



Toward accurate determination of surface soil organic matter: hyperspectral retrieval with physical, statistical, and deep learning

Chao Niu^{a,b,c,1}, Yuan Zhang^{d,1}, Yayu Yang^{a,b,c},
Kun Tan^{a,b,c,*}, Linya Zhao^{a,b,c}, Lijun Tian^{a,b,c}

^a Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China

^b Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China

^c School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China

^d School of Resources and Geosciences, China University of Mining and Technology, Xuzhou 221116, China

ARTICLE INFO

Handling Editor: Jingyi Huang

Keywords:

Soil organic matter

Hyperspectral

Hapke model

Triple collocation

ABSTRACT

Soil organic matter (SOM) is a key indicator of soil quality and ecosystem functioning, directly linked to crop productivity and sustainable development. Accurately characterizing its spatiotemporal dynamics is crucial for activities related to land use planning, environmental protection, and land degradation management. Soil imaging spectroscopy has emerged as a validated technique for estimating soil properties, particularly gaining traction for quantifying organic matter. In this study, a physical retrieval model for SOM content is developed using the radiative transfer model. By incorporating the effect of moisture on reflectance, this method offers a novel strategy for addressing the challenges in retrieving soil environmental quality parameters. The physical-based, feature-based, and deep neural network-based methods were compared and cross-validated within a unified SOM retrieval framework. The coefficients of determination (R^2) of the validation sets were 0.556 (physical-based model), 0.688 (feature-based statistical model), and 0.725 (deep neural network-based model), respectively. Correlation analysis results indicate good agreement among the three methods, with correlation coefficients greater than 0.6. By employing an extended triple collocation algorithm to assess the error distribution, it was found that the physical-based model achieved a squared correlation coefficient of 0.839 against the unknown true value, outperforming the other two models. The categorical triple collocation analysis revealed clear differences in the performance of the three methods at different organic matter gradients. Finally, the three SOM products were merged using TCA-derived weights to generate an integrated SOM map.

1. Introduction

In soils, most carbon occurs as a complex mixture of organic matter biopolymers. As the world aims for carbon neutrality, soil organic matter (SOM)—the largest terrestrial carbon reservoir after the oceans—has become an important research topic (Kogel-Knabner and Amelung 2021). Simultaneously, the spatial variability characteristics of regional SOM stocks can provide an important scientific basis for ecosystem carbon pool management. The development of fast, cost-effective, environmentally friendly, and generalizable SOM retrieval methods to support precision agriculture, digital soil mapping, and global soil carbon monitoring is important for maintaining agricultural

productivity and environmental carrying capacity (Angelopoulou et al. 2020; Meng et al. 2025; Padarian et al. 2022).

Hyperspectral imaging spectroscopy based on visible–near-infrared (VNIR) and shortwave-infrared (SWIR) reflectance spectra spanning 400–2500 nm has emerged as a powerful tool for quantitatively characterizing and mapping key soil environmental properties (Castaldi et al. 2016; Chen et al. 2025; Dutta and Kumar 2019). The soil matrix contains numerous chemical and physical chromophores, typically present at low concentrations, with high spatial variability and strongly overlapping absorption features. These factors complicate spectral measurements, resulting in highly nonlinear and spatially variable relationships between spectral reflectance and the target variable. Given

* Corresponding author.

E-mail address: tankuncu@gmail.com (K. Tan).

¹ Chao Niu and Yuan Zhang contributed equally to this work, and should be considered co-first authors.

this inherent complexity, the spectral signatures of individual chromophores frequently overlap, and the absence of a systematic interpretation framework makes it difficult to isolate and assess the specific contribution of each constituent. Research on SOM spectroscopy has progressed from empirical relationships between reflectance and SOM content toward the interpretation of characteristic absorption features. In the VNIR/SWIR region, SOM-related spectral responses are associated with overtone and combination absorptions of organic functional groups, including C–H, O–H, N–H, and C–O bonds (Kim et al. 2022; Stenberg et al. 2010). Informative spectral regions have frequently been reported around 1400, 1900, and 2200 nm (Dufr  chou et al. 2015). However, these absorption features are generally broad and weak, and may overlap with the spectral responses of soil moisture and minerals, making it difficult to attribute individual spectral features to specific SOM constituents (He et al. 2009). Recent studies have therefore explored advanced spectral-analysis methods to enhance weak and overlapping SOM-related signals (Xia et al. 2024). In addition to spectral information, geographic and environmental variables have also been widely used to improve SOM estimation within the digital soil mapping framework. Commonly used covariates include topographic attributes derived from digital elevation models (DEM), climatic variables such as temperature and precipitation, land-use information and vegetation indices (Lamichhane et al. 2019; Liu et al. 2023; Zeraatpisheh et al. 2022). These variables can provide complementary information for characterizing the spatial variation of SOM. However, the availability and spatial resolution of environmental covariates vary considerably across regions, whereas hyperspectral data provide spatially continuous information for SOM retrieval.

Currently, most imaging spectroscopy methods for SOM content retrieval are multivariate techniques that include data processing, variable selection, statistical regression, and other steps (Chen et al. 2019; Pudelko and Chodak 2020; Shepherd et al. 2025). Soil reflectance is influenced by numerous environmental factors in practical measurements, and the signal-to-noise ratio of spectral reflectance affects the predictive performance. Therefore, selecting appropriate preprocessing strategies and employing effective data-mining techniques is crucial for enhancing the quality of spectral data and improving model accuracy. Various spectral signal preprocessing strategies have been proposed; typical approaches include noise removal (e.g., Savitzky–Golay filter (Vivo-Truyols and Schoenmakers 2006), wavelet denoising (Liu et al. 2011)), spectral correction (e.g., baseline correction (Engel et al. 2013), multiple scattering correction (Dotto et al. 2018), and spectral derivatives (Hong et al. 2020)). The spectral transformation methods based on signal extraction can be used to explore spectral features that are sensitive to the target variables, such as the simple spectral index (Thaler et al. 2019) and discrete wavelet transform (Meng et al. 2020). Simple mathematical transformations, including derivatives, logarithms, and ratios, are widely applied, especially continuum removal transformations. Absorption features such as absorption depth, absorption area, and spectral ratios can characterize the shape and intensity of diagnostic spectral responses and provide more comprehensive information than reflectance values at individual wavelengths (Wang et al. 2022). These features may also reduce the influence of background reflectance variations and enhance weak SOM-related spectral signals (Stenberg et al. 2010). For feature selection, approaches such as competitive adaptive reweighted sampling (Li et al. 2009), variable importance in projection (Cecillon et al. 2008), successive projection algorithms (Araujo et al. 2001), uninformative variable elimination (Cai et al. 2008), and genetic algorithms (Leari 2000) are used to screen the best spectral predictor variables to eliminate redundant information from the original spectral matrix and suppress the overfitting phenomenon caused by spectral signal covariance. Due to the complicated nature of spectra, most feature selection tools were not constructed from fundamental studies of simple molecules, but empirical method development, and thus tend to lack interpretability. In terms of regression model construction, commonly used methods include multivariate

regression and local-type modeling (Briedis et al. 2020), such as principal component regression (Fox and Metla 2005) and partial least squares regression, which are capable of handling highly collinear data to robustly link target variables with reflectance spectra. Machine learning models demonstrate strong nonlinear modeling capabilities and are widely applied in soil property modeling (Ding et al. 2025; Lamichhane et al. 2019; Tziolas et al. 2020). However, the performance of traditional machine learning models depends on the use of spectral preprocessing techniques.

Deep learning approaches can discover complex structures in high-dimensional data. Some studies have successfully applied deep-learning-based regression models in hyperspectral soil property prediction (Dong et al. 2025; Niu et al. 2026; Padarian et al. 2019; Saberioon et al. 2024; Tsakiridis et al. 2020). Deep learning architectures are stacks of many nonlinear modules that transform original inputs into abstract, high-level feature representations (Niu et al. 2025). The improved feature extraction capability is less sensitive to preprocessing steps and can identify critical patterns directly from reflectance spectra. Deep learning-based regression models improve quantitative prediction accuracy and achieve lower RMSE values than traditional chemometric methods. Currently, interpretability and repeatability are critical factors limiting the application of deep learning approaches in spectral analysis.

Radiative transfer equations can relate reflectance to soil properties through absorption and scattering processes. On this basis, surface properties can be retrieved from remote sensing measurements. Thus far, researchers have mostly focused on analyzing the statistical relationship between SOM and reflectance in the VNIR spectral interval, while physical-based approaches for quantifying SOM remain relatively limited. Fox and Sabbagh (2002) integrated surface reflectance intensities in the red and NIR channels to formulate the concept of the soil line and derive predictive relationships between surface SOM content and Euclidean distance along the image soil line. Yuan et al. (2019) developed a semi-empirical model for SOM retrieval by incorporating SOM content into the Kubelka–Munk framework. Yuan et al. (2020) further used sensitivity analysis based on a radiative transfer model to identify informative wavelengths for SOM estimation. More recently, Ou et al. (2022) proposed a Kubelka–Munk-based spectral correction model to reduce the influence of soil moisture on SOM retrieval from airborne hyperspectral imagery. Wu et al. (2023b) developed a soil multifactor radiative transfer model to explain the effects of SOM, particle-size distribution, and iron oxide content on soil spectra. Compared with conventional statistical models, radiative-transfer-based semi-empirical models and physical-based models offer a stronger theoretical basis and improved interpretability. Nevertheless, existing radiative transfer models commonly rely on simplified assumptions and sample-calibrated parameters. Their application under field conditions may also be affected by soil moisture, mineral composition, particle-size distribution, and complex imaging conditions. Moreover, previous studies have generally focused on individual retrieval models. A systematic framework that jointly compares physical-based, feature-based statistical, and deep-learning approaches using the same hyperspectral imagery and field observations remains limited. In addition, point-scale validation alone may be insufficient for evaluating the reliability of spatially continuous SOM products, particularly when field samples are sparse.

To address these limitations, this study develops an integrated framework for SOM retrieval, evaluation, and product merging. First, a physical-based retrieval approach is constructed by integrating soil-moisture correction, Hapke-model-based conversion from reflectance to single-scattering albedo (SSA), and non-negative matrix factorization (NMF)-based spectral decomposition. This approach is designed to improve the physical interpretability of the SOM retrieval process. Second, the physical-based approach is compared with a feature-based statistical regression model and a semi-supervised deep neural network (DNN) model under a unified dataset and evaluation protocol. Third, triple collocation analysis (TCA) is introduced to characterize the relative error properties of the three spatial products and support their

uncertainty-weighted merging. By assigning higher weights to products with lower estimated uncertainties, the merging procedure reduces the influence of model-specific random errors and improves the robustness of the final SOM product. Accordingly, the objectives of this study are to: (1) construct SOM retrieval models representing three complementary modeling paradigms, namely physical-based retrieval, feature-based statistical regression, and semi-supervised deep learning; (2) compare their point-scale predictive performance and map-scale spatial characteristics using field validation and TCA-derived relative error indicators; and (3) generate an integrated SOM product through TCA-based merging. The technical workflow of this study is shown in Fig. 1.

2. Materials

2.1. Study region

The geographic focus of this work is Yitong County, south-central Jilin Province, China, shown in Fig. 2. The study area is a hilly area in the transition zone from the Changbai Mountain Range to the Songliao Plain. Yitong County is a key commercial grain-producing area in China, with cultivated farmland as the dominant land-use type and corn and rice as the main crops. The climatic characteristics can be broadly summarized as relatively low precipitation, high temperatures, and slightly more sunshine. The study area covers approximately 139 km², with an average elevation of 305 m (range: 262–446 m). The Yitong River traverses this region in a southeast–northwest direction.

The study area is rich in mineral resources, with more than 30 types of minerals. The soils are dominated by dark brown forest soils, meadow soils, black soils, and alluvial soils. The region is characterized by extensive arable land, several gold-mining areas, and scattered settlements.

2.2. Basic SOM statistics

The distribution of sampling points was set according to rules based on a 2 km spatial gridding, with the sampling locations adjusted according to the actual distribution of cultivated land. The mining area within the study region was sampled more densely to better capture potential effects of mining activities on soil organic matter and the associated spatial heterogeneity. At each site, five subsamples were taken within a 1-meter radius and composited into a single bulk sample. All soil samples were taken from the 0–20 cm surface layer. Ninety-three samples were obtained, and SOM was determined following the Chinese Forestry Industry Standard LY/T 1237-1999 (State Forestry Administration of China, 1999). Among these samples, 62 were used for calibration and 31 were used for validation. The calibration and validation sets were partitioned using a stratified sampling strategy based on the measured SOM contents. Specifically, all 93 soil samples were first ranked according to their measured SOM values. The ranked samples were then divided into groups of three according to adjacent SOM values. Within each group, two samples were randomly selected for calibration, and the remaining sample was used for validation. This strategy ensured that both the calibration and validation sets covered the full SOM concentration range and maintained similar statistical distributions. The same calibration and validation datasets were used for all three retrieval models to ensure a fair comparison of model performance. Fig. 3 shows histograms of the SOM content and their basic statistics, including the maximum value (Max), minimum value (Min), mean value (Mean), standard deviation (SD), coefficient of variation (CV), and number of samples (*n*). The SOM content ranged from 14.75 to 49.83 g/kg, with an average value of 30.50 g/kg. The similar mean values indicate comparable SOM distributions among the full, calibration, and validation datasets.

