Smart Agricultural Technology 14 (2026) 102035

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

A Multi-Scale Framework for Predicting Continuous Soil Properties Using Ranked Sentinel-2 Images and Zone-Level Soil Data: From a Farm Case Study to a Regional Application

Hamed Etezadi [a,*], Yacine Bouroubi [b], Viacheslav Adamchuk [a], Maxime Leduc [c], Marc-Olivier Gasser [d], Md Saifuzzaman [a,e]

[a] Department of Bioresource Engineering, McGill University, Montréal, QC, H9 × 3V9, Canada
[b] Département de Géomatique Appliquée, Université de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada
[c] Mon Système Fourrager, Montréal, QC, H1W 1M5, Canada
[d] Institut de recherche et développement en agroenvironnement, 2700 EInstein St., Québec City, QC, G1P 3W8, Canada
[e] Department of Biology, McGill University, Montréal, QC, H3A 1B1, Canada

ARTICLE INFO

Keywords:
Soil Organic Matter Prediction
Random Forest
LASSO
Informative Image
LightGBM
CatBoost
Province-Scale Application

ABSTRACT

Predicting continuous soil properties from limited field observations remains a central challenge in precision agriculture, particularly when models must operate across multiple spatial scales. This study develops a multi-scale framework that combines multi-temporal Sentinel-2 imagery with legacy soil maps to estimate spatial variation in soil organic matter (SOM). Bare-soil pixels were extracted and spectral indices calculated after cloud and snow masking, and a LASSO-based procedure was used to select informative images before modelling. Two strategies were evaluated: one using only satellite-derived indices, and another integrating soil texture information.

At the farm scale, twelve Sentinel-2 images yielded 422 field–image records. A simple field-level averaging baseline achieved RMSE = 0.14 log(%SOM) and $R^2 = 0.74$, while the hybrid model predicting at the zone level achieved RMSE = 0.16 log(%SOM) and $R^2 = 0.83$, capturing within-field variability despite slightly higher point-wise error. Remote-sensing-only models performed poorly (RMSE ≈ 0.35–0.37 log(%SOM) and $R^2 < 0.10$), demonstrating that spectral indices alone cannot represent subsurface conditions.

The framework was then scaled to the province of Québec. Multi-year images, soil texture, topography, and climate variables were combined, and Random Forest, LightGBM, and CatBoost were tested after feature screening. At the province scale, predictive performance decreased ($R^2 = 0.287$–0.364), reflecting the increased agroclimatic, edaphic, and management heterogeneity across Québec. However, the corresponding RMSE values (≈ 0.101–0.103 log(%SOM)) indicate that prediction errors remain quantitatively moderate after back-transformation. Therefore, model performance at this scale should be interpreted not only in terms of explained variance, but also in terms of operational prediction error and regional differentiation capability. At the regional level, the framework supports decision-oriented applications that rely on relative field differentiation rather than precise point estimation.

1. Introduction

As demand for high-resolution soil information grows, precision agriculture increasingly relies on spatially detailed predictions of continuous soil properties, such as pH, texture, bulk density, and organic matter. These properties govern key soil processes and directly influence crop productivity, nutrient cycling, and land management decisions. Traditional soil mapping methods, based on sparse sampling and laboratory analyses, often fall short of capturing within-field heterogeneity, which is critical for site-specific interventions [1,2]. In this context, recent advances in earth observation technologies and machine learning have opened new avenues for generating spatially continuous, scalable, and cost-effective estimates of soil attributes. By leveraging remote sensing data, legacy soil survey maps, and data-driven models, researchers can now predict complex soil variables over large areas with

* Corresponding author.
E-mail address: hamed.etezadi@mail.mcgill.ca (H. Etezadi).

https://doi.org/10.1016/j.atech.2026.102035
Received 3 February 2026; Received in revised form 2 March 2026; Accepted 21 March 2026
Available online 21 March 2026

Nomenclature

Abbreviation

BI	Brightness Index
BSI	Bare Soil Index
CatBoost	Categorical Boosting
CI	Coloration Index
CV	Cross-Validation
DEM	Digital Elevation Model
EVI	Enhanced Vegetation Index
GEE	Google Earth Engine
LASSO	Least Absolute Shrinkage and Selection Operator
LightGBM	Light Gradient Boosting Machine
LOI	Loss-On-Ignition
MAD	Median Absolute Deviation
ML	Machine learning
MNDVI	Modified Normalized Difference Water Index
MSI	MultiSpectral Instrument
NBR2	Normalized Burn Ratio 2
NDMI	Normalized Difference Moisture Index
NDSI	Normalized Difference Snow Index
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
OLI	Operational Land Imager
OOF	Out-Of-Fold
OMI	Organic Matter Index
R2	Coefficient of Determination
RF	Random Forest
RI	Redness Index
RMSE	Root Mean Square Error
RS	Remote Sensing
RVI	Ratio Vegetation Index
SAVI	Soil-Adjusted Vegetation Index
SI	Soil Index
SOC	Soil Organic Carbon
SOM	Soil Organic Matter
SWIR	Short-Wave Infrared
VI	Vegetation Indices

increasing accuracy and temporal resolution [3]. Among the many variables of interest, a few stand out due to their dual agronomic importance and environmental relevance. One such variable is soil organic matter (SOM), which plays a central role in both soil health and carbon sequestration [4,5].

SOM monitoring has traditionally been accomplished by sampling at high density and analyzing soil samples in the laboratory. Despite the possibility of obtaining high-precision mapping results, traditional SOM monitoring requires considerable financial and labour resources. Many scholars have focused on optical remote sensing (RS) as a powerful Earth observation technique. SOM content can be mapped efficiently and non-destructively using RS. With the development of quantitative remote sensing techniques, it has become possible to rapidly and cost-effectively determine SOM content at regional scales [6]. To determine whether SOM and proximally sensed Vis-NIR spectra are related, Krishnan [7] used a spectrometric approach. For estimating SOM content, they developed a multivariate regression model. Liu et al. [8] developed a similar model using spectral bands from 550 to 810 nm as input variables. A correlation between satellite spectral bands (e.g., Landsat-8, Sentinel-2, etc.) and SOM contents was also determined by Wang [9], Castaldi [4], Mohamed [10], and Luo [11]. To map the distribution pattern of SOM, a quantitative remote sensing model was developed based on the spectral absorption characteristics of SOM in Vis-NIR spectra [12]. Due to their refined spectral resolution, hyperspectral satellite images are usually considered ideal data for SOM mapping. Hyperspectral data, however, cannot be used in SOM mapping at a regional scale due to a lack of data and the high cost of acquisition and processing [4].

Multispectral optical reflectance is one of the most widely available and accessible satellite data types and is currently recognized as one of the most valuable sources of satellite measurements for mapping regional SOM content [13]. Landsat-8/9 OLI sensor, with its low temporal frequency (16 days) and medium spatial resolution (30 m), offers free access for estimating SOM content, thanks to its multiple Vis-NIR-SWIR spectral bands. Even better, Sentinel-2A&B MSI sensor provides 13 spectral bands (443 nm to 2190 nm) and covers key spectral features for SOM estimation (450, 590, and 665 nm). Furthermore, Sentinel-2 has a shorter revisit time (5 days) and higher spatial resolution (10-20-60 m), making it better suited for evaluating large areas of SOM and monitoring them at high spatial resolution. Thus, according to studies performed by Castaldi, multispectral Sentinel-2 imagery can be used for predicting SOM over a 10,000 km^2 cropland area in north-eastern Germany by calibrating soil organic carbon (SOC) indices from the LUCAS topsoil spectral library and applying them to carefully selected bare-soil pixels using NDVI, visible-based vegetation indices, and NBR2 moisture/residue thresholds [4,14]. Using ASD FieldSpec spectroradiometer measurements and airborne hyperspectral measurements, Gholizadeh compared Sentinel-2's capability to estimate SOM [15]. The results demonstrate that the spatial resolution and spectral characteristics of Sentinel-2 are sufficient to describe the variability of SOM across four agricultural fields in Czechia, which are dominated by Chernozems, Luvisols, Cambisols, and Stagnosols, where SOC was predicted from Sentinel-2 bands and 18 spectral indices using site-specific SVM regression models. Other studies have shown that vegetation indices (VIs) are important indicators of vegetation cover and growth status [16–18], and that the Normalized Difference Vegetation Index (NDVI) is an influential indicator of SOM. According to Bhunia [19], Ratio Vegetation Index (RVI) and the NDVI were used as auxiliary variables to predict SOM. Based on the results, VI and SOM are highly correlated. In addition to RVI and SAVI, Zeraatpisheh et al. [20] highlighted RVI as an equally important covariate. According to Gholizadeh et al. [15], the Sentinel-2 constellation can predict SOM using spectral indices such as NDVI and SAVI. Modified Normalized Difference Water Index (MNDWI) was also found to be sensitive to soil when Lu [21] used it as an ancillary variable in Landsat-based models to predict soil organic carbon distribution and its dynamic change between 2008 and 2013 in a mountainous hickory plantation region in western Lin-An District, Zhejiang Province, China.

Furthermore, to alleviate the effect of vegetation on remote sensing images, more research has focused on bare soil or low-density crop regions [22]. In an agricultural setting, the ideal time window for a bare-soil study varies with soil moisture, vegetation cover, and crop residue cover, limiting studies of SOM to small areas. Synthetic Soil Image - SYSI (Synthetic Bare Soil Image) was developed by Demattê [23] using multi-temporal satellite images for the purpose of constructing synthetic bare soil images. Using synthetic images, Gasmi [24] showed that bare soil pixel mean spectral reflectance improved the accuracy of the prediction model. It is important to note, however, that the optimal synthetic image differs for each soil property prediction. For digital soil mapping, Diek [25] analyzed Landsat time series to extract the barest pixels, and soil texture and soil organic matter were also mapped. When remote sensing data of bare soil is combined with measured soil property data, it is possible to develop more accurate soil property prediction models [26,27].

On the other hand, legacy soil series have garnered significant attention from researchers as a critical factor in determining SOM [28].

Organic matter stabilization, decomposition rates, and water retention are all affected by the distribution of sand, silt, and clay within the soil matrix. Several studies have consistently shown that soil texture and SOM content are correlated, emphasizing their importance in soil fertility prediction and management [29–33]. Despite their availability, soil survey maps are often underutilized in large-scale RS-based SOM frameworks. Researchers aim to enhance the accuracy and reliability of SOM estimation models by incorporating soil texture, providing valuable insights for sustainable soil management [34].

Machine learning (ML) is a conventional approach for estimating SOM from RS data. For RS-based SOM mapping, multivariate regression models were applied, such as partial least squares regression [4], multivariable linear regression [19], support vector machine [6], and random forest algorithms [11]. In remote sensing prediction, the random forest (RF) model is widely used due to its good performance and interpretability [35–38]. In many studies, RF regression has been utilized to estimate grain yields, predict surface temperatures, calculate water quality parameters, and predict soil physical and chemical properties [39,40]. RF algorithms have proven to be effective at predicting SOC and/or SOM [41,42]. Nevertheless, RF has known limitations when applied to large, highly heterogeneous datasets, where interactions and nonlinearities become more complex. Recent advances in gradient boosting decision tree algorithms, particularly LightGBM and CatBoost, address these challenges by offering superior scalability, better handling of categorical variables, and improved accuracy in remote sensing applications [43–45]. Despite their potential, their application to SOM prediction has been limited, highlighting a critical gap that motivates the present study.

In addition, cultivated regions exhibit strong temporal fluctuations in soil exposure driven by vegetation dynamics, crop residue, and management practices. The change in SOM content can, however, be ignored in the short term, i.e. over periods of several years up to about a decade, due to the slow pace at which it occurs over extended periods [46]. This stability provides an opportunity: multi-temporal satellite imagery can be combined with soil sampling to mitigate seasonal surface noise while still yielding reliable predictions of underlying SOM variability.

Despite increasing use of Sentinel-2 imagery for soil property prediction, several methodological limitations remain. Many studies rely primarily on spectral indices without systematically evaluating the added value of ancillary soil information, particularly across multiple spatial scales. In addition, image selection is often treated implicitly or based on predefined temporal composites, rather than being explicitly ranked and filtered according to predictive contribution. As a result, limited attention has been given to how image informativeness, scale transitions (from farm to region), and integration of categorical soil information jointly influence model robustness and generalization performance. Addressing these gaps requires a structured, scale-aware framework that explicitly evaluates the contribution of multi-temporal imagery and ancillary soil data under grouped validation schemes. Building on this premise, the present study pursued two complementary approaches for SOM spatialization based on Sentinel-2 data: (i) a farm-scale case study designed to test whether adding soil series/texture to remote sensing indices improves within-field prediction, and (ii) a province-scale study across Québec designed to evaluate scalability and algorithmic performance. The study focuses on SOM as a representative continuous soil property to evaluate how multi-temporal satellite imagery and ancillary soil data jointly influence predictive robustness. The central hypothesis is that explicitly ranking multi-year imagery and integrating soil information within a grouped validation framework can enhance predictive stability across spatial scales. Accordingly, the study pursues four primary objectives: (1) to construct a multi-temporal Sentinel-2–based SOM prediction framework using adaptive bare-soil extraction; (2) to implement and evaluate an image-ranking strategy based on predictive contribution; (3) to assess the added value of integrating categorical soil texture information with spectral features; and (4) to examine model generalization and performance transitions from farm to province scale.

