyal S, Purohit S, Chaurasia K, et al (2022) Quantification of carbon sequestration by urban forest using landsat 8 oli and machine learning algorithms in jodhpur, india. Urban Forestry & Urban Greening 67:127445

West TA, Bomfim B, Haya BK (2024) Methodological issues with deforestation baselines compromise the integrity of carbon offsets from redd+. Global Environmental Change 87:102863

World Bank Climate Change Knowledge Portal (2024) Era5 netcdf browser. https://climateknowledgeportal.worldbank.org/netcdf-browser?prefix=data/era5-x0.25/, accessed 29 April 2026

Wu J, Orlandi F, O'Sullivan D, et al (2022) Linkclimate: An interoperable knowledge graph platform for climate data. Computers & Geosciences 169:105215

Zakkour P, Cook G (2024) Measurement, reporting and verification (MRV) and accounting for carbon dioxide removal (CDR) in the context of both project-based approaches and national greenhouse gas inventories (NGHGI). Tech. Rep. 2024-09, IEAGHG, https://doi.org/10.62849/2024-09

Zhang T, Huang LM, Yang RM (2024) Evaluation of digital soil mapping projection in soil organic carbon change modeling. Ecological Informatics 79:102394. https://doi.org/10.1016/j.ecoinf.2023.102394