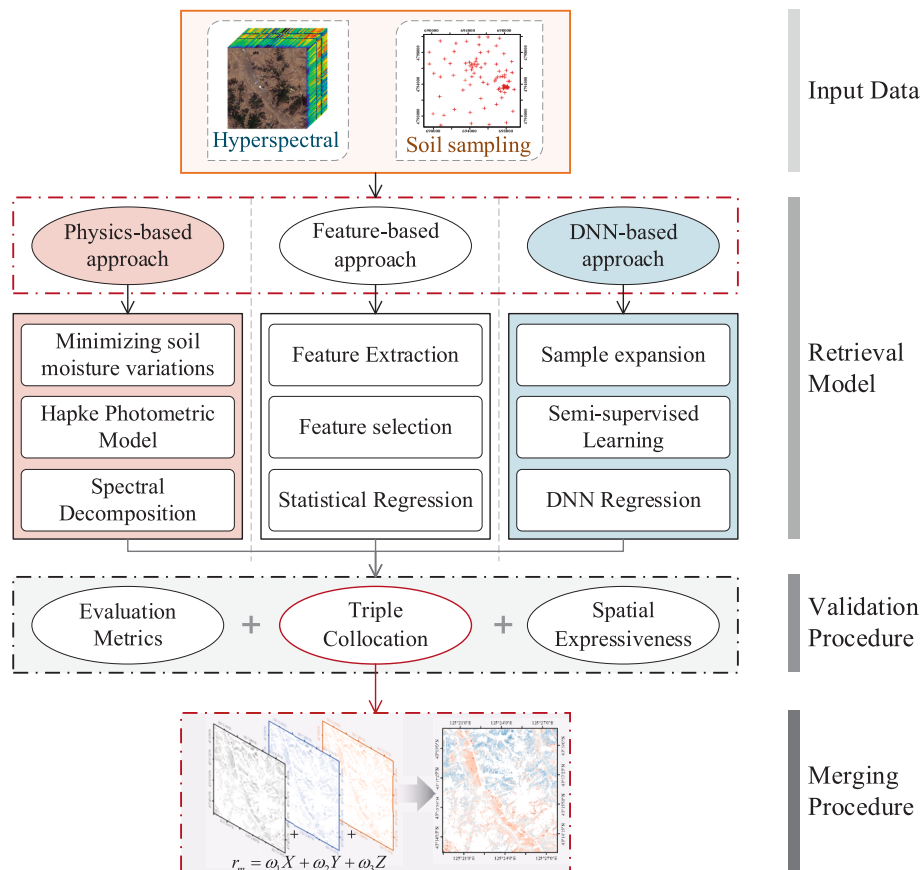


Fig. 1. Workflow overview.

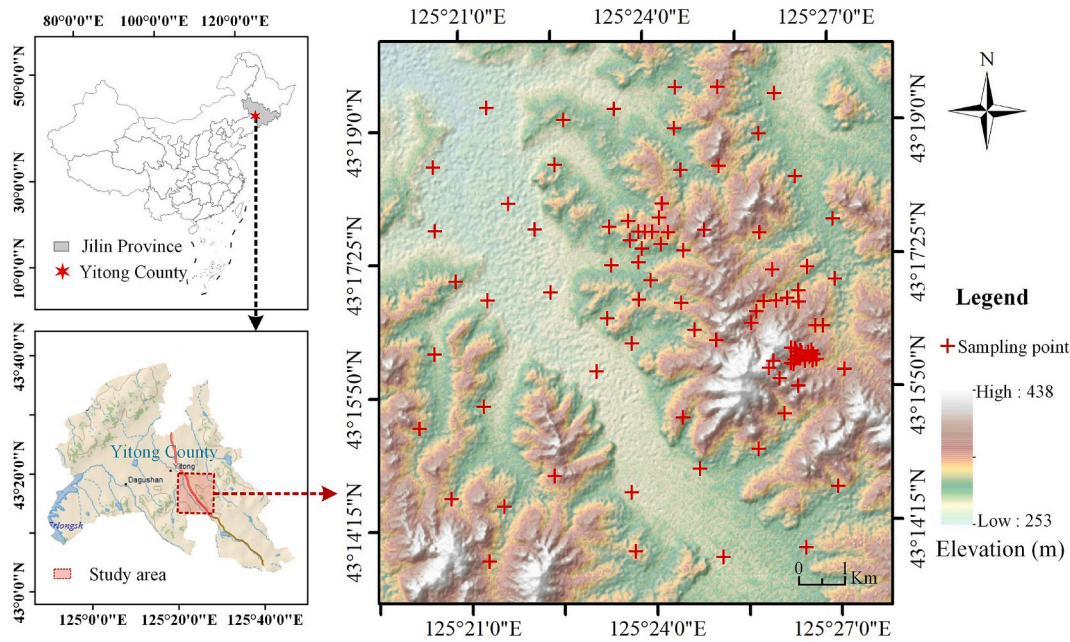


Fig. 2. The study region and field sampling points.

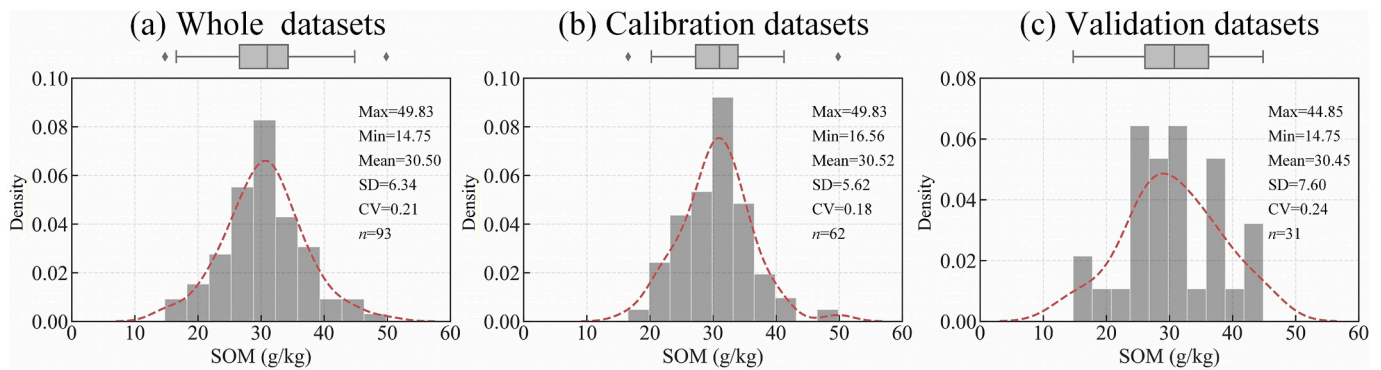


Fig. 3. Histograms of SOM in the whole, calibration, and validation datasets.

2.3. Airborne hyperspectral imagery and data preprocessing

The airborne hyperspectral imaging was synchronized with the surface soil sampling. Airborne hyperspectral images with a spatial resolution of 4.5×4.5 m were acquired using a HyMap imaging spectrometer between April 18 and April 22, 2017. The imagery was collected during the post-tillage period, when the soil surface was largely bare and the flight campaign was conducted under conditions of high visibility, which helped reduce variations in illumination and surface soil moisture during image acquisition. The spectral range was 466–2470 nm, with resolutions ranging from 13 to 17 nm. The georectified radiance data were atmospherically corrected to apparent surface reflectance using the HyCorr program based on the ATREM model (Center for the Study of Earth from Space 1999). Noisy bands from 1370 to 1514 nm and 1788 to 1996 nm were removed due to absorption by atmospheric water vapor, and 114 spectral bands were retained.

Since the main research target was bare soil, it was necessary to extract a soil mask. During field sampling, we identified five main feature types: soil, vegetation, water, buildings, and roads. Pure spectra of these five feature types were manually selected, and fully constrained least squares unmixing was used to perform spectral unmixing and calculate abundance information based on the selected endmembers. Through repeated tests, a threshold value of 0.7 was set, and pixels with

a soil abundance greater than this threshold value were extracted as the pure soil pixels.

3. Methodology

In Section 3.1, the implementation of SOM retrieval algorithms is first presented, using physical-based, feature-based statistical regression, and DNN techniques. The model evaluation and analysis method is described in Section 3.2. Finally, the triple collocation-based analysis method is introduced in Section 3.3. The main analyses were conducted using Python 3.10, with the scikit-learn, PyTorch, and GDAL libraries.

3.1. SOM retrieval model construction

3.1.1. Physical-based model

To relate soil reflectance to SOM content, we discuss the effect of soil moisture during radiative transfer, using a conceptual model that represents moisture as a thin water layer coating the soil surface. Then, the underlying theory of the Hapke photometric model is described. In this study, surface reflectance was transformed into single scattering albedo (SSA) using the Hapke equation, thereby approximating the originally nonlinear mixed signal as a linear combination of endmembers. Then, the constrained bilinear model was combined with a non-negative

matrix factorization algorithm to separate the SOM concentration information from the soil SSA matrix, and a physical-based model was constructed. The physical-based model procedure is outlined in Fig. 4.

3.1.1.1. Conversion of raw soil reflectance to underlying dry soil reflectance. Soil moisture variability in the field is unavoidable; field samples always exhibit small-scale moisture heterogeneity. Surface soil moisture content is known to impact soil reflectance at all wavelengths of the Vis-NIR solar spectrum range (400–2400 nm). Thus, the prediction accuracy of soil property information can be improved by minimizing the moisture effect in NIR spectra (Diek et al. 2019; Wijewardane et al. 2016).

Soil moisture is also a critical variable in radiative transfer model construction because moisture can alter the background reflectance intensity of bare soil and the apparent absorption depth in the spectrum. Water film coating models, which model wet soil as a rough dry soil substrate with a superficial liquid-water film, are widely used to model radiative transfer in soil moisture (Babiet et al. 2018). The total soil reflectance consists of two components: (1) surface Fresnel reflection from the water film and (2) multiple internal reflections at the water–soil interface. Based on this strategy, Yang et al. (2020) constructed a wet soil reflectivity model, assuming that the water film thickness follows an irregular Poisson spatial distribution.

The water film coating model is described as:

$$R_{\text{meas}} = R_{\text{dry}}P(k=0) + \sum_{k=1}^{\infty} R_{\text{wet}}(k)P(k) \quad (1)$$

where R_{dry} denotes the dry soil reflectance, and $P(k)$ is the fraction of the dry soil area, which is subject to $P(k) = e^{-\mu} \mu^k / (k!)^{-1}$. k is the number of elementary water films. μ is a function of soil moisture: $\mu = (SM_p - 5)/SM_c \cdot SM_p$ and SM_c are the soil moisture volume and soil moisture capacity, respectively. The successive interaction (R_{wet}) of the Fresnel reflection and refraction (r_{12} , r_{21}) at the liquid–soil interface is calculated by summing the amplitudes:

$$R_{\text{wet}}(k) = r_{12} + \frac{(1 - r_{12})\exp(-2\kappa_w k \Delta) R_b (1 - r_{21})}{1 - r_{21} \exp(-2\kappa_w k \Delta) R_b} \quad (2)$$

where R_b is the background reflectance attributable to the Fresnel reflection of the liquid–soil interface. κ_w is the water absorption coefficient, and Δ represents the optical thickness of the monolayer water film. r_{12} and r_{21} are the Fresnel reflectance from air to water and water to air, respectively. The complete analytical expressions of all terms involved are beyond the scope of this study. By coupling Eq. (1) and (2), we constructed a soil moisture spectral reflectance model capable of

forward and inverse modeling between dry soil spectra and moist soil spectra.

To convert the original soil spectrum to the dry soil spectrum, we must first obtain the moisture content of the soil. In this paper, we use the normalized soil moisture calculated by Optical TRapezoid Model (OPTRAM) (Babaeian et al. 2018) as a proxy for soil moisture. This approach is applicable without requiring prior knowledge or calibration and has shown reasonably good performance for soil moisture retrieval. As an alternative to the conventional triangle/trapezoid methods for estimating soil moisture from transformed reflectance data, OPTRAM leverages the linear physical correlation between normalized soil moisture (W) and shortwave STR.

$$W = \frac{i_d + s_d NDVI - STR}{i_d - i_w + (s_d - s_w) NDVI} \quad (3)$$

where STR is the SWIR-transformed reflectance: $STR = (1 - R_{\text{SWIR}})^2 / 2R_{\text{SWIR}}$. After the soil moisture content is determined, the corresponding dry soil spectrum can be derived from the original soil spectrum.

3.1.1.2. Conversion of reflectance to SSA. The Hapke model (Hapke 2008) offers an approximate solution to the radiative transfer equation with a sound mathematical basis and clear physical definition, describing the photometric scattering properties of the surface of a dense medium. The reflectance spectra must be converted to SSA spectra to derive modal abundances from Hapke model reflectance. The multiple scattered light rays emitted from the surface are modeled by $r(\mu_i, \mu_e, g, \lambda)$, which nonlinearly depends on the SSA (ω):

$$r(\mu_i, \mu_e, g, \lambda) = \frac{\omega(\lambda)}{4} \frac{1}{\mu_i + \mu_e} (P(g)[1 + B(g)] + [H(\mu_i)H(\mu_e) - 1]) \quad (4)$$

where μ_i and μ_e represent the cosines of the incidence and emergence angles, respectively. $P(g)$ represents the phase function of the scattering medium, and g indicates the phase angle. $H(\mu_i)$ and $H(\mu_e)$ denote the Chandrasekhar–Ambartsumian H-functions that describe multiple scattering, which is approximated by:

$$H(x) = \left\{ 1 - \omega x \left[r_0 + \frac{1 - 2r_0 x}{2} \ln \left(\frac{1+x}{x} \right) \right] \right\}^{-1} \quad (5)$$

where r_0 is the diffusive reflectance defined as: $r_0 = (1 - \sqrt{1 - \omega}) / (1 + \sqrt{1 - \omega})$. The opposition surge term $B(g)$ is neglected, as all measurements were taken at phase angles exceeding 15° . As an intrinsic property of the scattering medium, the SSA $\omega(\lambda)$ represents the ratio of total scattered light to the sum of scattered and

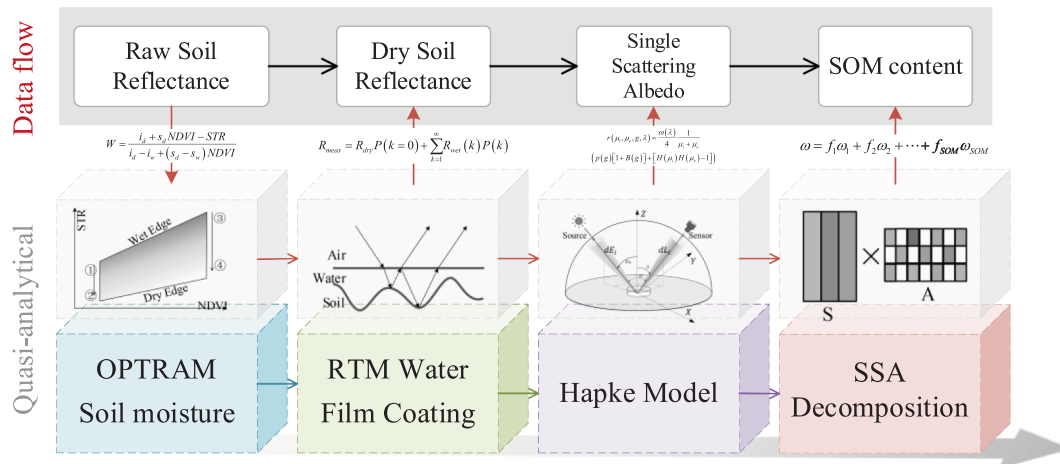


Fig. 4. Overview of the physical-based approach.

absorbed light by a particle at wavelength λ . SSA does not depend on the geometry of the scene or the photometry of the considered material.