2. Methodology

2.1. Modelling Framework Overview

This study presents a multi-scale and generalizable framework for predicting continuous soil attributes by integrating satellite-derived indices with zone-based categorical information. The framework is designed to operate under data-scarce conditions where only composite field-level soil samples and a corresponding zone map are available, making it suitable for both small-scale experiments and large-scale applications. Although soil organic matter is used as the primary case study in this work, the framework is generalizable to other continuous soil properties with comparable spatial behaviour. The modelling pipeline consists of three main components: 1) RS-based feature extraction using Google Earth Engine (GEE) and feature generation; 2) predictive modelling using spectral indices with and without categorical zone-level information; and 3) progressive evaluation of image informativeness to improve model performance. This modular design allows the framework to test the added value of legacy soil maps, quantify the role of image quality and temporal diversity, and evaluate its scalability from field-scale case studies to province-wide applications. The framework explicitly assumes the availability of categorical soil delineations of reasonable quality. While detailed legacy soil surveys can provide high local fidelity, the positional and thematic accuracy of soil maps typically decreases at broader spatial scales. This variability in map quality represents a fundamental source of uncertainty that propagates into zone-based modelling approaches. The progressive concept is used in two distinct roles: exploratorily at the farm scale (to quantify marginal image informativeness) and diagnostically at the province scale (training-only, out-of-fold (OOF)-based image cleaning), with a single leak-free test evaluation. After validating the framework at the farm scale, the same methodology was extended to a province-wide implementation to assess its generalization capacity across heterogeneous landscapes. Because large-scale modelling often exhibits weaker local correlations between spectral indices and soil properties, additional topographic (elevation, slope, aspect, etc.) and climatic (temperature, precipitation, etc.) predictors were incorporated only at the regional scale to represent spatial variability. Topographic attributes were chosen because terrain position strongly influences soil formation [47], water accumulation, and the redistribution of organic matter. In contrast, climatic variables capture long-term influences and moisture regimes that control decomposition rates, as well as the variability across the wide spatial scale of agro-climatic regions. Together, these predictor groups provide complementary, physically grounded information that cannot be reliably inferred from spectral indices alone, while remaining globally available and computationally scalable. This ensures that the framework remains both physically interpretable at the process level and transferable to other regions.

2.2. Remote Sensing Data Acquisition and Feature Generation

In the first component of the framework, satellite imagery was used to derive a set of spectral indices known to be informative for soil and vegetation analysis. A curated collection of Sentinel-2 surface reflectance images (COPERNICUS/S2_SR) was processed in GEE, with stringent cloud filtering applied for key agricultural periods. Imagery was restricted to April–June (spring, post-snowmelt) and September–December (post-harvest and residue-exposed periods), ensuring temporal consistency and minimizing vegetation interference. Sentinel-2 was selected because it provides high-quality Level-2A surface reflectance data, making it suitable for quantitative biophysical assessments. Its spatial resolution of up to 10 meters enables the detection of fine-scale soil surface variability and intra-field heterogeneity. Additionally,

Sentinel-2A&B constellation features a short revisit cycle (5 days) and a rich spectral configuration spanning 13 bands, including visible, red-edge, near-infrared (NIR), and short-wave infrared (SWIR) wavelengths. Together, these properties enable repeated observation of transient soil-exposure conditions rather than continuous spatial mapping.

For each image, multiple spectral indices, including vegetation metrics, bare soil indicators, and colour-based ratios, were computed and spatially aggregated within delineated field boundaries. These aggregations yielded image-level summaries (mean and standard deviation) that captured temporal variability in surface conditions rather than explicit spatial patterns. In addition to cloud masking (GEE default method) and to ensure only relevant surface conditions were analyzed, two critical masks were applied to the imagery: a snow mask and a bare-soil mask. The Normalized Difference Snow Index (NDSI) was used to identify and exclude snow-covered pixels, based on the reflectance contrast between green and short-wave infrared (SWIR) bands:

$$NDSI = \frac{B3 - B11}{B3 + B11} \tag{1}$$

where B3 corresponds to the green band and B11 to the SWIR band. Snow typically reflects visible light but absorbs SWIR, producing high NDSI values. Following ESA recommendations [48], pixels with NDSI > 0.42 were masked from further analysis. To isolate bare soil conditions, the Bare Soil Index (BSI) was applied:

$$BSI = \frac{(B11 + B4) - (B8 + B2)}{(B11 + B4) + (B8 + B2)} \tag{2}$$

where B2, B4, B8, and B11 are the blue, red, near-infrared (NIR), and SWIR bands, respectively. The BSI distinguishes exposed soil from vegetation and water using spectral brightness and moisture sensitivity. To account for field-level heterogeneity and varying acquisition conditions, a relative, field-adaptive thresholding strategy was employed. For each field, the mean BSI value was calculated, and pixels with BSI values exceeding this field-specific mean were retained as bare soil:

$$Bare\ Soil\ Mask = \begin{cases} 1\ if\ BSI_{pixel} > BSI_{mean\_field} \\ 0\ Otherwise \end{cases} \tag{3}$$

This approach prioritizes the barest soil fractions within each field and acquisition date rather than enforcing a fixed global threshold, ensuring consistent soil detection across heterogeneous surface conditions and imaging dates [49,50]. By defining bare-soil pixels relative to field-specific reflectance distributions, the method reduces sensitivity to absolute threshold selection and enhances robustness across agro-climatic gradients. Following masking, two groups of indices were calculated specifically over bare soil pixels. The first group focused on physical and chemical soil properties (Table 1), while the second captured vegetation or residue-related characteristics (Table 2).

Collectively, these indices provided a diverse set of predictors related to soil brightness, organic content, mineral composition, moisture, and vegetation status.

Table 1
Soil-related Indices calculated from Sentinel-2 surface reflectance.

Index Name	Formula/Equation	Description
Brightness Index (BI)	$\sqrt{\frac{B2^2 + B3^2 + B4^2}{3}}$	Measures the brightness of soil using visible bands.
Coloration Index (CI)	$\frac{B4 - B3}{B4 + B3}$	Distinguishes soil coloration based on the contrast between red and green bands.
Organic Matter Index (OMI)	$\frac{1}{B3^2}$	Soil-sensitive index derived from the green band (B3), evaluated empirically for SOM prediction.
Redness Index (RI)	$\frac{B4^2}{B2 \times B3^2}$	Highlights reddish soils, often linked to iron oxide content.
Soil Index (SI)	$\frac{B4 - B2}{B4 + B2}$	Enhances soil detection by highlighting contrast between red and blue bands.

Table 2
Vegetation-related Indices calculated from Sentinel-2 surface reflectance.

Index Name	Formula	Description
Normalized Difference Moisture Index (NDMI)	$\frac{B8 - B11}{B8 + B11}$	Reflects vegetation water content, utilizing near-infrared and short-wave infrared bands.
Normalized Difference Vegetation Index (NDVI)	$\frac{B8 - B4}{B8 + B4}$	Evaluates vegetation greenness and health by measuring the contrast between red and near-infrared bands.
Enhanced Vegetation Index (EVI)	$2.5 \times \frac{B8 - B4}{B8 + 6 \times B4 - 7.5 \times B2 + 1}$	Reduces atmospheric effects and improves sensitivity to vegetation.
Soil-Adjusted Vegetation Index (SAVI)	$\frac{B8 - B4}{B8 + B4 + 0.5} \times 1.5$	Enhances vegetation detection in areas with significant bare soil.

To represent broader environmental gradients that influence soil development, a set of topographic and climatic descriptors was prepared exclusively for the province-scale analysis, Table 3. Topographic variables were derived from the NASADEM HGT digital elevation model (NASA/NASADEM_HGT/001). Elevation, slope, and aspect were computed from the Digital Elevation Model (DEM) using standard geomorphometric operators and summarized for each field polygon using combined mean–standard-deviation statistics at the native 30 m resolution. These summaries describe local relief, surface inclination, and landscape orientation at the field scale. Climatic conditions were characterized using long-term bioclimatic normals from the WorldClim v1 BIO dataset (WORLDCLIM/V1/BIO), selected for consistency with established GEE implementations and because its 1 km resolution is sufficient for capturing regional climatic gradients relevant to SOM formation. Two variables, BIO1 (annual mean temperature, scaled by 10) and BIO12 (yearly precipitation), were aggregated over each field boundary at a 1 km target resolution using the same summary statistics. Given their coarse spatial resolution relative to Sentinel-2 imagery, these variables related to the climate are not intended to resolve within-field variability. Instead, they serve as contextual covariates capturing regional climatic gradients and long-term constraints on SOM formation. Because these layers represent multi-decadal climatic averages, the resulting descriptors are temporally invariant and remain constant across all Sentinel-2 acquisitions associated with a given field. Accordingly, topographic and climatic predictors were treated as static ancillary variables for regional-scale modelling, whereas the farm-scale analysis relied solely on spectral indices and categorical soil information.

2.3. Data Augmentation through Temporal Variation

To enrich the dataset and improve model generalization, temporal variability in satellite imagery was explicitly leveraged as a form of data augmentation. It was assumed that soil organic matter content remains relatively stable over short time periods, especially in mineral soils under consistent management [51,52]. This assumption enabled the use of multiple satellite images per field without requiring repeated field/-soil sampling. Each field was paired with several Sentinel-2 images, each providing a distinct snapshot of surface conditions. While the SOM label remained fixed within a given field, the satellite-derived features varied over time, effectively creating unique training instances. Importantly, this strategy does not imply temporal prediction of SOM change but rather exploits temporal diversity to better characterize the underlying soil signal. This strategy increased the size and diversity of the training dataset, capturing the variability introduced by changing surface states such as soil moisture, residue cover, and transient vegetation, thereby reducing overfitting risks and enhancing model robustness.

Table 3
Summary of topographic and climatic ancillary covariates used in the province-scale analysis.

Covariate Group	Variable Name	Source Dataset	Resolution	Description	Field-level Aggregation
Topography	Elevation	NASADEM HGT (NASA/NASADEM_HGT/001)	30 m	Represents terrain height and broad-scale relief	Mean, Standard Deviation
Topography	Slope	Derived from NASADEM	30 m	Indicates surface steepness and influences drainage and erosion	Mean, Standard Deviation
Topography	Aspect	Derived from NASADEM	30 m	Describes landscape orientation and potential exposure effects	Mean, Standard Deviation
Climate	BIO1 (Annual Mean Temperature)	WorldClim v1 – BIO	1 km	Long-term average temperature (scaled by 10), influencing decomposition and SOM stabilization	Mean, Standard Deviation
Climate	BIO12 (Annual Precipitation)	WorldClim v1 – BIO	1 km	Long-term annual precipitation controlling moisture regime	Mean, Standard Deviation

2.4. Informative Image Identification Using LASSO

To reduce dimensionality and identify the most informative predictors, the modelling framework employed the Least Absolute Shrinkage and Selection Operator (LASSO) regression. LASSO is a linear regression method that introduces L_1 regularization, which penalizes the absolute size of regression coefficients. As a result, it simultaneously performs variable selection and regularization by shrinking less important coefficients to exactly zero, thereby yielding a sparse model that retains only the most relevant features [53]. This property makes LASSO particularly suitable for high-dimensional datasets with multicollinearity, a frequent condition in remote sensing where numerous correlated indices are available. In this study, LASSO was used to select the most influential spectral index from the satellite-derived feature set for predicting the target value, here, the SOM. By filtering out redundant or irrelevant variables, LASSO enhances both the efficiency and interpretability of the subsequent modelling stages.

The response variable, SOM content, was log-transformed using a base-10 logarithm before model training [54,55].This transformation was applied to reduce skewness, stabilize variance, and mitigate the impact of extreme values in the data. By bringing the distribution closer to normality, the log transformation also helped satisfy the assumptions of linear regression models such as LASSO and facilitated more stable coefficient estimation. Overall, the application of LASSO served two key purposes: (1) improving prediction accuracy by focusing on the most relevant features, and (2) increasing interpretability by simplifying the model structure. These advantages align with the broader goal of precision soil mapping: developing data-efficient, transparent, and reproducible models. In this study, LASSO is not the final predictive model; it is used solely as a leak-free, linear screening tool to suppress multicollinearity and reduce the candidate set of image-derived indices before training the non-linear learners (RF/LightGBM/CatBoost). L_1-penalization yields sparse and reproducible subsets under group-aware cross-validation, which simplifies downstream modeling and improves interpretability without imposing linearity on the final predictor. This design separates feature screening (linear, transparent, fast) from function learning (non-linear, flexible), thereby retaining the capacity to model complex soil-spectral relationships while keeping the pipeline auditable and reproducible. Wrapper-based feature selection methods directly coupled to non-linear learners (e.g., recursive feature elimination with Random Forest) were intentionally not adopted. While potentially capable of capturing complex interactions, such approaches substantially increase computational cost, reduce reproducibility, and entangle feature selection with model-specific behavior. The adopted decoupled screening strategy therefore prioritizes stability, transparency, and scalability across modelling scenarios.