3.1.1.3. SSA spectra decomposition. The interaction of the incident light rays with the particulate surfaces results in a nonlinear relationship between the reflectance spectrum and the individual components. The reflectance signal captured at the sensor level represents the contributions of each material. According to the Hapke bidirectional reflectance model, the SSA enables the nonlinear reflectance mixture to be converted into a linear mixture:

$$\omega = R^{-1}(r) = f_1\omega_1 + f_2\omega_2 + \dots + f_N\omega_N \quad (6)$$

where f_i represents the weight of the i^{th} endmember SSA. In Eq. (6), the soil responses are treated as a weighted sum of various soil spectral properties, thus decomposing the mixed multi-component information in the SSA matrix ω into a bilinear representation, in which each term represents a latent spectral component contributing to the mixed SSA signal. The constituents of the observed spectrum are commonly called “endmembers.” However, given the complexity and prevalence of soil chromophores (e.g., iron oxides, humic substances), soils are rarely pure, making it difficult to determine the appropriate composition or endmember spectrum. Additionally, the mechanism by which organic matter affects the soil spectrum is often obscured by the combined effects of all interactions, and the underlying physics are not well understood or completely represented. Nevertheless, spectral decomposition methods and mixture modeling can provide a promising exploratory approach. This parametric approach models soil SSA as a linear combination of several spectral vectors, involving the decoupling of soil observation signals into a set of simplified parameters related to specific soil properties. Following Eq. (6), if the contribution of SOM to the soil SSA is assumed to be the endmember ω_{SOM} , ω is defined as:

$$\omega = f_1\omega_1 + f_2\omega_2 + \dots + f_{\text{SOM}}\omega_{\text{SOM}} \quad (7)$$

For the SSA matrix $\mathbf{W}$ containing m soil samples:

$$\mathbf{W} = \begin{bmatrix} f_1^1 & f_1^2 & \dots & f_1^{\text{SOM}} \\ f_2^1 & f_2^2 & \dots & f_2^{\text{SOM}} \\ \vdots & \vdots & \ddots & \vdots \\ f_m^1 & f_m^2 & \dots & f_m^{\text{SOM}} \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_{\text{SOM}} \end{bmatrix} = [\mathbf{F}_1 \quad \mathbf{F}_2 \quad \dots \quad \mathbf{F}_{\text{SOM}}] \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_{\text{SOM}} \end{bmatrix} = \mathbf{F}\mathbf{S} \quad (8)$$

where $\mathbf{F}$ represents the weight matrix, and $\mathbf{S}$ represents the underlying “basic” SSA vector. $\mathbf{F}_{\text{SOM}}$ is the corresponding abundance vector of SOM. In this study, the measured SOM content was considered to be equal to $\mathbf{F}_{\text{SOM}}$. Following Eq. (8), SOM retrieval is transformed into a blind source separation problem.

There are still theoretical gaps in the a priori knowledge of optical constants in the electronic absorption region of SOM, which hinder the accurate estimation of the corresponding exact SSA spectra of SOM. In this study, a non-negative matrix factorization (NMF) algorithm was implemented for the SSA spectral decomposition analysis. NMF utilizes an iterative algorithm to provide a linear, low-rank approximation of non-negative matrix data, decomposing the signal into an endmember matrix and an abundance map matrix while minimizing Euclidean distance error.

$$\min_f(\mathbf{S}, \mathbf{A}) = \frac{1}{2} \|\mathbf{W} - \mathbf{S}\mathbf{A}\|_F^2 \quad (9)$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm of the matrix $\mathbf{W}$. For Eq. (9), the solution is obtained by iteratively combining elementwise multiplicative update rules. Although the NMF problem is an NP-hard, incorporating suitable regularizers can enhance solution uniqueness. In this study, we included the following constraints:

(1) To enforce physical consistency, the calculated abundances must

follow the abundance sum-to-one constraints. $\mathbf{W}$ and $\mathbf{S}$ in Eq. (8) are replaced by $\mathbf{W}_a$ and $\mathbf{S}_a$ in each iteration, respectively:

$$\mathbf{W}_a \leftarrow \begin{bmatrix} \mathbf{W}_{b \times m} \\ \delta \mathbf{1}_m \end{bmatrix}, \mathbf{S}_a \leftarrow \begin{bmatrix} \mathbf{S}_{b \times n} \\ \delta \mathbf{1}_n \end{bmatrix} \quad (10)$$

(2) During the iteration, $\mathbf{F}_{\text{SOM}}$ remains constant. Therefore, $\mathbf{W}_a$ and $\mathbf{S}_a$ are augmented as:

$$\widehat{\mathbf{W}}_a \leftarrow \begin{bmatrix} \mathbf{W}_a \\ \mathbf{SOM}_m \end{bmatrix}, \widehat{\mathbf{S}}_a \leftarrow \begin{bmatrix} \mathbf{S}_a \\ 0_{n-1}, 1 \end{bmatrix} \quad (11)$$

where b , m , and n are the band number, sample size, and the number of “basis curves,” respectively. $\mathbf{SOM}_m$ is the actual measured SOM content data. The SOM content retrieval was performed pixel by pixel using the fully constrained least squares method combined with the basis vector, $\mathbf{S}$, to finally achieve digital mapping of SOM in the study area.

3.1.2. Feature-based statistical model

Soil absorption features in the Vis-NIR spectrum are typically weak and broad rather than sharp, and no absorption peaks are directly related to SOM. Furthermore, soil hyperspectral data tend to have high noise levels and overlapping absorptions due to the narrow spectral channel bandwidth, interfering with the effectiveness of spectral modeling. Thus, preprocessing techniques and feature mining approaches are needed to extract hidden effective information from the original soil spectral data to improve the modeling performance.

Advances in multivariate statistics and chemometrics over recent decades have significantly bolstered quantitative soil spectroscopy research for property prediction. Meanwhile, developments in feature engineering and spectral conversion techniques have made it possible to precisely model the physical and chemical components of the soil. Extracting spectral features with enhanced information by reducing unmodeled variations (e.g., instrumental and experimental artifacts) is an important step. Spectral transformation methods based on signal extraction can detect spectral features sensitive to the target variable. Such methods preprocess the data from the reflected signal to improve the effective information content of the spectral data. The main parameters of the spectral absorption properties utilized are the absorption depth, absorption area, spectral ratio, and the original reflectance of certain wavelengths. The absorption depth D is the vertical distance between the position of the absorption peak and the transverse axis:

$$D(\lambda) = 1 - R_m(\lambda)/R_c(\lambda) \quad (12)$$

where R_m is the reflectance at the minimum of the absorption feature, and R_c is the continuum at a single wavelength, λ . The absorption area A is the area between the absorption peak and the transverse axis, obtained by integration:

$$A = \int_{\lambda_1}^{\lambda_2} [1 - R(\lambda)/R_c(\lambda)] \quad (13)$$

In the feature extraction stage, the spectral data were first preprocessed using the continuum removal method, and the effective spectral absorption feature parameters were extracted by calculating the mathematical transformation features, such as the absorption depth, absorption area, and spectral logarithm (Fig. 5). The effective features were then filtered by the threshold method according to Pearson’s correlation coefficient. In the retrieval stage, a bridge was constructed between the soil spectral characteristics and the SOM content in combination with the ExtraTrees regression (ETR) algorithm. The ETR algorithm is a variant of random forest with more randomization at each step to select an optimal cut/split or decision boundary. ETR uses the full training set and heavily randomized splits to build an ensemble of trees, making the model more computationally efficient and generally less prone to overfitting and high variance than standard random forests.

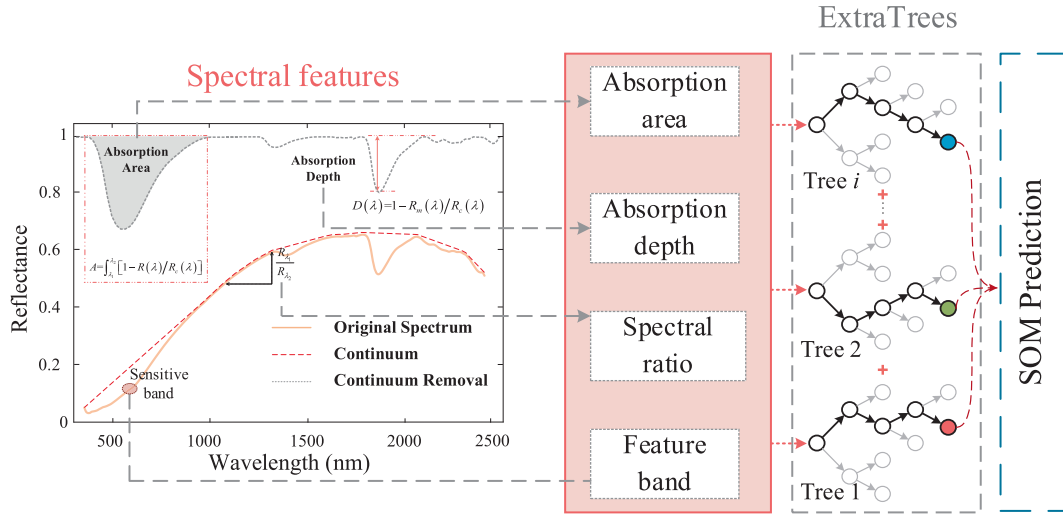


Fig. 5. Schematic diagram of the feature-based statistical approach.

3.1.3. Semi-supervised DNN model

The emerging technology of deep learning models has proven effective for processing complex multidimensional and multivariate nonlinear data. In the field of spectral analysis, DNNs can extract discriminative features from reflectance spectra, thereby enhancing feature representation and prediction accuracy, providing a new approach for hyperspectral retrieval analysis. However, their typically complex architectures require large training sets, whereas soil spectroscopy studies often provide only a limited number of matched samples. Ou et al. (2021) proposed a semi-supervised DNN retrieval method, which addresses the problems of weak expression, model overfitting, and inadequate training of DNNs in small-sample learning (Fig 6). In accordance with the first law of geography, eight nearest neighbors within the study image were identified for each labeled sample. These neighboring samples were added to a candidate pool as potential pseudo-labeled training data. Δ was defined as an incremental selection criterion for the pseudo-labeled samples:

$$\Delta = SA^* \text{dist}^* \frac{D_{rms}}{V} \quad (14)$$

Where SA , dist , and D_{rms} are the spectral angle, Euclidean distance, and predicted difference between the unlabeled sample X and the adjacent true labeled sample X^* , respectively, and can be expressed as in Eq. (15). Only samples satisfying the Δ were retained as pseudo-labeled samples and incrementally incorporated into the training set. The DNN was subsequently retrained using the expanded dataset, thereby enabling the model to exploit both labeled and pseudo-labeled samples.

$$\begin{cases} \text{dist}(X^*, X_i) = \sqrt{|X^* - X_i|^2} \\ D_{rms} = \sqrt{0.5(V_{m1} - V)^2 + (V_{m2} - V)^2} \\ SA(X^*, X_i) = \cos^{-1} \left[(X^*)^T X_i / (X^*)^T (X^*)^{1/2} (X_i^T X_i)^{1/2} \right] \end{cases} \quad (15)$$

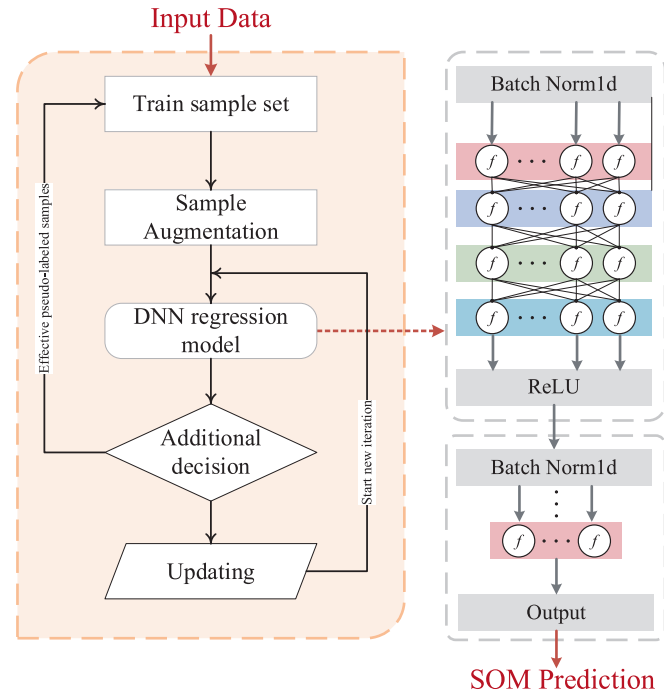


Fig. 6. The DNN approach framework.

3.2. Evaluation of model performances and spatial expressiveness

In this study, a collaborative validation strategy was used to evaluate the SOM maps, which contain two dimensions of retrieval precision and spatial expressiveness.

3.2.1. Regression evaluation metrics

Retrieval performance was evaluated using the coefficient of determination (R^2), RMSE, mean absolute error (MAE), and the ratio of performance to interquartile range (RPIQ). These formulas are expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^N [z(x_i) - \hat{z}(x_i)]^2}{\sum_{i=1}^N [z(x_i) - \bar{z}]^2} \quad (16)$$

$$RMSE = \left[\frac{1}{N} \sum_{i=1}^N (z(x_i) - \hat{z}(x_i))^2 \right]^{1/2} \quad (17)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |z(x_i) - \hat{z}(x_i)| \quad (18)$$

$$RPIQ = \frac{IQ}{RMSE} \quad (19)$$

where $z(x_i)$ is the observed value and $\hat{z}_i$ is the estimated value of sample i . N is the number of observations, and $\bar{z}$ is the mean of the observed

values. IQ denotes the interquartile range.

3.2.2. Spatial distribution patterns

To analyze spatial representativeness, the semivariogram function was employed to characterize spatial structure (Ortiz et al. 2010). The function provides a quantitative description of the scale and pattern of spatial heterogeneity and is a fundamental component of regionalized variable theory.

For a one-dimensional transect sampled at regular positions ($x = 1, 2, \dots, n$) from a spatial random field with pixel values $z(x)$, the spatial relationship for a given lag h is summarized by the average semivariance $\gamma(h)$ computed over all pairs of pixels separated by distance h :

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z(x_i) - Z(x_{i+h}))^2 \quad (20)$$

where $N(h)$ is the number of observation pairs in the h -lag class, subject to $N(h) = \{(i, j) | x_i - x_j = h\} h |N_h|$. Z is the value of the variable of interest at spatial position x_i . The quantity $Z(x_{i+h})$ is the variable value at distance h from x_i . The optimal semivariogram model was selected by maximizing the coefficient of determination and minimizing the residual sum of squares. In this study, the stable covariance model was used as the semivariance function, which can be expressed as:

$$\gamma(h) = \exp\left(-\left(\frac{ch}{r}\right)^\alpha\right) \quad (21)$$

where C_0 is the nugget, c is the partial sill, h is the input distance lag, and r is the effective range. These three parameters are usually utilized to measure the variogram. α is a shape parameter with $\alpha \in (0, 2)$.

3.3. Triple collocation analysis

When validating quantitative remote sensing products, it is practically infeasible to obtain the corresponding true values of individual pixels to be evaluated due to the limitation of sample representativeness, which poses a problem for validation. To address this challenge, triple collocation analysis (TCA), which allows the random error variance of three collocated datasets representing the same continuous geophysical variables to be verified, is considered a promising method when the error-free true value is unknown (Gruber et al. 2016). TCA has been widely adopted to validate quantitative remote sensing products in the fields of hydrometeorology, oceanography, and ecology (Gruber et al. 2017). TCA is subject to the following assumptions: (1) the three collocated datasets are linearly related to the unknown true value, and their errors are uncorrelated; (2) the datasets are not biased, and the expectation of error is treated as zero; (3) each dataset has the same number of observations for the geophysical parameters N .

3.3.1. Extended triple collocation

The extended triple collocation (ETC) method (McColl et al. 2014) builds upon the original mathematical hypotheses. It simultaneously estimates the RMSE and correlation coefficients (R). ETC employs a linear function to represent the relationship between the SOM products and the “unknown truth” SOM:

$$x_i = \alpha_i + \beta_i \theta + \varepsilon_i, i \in [X, Y, Z] \quad (22)$$

where α_i and β_i denote the additive and multiplicative systematic error corresponding to the i^{th} dataset, respectively. ε_i is the residual of the estimated and true values. The calculation of error covariance between the i^{th} and j^{th} input, defined as C_{ij} , is given by:

$$C_{ij} = \text{Cov}(x_i, x_j) = \begin{cases} \beta_i \beta_j \sigma_\theta^2 & (i \neq j) \\ \beta_i^2 \sigma_\theta^2 + \sigma_{\varepsilon_i}^2 & (i = j) \end{cases} \quad (23)$$

The RMSE (σ) and correlation coefficient (R) of each product can be

derived as follows:

$$\sigma = \begin{cases} \sigma_1 = \sqrt{C_{11} - \frac{C_{12}C_{13}}{C_{23}}} \\ \sigma_2 = \sqrt{C_{22} - \frac{C_{12}C_{23}}{C_{13}}} \\ \sigma_3 = \sqrt{C_{33} - \frac{C_{13}C_{23}}{C_{12}}} \end{cases}, R = \begin{cases} R_{\theta X} = \sqrt{\frac{C_{12}C_{13}}{C_{11}C_{23}}} \\ R_{\theta Y} = \text{sign}(C_{13}C_{23}) \sqrt{\frac{C_{12}C_{23}}{C_{22}C_{13}}} \\ R_{\theta Z} = \text{sign}(C_{12}C_{23}) \sqrt{\frac{C_{13}C_{23}}{C_{33}C_{12}}} \end{cases} \quad (24)$$

3.3.2. Categorical triple collocation

Classical TCA is no longer applicable when the object of study is defined as a categorical variable. To address this problem, McColl et al. (2016) proposed the categorical triple collocation (CTC) algorithm. The CTC method defines a binary classifier that uses a categorical framework instead of the linear regression framework used in classical TCA and calculates the error distribution of the categorical variables relative to the unknown true values. CTC has been successfully applied in studies of freeze-thaw cycle monitoring and earthquake damage classification. Instead of using the error variance as an evaluation metric, the CTC method redefines a feature vector and ranks the classification accuracy of different methods in the following form:

$$V = \begin{cases} v_1 = \sqrt{\frac{C_{12}C_{13}}{C_{23}}} \\ v_2 = \sqrt{\frac{C_{12}C_{23}}{C_{13}}} \\ v_3 = \sqrt{\frac{C_{23}C_{13}}{C_{12}}} \end{cases} \quad (25)$$

3.3.3. Merging approach

Merging aims to produce estimates with superior detection capabilities. TCA is utilized to merge different datasets, which can be achieved by linearly combining all three products using the least squares method:

$$r_m = \omega_1 X + \omega_2 Y + \omega_3 Z \quad (26)$$

Based on the error variance calculated using the ETC, the merging weights of the three different inputs can be calculated as follows:

$$w_{ETC} = \begin{cases} \omega_1 = \frac{\sigma_2^2 \sigma_3^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2} \\ \omega_2 = \frac{\sigma_1^2 \sigma_3^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2} \\ \omega_3 = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 \sigma_2^2 + \sigma_1^2 \sigma_3^2 + \sigma_2^2 \sigma_3^2} \end{cases} \quad (27)$$

Thus, the merging of three input products is realized. In this study, the SOM retrieval results were first considered as continuous variables, and the spatial distribution of errors for the three SOM retrieval methods was calculated using the classical TCA method. The threshold method was used to classify each of the three results into several classes, which were converted into category variables, and the CTC method was used to calculate the error profile of each category. Finally, TCA-based merging was implemented on the three SOM datasets to obtain SOM maps closest to the ground truth values.

4. Results and discussion

4.1. Model parameterization

4.1.1. Physical-based model

In summary, the physical-based method for retrieving SOM involves three steps. (1) Correction step: The original spectral data were

corrected to the underlying dry soil spectral reflectance data using the soil-moisture correction model. (2) Transformation step: Reflectance was converted to SSA using the Hapke radiative transfer model. (3) Decomposition step: SOM was assessed using SSA spectral decomposition techniques.

4.1.1.1. Soil moisture. In the correction step, the spatial distribution of soil moisture shown in Fig. 7 was retrieved using the OPTRAM algorithm, with model parameters determined from the distribution of pixels in the STR-NDVI feature space.

The soil moisture calculated using OPTRAM was used as input to the water-film coating model to correct the reflectance spectra pixel by pixel. The uncorrected and soil moisture-corrected spectra were grouped into four moisture gradients according to soil moisture content. For each gradient, the mean reflectance spectrum was calculated from multiple samples, with the shaded band representing the standard deviation. As shown in Fig. 8(a) and (b), the corrected spectra generally exhibited higher reflectance than the original spectra, indicating that the correction model effectively reduced the suppressive effect of soil moisture on soil reflectance. The enhancement in the Vis region was slight due to the insensitivity of the Vis region to changes in soil moisture. In the SWIR region, the corrected spectrum exhibited a greater enhancement relative to the original spectrum.

Fig. 8(c) shows the correlation coefficient curves between the raw and corrected soil spectra and the SOM content. The correlation coefficients generally improved with moisture correction, especially in the Vis (500–800 nm) and SWIR (2,200–2500 nm) ranges, where valuable information related to SOM was enhanced.

4.1.1.2. SSA basis vector. Because the HyMap imagery was acquired under a near-nadir viewing configuration, the emission angle was approximated as 0° for the SSA conversion. Since the solar zenith angles varied only slightly among the flight strips, the mean solar zenith angle, approximately 30° , was adopted as the representative value for the SSA conversion to simplify the calculation process. The decomposition rank

N , representing the number of basis vectors, was selected empirically by repeating the decomposition and SOM retrieval with different candidate N values while keeping the remaining settings unchanged. $N = 4$ yielded the best retrieval performance within the calibration dataset and was therefore adopted. Fig. 9(a) represents the corresponding basis vector V_{som} for SOM. Fig. 9(b), (c), and (d) show the other three latent basis vectors V_1 , V_2 , and V_3 . The four vector curves have different shapes and variation ranges. A clear trough can be observed near 600–700 nm in V_{som} , which is probably due to the strong absorption of SOM in this band interval. V_1 and V_2 have more obvious peaks near 700 nm, and V_3 shows a trend similar to the original soil spectral curve.

The morphological structure of SOM is extremely complex as it represents a variety of physical and chemical forms. Due to this amorphous characteristic, it is challenging to directly derive the endmember SSA of “pure organic matter.” To clarify, the basis vectors generated by the iterative NMF method in this paper do not represent the SSA spectra of pure organic matter in the actual physical sense but rather represent the contribution of the mixed SSA matrix of the SOM pairs separated by the spectral decomposition method within the constraints of the available measured data. For the four basis vector curves solved by the NMF algorithm, the detailed physical sense analysis and authenticity verification of their morphological characteristics require several geochemical simulation experiments, which lie outside the scope of the present work and will be explored in future studies.

4.1.2. Feature-based model

Combined with Pearson's correlation coefficients, the parameters of soil reflectance spectral absorption properties, including the absorption area, absorption depth, and spectral ratio, were analyzed and calculated. Because it is difficult to establish a stable SOM retrieval model due to the limited absorption characteristic parameters of soil spectra, the wavebands that were strongly correlated with SOM were added to improve the retrieval accuracy. The final parameters of the spectral features included in the retrieval modeling are shown in Table 1.

The three ETR model parameters include random inputs per node, tree count, and minimum split size. To ensure robustness, each model ran 100 times with varying random states, using the modal value as the final prediction.

4.1.3. DNN-based model

The DNN-based model used the preprocessed surface reflectance spectra as input. Before model training, noisy bands and atmospheric water-vapor absorption bands were removed. The DNN adopted in this study consists of six hidden layers and a single output layer, using batch normalization before each linear layer, using rectified linear units (ReLU) as the activation function, and incorporating a dropout function with a dropout rate of 0.5. The Adam optimizer was introduced as the optimizer with a step size of $2e^{-4}$. The mean square error was used as the loss function.

4.2. Model evaluation

Combined with the collaborative validation strategy, the in-situ prediction accuracy and the spatial patterns of the three SOM retrieval models were separately compared and evaluated.

4.2.1. Comparison of model evaluation metrics

As shown in Table 2, the physical-based, feature-based, and DNN-based models were evaluated using the same calibration and validation datasets. Therefore, the differences in model accuracy were attributable to the retrieval methods themselves rather than to differences in sample partitioning. It should also be noted that R^2 and RMSE describe different aspects of model performance. R^2 reflects the proportion of observed SOM variance explained by the model, whereas RMSE represents the absolute magnitude of prediction errors in g/kg. Therefore, these two metrics are not necessarily strictly proportional,

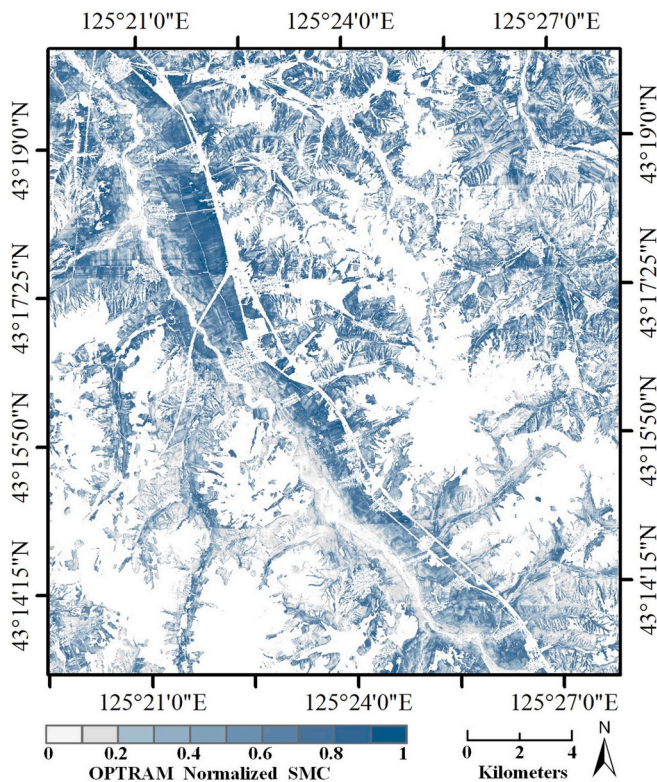


Fig. 7. Soil moisture using the OPTRAM algorithm.