2.5. Integration of Cross-Validation in LASSO

To ensure reliable feature selection and prevent overfitting, cross-validation (CV) was incorporated into the LASSO regression process. The effectiveness of LASSO depends critically on the choice of the regularization parameter α (alpha), which controls the strength of coefficient penalization. Selecting an appropriate α balances model sparsity with predictive accuracy [56].

In this study, a five-fold group-based CV scheme was applied to identify the optimal α that minimized prediction error. Empirically, five-fold provides a strong compromise between variance and bias: they avoid the instability of leave-one-out CV while reducing the potential bias of fewer folds [57]. The dataset was partitioned into five equal subsets (folds defined by field groups); during each iteration, four folds were used for training and one for validation. The process was repeated five times so that each fold served as a validation set once, and the average error across folds provided a robust estimate of generalization performance under a grouped data structure. This grouped structure also prevents statistical pseudo-replication, since multiple temporal image records sharing the same SOM label are assigned to the same fold rather than being split across training and validation sets.

This integration of CV served three key purposes: 1) Hyperparameter tuning: it enabled the selection of an optimal regularization parameter that balances model complexity and predictive accuracy; 2) Generalization control: by evaluating performance across different data splits, it ensured that the selected features generalized well to unseen data; 3) Stability and robustness: averaging performance over folds reduced the variance in feature selection outcomes, especially in the presence of correlated predictors or noise. By embedding a group-aware CV directly into feature selection, the final set of spectral indices retained for downstream modelling was both parsimonious and robust to field-level dependence, providing a stable and reproducible basis for soil organic matter estimation.

2.6. Train–Test Structure and Prevention of Spatial and Temporal Leakage

To ensure unbiased evaluation and prevent spatial or temporal leakage, the dataset was partitioned into 80% training and 20% testing fields, based solely on FIELD_ID. All samples, soil zones, and image–field combinations associated with a given field were assigned entirely to either the training or test subset. This guarantees that: 1) no soil zone or polygon from a test field is ever observed during model development, 2) no temporal record from a test field contributes to image ranking or feature selection, and 3) all hyperparameter tuning and LASSO-based feature screening remain fully leak-free. All internal validation within the training set used GroupKFold cross-validation, with FIELD_ID as the grouping variable. This ensured that individual folds never share spatial information and that cross-validation reflects the model's ability to generalize across independent fields rather than across repeated observations of the same field. Within this structure, all model-building operations were performed exclusively on the training fields, including: 1) LASSO-based feature selection, 2) per-image OOF error profiling, 3) progressive image accumulation, 4) robust ΔR^2 filtering, and 5) Random Forest hyperparameter tuning.

The held-out test fields were evaluated once, after all modelling steps were finalized, using the final model trained on the cleaned training subset. This strict separation preserves the integrity of the evaluation and ensures that no information from the test fields influences any intermediate stage of the modelling pipeline.

2.7. Image-Based Feature Ranking

Building on the augmented Sentinel-2 dataset described earlier, this study introduces a modelling strategy that evaluates the predictive utility of individual satellite images for estimating a target variable. Each satellite image, previously linked to field-specific observations via multi-temporal augmentation, is treated as a distinct data source that captures unique spectral and surface conditions. The resulting image-field records, each corresponding to a specific field observed on a specific date, serve as the analytical basis for model training and performance evaluation. This formulation allows the informativeness of each image to be assessed both independently and in combination with others.

Following feature selection using LASSO regression with integrated cross-validation, each image was independently evaluated using only the training subset of fields to prevent leakage. Specifically, a model was trained using only the records associated with a given image, and the root-mean-square error (RMSE) of the predictions was computed under a group-aware validation structure. Images were then ranked by average RMSE to determine their relative informativeness. The progressive modelling framework iteratively accumulated images according to this ranking. For each top-ranked image subset, a Random Forest model was trained on the corresponding training records. At the same time, the held-out test fields (defined earlier in the 80/20 split) were used for final performance evaluation. This strategy (1) quantifies the incremental contribution of each image to predictive accuracy, and (2) identifies a subset of temporally diverse and reliable satellite acquisitions that provide the strongest foundation for SOM prediction. By emphasizing the informativeness of images rather than their quantity, the method enables more efficient, targeted remote-sensing workflows, particularly in data-constrained agricultural systems.

2.8. Progressive Modelling Approaches

The impact of input data type on model performance was assessed via two approaches (Fig. 1). Both share a common framework in which satellite images are ranked by informativeness and models are trained progressively on larger subsets. The key difference is the inclusion or exclusion of soil categorical data, here, soil texture, in the feature set. Notably, the LASSO step serves exclusively as a linear screening tool for feature/image selection; the final reported accuracies derive from non-linear ensemble models trained on these features. This ensures complex non-linearities and interactions are captured during final modelling, rather than being restricted by the linear selector. During progressive modelling, images were added sequentially in order of their ranking. At each iteration, models were trained only on the training fields, while the held-out test fields remained fixed.

2.8.1. Approach 1: Spectral Indices Only

This approach investigates how predictive accuracy improves as

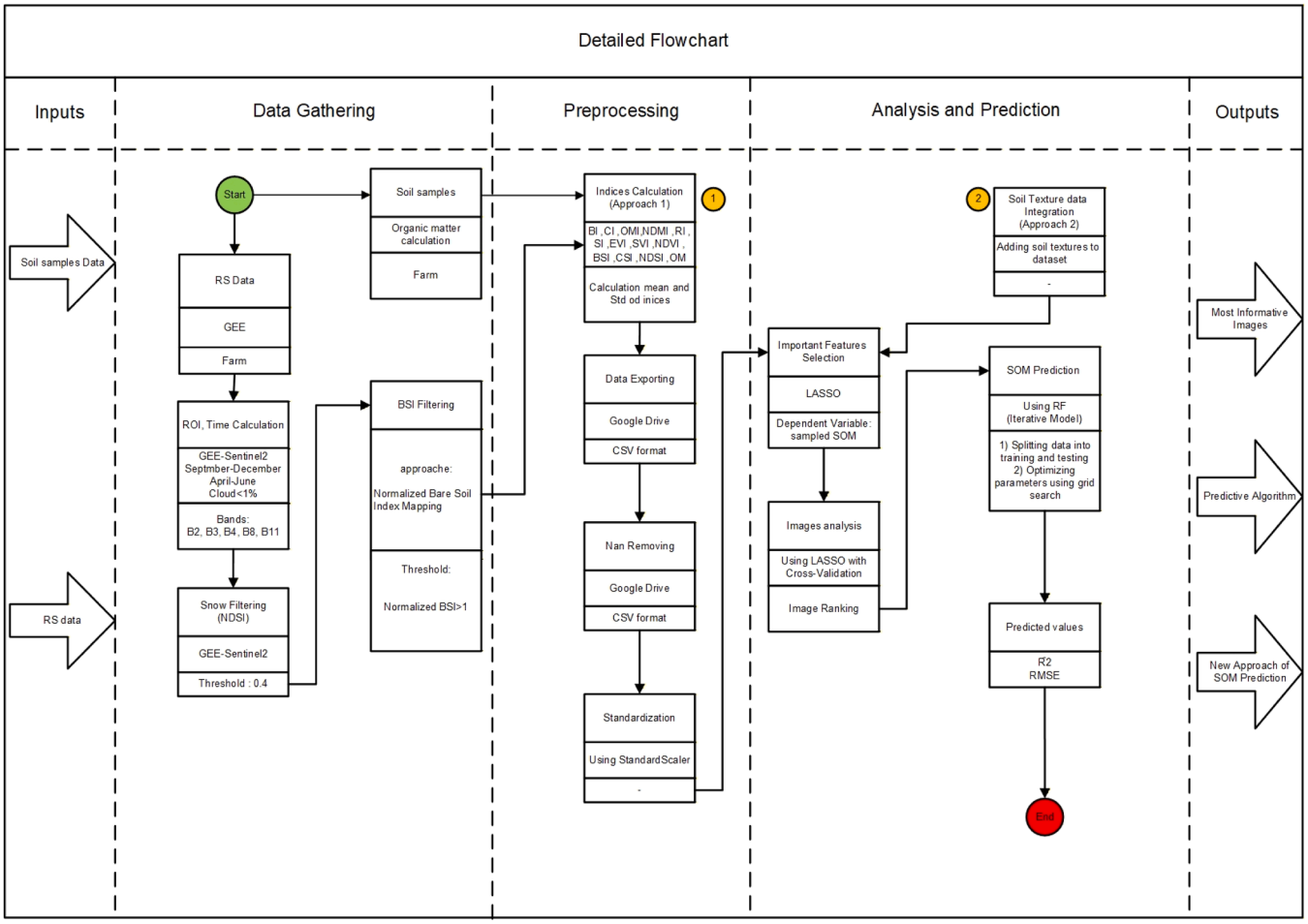

Fig. 1. Detailed flowchart of the method for predicting SOM (Farm Scale).

more informative satellite images are incorporated into the training data. The analysis begins by selecting features via LASSO regression with cross-validation, yielding a subset of spectral indices with nonzero coefficients. These selected features are used in all iterations of model training. An iterative, scenario-based framework is employed. In the first scenario, only the top-ranked image (based on training RMSE) is used for training. In the second scenario, the top two images are combined. This continues incrementally until all ranked images are included. For each scenario: 1) the dataset is filtered to include only records corresponding to the selected images; 2) the model is trained using a Random Forest Regressor on the filtered training set; and 3) the test set remains fixed across all iterations, consisting of a set of fields withheld entirely from the training process.

The objective of this approach is to quantify the marginal value of each image and determine how the progressive inclusion of high-ranked satellite observations improves target value prediction. By systematically prioritizing informative inputs, the method promotes efficient data use and enhances modelling outcomes in precision agriculture.

2.8.2. Approach 2: Integrating Soil Texture with Spectral Indices

In the second approach, spatially explicit categorical information, namely soil texture classes derived from legacy soil series maps, is incorporated alongside the satellite-derived spectral indices. This reflects a more realistic scenario in which both remote sensing data and zone-based field characteristics are available for modelling. After one-hot encoding the soil texture variable, the expanded feature set (spectral indices + soil Indices) is subjected to LASSO regression with cross-validation. This process identifies a refined set of relevant features from both data sources. The same iterative training procedure is then applied: 1) in each scenario, progressively ranked images are added to the training set; 2) the selected features (including soil texture if retained by LASSO) are used for model fitting; and 3) a fixed test set of unseen fields is used for evaluation. This approach aims to determine whether incorporating soil-specific categorical information improves predictive performance relative to using spectral data alone. It also provides insight into the relative contributions of spatial context and surface reflectance in predicting SOM at the field scale.

2.9. Machine Learning Algorithms for Province-Scale Modelling

The modelling workflow followed a stepwise structure designed to distinguish local, within-field behaviour from broader province-scale drivers. The process progressed from farm-scale exploratory modelling to the construction of hierarchical province-scale scenarios, with each stage clarifying how different groups of covariates should be incorporated into the overall framework.

1. Farm-scale exploratory phase: The modelling process began with farm-level experiments whose purpose was to characterize the roles of bare-soil spectral reflectance and soil texture zones in explaining within-field spatial patterns of SOM. This stage was used to understand the information content of Sentinel-2 indices under controlled, limited spatial heterogeneity and to determine whether legacy soil map delineations should be explicitly represented in subsequent large-scale formulations.
2. Establishing the province-scale baseline (S1): To ensure that province-scale modelling incorporated mapped pedological structure, the baseline scenario (S1) was defined as a combination of remote-sensing spectral indices and soil-texture categories. At this stage, RS-only formulations were not extended to the provincial context, as they had already served their exploratory role in the farm-scale phase of the workflow.
3. Progressive covariate enrichment (S2 and S3): Because SOM at regional scales is influenced by environmental gradients that extend beyond local soil units, two structured extensions were introduced to assess the contribution of additional covariate groups: 1) S2: incorporation of topographic descriptors (elevation, slope, aspect) to represent geomorphological variation across Québec; 2) S3: addition of climatic covariates (mean annual temperature and annual precipitation) to account for broader biophysical controls operating at regional scales. In this scenario, climate covariates primarily act as regional stratification variables rather than local predictors, allowing the model to adjust baseline SOM expectations across climatic zones while finer-scale variability is governed by soil texture, topography, and spectral information.
4. Final hierarchical structure: This stepwise design resulted in three province-scale scenarios, S1 (RS indices + Soil indices), S2 (S1 + Topography indices), and S3 (S2 + Climate indices), forming a hierarchical modelling framework for evaluating how different covariate groups contribute to SOM prediction at progressively larger spatial scales.

Three tree-based ensemble models were evaluated at the province scale: Random Forest (RF), LightGBM, and CatBoost (Table 4). To ensure comparability and prevent overfitting with limited samples per image, all models were trained using compact, conservative hyperparameter grids. This design prioritizes generalization and stability over aggressive optimization, reflecting realistic operational constraints in large-scale soil modelling. The two boosting models were included to assess whether sequential error-correction and enhanced handling of heterogeneous predictors could improve performance beyond that of bagging-based RF under province-wide conditions. However, boosting approaches can be more sensitive to hyperparameter settings, class imbalance in categorical encodings, and noise, which may reduce robustness under conservative tuning and highly heterogeneous regional data. The motivation for including these methods was to examine whether more recent boosting approaches, designed for scalability and heterogeneous data, can effectively enhance soil property prediction performance at larger spatial scales. All three algorithms were embedded within the same modelling pipeline:

1. Feature selection using LASSO regression with cross-validation to identify the most informative predictors from spectral indices and categorical soil variables.
2. Progressive image ranking was performed using OOF residuals from a group-aware Random Forest model trained with GroupKFold. Rather than training a separate model for each image, the code generated global OOF predictions and then computed RMSE per Image_ID by grouping residuals. Images were ranked from lowest to

Table 4
Key characteristics of LightGBM and CatBoost algorithms and their role in the present study.