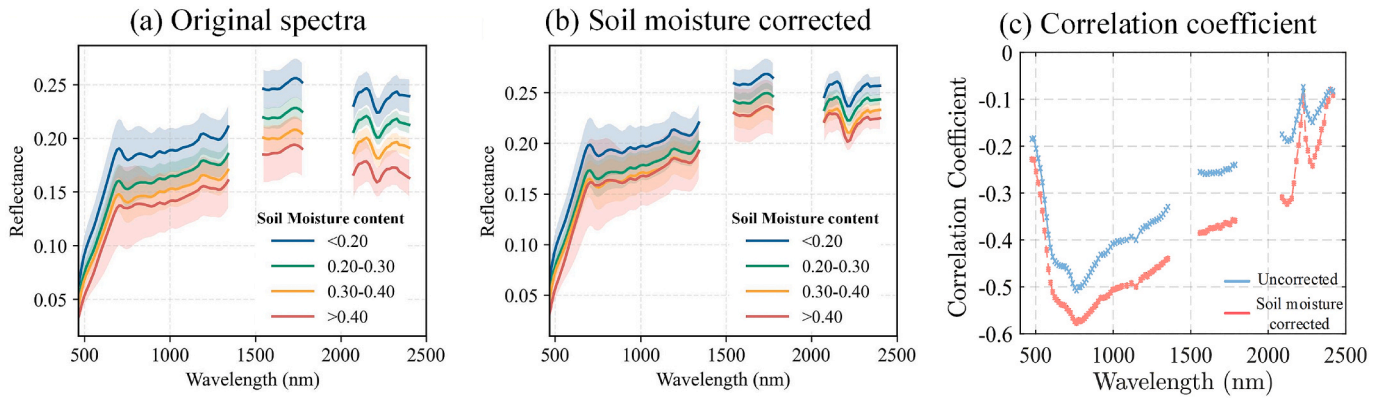


Fig. 8. Original spectra of the sampling sites (a), soil moisture-corrected spectra (b), and correlation coefficients between SOM content and the uncorrected and soil moisture-corrected spectra (c).

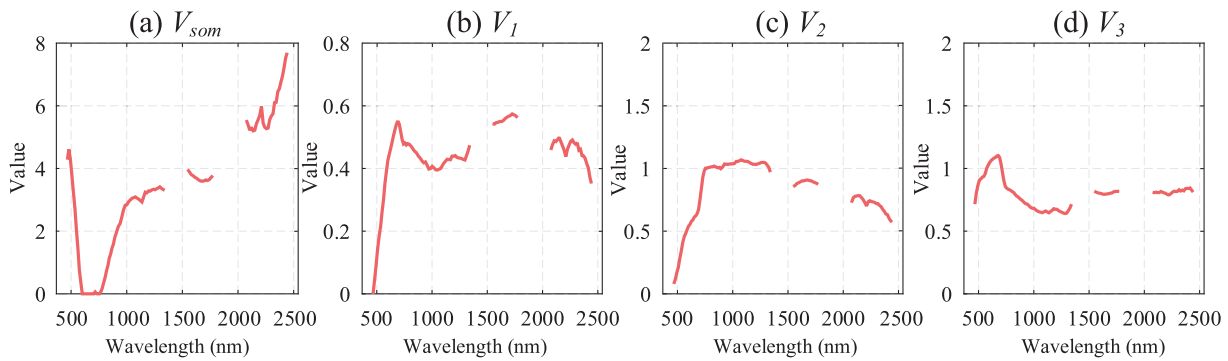


Fig. 9. Four basis vectors following SSA matrix decomposition.

Table 1
Feature combinations.

Features	Wavelength (nm)
Absorption area	2200
Absorption depth	630, 1300, 1630, 2000, 2200
Spectral ratio	$\frac{R_{613}}{R_{507}}$, $\frac{R_{2194}}{R_{2120}}$, $\frac{R_{613} - R_{507}}{R_{613} + R_{507}}$, $\frac{R_{2194} - R_{2120}}{R_{2194} + R_{2120}}$
Selected Spectral	494, 508, 522, 535, 548, 588, 601, 614, 626, 640, 653, 665, 729, 742, 755, 767, 780, 793, 805, 1355, 2075, 2090, 2105, 2121, 2136, 2166, 2180, 2195, 2211, 2226

particularly when the calibration and validation sets have slightly different SOM distributions and variances. For this reason, R^2 , RMSE, MAE, and RPIQ were jointly considered to provide a more comprehensive evaluation of model performance. To further evaluate the effectiveness of the semi-supervised sample expansion strategy, the DNN model was additionally trained using only the original 62 labeled samples. The labeled-only DNN achieved an R^2 of 0.632 on the validation

set. After incorporating the pseudo-labeled samples using semi-supervised strategy, the R^2 value improved to 0.725. These results indicate that the semi-supervised expansion improved the validation performance of the DNN under the limited labeled-sample condition. The labeled-only DNN was included only as a comparative baseline to assess the contribution of pseudo-label expansion, whereas the semi-supervised DNN was used for subsequent spatial analyses.

Fig. 10 presents the scatter diagrams of the measured and estimated SOM content. Comparing the statistical metrics of the models, the DNN-based model achieved the best performance, with an R^2 of 0.841 and 0.725 on the calibration set and validation set, respectively. The physical-based model showed lower predictive accuracy than the other two models, with R^2 values of 0.452 for calibration and 0.556 for validation.

Moreover, a two-dimensional Taylor diagram was used to summarize the comparisons of the results from the three models with observations. Taylor diagrams visually compare variables across test and reference datasets. Fig. 11 is a Taylor diagram comparing the three models in terms of Pearson's correlation coefficient, centered root-mean-square difference (RMSD), and the standard deviation. RMSD quantifies the

Table 2
Model evaluation indexes.

Approach	Calibration sets				Validation sets			
	R_c^2	RMSE _c (g/kg)	MAE _c (g/kg)	RPIQ _c	R_v^2	RMSE _v (g/kg)	MAE _v (g/kg)	RPIQ _v
Physical-based	0.452	4.711	3.663	1.59	0.556	4.460	3.740	1.190
Feature-based	0.631	3.733	2.793	2.010	0.688	3.543	2.501	1.503
DNN-based (Semi-supervised)	0.841	2.449	1.697	3.064	0.725	3.328	2.286	1.601
DNN-based (Labeled only)	0.786	3.018	2.143	2.486	0.632	3.925	3.177	1.357

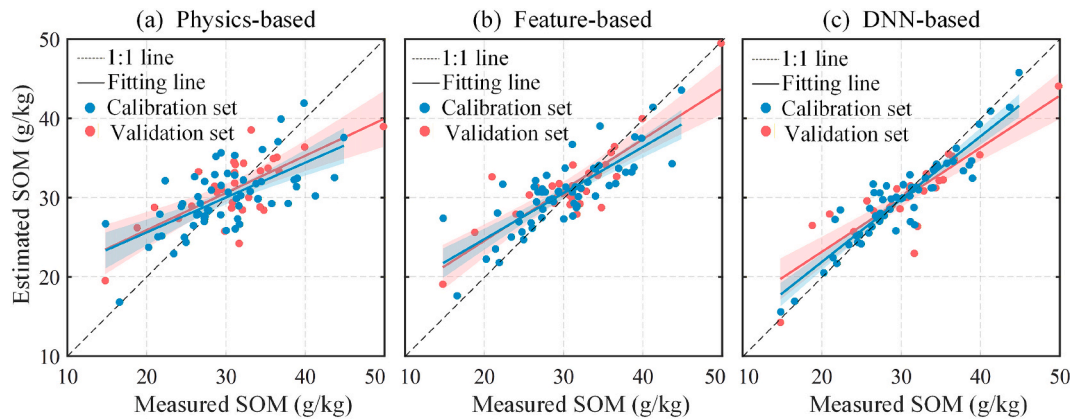


Fig. 10. Scatter diagrams of the predicted and measured SOM of the physical-based (a), feature-based (b), and DNN-based (c) models.

difference between the retrieved and observed SOM patterns after removing the effect of the mean bias.

The Taylor diagram in Fig. 11 summarizes the whole set of statistical measures and relationships between the retrieved SOM and in-situ measurements. Considering the in-situ data with $p = 1$ and $SD = 1$ on the Taylor diagram as the reference point (asterisk on the figure), closer proximity to this point indicates better model performance. According to the Taylor diagram, the correlation coefficient of the DNN model was higher than those of the other two models.

4.2.2. Spatial distribution pattern of SOM

To verify the spatial consistency of the SOM retrieval results, the three retrieval models proposed above were applied to all spatial data (masked by bare soil), and the SOM mapping results are shown in Fig. 12.

According to the retrieval results in Fig. 12, areas with high SOM content are mainly distributed in large cultivated fields along the Yitong River, which is likely due to human activities, such as cultivation practices and fertilizer application. Soils with lower SOM content are mainly found in uncultivated bare hilly areas possibly because of sparse vegetation cover, limited agricultural inputs, and enhanced soil erosion and downslope transport by runoff. The retrieval results of the three methods were basically consistent, and the spatial variation in SOM content showed similar distribution trends. To further evaluate the environmental plausibility of the SOM retrieval results, spatial autocorrelation and terrain-related spatial association analyses were conducted for the three SOM products. The Global Moran's I values of the

physical-based, DNN-based, and feature-based products were 0.748, 0.600, and 0.532, respectively, indicating that all three SOM products exhibited positive spatial autocorrelation and preserved coherent spatial structures. In addition, the bivariate Moran's I values between the SOM products and neighboring DEM values were -0.640 , -0.427 , and -0.226 , respectively. The consistently negative values indicate that high SOM values tended to be spatially associated with lower-elevation neighborhoods, whereas low SOM values tended to be associated with higher-elevation neighborhoods. This terrain-related pattern is consistent with the expected influence of topography on soil organic matter redistribution, erosion, and deposition processes. Although the strength of the association differed among the three models, the consistent negative spatial relationship with DEM suggests that the retrieved SOM products captured environmentally meaningful terrain-related gradients.

Fig. 13 presents the variation in the average SOM content with distance across the west–east and north–south directions. The mean variations of the three products in the two directions are comparable, demonstrating good agreement.

Some representative areas in the study region were selected and the details of the retrieval results of the three methods were compared. Fig. 14 shows that all three products generated detailed textures and consistent distribution trends.

Histograms were used to show the probability distribution of the whole image retrieval results for the three methods. As shown in Fig. 15, the gray background histogram is the probability distribution of the measured SOM values at the sampling sites, used as reference observations for comparison with the three retrieval results. The superimposed curve represents the probability density function (PDF). The closer the predicted SOM probability density function is to that of the observed samples, the better the prediction performance. The mean SOM values of the physical-based, feature-based, and DNN-based models are 31.78, 33.36, and 31.46 g/kg, respectively, with standard deviations of 4.42, 3.82, and 4.38 g/kg.

The histogram analysis shows that the dominant peak (~ 30 g/kg) in the observed SOM distribution coincides with that of the physical-based model predictions, whereas the feature-based and DNN-based predictions show much higher peaks. According to Fig. 15, the full SOM map retrieval results of the three methods had skewed distributions that deviated from the distribution of the measured SOM values at the sampling points but shared a similar overall pattern. Both feature-based and DNN-based methods showed significant aggregation around 30 g/kg and were lower in other intervals. The feature-based method adopts the ETR method for modeling and combines multiple decision trees together, which can produce good prediction results but has difficulty predicting data beyond the range of the training set and is prone to overfitting. For the DNN-based method, the learning ability of DNN improves with sample expansion but may have difficulties with

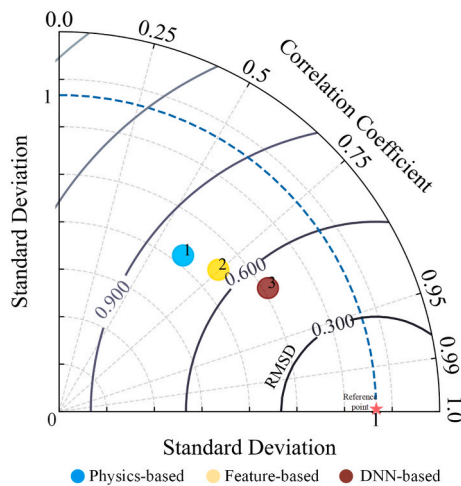


Fig. 11. Taylor diagram of the physical-based, feature-based, and DNN-based models.

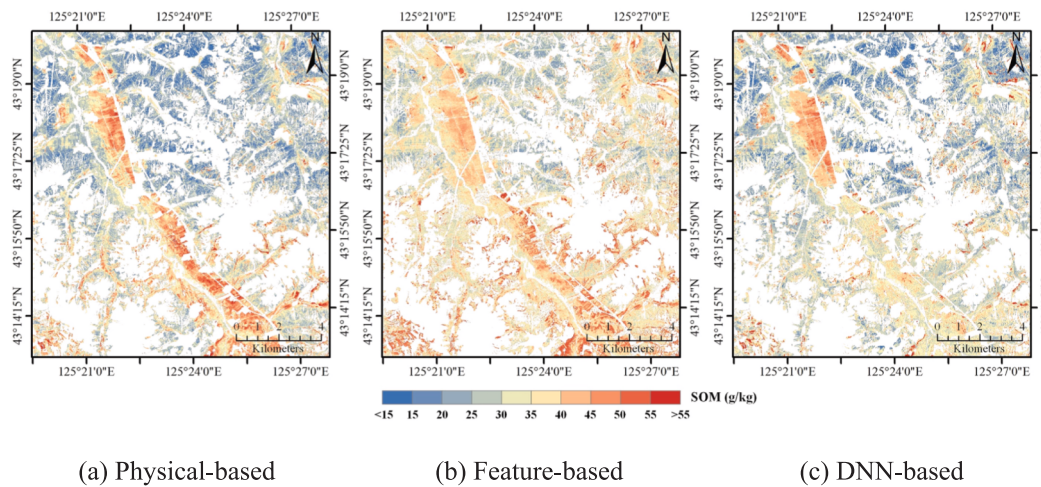


Fig. 12. Maps of SOM content estimated using the physical-based model (a), feature-based model (b), and DNN-based model (c).