Characteristic	LightGBM	CatBoost
Core principle	Histogram-based gradient boosting with leaf-wise tree growth	Ordered boosting with categorical feature handling
Handling of large datasets	Highly efficient due to histogram binning and leaf-wise split selection	Efficient, but optimized for smaller to medium-sized heterogeneous data
Categorical variables	Requires explicit one-hot or label encoding	Native treatment of categorical variables (ordered statistics)
Regularization	Supports L1/L2 regularization and constraints on leaves	L2 regularization built-in; robust to overfitting with smaller datasets
Training speed	Very fast on large, high-dimensional datasets	Slower than LightGBM but more stable with complex categorical interactions
Strengths	Scalability; good performance with large-scale, high-dimensional RS data	Robustness; reduces target leakage; improved accuracy in heterogeneous data
Limitations	May be sensitive to noisy categorical encodings	Longer training time; more computationally demanding

highest OOF RMSE to form the initial ordering. The ranking procedure was reproducible because the GroupKFold strategy is deterministic (no shuffling), and model stochasticity was controlled through fixed initialization parameters.

3. Group-aware data partitioning, with fields used as the grouping unit, to ensure spatial independence between training and testing subsets and to avoid information leakage.

4. Robust removal of detrimental images was implemented using a cumulative OOF-based evaluation. Images were progressively accumulated in ranked order, and for each cumulative set of top-k images, a new OOF prediction was computed to obtain the cumulative performance measure $R^2(k)$. This progressive ΔR^2-based filtering procedure was applied only at the province scale and was not part of the farm-scale workflow described in Fig. 1. The first-difference series $\Delta R^2(k) = R^2(k) - R^2(k-1)$ was then calculated, and robust z-scores of these differences were derived using the median and median absolute deviation (MAD). Unlike the earlier description, the robust filtering in the final code was applied to $\Delta R^2(k)$ rather than to raw R^2 or RMSE values, making the detection directly sensitive to harmful performance drops. Images yielding $Z < -2.5$ were flagged and removed. The robust z-score was computed using the median and median absolute deviation (MAD), scaled by 1.4826 to approximate standard deviation. A cutoff of -2.5 was selected as a conservative outlier-detection threshold to identify substantial negative ΔR^2 deviations while avoiding excessive removal of marginally informative images. This ΔR^2-based robust filtering step was used only at the province scale, where the larger and more heterogeneous image set required a formal detection rule. In the farm-scale case study, where fewer images were available and conditions were more homogeneous, a simple RMSE-based ranking was sufficient, and no ΔR^2-based filtering was applied. The robust z-score was calculated as:

$$Z_i = \frac{d_i - Median(d)}{1.4826 \times MAD(d)} \tag{4}$$

where d_i represents the successive difference in cumulative, $Median(d)$ is the median of all differences, $MAD(d)$ is the median absolute deviation, and the constant 1.4826 is used to make the scale consistent with the standard deviation under normality.

5. Performance evaluation was implemented twice: first in log-transformed space (log10(SOM)), and second in the original SOM scale using Duan's smearing factor. In the final pipeline, the smearing factor was computed from aggregated out-of-fold (OOF) residuals across all training folds for each progressive subset of clean images, and again for the final one-shot model trained on all cleaned images, ensuring unbiased and stable back-transformation. This adjustment allowed comparability across models and interpretability in the original units. For the farm-scale case, evaluation in log space alone was sufficient given the narrower SOM distribution. The smearing factor was defined using:

$$\widehat{S} = \frac{1}{n} \sum_{j=1}^{n} 10^{\epsilon_j} \tag{5}$$

where n is the total number of training observations, ϵ_j is the residual for observation j in log-transformed space, defined as:

$$\epsilon_j = y_{log,j} - \widehat{y}^{OOF}_{log,\ j} \tag{6}$$

where $y_{log,j}$ is the observed SOM in log space and $\widehat{y}^{OOF}_{log,\ j}$ is its out-of-fold prediction. 10^{ϵ_j} represents the back-transformed residual, converting the error from log space (base 10) back to the original scale.

Hyperparameter tuning was performed once per algorithm using group-aware cross-validation on the training set, with groups defined at the field level. The optimization was conducted under a fixed train/test partition and controlled model initialization to ensure reproducibility; no repeated multi-seed tuning was performed. For Random Forest, the tuning focused on tree depth and the number of estimators. For LightGBM, the search explored learning rate, number of leaves, maximum depth, and regularization terms. For CatBoost, tuning was performed on the number of boosting iterations, learning rate, tree depth, and L2 regularization. These compact yet representative grids ensured a fair comparison across algorithms without requiring repeated searches. For the final evaluation, models were trained on the cleaned image set and all training fields, then tested once on the held-out fields. Predictions in the original SOM scale were obtained using Duan's smearing correction to back-transform log-space outputs. To isolate the contribution of ancillary soil information, models trained on spectral indices alone were directly compared with those trained on spectral indices combined with soil-texture classes.

Model evaluation was conducted at two complementary levels: the image level, assessing the predictive informativeness of individual satellite acquisitions, and the field level, evaluating overall model accuracy for independent agricultural fields. This dual assessment enabled differentiation between temporal variability associated with image quality and spatial variability driven by soil and management differences. Following the comparative analysis of the three machine-learning algorithms, the best-performing model, which achieved the highest predictive accuracy and stability across both evaluation levels, was selected for further enhancement.

6. In the final stage, topographic (elevation, slope, aspect) and climatic (BIO1, BIO12) variables were incorporated into this optimal model to test whether integrating physically meaningful environmental predictors could improve its performance. These variables were chosen because terrain position governs soil moisture and organic-matter redistribution, while climatic gradients control temperature and precipitation regimes that influence organic-matter decomposition. By augmenting the best-performing algorithm with these complementary covariates, the framework sought to strengthen its predictive capacity and assess its generalizability under broader environmental conditions.

2.10. Model Evaluation and Comparison with Baseline

Model performance was assessed using two standard regression metrics: Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). RMSE quantifies the average magnitude of prediction error in the same scale as the response variable (log-transformed SOM), while R^2 reflects the proportion of variance explained by the model. To benchmark the results, a baseline model was implemented using field-level averaging. In this traditional method, a single SOM value is calculated for each field, typically as a composite or arithmetic average, and then assigned to all zones or sampling points within the field. This approach, while commonly used in agricultural practice, does not account for intra-field variability or spatial differentiation among zones.

$$RMSE_{base} = \sqrt[2]{\frac{\sum_{i=0}^{n} \left(Log(SOM)_{point} - \overline{Log(SOM)}_{Field}\right)^2}{n}} \tag{7}$$

where $Log(SOM)_{point}$ is the log-transformed SOM value measured at a sampling point within a given field, $\overline{Log(SOM)}_{Field}$ is the average of the log-transformed SOM values across all sampling points in that field, and n is the number of samples collected within the field.

For completeness, the formula for R^2 was also used to quantify the

proportion of explained variance in both the baseline and the proposed models:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \quad (8)$$

where x_i are observed log-transformed SOM values, $\hat{x}_i$ are predicted values, and $\bar{x}_i$ is the mean of the observed values.

The baseline serves as a reference point to assess whether the proposed model, which integrates satellite-derived indices and, optionally, soil zone information, offers meaningful improvements in predictive accuracy. By directly comparing both RMSE and R^2 values between the baseline and the modelling approaches, the analysis quantifies the added value of incorporating spatial and temporal information. This comparison enables a clear assessment of whether using remotely sensed data and feature selection strategies justifies the additional complexity introduced, especially in the context of site-specific soil management.

The farm-scale baseline provides a reference point to assess whether the proposed modelling approach captures meaningful within-field variation beyond a simple field-average estimate. At the province scale, no baseline averaging was used because the evaluation focuses on cross-field generalization, and assigning a single average SOM value to entire fields would not constitute a meaningful comparator. At the regional scale, however, constructing a single temporal composite from heterogeneous fields with asynchronous bare-soil exposure was considered inappropriate. A province-wide median composite would primarily encode broad spatial gradients rather than field-specific soil signals and therefore does not represent a meaningful baseline for the decision context of this study. By comparing RMSE and R^2 between the farm-scale baseline and the farm-scale modelling results, the analysis quantifies the added value of incorporating remotely sensed information and feature-selection strategies in site-specific soil management.

2.11. Study Site and Data Description

2.11.1. Farm-scale case study

The study was conducted at McGill University's Macdonald Campus Farm, located in Sainte-Anne-de-Bellevue, Québec, Canada. The site encompasses 30 agricultural fields totalling 193 ha and exhibits a wide range of soil types, from deep organic to mineral soils, including clay, sand, and intermediate textures. A detailed soil survey conducted in 1971, using the Canadian System of Soil Classification, documented 19 distinct soil series across the landscape [58]. In 2020, a 1-hectare center-point grid sampling design was implemented, yielding 184 soil samples from the 0–15 cm depth interval. Each sample was collected as a composite of 5-10 cores collected within a 3 × 3 m area. All samples were georeferenced and subsequently assigned to both a field and a soil-type zone based on the legacy soil series map and analyzed in the laboratory for SOM using the loss-on-ignition method. This grid-based approach ensured systematic coverage of all fields and soil zones while maintaining a practical sampling effort consistent with standard agronomic monitoring practices. The SOM measurements were log-transformed to mitigate skewness and stabilize variance. These transformed values are used as the response variable. The SOM dataset exhibited substantial variability (Table 5). Soil type, derived from the legacy map, served as the primary categorical predictor in the modelling framework. No independent quantitative assessment of positional or thematic accuracy of the legacy soil maps was conducted in this study. However, soil information was aggregated within field boundaries and used as categorical contextual variables, thereby reducing sensitivity to pixel-level boundary uncertainty. At the farm scale, the availability of a detailed, field-verified soil survey provides relatively high confidence in zone delineation. This setting represents a best-case scenario for assessing the potential of zone-based categorical information to support SOM modelling.

Table 5
Summary statistics of SOM values (raw and log-transformed).

Statistic	%SOM	log(%SOM)
Minimum	1.40	0.14
Maximum	61.00	1.78
Mean	7.29	0.73
Median	4.60	0.66
Standard Deviation	8.45	0.27

To complement the ground-based measurements, a multi-temporal collection of Sentinel-2 Level-2A surface reflectance (S2_SR) images was obtained from Google Earth Engine (GEE) and covered the 2019–2023 growing seasons. To ensure comparability across years, imagery was restricted to two consistent seasonal windows: April–June (spring, post-snowmelt bare-soil conditions) and September–December (post-harvest and residue-exposed periods) to maintain temporal consistency and maximize bare-soil detection. Images with less than 1% cloud cover were retained, and spectral indices were computed after applying snow and bare-soil masks as described. Image-derived features were spatially aggregated at the field level and paired with soil observations to construct the farm-scale modelling dataset (Fig. 2).

2.11.2. Province-scale dataset

To extend the analysis beyond a single farm and evaluate scalability, a larger dataset was compiled for the province of Québec. A total of 1412 soil samples were collected from 440 agricultural fields across the province during 2021-2022 (Fig. 3). Sampling points were georeferenced using an iPad GPS to ensure accurate positioning. Each soil sample was obtained as a composite of 8–10 soil cores taken to a depth of 17 cm (≈7 in.), thoroughly mixed in a bucket, and subsequently processed for laboratory analysis using the same loss-on-ignition (LOI) method applied in the farm-scale case study. For this province-wide dataset, a multi-temporal collection of Sentinel-2 Level-2A surface reflectance (S2_SR) imagery was acquired for the 2019–2023 growing seasons using the same Google Earth Engine (GEE) workflow, assuming short-term SOM stability over this multi-year window as justified in the Introduction [46]. As in the farm-scale analysis, imagery was restricted to the April–June and September–December windows to maximize bare-soil visibility and ensure temporal consistency across datasets. The same adaptive cloud, snow, and field-specific BSI filtering procedure described for the farm-scale workflow was consistently applied at the province scale. The same set of spectral indices used in the case study was computed to ensure comparability. Importantly, the farm-scale and province-scale datasets are both temporally and spatially independent, with no overlap in sampling locations or imagery coverage. Accordingly, results from the two analyses are interpreted separately, with the farm-scale case study exploring within-field behaviour under high-quality soil mapping, and the province-scale dataset used to evaluate generalization and robustness under heterogeneous regional conditions.

3. Results and Discussion

The primary objective of this study is to evaluate the predictive capability of different data configurations and image-selection strategies, rather than to generate a final spatial prediction map. While spatial visualization can be valuable in applied mapping contexts, such maps are only meaningful when prediction uncertainty and spatial resolution are well aligned with the decision scale. The focus here is therefore on model performance, generalization, and robustness across scales. Accordingly, results are assessed using quantitative validation metrics rather than visual inspection of mapped outputs.

3.1. Farm-scale performance

The SOM data showed substantial variability across fields and among soil types, as summarized in the descriptive statistics (Table 5) and

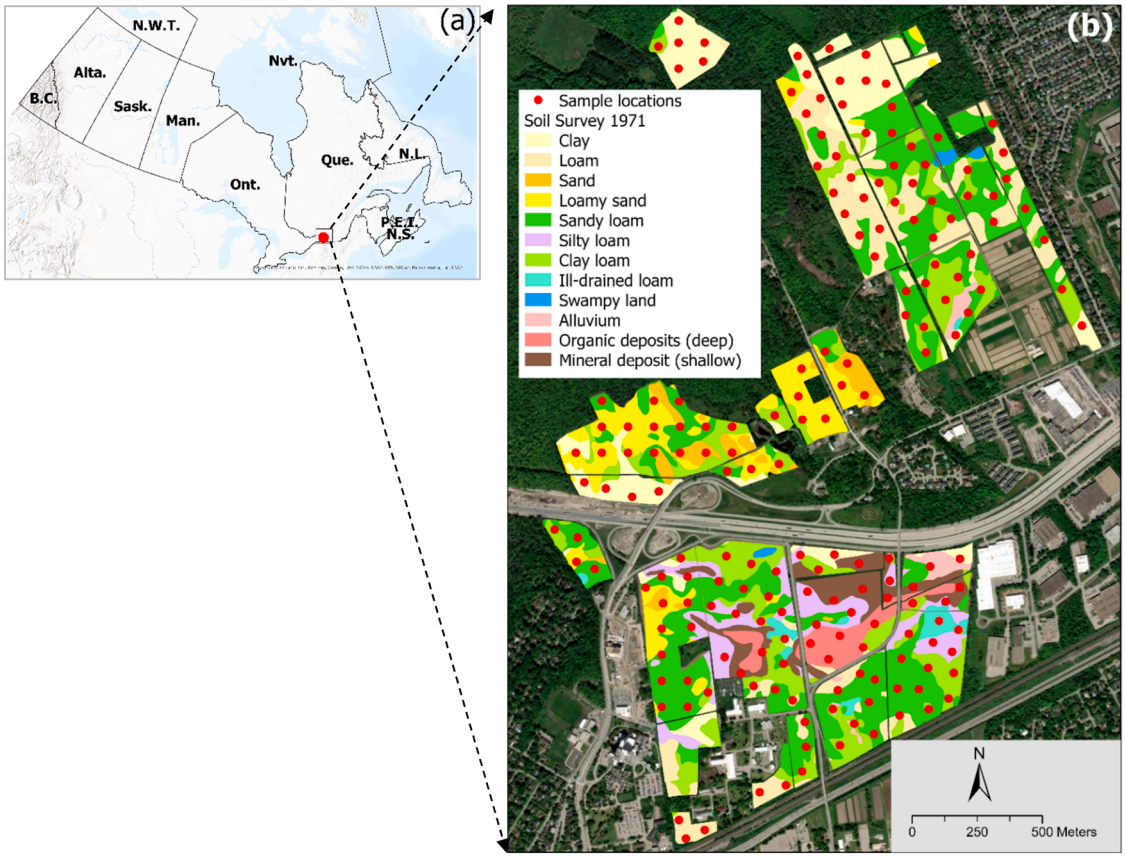

Fig. 2. (a) Map of Canadian provinces showing southern Quebec, where the field survey was conducted; and (b) map of agricultural fields showing soil sampling locations and the dominant loamy sand and sandy loam soils from the 1971 soil survey.

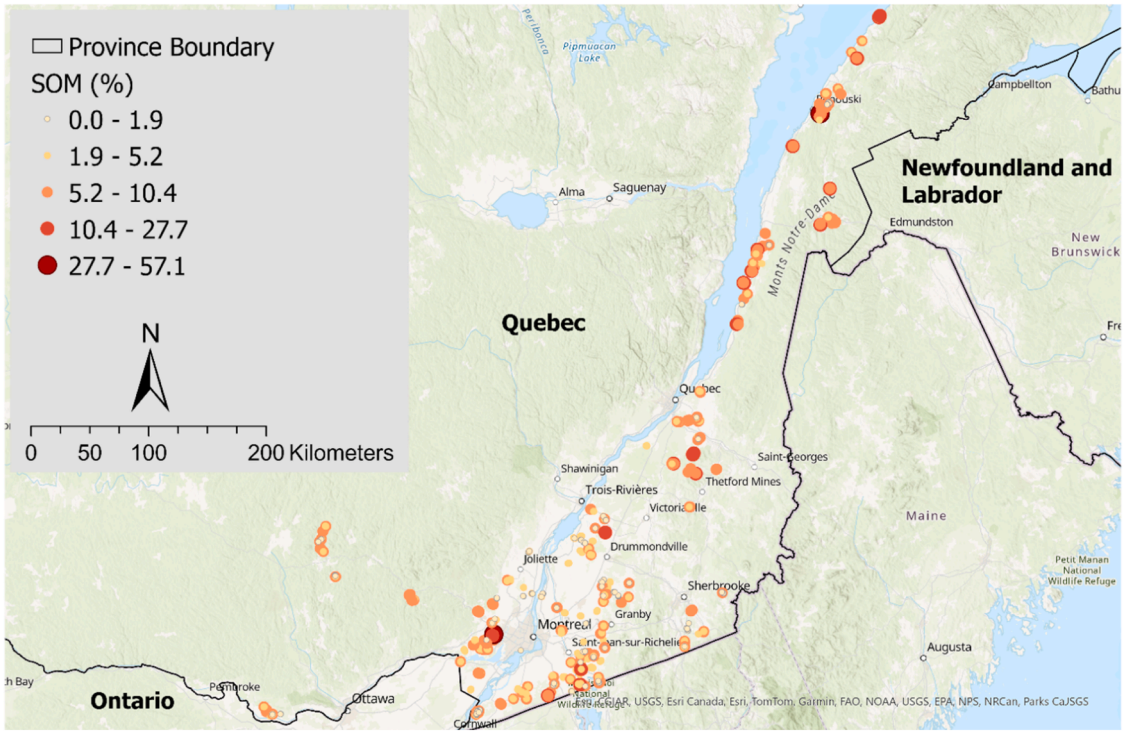

Fig. 3. Map of soil sampling locations in Quebec illustrating soil organic matter dynamics, with values ranging from 0% to 57%.

illustrated in the distribution plots (Figs. 4). To reduce skewness and stabilize variance, SOM values were log-transformed prior to analysis,

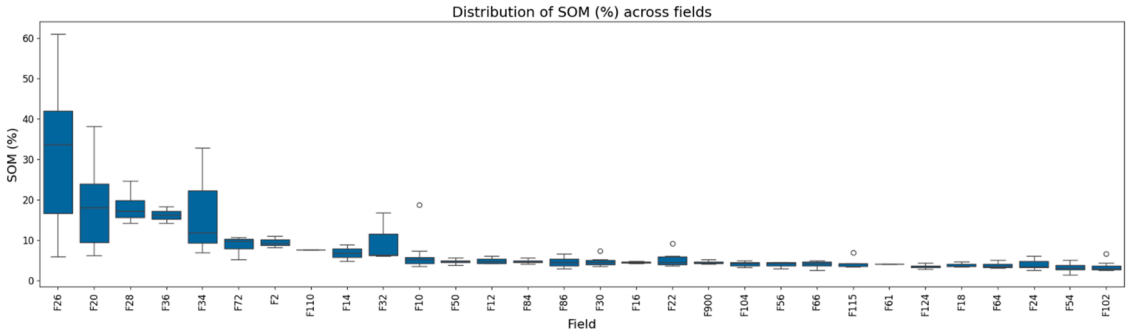

Fig. 4. Boxplot of measured SOM (%) values by field ID.

and the transformed values were used as the response variable. Soil type delineated from the legacy soil-series map served as the only categorical predictor in this phase.

A total of twelve cloud-free Sentinel-2 images produced 422 augmented field–image records across all sampled fields. To establish a benchmark, a simple field-level averaging model was constructed in which all soil samples within a field were assigned a single mean SOM value. This baseline, although producing a low error (RMSE = 0.14 log (%SOM); R^2 = 0.74), is inherently limited because it assumes spatial homogeneity within fields and collapses all soil zones to a constant value. In contrast, the proposed RS+soil model operated at the zone level by integrating spectral indices with categorical soil-texture classes derived from the legacy soil series map. This model achieved RMSE = 0.16 log(%SOM) with R^2 = 0.83. While the RMSE was slightly higher than the baseline, an expected outcome due to increased prediction granularity, the large improvement in R^2 demonstrates substantially stronger explanatory power and accurate differentiation among soil zones within the same field. Models relying solely on spectral indices performed poorly (RMSE ≈ 0.35–0.37 log(%SOM); $R^2 < 0.10$). These results confirm that multi-temporal remote-sensing signals alone are insufficient for representing subsurface SOM variation, and that zone-level soil delineations provide essential structural information.

To interpret these results correctly, it is important to recognize that the baseline and hybrid models operate at different conceptual levels. The baseline is evaluated at the field level and minimizes error by design, whereas the hybrid RS+soil model is evaluated at the zone level and explicitly resolves within-field variability. As a result, RMSE alone is not a sufficient indicator of model usefulness in this setting. Instead, R^2 provides a more meaningful measure of explanatory capacity by quantifying how well the model captures relative variability among zones. From a precision agriculture perspective, the ability to discriminate spatial patterns within fields is more informative for management decisions than minimizing point-wise error through spatial averaging.

To further support the quantitative evaluation, spatial maps were produced to visualize the observed field-mean SOM, the model-predicted SOM, and the associated prediction uncertainty. The observed and predicted SOM maps (Fig. 5) demonstrate a strong spatial correspondence at the field scale, confirming that the RS+soil model successfully captures the dominant spatial patterns of SOM across the study area, including areas with relatively higher SOM content. In addition, uncertainty maps (Fig. 6), quantified as the interquartile range (IQR) and standard deviation (STD) of predictions derived from multi-temporal Sentinel-2 imagery, identify fields with higher prediction variability, indicating locations of reduced model confidence. Together, these visual diagnostics provide an essential spatial interpretation of model behaviour, illustrating not only agreement between observations and predictions but also the spatial reliability and temporal stability of predictions. Such spatially explicit assessment complements RMSE and R^2 metrics and reinforces the practical relevance of the proposed approach for precision agriculture decision-making.

3.2. Province-scale performance

3.2.1. Baseline performance (Remote sensing+ Soil indices, Scenario S1)

Province-scale modelling combined multi-year Sentinel-2 indices with soil-texture categories across 440 agricultural fields and 1412 SOM samples. Under Scenario S1, which included only spectral indices and soil-texture categories, Random Forest achieved moderate but stable performance. At the image level, the model obtained RMSE = 0.1135 log (%SOM) and R^2_log = 0.131, while aggregation at the field level improved performance to RMSE = 0.1116 log(%SOM) and R^2_log = 0.218 (Table 6).

Fig. 5. The observed and predicted SOM maps.

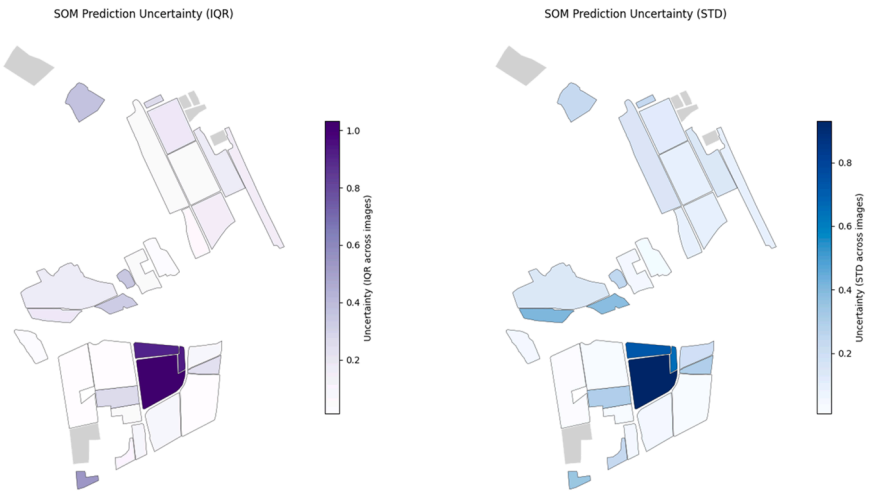

Fig. 6. Uncertainty maps.

Table 6
Performance of the three algorithms under RS + Soil (S1) at the province scale.