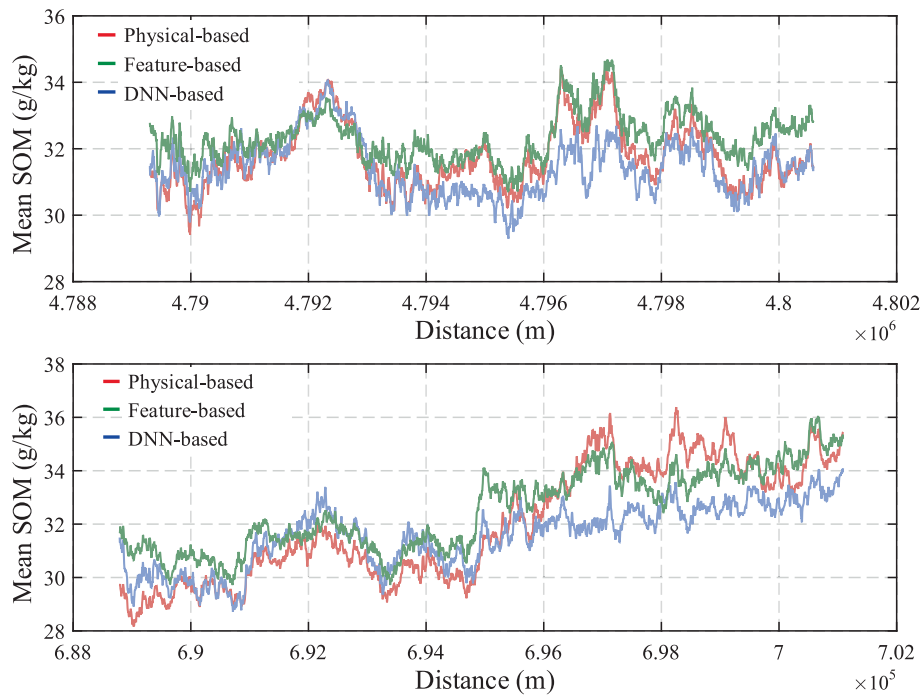


Fig. 13. The average SOM content from west to east (a) and north to south (b).

prediction intervals with fewer samples due to the deactivation phenomenon of neurons. The physical-based model had a more uniform and smooth distribution, which is due to the clear physical basis of the model. The spectral decomposition method maximized the separation of organic responses to the SSA matrix.

The cumulative distribution function (CDF) is used to describe the probability distribution of a real random variable. Comparing the CDF curve trends can reflect the distribution of results of different retrieval methods, allowing us to measure the consistency of the retrieval results. The gray curves in Fig. 15(d) represent the distribution of the measured values of the original sampling points, and the colored curves represent the distribution of the retrieval models. The CDF curves of the three models showed a similar trend to the CDF curves of the sampling points, indicating broadly similar distributional trends between the retrieved products and the field observations. Each model has different learning abilities for each data interval and can be affected by unbalanced samples with high leverage values; however, in sampling experiments, the

set sampling points may not cover some high- or low-value areas and cannot fully represent the entire study area, causing deviations between the retrieved values and the measured values.

Density scatter diagrams of the three methods were plotted to quantitatively express the correlations between the retrieval results of the three models (Fig. 16). The overall retrieval results of the three models were in good agreement, with correlation coefficients higher than 0.6.

4.2.3. Semivariogram analysis

The semivariogram encapsulates key information on spatial structure. The SOM semivariograms were obtained by using the three models to fit the semivariance, expressed by three key parameters: effective range, nugget, and partial sill. The range indicates the spatial scale of heterogeneity, while the sill signifies the total heterogeneity. The nugget, representing variance at zero lag, accounts for microscale variability occurring at distances smaller than the sampling interval,

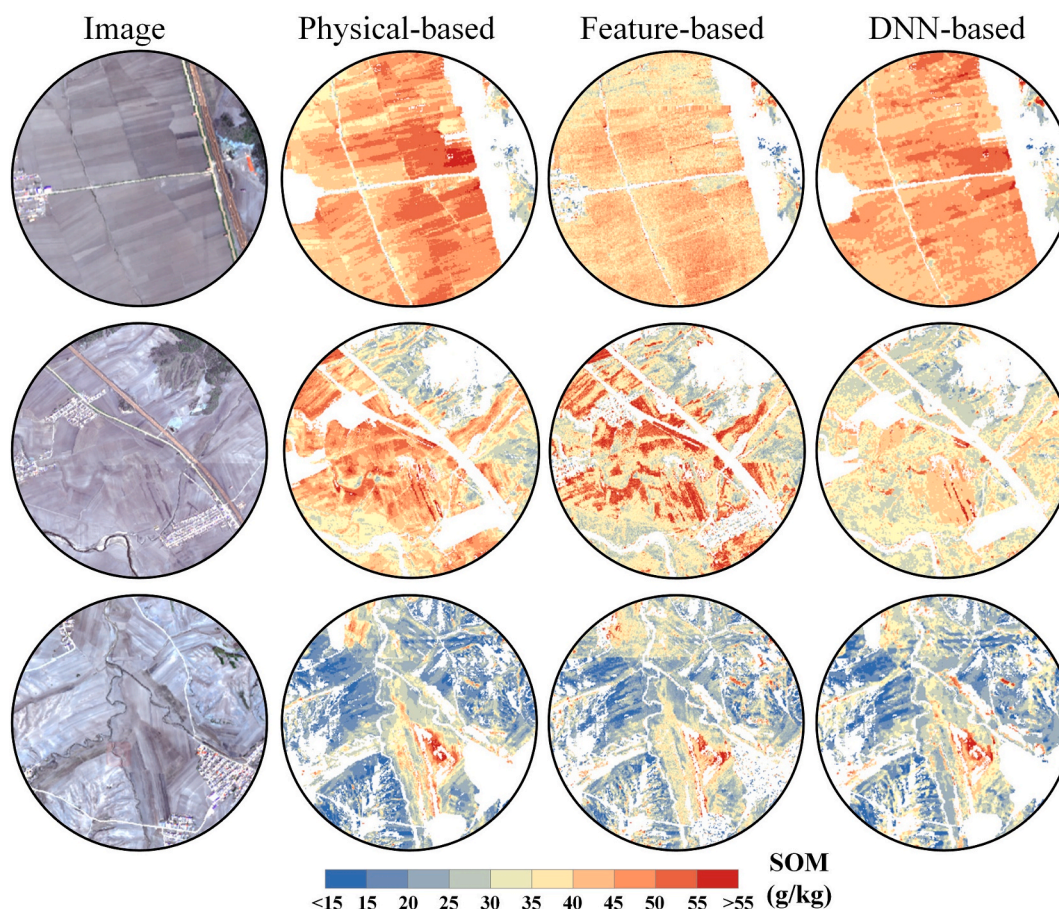


Fig. 14. Details of the SOM spatial variability.

along with any sampling errors. The inferred variogram models and the related SOM parameters are presented in Fig. 17. In general, the three SOM products yielded asymptotic semivariograms. The fitted nugget values were close to 0, indicating that microscale variation and measurement errors were relatively small compared with the modeled spatial variation. The stable covariance model closely fitted the experimental variograms calculated for SOM, and the spatial distributions of the predicted SOM were similar among the three models, which was confirmed by the similarities of their semivariogram parameters.

4.3. TCA-based cross-evaluation and merging

By combining classical ETC and CTC algorithms, the three SOM products were cross-validated to achieve the best merging. The global error of the SOM product was first evaluated by the classical TCA method. The calculation procedure is as follows: each SOM map was collapsed into a $1 \times N$ one-dimensional vector, where N is the total number of pixels; the three results were combined into a $3 \times N$ vector set, and pixels containing null values in any product were excluded, and the remaining collocated triplets were input into the TCA algorithm. To assess the applicability of TCA, the mean prediction biases and pairwise residual correlations were examined at the independent validation sites. The mean residuals of the physical-based, feature-based, and DNN-based models were 0.022, 0.569, and -0.636 g/kg, respectively. Their 95% confidence intervals are shown in Table 3. None of the mean residuals differed significantly from zero (all $p > 0.05$), indicating that no evident systematic overestimation or underestimation was detected.

To further examine residual dependence while reducing potential point-to-pixel mismatch and local geolocation effects, a 3×3 spatial window centered on each of the validation sites was used. For each SOM

product, the mean prediction within the neighborhood was calculated and compared with the corresponding field measurement to obtain one site-level residual. Pairwise Pearson correlations among the residuals were 0.25 for the physical-based versus feature-based models, 0.35 for the physical-based versus DNN-based models, and 0.38 for the feature-based versus DNN-based models. These weak-to-moderate correlations indicate that strong cross-error dependence was not evident among the three retrieval methods.

The ETC-based evaluation results are shown in Table 4. The physical-based model achieved an R^2 of 0.839 and a σ of 3.163 g/kg, outperforming the other two methods and indicating the highest global agreement with the unknown true value. The ETC-based evaluation results differ from the point-scale validation results in Table 2, where the DNN-based model achieved the highest predictive accuracy. This difference arises because the two evaluation strategies characterize different aspects of model performance. Field-based validation primarily reflects predictive accuracy at the sampling locations, whereas ETC evaluates the relative error characteristics and map-scale consistency of spatially continuous products. The DNN-based model can effectively capture nonlinear relationships at the validation sites, but its spatial predictions may be more sensitive to the limited number and spatial distribution of field samples. In contrast, the physical-based model, although less accurate at individual validation sites, is constrained by radiative-transfer principles and may provide relatively stable spatial patterns.

In practical digital soil mapping studies, it is common practice to classify soils according to their organic matter content. According to the classification rules of the Second National Soil Survey of China (Yang Fan et al. 2017), SOM content is classified into six levels, as listed in Table 5. Because pixels with SOM contents below 10 g/kg accounted for

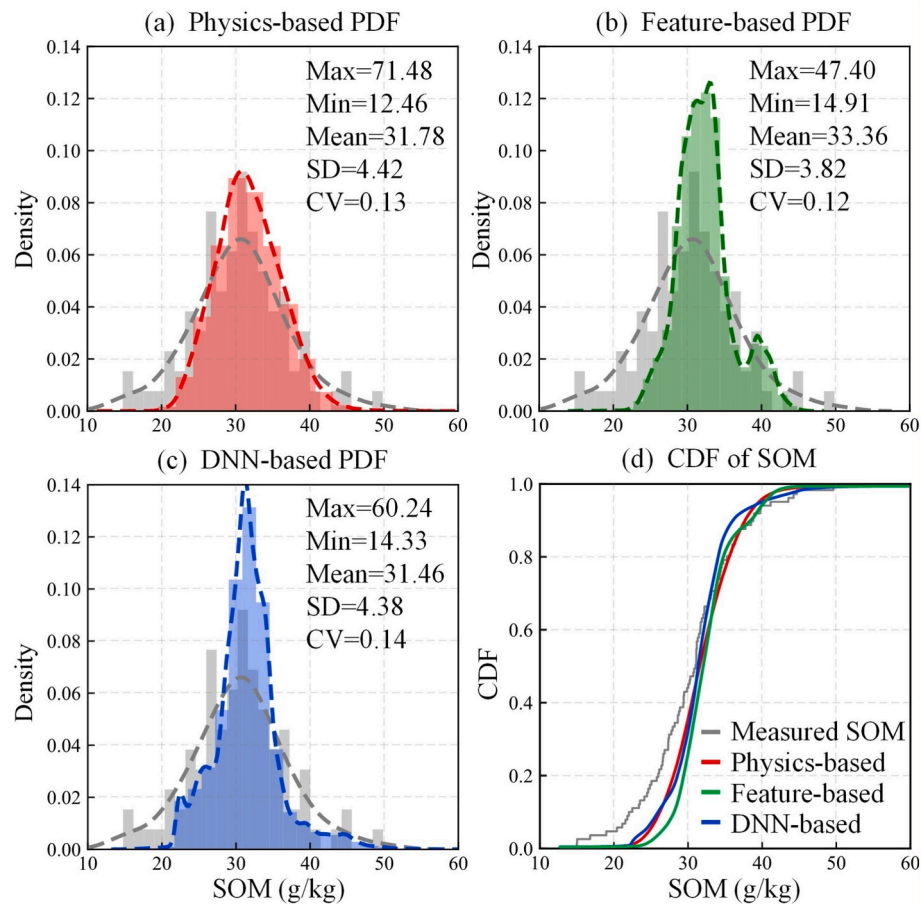


Fig. 15. Histograms of the original soil sample SOM and the SOM predicted by the physical-based (a), feature-based (b), and DNN-based (c) models. The superimposed curve represents the PDF, and the dark gray bars show the original SOM values. (d) The cumulative distribution function (CDF).

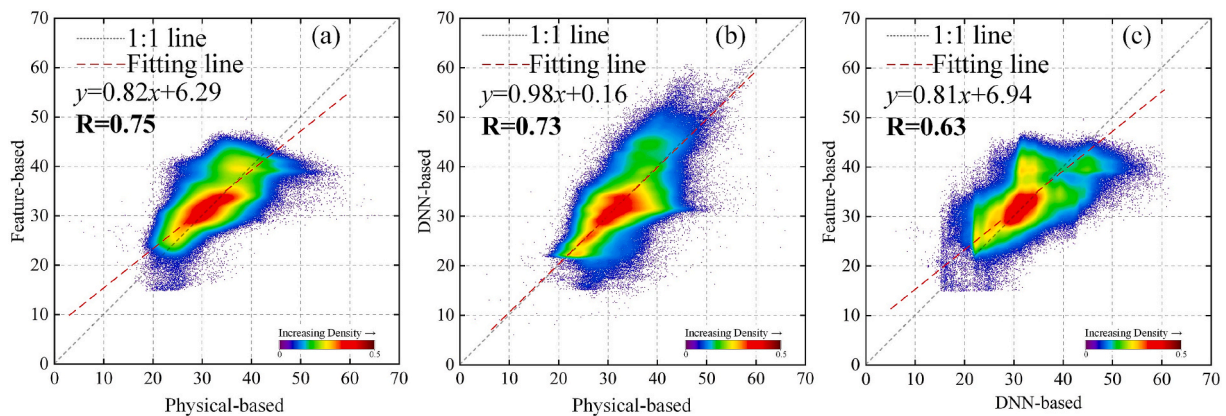


Fig. 16. Correlation analysis and comparison of physical-based, feature-based, and DNN-based models.

only a small proportion of the study area, Levels 4–6 were combined into a single category of SOM < 20 g/kg for the CTC analysis. The specific procedure is as follows: for each organic matter grade, the pixels in the range of the grade were redefined as 1, and the remaining pixels were set to -1. Then, the three products were combined into a $3 \times N$ vector group containing only 1 and -1, which was input into the classification TCA algorithm to obtain the evaluation matrix V . The same procedure was applied to the four SOM categories used in this study, and the results are shown in Table 5.