Algorithm	Level	RMSE log(% SOM)	R^2 log(% SOM)	RMSE_linear	R^2_linear
Random Forest	Image	0.1135	0.131	1.245	0.11
Random Forest	Field	0.1116	0.218	1.287	0.211
LightGBM	Image	0.1188	0.05	1.317	0.003
LightGBM	Field	0.1116	0.217	1.293	0.203
CatBoost	Image	0.1167	0.083	1.296	0.035
CatBoost	Field	0.1145	0.176	1.333	0.153

These values are lower than those observed at the farm scale, which is expected given the substantial increase in spatial extent, pedological diversity, and management heterogeneity across Québec. At this broader scale, many sources of spectral variability (e.g., moisture, crop residue, roughness, and tillage timing) are only partially related to SOM, making optical reflectance a noisy and indirect proxy for subsurface carbon content. Random Forest consistently outperformed LightGBM and CatBoost under S1, particularly at the image level, suggesting that RF was more robust to heterogeneous image–field combinations and noisy spectral observations in this baseline configuration. This performance gap does not imply an inherent superiority of RF over boosting algorithms. Rather, it reflects the interaction between model structure, conservative hyperparameter tuning, and the nature of the input data. In this study, province-scale modelling involves heterogeneous fields, noisy categorical soil information, and relatively limited effective sample sizes per image. Under these conditions, RF's bagging-based aggregation appears less sensitive to local noise and misclassification, whereas boosting methods may amplify unstable patterns when aggressive regularization is intentionally avoided.

3.2.2. Province-scale progressive covariate enrichment (Scenarios S1 → S2 → S3)

Because Random Forest provided the strongest and most stable baseline, it was used to evaluate whether adding environmental covariates improves SOM predictability at large scale. Incorporating topographic attributes (Scenario S2) produced a modest but systematic improvement over S1, increasing image-level R^2_log from 0.131 → 0.137 and field-level R^2_log from 0.218 → 0.239 (Table 7). These improvements reflect the influence of slope and elevation on water redistribution, drainage, and organic matter accumulation across Québec. The largest improvement occurred when both topography and climate were added (Scenario S3). Under this full configuration, Random Forest achieved: Image level: RMSE = 0.1029 log(%SOM), R^2_log = 0.287, and Field level: RMSE = 0.1006 log(%SOM), R^2_log = 0.364. This represents a ~9% reduction in RMSE log(%SOM) relative to S1 and more than a doubling of R^2_log at the image level (0.131 → 0.287). At the field level, where noise is reduced through aggregation, the model reached R^2_log = 0.364, the strongest performance observed in the study. This improvement reflects the ability of coarse-resolution climate variables to capture broad geographic clustering and regional environmental gradients, rather than an enhancement of within-field spatial detail.

Table 7
Performance of the three scenarios using RF.

Scenario	Level	RMSE log (%SOM)	R^2 log(% SOM)	RMSE_linear	R^2_linear
S1: RS + Soil	Image	0.1135	0.131	1.245	0.11
S1: RS + Soil	Field	0.1116	0.218	1.287	0.211
S2 = S1 + Topography	Image	0.1132	0.137	1.214	0.153
S2 = S1 + Topography	Field	0.11	0.239	1.245	0.261
S3 = S2 + Climate	Image	0.1029	0.287	1.133	0.262
S3 = S2 + Climate	Field	0.1006	0.364	1.151	0.368

Although the R^2 values are moderate, such reductions are expected when scaling from farm-level modelling to province-wide prediction, given increased spatial heterogeneity and unobserved drivers. In large-scale environmental modelling, RMSE and predictive stability across folds often provide a more meaningful indicator of practical reliability than variance explanation alone. When interpreted in this context, the observed error magnitude supports the framework's applicability for field-level ranking and regional stratification. Achieving this level of explanatory power at the provincial scale, using only freely available remote-sensing data, legacy soil maps, and coarse environmental covariates, demonstrates the practical transferability of the framework under realistic data constraints. The performance gains observed in S3 further confirm that SOM variability at large scales is driven primarily by long-term ecological gradients rather than instantaneous surface reflectance. Climate and topography, therefore, provide essential complementary information that remote sensing alone cannot capture, making the improved accuracy under S3 a more realistic representation of landscape-level soil-forming processes. From a decision-support perspective, the framework is not intended to provide precise point estimates of SOM. Rather, it supports relative differentiation among fields

and zones, which is sufficient for applications such as stratified management, sampling prioritization, and regional-scale agronomic planning.

3.2.3. Cross-validation stability and error behaviour across scenarios

Beyond average accuracy metrics, model robustness and stability across cross-validation folds are critical for evaluating large-scale soil prediction frameworks. To assess the consistency of model performance under different covariate configurations, the distribution of cross-validation RMSE values across folds was compared for Scenarios S1, S2, and S3 (Fig. 7).

The boxplots reveal a clear reduction in both median RMSE and interquartile range as additional environmental covariates are incorporated. Scenario S1 (spectral indices + soil texture) exhibits the largest spread, indicating higher sensitivity to fold composition and greater instability under heterogeneous conditions. Incorporating topographic variables in S2 reduces both central error and variability, while the full configuration (S3) shows the lowest RMSE values and the narrowest dispersion across folds.

This trend is further reinforced by the distribution of individual cross-validation realizations. The tighter clustering of individual fold RMSE values (grey points) in S3, compared with the more scattered pattern observed in S1, indicates improved stability and reduced sensitivity to fold-specific data composition. This progressive tightening of the RMSE distribution demonstrates that adding topographic and climatic information does not merely improve average predictive accuracy but also enhances model stability and generalization. In the context of province-scale modelling, where training data span diverse soil–climate regimes, this reduction in variance across folds is as important as gains in mean performance.

Complementary insight is provided by examining the distribution of prediction errors on the held-out test fields (Fig. 8). Kernel density estimates of log-space prediction errors show that all scenarios are broadly centred around zero, indicating limited systematic bias after model calibration. However, a slight rightward shift in the error peaks is observed, suggesting a weak tendency toward overestimation of log-transformed SOM values. This slight positive bias is most pronounced in Scenario S1 and progressively diminishes as additional environmental covariates are included, reaching its minimum extent under Scenario S3.

In addition to biased behaviour, marked differences are observed in dispersion and tail behaviour. Scenario S1 displays the widest error distribution, with heavier tails reflecting sensitivity to image-specific noise and unmodelled environmental gradients. Scenario S2 shows a modest contraction of the error spread, whereas Scenario S3 exhibits the most concentrated distribution around zero, with reduced variance and fewer extreme errors.

Notably, the narrowing of the error distribution under S3 indicates

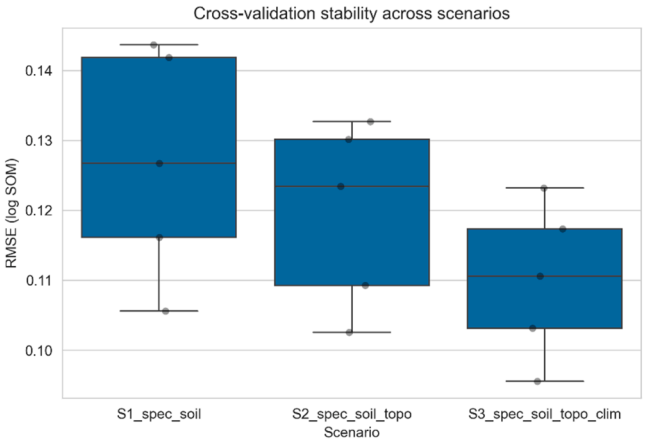

Fig. 7. Cross-validation stability across modelling scenarios.

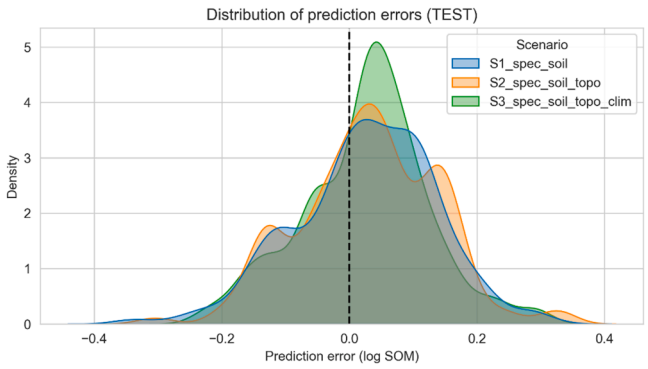

Fig. 8. Distribution of prediction errors on the independent test set.

that performance gains are not driven by a small subset of well-predicted fields, but rather by a global reduction in uncertainty across the test set. This behaviour confirms that the inclusion of climatic covariates primarily improves large-scale consistency rather than fine-scale point accuracy, aligning with the interpretation that climate acts as a regional stratification variable rather than a local predictor.

3.2.4. Progressive image accumulation: identifying the optimal number of clean images

Province-scale modelling relied on a ranked set of 78 "clean" Sentinel-2 images, identified using the LASSO-based, training-only screening procedure and the ΔR^2-based robust filtering. Progressive modelling curves (Fig. 9) reveal a consistent pattern:

1) When only a few images are used, the model is unstable because each acquisition reflects surface conditions.
2) Performance improves rapidly between approximately 10 and 25 images, where temporal diversity becomes sufficient to smooth out transient moisture and residue effects.
3) Beyond about 45–55 images, the gains plateau and additional images contribute only marginal improvements.
4) The best performance in Scenario S3 is obtained with roughly 47 images, near the onset of the plateau.

This behaviour indicates that SOM prediction is not driven by a single "optimal" image or narrow seasonal window. Inspection of the temporal distribution of these high-ranking images indicates that most were acquired during early spring and late autumn, when vegetation cover is minimal, and soil exposure is highest. Mid-summer acquisitions, typically dominated by dense vegetation, were rarely retained among the top-ranked images.

Rather than relying on any individual image, the model benefits from aggregating multiple clean acquisitions that collectively sample a wide range of surface states. Each image provides a partial, noisy view of the underlying soil condition, influenced by transient factors such as moisture, residue, roughness, and viewing geometry. By progressively combining images, the framework effectively suppresses these transient effects and reinforces the persistent soil signal associated with SOM. This result provides a mechanistic explanation for why single-date or narrow-window approaches often underperform in regional SOM mapping and highlights the importance of multi-temporal strategies for robust digital soil assessment.

3.2.5. Field-level prediction behaviour and sources of uncertainty

Field-level parity plots (Fig. 10) show that the Random Forest model under Scenario S3 reasonably captures the overall SOM gradient across fields but exhibits systematic, scale-dependent biases. High-SOM fields, often associated with organic-rich or poorly drained soils, tend to be underpredicted, whereas coarse-textured, low-SOM fields are mildly overpredicted. These patterns reflect both data limitations and intrinsic

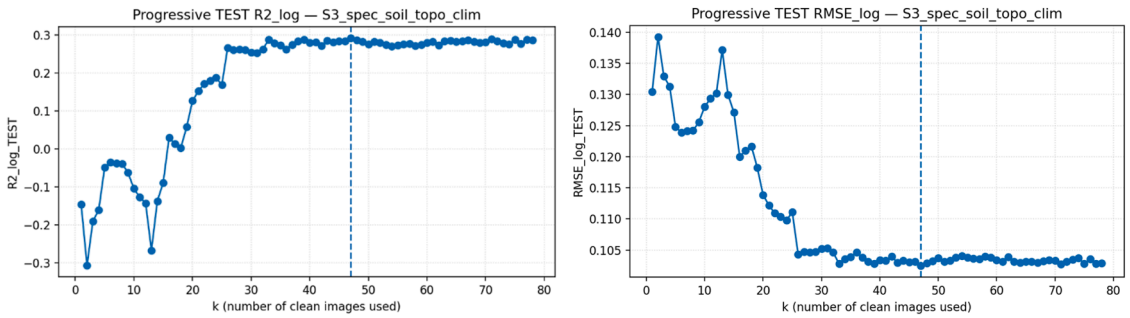

Fig. 9. Progressive image accumulation (Image Level Prediction).

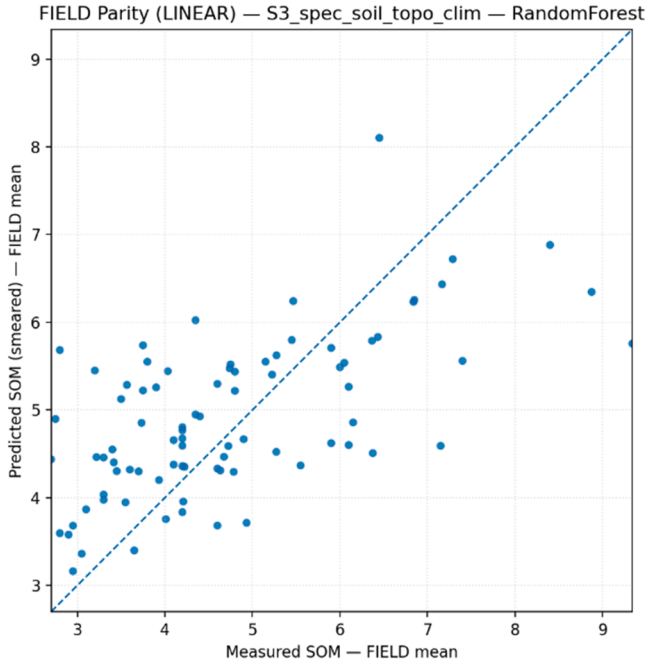

Fig. 10. Predicted vs Measured SOM.

ambiguities in large-scale soil prediction, rather than solely random model error. Several factors contribute to these patterns:

1) Uncertainty in legacy soil maps. The province-level soil series map was not explicitly designed for quantitative SOM prediction. At this scale, soil unit boundaries are necessarily generalized, and positional mismatches between mapped polygons and sampling locations are more likely. Such misclassification introduces noise into the categorical predictors and weakens the statistical link between soil classes and observed SOM. This effect represents a structural limitation of zone-based modelling approaches that rely on legacy soil information.
2) Spectral ambiguity under bare-soil conditions. Dark organic soils and moist mineral soils can produce similar reflectance signatures, especially in the visible and near-infrared bands. Even with SWIR-based indices and strict bare-soil masking, complete separation of organic matter effects from moisture and residue effects is not achievable, leading to uncertainty in residual prediction at the field level.
3) Missing management information. Tillage regime, crop rotation, residue management, and manure application history are not included in the covariate set. These management factors can substantially affect SOM and surface conditions, but they are treated as unobserved sources of variability in the current framework.

Despite these limitations, the model preserves broad class separability at the field level, enabling consistent differentiation among low-, medium-, and high-SOM fields. This level of relative ordering is sufficient for many decision-support workflows, including stratified soil sampling, prioritization of soil testing, and preliminary delineation of management zones, where ranking stability is more critical than minimizing absolute prediction error.

From an uncertainty perspective, the combination of group-aware cross-validation, out-of-fold residual analysis, and Duan's smearing correction mitigates over-optimism in performance estimates and yields realistic error metrics in both log-transformed and original SOM units. Nevertheless, the current framework reports global summary statistics rather than spatially explicit uncertainty measures. Extending the approach to include prediction intervals or probabilistic outputs, such as quantile regression forests or ensemble-based variance estimates, represents a logical next step toward risk-aware and operational soil management applications.

3.2.6. *Comparison with previous studies*

The performance observed in this study is broadly consistent with, and in several respects complementary to, findings from previous SOM mapping efforts using multispectral and synthetic bare-soil imagery. At the farm scale, performance ($R^2 = 0.83$ at the zone level) falls within the upper range of values reported under high-quality soil delineations and controlled conditions. Studies using Sentinel-2 or Landsat-based indices together with detailed local covariates have typically reported strong within-field fits, especially when SOM variation is large, and sampling density is high [59–61]. In this context, recovering fine-scale variability using only a single legacy soil map, multi-temporal imagery, and a limited number of samples per field highlights the structural value of zone-level categorical information, rather than relying on dense sensing alone.

At regional scales, several authors have reported that multispectral SOM prediction rarely exceeds moderate explanatory power, particularly when models are calibrated across diverse soil types and land uses. Reported R^2 values for Sentinel-2–based SOM or SOC mapping at landscape to regional scales often range from low (≈0.1) to moderate (≈0.4), depending on the richness of auxiliary data and the strength of moisture and residue artifacts [62–64]. In comparison, the province-scale results under Scenario S3 (R^2_log = 0.287 at the image level and 0.364 at the field level) place this framework near the upper portion of that spectrum, despite relying only on freely available imagery, a legacy soil map, and coarse topographic and climatic layers. This suggests that the combination of LASSO-based image ranking, robust image cleaning, and explicit integration of soil texture, topography, and climate is at least competitive with, and in some cases more efficient than, approaches based solely on synthetic bare-soil composites or single-date acquisitions. The fact that downstream Random Forest and boosting models consistently improved with LASSO-selected images suggests that the screening step did not suppress critical non-linear signal but rather removed redundant or noisy acquisitions that would otherwise degrade

model stability.

The relative advantage of Random Forest observed here is consistent with findings from other regional-scale soil and environmental studies, where RF often outperforms boosting models under noisy or weakly informative predictor sets. In such contexts, the variance-reduction mechanism of bagging can be more effective than sequential error correction, particularly when categorical predictors carry positional uncertainty, as is the case with legacy soil maps.

An additional contributor to the reduced explanatory power at the provincial scale is the variable quality of legacy soil maps. While such maps provide valuable categorical structure, they were not originally designed for quantitative digital soil modelling. Errors in boundary placement and thematic classification can attenuate model performance, particularly when predictions are evaluated at the field level. Rather than invalidating the approach, this limitation defines the operational envelope of zone-based frameworks: achievable accuracy is bounded by the fidelity of the underlying soil delineation.

Another distinction from prior work lies in the explicit quantification of image informativeness. Many studies implicitly assume that carefully chosen dates or seasonal windows are sufficient to obtain representative bare-soil conditions [65]. In contrast, the present framework treats each image as a candidate information source, ranks images based on out-of-fold performance, and discards those that demonstrably harm the model. This design directly links the quality of multi-temporal imagery to prediction accuracy and provides a transferable recipe for practitioners working with heterogeneous archives and limited ground truth.

4. Conclusion

This study developed a scalable and interpretable framework for predicting soil organic matter (SOM) by integrating multi-temporal Sentinel-2 imagery, legacy soil maps, and structured image ranking. At the farm scale, combining spectral indices with soil categorical information resulted in strong predictive performance (RMSE = 0.16 log (%SOM), $R^2 = 0.83$), demonstrating that fine-scale SOM variability can be reliably recovered when local soil-type structure is explicitly incorporated. At the province scale, model performance improved substantially once topographic and climatic covariates were added: under the full-feature configuration (Scenario S3), the Random Forest model achieved RMSE = 0.1029 log(%SOM) and R^2_log = 0.287 at the image level, and RMSE = 0.1006 log(%SOM) and R^2_log = 0.364 at the field level. Although an R^2 of approximately 0.36 at the field level is moderate in absolute terms, its value is best interpreted in a decision-oriented rather than purely statistical context. At regional scales, SOM variability is influenced by partially unobserved drivers such as long-term management history, drainage modification, and manure application, as well as uncertainties in legacy soil delineations. These factors impose an upper bound on the achievable explanatory power when relying exclusively on open and spatially consistent covariates. Within this constraint, explaining roughly one-third of the variance remains informative because many agronomic workflows rely on relative spatial differentiation, such as separating low-, medium-, and high-SOM fields for targeted sampling, amendment prioritization, or preliminary management zone delineation, rather than on exact absolute SOM estimates. Importantly, the decoupled use of linear screening and non-linear prediction provides a transparent and computationally tractable pathway for large-scale SOM modelling.

In practical terms, the achieved accuracy is sufficient for decision-support tasks that rely on relative spatial differentiation, such as delineating management zones, prioritizing sampling locations, or guiding adaptive nutrient experiments, rather than for precise estimation of absolute SOM concentrations. Overall, the framework demonstrates both the potential and the conditional limitations of scalable SOM modelling, with its performance at broader scales constrained by the quality of available legacy soil delineations. It generalizes effectively beyond individual farms while remaining interpretable. Yet its explanatory capacity at broader scales is constrained by uncertainties in legacy soil maps, unobserved management practices, and residual environmental gradients.

By explicitly ranking and filtering multi-temporal satellite images, and by integrating soil texture, topography, and climate, the framework provides a transparent recipe for constructing SOM prediction models that are both data-efficient and operationally realistic. In this sense, it offers a bridge between local case studies and province-scale digital soil assessments built entirely on open data.

Future research should focus on three main directions. First, systematic offset-based correction strategies could be introduced to reduce persistent bias after model fitting. Second, a residual-informed adaptive correction layer could be developed to dynamically model structured prediction errors using flexible learners. Third, extending the framework toward probabilistic mapping with spatially explicit uncertainty estimates and region-stratified modelling would enhance its robustness and usefulness for risk-aware decision-making. Collectively, these extensions would strengthen the framework as a practical foundation for multi-scale digital soil assessment across heterogeneous agricultural landscapes.

Data availability statement

The data used in this study are confidential and cannot be publicly shared due to privacy and contractual restrictions. Access may be granted upon reasonable request and with appropriate authorization from the data provider.

Statement on the Use of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Grammarly and WordTune to enhance academic phrasing and improve the fluency of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Ethical Statement

This manuscript complies with Elsevier's Publishing Ethics Policy.

All authors confirm that

- ✓ The manuscript is original, has not been published before, and is not under consideration elsewhere.
- ✓ All authors have contributed substantially to the conception, data collection, analysis, drafting, or critical revision, and approve the final version.
- ✓ Proper acknowledgment has been given to any work, data, or ideas of others through citations.
- ✓ There is no conflict of interest or competing financial interest to declare.
- ✓ if accepted, the article will not be published elsewhere in the same form, in English or in any other language, including electronically, without the written consent of the copyright-holder.

CRediT authorship contribution statement

Hamed Etezadi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Yacine Bouroubi:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Viacheslav Adamchuk:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition. **Maxime Leduc:** Writing – review & editing, Funding acquisition, Data curation. **Marc-Olivier Gasser:** Writing – review & editing,

Methodology, Data curation. **Md Saifuzzaman:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partly supported through the Quebec Research Fund – Nature and Technologies (FRQNT) Partnership Research Program – Sustainable Agriculture (RQRAD).