Fig. 18 visualizes the CTC calculation results of the three models at different organic matter classes. The results showed that the V of the

physical-based model was higher than those of the other two models at Level 2 ($30 \leq \text{SOM} < 40$ g/kg), Level 3 ($20 \leq \text{SOM} < 30$ g/kg), and the combined Level 4–6 category (SOM < 20 g/kg), indicating that the physical-based model performed better in these intervals. In comparison, the DNN-based model achieved the highest V value at Level 1 (SOM ≥ 40 g/kg).

The weights of the three products were calculated according to Eq. (26) and (27). The TCA-based approach is preferred over simple averaging because it can assign optimal weights to the input products and produce a better merged result in regions where the parent datasets differ substantially. The SOM merged result is presented in Fig. 19.

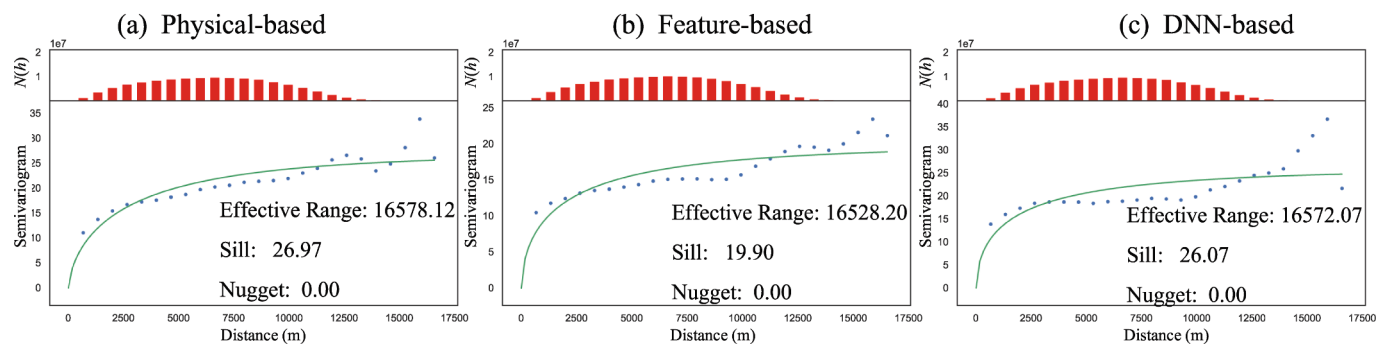


Fig. 17. The theoretical (line) and experimental (marker) semivariograms of SOM.

Table 3			
Residual bias diagnostics for the three SOM models.			
Model	Mean residual (g/kg)	95 % confidence interval (g/kg)	p-value
Physical-based	0.022	(−1.672, 1.716)	0.979
Feature-based	0.569	(−0.759, 1.897)	0.388
DNN-based	−0.636	(−1.876, 0.605)	0.303

Table 4			
ETC-based evaluation.			
Evaluation index	Physical-based	Feature-based	DNN-based
R^2	0.839	0.651	0.619
σ (g/kg)	3.163	5.085	7.315

Table 5				
CTC-based evaluation.				
Organic matter level SOM (g/kg)	Level 1 ≥ 40	Level 2 30–40	Level 3 20–30	Level 4–6 < 20
Physical-based	0.223	0.713	0.752	0.053
Feature-based	0.174	0.519	0.587	0.040
DNN-based	0.258	0.668	0.697	0.023

Although the three products were generated using different retrieval methods, they were derived from the same HyMap imagery and shared several preprocessing steps, including radiometric calibration, atmospheric correction, geometric registration, and soil mask extraction. Previous studies have shown that correlated errors and shared structural error components can introduce systematic deviations into TCA error estimates and subsequently affect the derived fusion weights (Pan et al. 2015; Yilmaz and Crow 2014). Therefore, common uncertainties introduced by the shared imagery and preprocessing may lead to cross-product error correlations, causing part of the common-mode error to be interpreted as shared signal rather than product-specific error. Because an additional fully independent observation is unavailable, the magnitude of this effect cannot be explicitly quantified and remains a source of uncertainty in the current TCA-based analysis. Future work should incorporate more independent observations and explore generalized collocation approaches that can explicitly account for correlated error components.

4.4. SOM quantitative spectral analysis: physical, statistical, or synergistic?

According to Stevenson (Stevenson 1994), SOM is divided into two main categories: (1) unaltered substances, which mainly include plant and animal residues; and (2) transformation products, consisting primarily of humus, semi-decomposition products, and microbial

metabolites. In this regard, SOM is described as a heterogeneous mixture of microbial and chemical transformation products largely composed of organic detritus. Its visible and near-infrared (Vis–NIR) spectral response is also complex because soil reflectance is influenced by overlapping absorption features and interactions among multiple soil constituents (Stenberg et al. 2010). In particular, the broad and weak spectral responses associated with organic matter make it difficult to assign individual absorption features to specific SOM functional groups (Wu et al. 2023b).

A rising trend in quantitative soil property modeling is the use of data-driven machine learning for regression, which often lacks explicit physical theory (Wadoux et al. 2020). Consequently, data-driven methods utilizing remote sensing of soil reflectance properties are gaining popularity for quantifying soil properties. Many of these approaches are “black box” predictive models, lacking explanatory physical mechanisms and explicit covariate relationships. These models are based on statistical learning theory, which directly builds prediction models from spectral data to predict physicochemical properties by mining soil spectral features in a statistical sense, ignoring the radiative transfer process and physical mechanisms of reflectance spectra. In addition, the lack of a physical basis hinders empirical statistical models from deducing essential soil spectral characteristics through remote sensing physics, causing confusion in spectral features during retrieval and introducing uncertainties in SOM retrieval (Ma et al. 2023; Minasny et al. 2024). In comparison, traditional physical models offer explicit structures that clarify covariate relationships, providing inherent explanatory power. The physical model based on the radiative transfer equation has a clear mathematical basis and physical significance, creating a new avenue for SOM retrieval (Ou et al. 2022; Wu et al. 2023a). In this study, the physical-based model achieved lower point-scale R^2 values than the feature-based and DNN-based models, whereas the ETC results indicated relatively high map-scale consistency. However, several limitations should be acknowledged. The SSA basis vectors derived from NMF should not be interpreted as the spectra of strictly pure soil constituents. Instead, they represent constrained spectral components extracted from the mixed SSA matrix under the available observations. In addition, the physical-based model involves multiple parameters related to soil-moisture correction, observation geometry, SSA conversion, and spectral decomposition, which increases the complexity of model implementation and calibration (Wu et al. 2023a). Future studies should conduct controlled laboratory experiments to validate the physical interpretation of the SSA basis vectors, improve parameter optimization through additional physical constraints and regularization, develop regional calibration strategies, and explore integration with data-driven methods to improve predictive accuracy, robustness, and interpretability.

Consequently, machine learning cannot completely replace physical models, and combining both may provide new solutions for the hyper-spectral retrieval of SOM. Instead of constructing the ideal model for every possible situation, future work should concentrate on exploiting the advantages and weaknesses of different modeling approaches in a

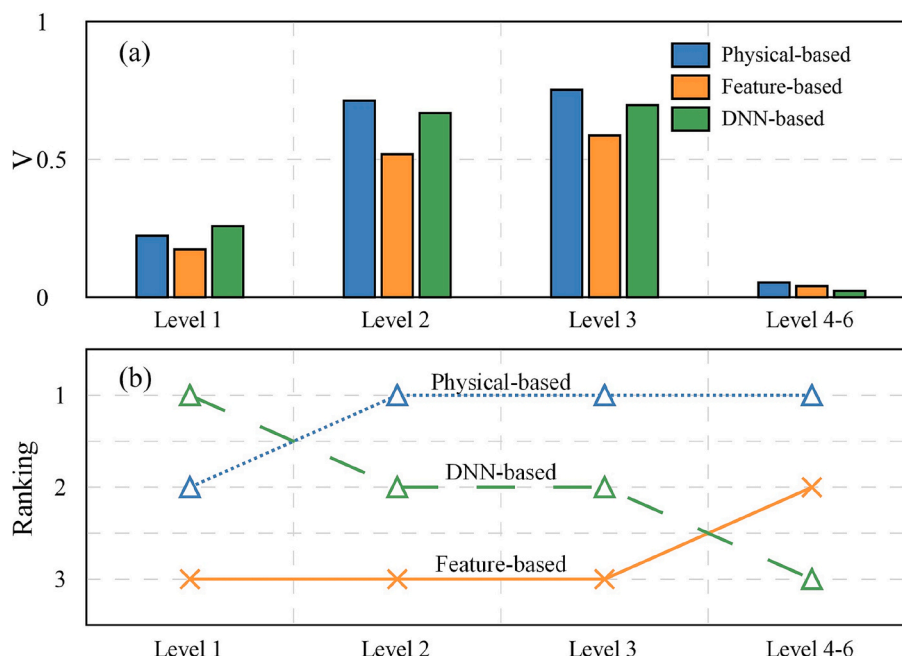


Fig. 18. The estimation of V (a) and rankings (b).

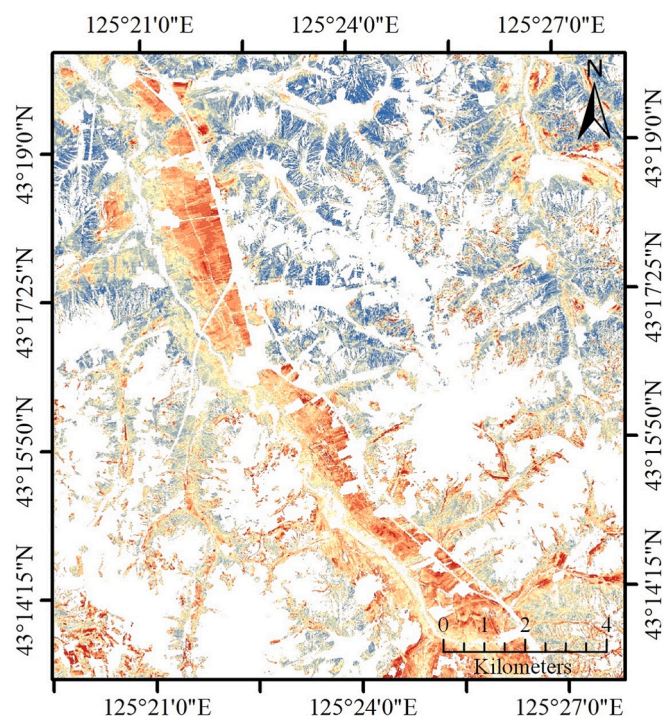


Fig. 19. TCA-based merged SOM map generated from the three model products.

synergistic manner.

5. Conclusion

Remote sensing techniques are being rapidly advanced to retrieve information on soil organic matter (SOM) over large areas. Most current quantitative SOM retrieval methods based on hyperspectral data rely on purely data-driven models, whose performance and interpretability are constrained by the lack of explicit physical mechanisms. In contrast, physical-based models derived from the radiative transfer equation

possess a clear mathematical foundation and physical meaning, and therefore offer strong potential for more robust and interpretable quantitative spectral analysis of SOM.

The main contributions of this study are summarized as follows:

(1) This study developed an SOM retrieval method based on radiative-transfer modeling. We discussed the effect of soil moisture during radiative transfer based on a water film coating approach. The Hapke model was used to convert the nonlinear soil spectra into a linear combination of single scattering albedo (SSA) based on the nonlinear mixing theory of soil spectra. Next, we combined a nonnegative matrix factorization algorithm and a constrained bilinear model to decompose the organic matter basis vector and mass fraction vector from the soil SSA matrix. Finally, SOM retrieval was achieved using a fully constrained least squares algorithm.

(2) A triple collocation analysis (TCA) algorithm was introduced to evaluate and analyze the error distribution of the three models. The methods based on spectral absorption features and semi-supervised deep neural network were utilized as counterparts. The results of the three methods were consistent, with correlation coefficients greater than 0.6. The semivariogram analysis demonstrated that the spatial distribution of the predicted SOM was similar among the three models. The squared correlation coefficient between the physical-based model and the unknown true value was 0.839, outperforming the other two models. The relative accuracy of each model within different organic matter classes was verified using a categorical triple collocation algorithm, which demonstrated the superior performance of the physical-based model. The TCA evaluation algorithm provides a novel approach for the comprehensive evaluation and integration of quantitative remote sensing SOM products.

CRediT authorship contribution statement

Chao Niu: Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation. **Yuan Zhang:** Writing – original draft, Visualization, Software, Methodology. **Yayu Yang:** Visualization, Validation. **Kun Tan:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition, Conceptualization. **Linya Zhao:** Visualization, Validation. **Lijun Tian:** Visualization, Software.

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.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 42171335, 42501494), the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (No.2024CSJZN1300), the Shanghai Municipal Education Commission Science and Technology Project (2024A102002), the National Civil Aerospace Project of China (No. D040102), the China Postdoctoral Science Foundation (2024M760925, 2025T180089).

Data availability

The data that has been used is confidential.