References

[1] A.B. McBratney, M.M. Santos, B. Minasny, On digital soil mapping, Geoderma 117 (1-2) (2003) 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4.
[2] S. Rezapour, S. Siavash Moghaddam, A. Nouri, K. Khosravi Aqdam, Urbanization influences the distribution, enrichment, and ecological health risk of heavy metals in croplands, Scientific Reports 12 (1) (2022) 3868.
[3] K.K. Aqdam, S. Rezapour, F. Asadzadeh, A. Nouri, An integrated approach for estimating soil health: Incorporating digital elevation models and remote sensing of vegetation, Computers and Electronics in Agriculture 210 (2023) 107922.
[4] F. Castaldi, A. Hueni, S. Chabrillat, K. Ward, G. Buttafuoco, B. Bomans, K. Vreys, M. Brell, B. van Wesemael, Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands, ISPRS Journal of Photogrammetry and Remote Sensing 147 (2019) 267–282, https://doi.org/10.1016/j.isprsjprs.2018.11.026.
[5] X. Xu, X. Zhai, Mapping Soil Organic Matter Content during the Bare Soil Period by Using Satellite Data and an Improved Deep Learning Network, Sustainability 15 (1) (2022) 323, https://doi.org/10.3390/su15010323.
[6] M. Zhang, M. Zhang, H. Yang, Y. Jin, X. Zhang, H. Liu, Mapping regional soil organic matter based on sentinel-2a and modis imagery using machine learning algorithms and google earth engine, Remote Sensing 13 (15) (2021) 2934, https://doi.org/10.3390/rs13152934.
[7] P. Krishnan, J.D. Alexander, B. Butler, J.W. Hummel, Reflectance technique for predicting soil organic matter, Soil Science Society of America Journal 44 (6) (1980) 1282–1285, https://doi.org/10.2136/sssaj1980.03615995004400060030x.
[8] H. Liu, Y. Zhang, B. Zhang, Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China, Environmental monitoring and assessment 154 (2009) 147–154, https://doi.org/10.1007/s10661-008-0385-4.
[9] X. Wang, F. Zhang, V.C. Johnson, New methods for improving the remote sensing estimation of soil organic matter content (SOMC) in the Ebinur Lake Wetland National Nature Reserve (ELWNNR) in northwest China, Remote Sensing of Environment 218 (2018) 104–118, https://doi.org/10.1016/j.rse.2018.09.020.
[10] E.S. Mohamed, A.A.E. Baroudy, T. El-Beshbeshy, M. Emam, A. Belal, A. Elfadaly, A. A. Aldosari, A.M. Ali, R. Lasaponara, Vis-nir spectroscopy and satellite landsat-8 oli data to map soil nutrients in arid conditions: A case study of the northwest coast of egypt, Remote Sensing 12 (22) (2020) 3716, https://doi.org/10.3390/rs12223716.
[11] C. Luo, Y. Wang, X. Zhang, W. Zhang, H. Liu, Spatial prediction of soil organic matter content using multiyear synthetic images and partitioning algorithms, Catena 211 (2022) 106023, https://doi.org/10.1016/j.catena.2022.106023.
[12] Y. Guo, W.-J. Ji, H.-H. Wu, Z. Shi, Estimation and mapping of soil organic matter based on Vis-NIR reflectance spectroscopy, Spectroscopy and spectral analysis 33 (4) (2013) 1135–1140, https://doi.org/10.3964/j.issn.1000-0593(2013)04-1135-06.
[13] X. Xu, Y. Chen, M. Wang, S. Wang, K. Li, Y. Li, Improving estimates of soil salt content by using two-date image spectral changes in Yinbei, China, Remote Sensing 13 (20) (2021) 4165, https://doi.org/10.3390/rs13204165.
[14] F. Castaldi, S. Chabrillat, A. Don, B. van Wesemael, Soil organic carbon mapping using LUCAS topsoil database and Sentinel-2 data: An approach to reduce soil moisture and crop residue effects, Remote Sensing 11 (18) (2019) 2121, https://doi.org/10.3390/rs11182121.
[15] A. Gholizadeh, D. Žižala, M. Saberioon, L. Borůvka, Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging, Remote Sensing of Environment 218 (2018) 89–103, https://doi.org/10.1016/j.rse.2018.09.015.
[16] K.K. Aqdam, F. Asadzadeh, S. Rezapour, A. Nouri, F. Shabani, An integrated soil health and machine learning framework for quantifying soil degradation in semi-arid agricultural lands, Soil and Tillage Research 259 (2026) 107099.
[17] L. Guo, H. Zhang, T. Shi, Y. Chen, Q. Jiang, M. Linderman, Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images, Geoderma 337 (2019) 32–41, https://doi.org/10.1016/j.geoderma.2018.09.003.
[18] M.J. Hill, Vegetation index suites as indicators of vegetation state in grassland and savanna: An analysis with simulated SENTINEL 2 data for a North American transect, Remote Sensing of Environment 137 (2013) 94–111, https://doi.org/10.1016/j.rse.2013.06.004.
[19] G.S. Bhunia, P. Kumar Shit, H.R. Pourghasemi, Soil organic carbon mapping using remote sensing techniques and multivariate regression model, Geocarto International 34 (2) (2019) 215–226, https://doi.org/10.1080/10106049.2017.1381179.
[20] M. Zeraatpisheh, S. Ayoubi, A. Jafari, S. Tajik, P. Finke, Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran, Geoderma 338 (2019) 445–452, https://doi.org/10.1016/j.geoderma.2018.09.006.
[21] W. Lu, D. Lu, G. Wang, J. Wu, J. Huang, G. Li, Examining soil organic carbon distribution and dynamic change in a hickory plantation region with Landsat and ancillary data, Catena 165 (2018) 576–589, https://doi.org/10.1016/j.catena.2018.03.007.
[22] J.K.M. Biney, J. Houška, J. Volánek, D.K. Abebrese, J. Cervenka, Examining the influence of bare soil UAV imagery combined with auxiliary datasets to estimate and map soil organic carbon distribution in an erosion-prone agricultural field, Science of the total environment 870 (2023) 161973, https://doi.org/10.1016/j.scitotenv.2023.161973.
[23] J.A.M. Demattê, C.T. Fongaro, R. Rizzo, J.L. Safanelli, Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images, Remote Sensing of Environment 212 (2018) 161–175, https://doi.org/10.1016/j.rse.2018.04.047.
[24] A. Gasmi, C. Gomez, P. Lagacherie, H. Zouari, A. Laamrani, A. Chehbouni, Mean spectral reflectance from bare soil pixels along a Landsat-TM time series to increase both the prediction accuracy of soil clay content and mapping coverage, Geoderma 388 (2021) 114864, https://doi.org/10.1016/j.geoderma.2020.114864.
[25] S. Diek, F. Fornallaz, M.E. Schaepman, R. De Jong, Barest pixel composite for agricultural areas using landsat time series, Remote Sensing 9 (12) (2017) 1245, https://doi.org/10.3390/rs9121245.
[26] W. de Sousa Mendes, J.A. Demattе, N.E.Q. Silvero, L.R. Campos, Integration of multispectral and hyperspectral data to map magnetic susceptibility and soil attributes at depth: A novel framework, Geoderma 385 (2021) 114885, https://doi.org/10.1016/j.geoderma.2020.114885.
[27] B. Minasny, A.B. McBratney, Digital soil mapping: A brief history and some lessons, Geoderma 264 (2016) 301–311, https://doi.org/10.1016/j.geoderma.2015.07.017.
[28] A.J. Franzluebbers, Texture and organic matter associations with soil functional properties in crop and conservation land uses in North Carolina, Soil Science Society of America Journal 88 (2) (2024) 449–464, https://doi.org/10.1002/saj2.20620.
[29] A. Don, T. Scholten, E.D. Schulze, Conversion of cropland into grassland: Implications for soil organic-carbon stocks in two soils with different texture, Journal of Plant Nutrition and Soil Science 172 (1) (2009) 53–62, https://doi.org/10.1002/jpln.200700158.
[30] O.K. Hounkpatin, F.O. de Hipt, A.Y. Bossa, G. Welp, W. Amelung, Soil organic carbon stocks and their determining factors in the Dano catchment (Southwest Burkina Faso), Catena 166 (2018) 298–309, https://doi.org/10.1016/j.catena.2018.04.013.
[31] S. Mirzaee, S. Ghorbani-Dashtaki, J. Mohammadi, H. Asadi, F. Asadzadeh, Spatial variability of soil organic matter using remote sensing data, Catena 145 (2016) 118–127, https://doi.org/10.1016/j.catena.2016.05.023.
[32] C. Schillaci, M. Acutis, L. Lombardo, A. Lipani, M. Fantappie, M. Märker, S. Saia, Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling, Science of the total environment 601 (2017) 821–832, https://doi.org/10.1016/j.scitotenv.2017.05.239.
[33] S. Zhang, Y. Huang, C. Shen, H. Ye, Y. Du, Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information, Geoderma 171 (2012) 35–43, https://doi.org/10.1016/j.geoderma.2011.07.012.
[34] H. Etezadi, V. Adamchuk, Y. Bouroubi, M. Leduc, M. Gasser, D. TitleyPeloquin, Quantifying intra-field soil variability using categorical data: A case study of predicting soil organic matter using soil survey maps, Smart Agricultural Technology (2025) 101503, https://doi.org/10.1016/j.atech.2025.101503.
[35] C. Luo, B. Qi, H. Liu, D. Guo, L. Lu, Q. Fu, Y. Shao, Using time series sentinel-1 images for object-oriented crop classification in google earth engine, Remote Sens 13 (4) (2021) 561.
[36] R. Taghizadeh-Mehrjardi, K. Schmidt, A. Amirian-Chakan, T. Rentschler, M. Zeraatpisheh, F. Sarmadian, R. Valavi, N. Davatgar, T. Behrens, T. Scholten, Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space, Remote Sensing 12 (7) (2020) 1095, https://doi.org/10.3390/rs12071095.
[37] X. Xu, Z. Wang, X. Song, W. Zhan, S. Yang, A remote sensing-based strategy for mapping potentially toxic elements of soils: Temporal-spatial-spectral covariates combined with random forest, Environmental Research 240 (2024) 117570.
[38] F. Zhang, X. Yang, Improving land cover classification in an urbanized coastal area by random forests: The role of variable selection, Remote Sensing of Environment 251 (2020) 112105, https://doi.org/10.1016/j.rse.2020.112105.
[39] M. Chen, X. Qiu, W. Zeng, D. Peng, Combining sample plot stratification and machine learning algorithms to improve forest aboveground carbon density estimation in northeast China using airborne LiDAR data, Remote Sensing 14 (6) (2022) 1477, https://doi.org/10.3390/rs14061477.
[40] M. Marshall, M. Belgiu, M. Boschetti, M. Pepe, A. Stein, A. Nelson, Field-level crop yield estimation with PRISMA and Sentinel-2, ISPRS Journal of Photogrammetry and Remote Sensing 187 (2022) 191–210, https://doi.org/10.1016/j.isprsjprs.2022.03.008.

[41] K.K. Aqdam, N.Y. Mahabadi, H. Ramezanpour, S. Rezapour, Z. Mosleh, E. Zare, Comparison of the uncertainty of soil organic carbon stocks in different land uses, Journal of Arid Environments 205 (2022) 104805.
[42] E. Duarte, E. Zagal, J.A. Barrera, F. Dube, F. Casco, A.J. Hernández, Digital mapping of soil organic carbon stocks in the forest lands of Dominican Republic, European journal of remote sensing 55 (1) (2022) 213–231, https://doi.org/10.1080/22797254.2022.2045226.
[43] J.M. Ahn, J. Kim, K. Kim, Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting, Toxins 15 (10) (2023) 608, https://doi.org/10.3390/toxins15100608.
[44] A. Elmotawakkil, A. Moumane, A. Ait Youssef, N. Enneya, Machine Learning and Remote Sensing for Modeling Groundwater Storage Variability in Semi-Arid Regions, Intelligent Geoengineering (2025), https://doi.org/10.1016/j.ige.2025.08.001.
[45] S.V. Razavi-Termeh, A. Sadeghi-Niaraki, S.I. Abba, J. Hussain, S.-M. Choi, Flood-prone area mapping using a synergistic approach with swarm intelligence and gradient boosting algorithms, Scientific Reports 15 (1) (2025) 27924, https://doi.org/10.1038/s41598-025-12022-6.
[46] X. Dou, X. Wang, H. Liu, X. Zhang, L. Meng, Y. Pan, Z. Yu, Y. Cui, Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China, Geoderma 356 (2019) 113896, https://doi.org/10.1016/j.geoderma.2019.113896.
[47] S. Faramarzi, E. Pazira, M. Masihabadi, A. Mohammadi Torkashvand, B. Motamedvaziri, Modeling and estimating the spatial distribution of soil organic matter content in irrigated lands, International Journal of Environmental Science and Technology 19 (8) (2022) 7399–7410.
[48] ESA, European Space Agency, in: https://sentiwiki.copernicus.eu/web/s2-processing, 2025.
[49] U. Heiden, P. d'Angelo, P. Schwind, P. Karlshöfer, R. Müller, S. Zepp, M. Wiesmeier, P. Reinartz, Soil reflectance composites—improved thresholding and performance evaluation, Remote Sensing 14 (18) (2022) 4526, https://doi.org/10.3390/rs14184526.
[50] J. Xue, X. Zhang, Y. Huang, S. Chen, L. Dai, X. Chen, Q. Yu, S. Ye, Z. Shi, A two-dimensional bare soil separation framework using multi-temporal Sentinel-2 images across China, International Journal of Applied Earth Observation and Geoinformation 134 (2024) 104181, https://doi.org/10.1016/j.jag.2024.104181.
[51] D.S.E. Bramble, I. Schöning, L. Brandt, C. Poll, E. Kandeler, S. Ulrich, R. Mikutta, C. Mikutta, W.L. Silver, K.U. Totsche, Land use and mineral type determine stability of newly formed mineral-associated organic matter, Communications Earth & Environment 6 (1) (2025) 415, https://doi.org/10.1038/s43247-025-02400-3.
[52] M.L. Carvalho, V.F. Maciel, R.d.O. Bordonal, J.L.N. Carvalho, T.O. Ferreira, C.E. P. Cerri, M.R. Cherubin, Stabilization of organic matter in soils: drivers, mechanisms, and analytical tools–a literature review, Revista Brasileira de Ciência do Solo 47 (2023) e0230130, https://doi.org/10.36783/18069657rbcs20220130.
[53] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society Series B: Statistical Methodology 58 (1) (1996) 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
[54] O. Odebiri, O. Mutanga, J. Odindi, R. Slotow, P. Mafongoya, R. Lottering, R. Naicker, T.N. Matongera, M. Mngadi, Remote sensing of depth-induced variations in soil organic carbon stocks distribution within different vegetated landscapes, Catena 243 (2024) 108216, https://doi.org/10.1016/j.catena.2024.108216.
[55] C. Yang, L. Yang, L. Zhang, C. Zhou, Soil organic matter mapping using INLA-SPDE with remote sensing based soil moisture indices and Fourier transforms decomposed variables, Geoderma 437 (2023) 116571, https://doi.org/10.1016/j.geoderma.2023.116571.
[56] L. Freijeiro-González, M. Febrero-Bande, W. González-Manteiga, A critical review of LASSO and its derivatives for variable selection under dependence among covariates, International Statistical Review 90 (1) (2022) 118–145, https://doi.org/10.1111/insr.12469.
[57] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, 112, Springer, 2013.
[58] C.A.S.C.C.S.C.W. Group, N.R.C. Canada, C. Agriculture, A.-F.C.R. Branch, The Canadian system of soil classification, NRC Research Press, 1998.
[59] Z. Cui, S. Chen, B. Hu, N. Wang, C. Feng, J. Peng, Mapping Soil Organic Carbon by Integrating Time-Series Sentinel-2 Data, Environmental Covariates and Multiple Ensemble Models, Sensors 25 (7) (2025) 2184, https://doi.org/10.3390/s25072184.
[60] Y. Deng, X. Zhao, Y. Tian, X. Zhang, J. Cao, L. Yin, B. Zhang, Impact of different environmental covariate selection strategies on mapping accuracy of soil organic carbon in salt-affected coastal farmland, Ecological Indicators 178 (2025) 113956, https://doi.org/10.1016/j.ecolind.2025.113956.
[61] N.E.Q. Silvero, J.A.M. Demattê, M.T.A. Amorim, N.V. dos Santos, R. Rizzo, J. L. Safanelli, R.R. Poppiel, W. de Sousa Mendes, B.R. Bonfatti, Soil variability and quantification based on Sentinel-2 and Landsat-8 bare soil images: A comparison, Remote Sensing of Environment 252 (2021) 112117, https://doi.org/10.1016/j.rse.2020.112117.
[62] D. Charishma, V. Kuligod, S. Gundlur, M. Potdar, M. Doddamani, H. Nagaveni, Estimation of top soil properties by Sentinel-2 imaging, Geology, Ecology, and Landscapes (2024) 1–10, https://doi.org/10.1080/24749508.2024.2392920.
[63] Q. Chen, Y. Wang, X. Zhu, Soil organic carbon estimation using remote sensing data-driven machine learning, PeerJ 12 (2024) e17836, https://doi.org/10.7717/peerj.17836.
[64] C. Luo, W. Zhang, X. Zhang, H. Liu, Mapping soil organic matter content using Sentinel-2 synthetic images at different time intervals in Northeast China, International Journal of Digital Earth 16 (1) (2023) 1094–1107, https://doi.org/10.1080/17538947.2023.2192005.
[65] B. Delaney, K. Tansey, M. Whelan, Satellite remote sensing techniques and limitations for identifying bare soil, Remote Sensing 17 (4) (2025) 630, https://doi.org/10.3390/rs17040630.