References

- Angelopoulou, T., Balafoutis, A., Zalidis, G., Bochtis, D., 2020. From Laboratory to Proximal Sensing Spectroscopy for Soil Organic Carbon Estimation-A Review. *Sustainability* 12.
- Araujo, M.C.U., Saldanha, T.C.B., Galvao, R.K.H., Yoneyama, T., Chame, H.C., Visani, V., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intel. Lab. Syst.* 57, 65–73.
- Babaeian, E., Sadeghi, M., Franz, T.E., Jones, S., Tuller, M., 2018. Mapping soil moisture with the OPTical TRapezoid Model (OPTRAM) based on long-term MODIS observations. *Remote Sens. Environ.* 211, 425–440.
- Bablet, A., Vu, P.V.H., Jacquemoud, S., Viallefont-Robinet, F., Fabre, S., Briottet, X., Sadeghi, M., Whiting, M.L., Baret, F., Tian, J., 2018. MARMIT: A multilayer radiative transfer model of soil reflectance to estimate surface soil moisture content in the solar domain (400–2500 nm). *Remote Sens. Environ.* 217, 1–17.
- Briedis, C., Baldock, J., Sa, J.C.D., dos Santos, J.B., Milori, D.M.B.P., 2020. Strategies to improve the prediction of bulk soil and fraction organic carbon in Brazilian samples by using an Australian national mid-infrared spectral library. *Geoderma* 373.
- Cai, W.S., Li, Y.K., Shao, X.G., 2008. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra. *Chemom. Intel. Lab. Syst.* 90, 188–194.
- Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R., 2016. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sens. Environ.* 179, 54–65.
- Cecillon, L., Cassagne, N., Czarnes, S., Gros, R., Brun, J.J., 2008. Variable selection in near infrared spectra for the biological characterization of soil and earthworm casts. *Soil Biol. Biochem.* 40, 1975–1979.
- Center for the Study of Earth from Space (1999). *ATmosphere REMoval Program (ATREM) User's Guide, Version 3.1*. In: Boulder, CO, USA: University of Colorado.
- Chen, D., Chang, N.J., Xiao, J.F., Zhou, Q.B., Wu, W.B., 2019. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. *Sci. Total Environ.* 669, 844–855.
- Chen, X., Yuan, F., Ata-Ul-Karim, S.T., Liu, X., Tian, Y., Zhu, Y., Cao, W., Cao, Q., 2025. A bibliometric analysis of research on remote sensing-based monitoring of soil organic matter conducted between 2003 and 2023. *Artif. Intell. Agric.*
- Diek, S., Chabrilat, S., Nocita, M., Schaepman, M.E., de Jong, R., 2019. Minimizing soil moisture variations in multi-temporal airborne imaging spectrometer data for digital soil mapping. *Geoderma* 337, 607–621.
- Ding, Z., Liu, K., Grunwald, S., Smith, P., Ciais, P., Wang, B., Wadoux, A.M.C., Ferreira, C., Karunarathne, S., Shurpali, N., 2025. Advancing Soil Organic Carbon Prediction: A Comprehensive Review of Technologies, AI. *Process-Based and Hybrid Modelling Approaches*, Advanced Science.
- Dong, C., Meng, X., Ruan, W., Cui, J., Zhang, X., Liu, H., 2025. An innovational hyperspectral prediction model for soil organic matter in croplands of the Northeast China Mollisols Region. *Soil Tillage Res.* 253, 106666.
- Dotto, A.C., Dalmolin, R.S.D., ten Caten, A., Grunwald, S., 2018. A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra. *Geoderma* 314, 262–274.
- Dufréhou, G., Grandjean, G., Bourguignon, A., 2015. Geometrical analysis of laboratory soil spectra in the short-wave infrared domain: Clay composition and estimation of the swelling potential. *Geoderma* 243, 92–107.
- Dutta, D., Kumar, P., 2019. A Framework for Global Characterization of Soil Properties Using Repeat Hyperspectral Satellite Data. *IEEE Trans. Geosci. Remote Sens.* 57, 3308–3323.
- Engel, J., Gerretzen, J., Szymanska, E., Jansen, J.J., Downey, G., Blanchet, L., Buydens, L. M.C., 2013. Breaking with trends in pre-processing? Trac-Trends in Analytical Chemistry 50, 96–106.
- Fox, G.A., Metla, R., 2005. Soil property analysis using principal components analysis, soil line, and regression models. *Soil Sci. Soc. Am. J.* 69, 1782–1788.
- Fox, G.A., Sabbagh, G.J., 2002. Estimation of soil organic matter from red and near-infrared remotely sensed data using a soil line Euclidean distance technique. *Soil Sci. Soc. Am. J.* 66, 1922–1929.
- Gruber, A., Dorigo, W.A., Crow, W., Wagner, W., 2017. Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals. *IEEE Trans. Geosci. Remote Sens.* 55, 6780–6792.
- Gruber, A., Su, C.-H., Zwieback, S., Crow, W., Dorigo, W., Wagner, W., 2016. Recent advances in (soil moisture) triple collocation analysis. *Int. J. Appl. Earth Obs. Geoinf.* 45, 200–211.
- Hapke, B., 2008. Bidirectional reflectance spectroscopy - 6. Effects of porosity. *Icarus* 195, 918–926.
- He, T., Wang, J., Lin, Z., Cheng, Y., 2009. Spectral features of soil organic matter. *Geo-spatial Inf. Sci.* 12, 33–40.
- Hong, Y.S., Guo, L., Chen, S.C., Linderman, M., Mouazen, A.M., Yu, L., Chen, Y.Y., Liu, Y. L., Liu, Y.F., Cheng, H., Liu, Y., 2020. Exploring the potential of airborne hyperspectral image for estimating topsoil organic carbon: Effects of fractional-order derivative and optimal band combination algorithm. *Geoderma* 365.
- Kim, M.-J., Lee, H.-I., Choi, J.-H., Lim, K.J., Mo, C., 2022. Development of a soil organic matter content prediction model based on supervised learning using vis-NIR/SWIR spectroscopy. *Sensors* 22, 5129.
- Kogel-Knabner, I., Amelung, W., 2021. Soil organic matter in major pedogenic soil groups. *Geoderma* 384.
- Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma* 352, 395–413.
- Leardi, R., 2000. Application of genetic algorithm-PLS for feature selection in spectral data sets. *J. Chemom.* 14, 643–655.
- Li, H.D., Liang, Y.Z., Xu, Q.S., Cao, D.S., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 648, 77–84.
- Liu, W., Chang, Q.R., Guo, M., Xing, D.X., Yuan, Y.S., 2011. Extraction of First Derivative Spectrum Features of Soil Organic Matter via Wavelet De-Noising. *Spectrosc. Spectr. Anal.* 31, 100–104.
- Liu, Y., Chen, S., Yu, Q., Cai, Z., Zhou, Q., Bellingrath-Kimura, S.D., Wu, W., 2023. Improving digital mapping of soil organic matter in cropland by incorporating crop rotation. *Geoderma* 438, 116620.
- Ma, Y., Minasny, B., Dematté, J.A., McBratney, A.B., 2023. Incorporating soil knowledge into machine-learning prediction of soil properties from soil spectra. *Eur. J. Soil Sci.* 74, e13438.
- McColl, K.A., Roy, A., Derksen, C., Konings, A.G., Alemohammed, S.H., Entekhabi, D., 2016. Triple collocation for binary and categorical variables: Application to validating landscape freeze/thaw retrievals. *Remote Sens. Environ.* 176, 31–42.
- McColl, K.A., Vogelzang, J., Konings, A.G., Entekhabi, D., Piles, M., Stoffelen, A., 2014. Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophys. Res. Lett.* 41, 6229–6236.
- Meng, X., Bao, Y., Zhang, X., Luo, C., Liu, H., 2025. A long-term global Mollisols SOC content prediction framework: Integrating prior knowledge, geographical partitioning, and deep learning models with spatio-temporal validation. *Remote Sens. Environ.* 318, 114592.
- Meng, X.T., Bao, Y.L., Liu, J.G., Liu, H.J., Zhang, X.L., Zhang, Y., Wang, P., Tang, H.T., Kong, F.C., 2020. Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *Int. J. Appl. Earth Obs. Geoinf.* 89.
- Minasny, B., Bandai, T., Ghezzehei, T.A., Huang, Y.-C., Ma, Y., McBratney, A.B., Ng, W., Norouzi, S., Padarian, J., Sharififar, A., 2024. Soil science-informed machine learning. *Geoderma* 452, 117094.
- Niu, C., Tan, K., Wang, X., Pan, C., 2025. Mapping nutrient pollution in inland water bodies using multi-platform hyperspectral imagery and deep regression network. *J. Hazard. Mater.* 488, 137314.
- Niu, C., Zhu, L., Tan, K., Wang, X., Niu, R., Xu, X., 2026. Large-scale monitoring of soil organic matter and salinization using satellite hyperspectral imagery and a deep residual attention network. *IEEE Trans. Geosci. Remote Sens.*
- Ortiz, B.V., Perry, C., Goovaerts, P., Vellidis, G., Sullivan, D., 2010. Geostatistical modeling of the spatial variability and risk areas of southern root-knot nematodes in relation to soil properties. *Geoderma* 156, 243–252.
- Ou, D., Tan, K., Wang, X., Wu, Z., Li, J., Ding, J., 2022. Modified soil scattering coefficients for organic matter inversion based on Kubelka-Munk theory. *Geoderma* 418, 115845.
- Ou, D.P., Tan, K., Lai, J., Jia, X.P., Wang, X., Chen, Y., Li, J., 2021. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. *Geoderma* 385.
- Padarian, J., Minasny, B., McBratney, A.B., 2019. Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma* 340, 279–288.
- Padarian, J., Stockmann, U., Minasny, B., McBratney, A., 2022. Monitoring changes in global soil organic carbon stocks from space. *Remote Sens. Environ.* 281, 113260.
- Pan, M., Fisher, C.K., Chaney, N.W., Zhan, W., Crow, W.T., Aires, F., Entekhabi, D., Wood, E.F., 2015. Triple collocation: Beyond three estimates and separation of structural/non-structural errors. *Remote Sens. Environ.* 171, 299–310.
- Pudelko, A., Chodak, M., 2020. Estimation of total nitrogen and organic carbon contents in mine soils with NIR reflectance spectroscopy and various chemometric methods. *Geoderma* 368.
- Saberioon, M., Gholizadeh, A., Ghaznavi, A., Chabrilat, S., Khosravi, V., 2024. Enhancing soil organic carbon prediction of LUCAS soil database using deep learning and deep feature selection. *Comput. Electron. Agric.* 227, 109494.

- Shepherd, J.E., Kanner, O., Amir, O., Ben-Zion, K., Ben-Dor, E., 2025. Rapid retrieval of soil surface aggregation as a joint attribute to soil spectral libraries. *Geoderma* 462, 117522.
- State Forestry Administration of China (1999). LY/T 1237-1999. Determination of organic matter in forest soil. In: Beijing, China: Standards Press of China.
- Stenberg, B., Rossel, R.A.V., Mouazen, A.M., Wetterlind, J., 2010. Visible and near infrared spectroscopy in soil science. *Adv. Agron.* 107, 163–215.
- Stevenson, F.J., 1994. *Humus chemistry: genesis, composition, reactions*. John Wiley & Sons.
- Thaler, E.A., Larsen, I.J., Yu, Q., 2019. A New Index for Remote Sensing of Soil Organic Carbon Based Solely on Visible Wavelengths. *Soil Sci. Soc. Am. J.* 83, 1443–1450.
- Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* 367.
- Tziolas, N., Tsakiridis, N., Ogen, Y., Kalopesa, E., Ben-Dor, E., Theocharis, J., Zalidis, G., 2020. An integrated methodology using open soil spectral libraries and Earth Observation data for soil organic carbon estimations in support of soil-related SDGs. *Remote Sens. Environ.* 244.
- Vivo-Truyols, G., Schoenmakers, P.J., 2006. Automatic selection of optimal Savitzky-Golay smoothing. *Anal. Chem.* 78, 4598–4608.
- Wadoux, A.-M.-C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth Sci. Rev.* 210, 103359.
- Wang, X., Li, L., Liu, H., Song, K., Wang, L., Meng, X., 2022. Prediction of soil organic matter using VNIR spectral parameters extracted from shape characteristics. *Soil Tillage Res.* 216, 105241.
- Wijewardane, N.K., Ge, Y.F., Morgan, C.L.S., 2016. Moisture insensitive prediction of soil properties from VNIR reflectance spectra based on external parameter orthogonalization. *Geoderma* 267, 92–101.
- Wu, F., Tan, K., Wang, X., Ding, J., Liu, Z., 2023a. A novel semi-empirical soil multi-factor radiative transfer model for soil organic matter estimation based on hyperspectral imagery. *Geoderma* 437, 116605.
- Wu, F., Tan, K., Wang, X., Ding, J., Liu, Z., Han, B., 2023b. A semi-analytical radiative transfer model for explaining soil spectral features. *Int. J. Appl. Earth Obs. Geoinf.* 118, 103250.
- Xia, K., Wu, T., Zhang, S., Wang, S., Li, X., Shen, Q., Li, T., 2024. A new method for high-precision estimation of soil organic matter using two-dimensional correlation spectroscopy—to support collaborative use of global open soil spectral libraries. *Geoderma* 445, 116877.
- Yang Fan, Y.F., Xu Yang, X.Y., Cui Yong, C.Y., Meng YuanDuo, M.Y., Dong Yan, D.Y., Li Rong, L.R., Ma YiBing, M.Y., 2017. Variation of soil organic matter content in croplands of China over the last three decades. *Acta Pedol. Sin.* 54, 1047–1056.
- Yang, P.Q., van der Tol, C., Yin, T.G., Verhoef, W., 2020. The SPART model: A soil-plant-atmosphere radiative transfer model for satellite measurements in the solar spectrum. *Remote Sens. Environ.* 247.
- Yilmaz, M.T., Crow, W.T., 2014. Evaluation of assumptions in soil moisture triple collocation analysis. *J. Hydrometeorol.* 15, 1293–1302.
- Yuan, J., Hu, C.H., Yan, C.X., Li, Z.Z., Chen, S.B., Wang, S.R., Wang, X., Xu, Z.Y., Ju, X.P., 2019. Semi-Empirical Soil Organic Matter Retrieval Model With Spectral Reflectance. *IEEE Access* 7, 134164–134172.
- Yuan, J., Wang, X., Yan, C., Chen, S., Wang, S., Zhang, J., Xu, Z., Ju, X., Ding, N., Dong, Y., 2020. Wavelength selection for estimating soil organic matter contents through the radiative transfer model. *IEEE Access* 8, 176286–176293.
- Zeraatpisheh, M., Garosi, Y., Owliaie, H.R., Ayoubi, S., Taghizadeh-Mehrjardi, R., Scholten, T., Xu, M., 2022. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. *Catena* 208, 105723.