



1 **Quantifying multiple sources of uncertainty in national-scale soil organic matter prediction**

2

3 **Juhwan Lee^{1,*}, Woosik Lee^{2,*}, and Sumin Hwang¹**

4 ¹Department of Smart Agro-Industry, Gyeongsang National University, Jinju 52725, Republic of
5 Korea

6 ²Department of Smart Distribution and Logistics, Gyeongsang National University, Jinju 52725,
7 Republic of Korea

8

9 *Correspondence to:* Juhwan Lee (juhwan.lee@gnu.ac.kr)

10 (will be submitted to Biogeosciences)

11

12 * These authors contributed equally to this work.

13

14



15 **Abstract**

16

17 Large-scale assessment of soil organic matter (SOM) relies on spatial prediction to extend sparse
18 observations to unsampled locations, but extensive soil databases do not necessarily ensure
19 reliable predictions. Predictive uncertainty can arise from limitations in sampling, response-
20 variable perturbation, and model specification, yet their relative contributions are rarely
21 quantified. Here, we developed nationwide models of SOM in paddy soils across South Korea
22 and used an ensemble random forest framework to decompose predictive uncertainty into
23 sampling, response-variable perturbation, and model-specification components. We compiled
24 283,618 observations with soil, terrain, climate, and vegetation predictors. The model explained
25 50% of the variation in independent SOM observations but underpredicted localized areas with
26 high SOM. Model specification was the largest contributor to total predictive variance (41.9%)
27 and was the dominant uncertainty source at 55.8% of locations, followed by sampling (29.3%)
28 and response-variable perturbation uncertainty (28.8%). Exchangeable calcium and annual
29 precipitation were the most influential predictors, although their importance rankings varied
30 across model configurations. These results show that predictive performance alone does not
31 reveal the sources of uncertainty in large-scale SOM assessment. Large datasets do not
32 necessarily provide equal predictive information across the response distribution. Explicit
33 decomposition of predictive uncertainty can identify whether improvements should prioritize
34 additional sampling, improved response-variable measurements, or alternative model
35 specifications, providing a basis for more targeted strategies to improve the reliability of large-
36 scale SOM predictions.

37

38 **Keywords:** soil organic matter; predictive uncertainty; paddy soils; random forest; uncertainty
39 decomposition

40

41

42



43 1 Introduction

44

45 Soil organic matter (SOM) consists of plant and animal residues returned to the soil at various
46 stages of decomposition, together with associated organic compounds, water, and inorganic
47 nutrients. It contributes to the formation of stable soil aggregates, supports microbial activity,
48 facilitates carbon sequestration, and plays a fundamental role in nutrient cycling, soil structure,
49 and water retention (Reeves, 1997; Hartmann and Six, 2023). Both the quantity and quality of
50 SOM underpin a wide range of ecosystem services essential to agricultural productivity and soil
51 quality. Consequently, maintaining sufficient SOM has become increasingly important under
52 global environmental change, particularly in the face of rapid climate shifts (Amundson et al.,
53 2015).

54 Because SOM is not controlled by a single process but emerges from the interactions
55 among organic inputs, decomposition, stabilization, and environmental controls (Six et al., 2002;
56 Cotrufo and Lavalley, 2022), uncertainty in its spatial prediction has implications for both
57 estimation accuracy and the interpretation of broad-scale biogeochemical patterns and inferred
58 environmental controls. Agricultural management can modify these processes through organic
59 amendments, fertilizer application, residue retention, reduced tillage, crop rotations, cover crops,
60 and other management practices (Lundy et al., 2015; Pittelkow et al., 2015a; Pittelkow et al.,
61 2015b; Young et al., 2021). Such management affects both the quantity and quality of SOM and
62 can produce cumulative effects over time. As a result, SOM reflects both current environmental
63 conditions and the cumulative legacy of past management, which vary across the landscape and
64 contribute to substantial spatial and temporal variability, complicating large-scale SOM
65 prediction. Although detailed process-level studies clarify how SOM formed and stabilized at the
66 plot or profile scale, extending this understanding to large spatial domains remains challenging.
67 Consequently, large-scale SOM assessment commonly relies on measured soil properties and
68 environmental variables as proxies for the drivers of SOM dynamics.

69 Large soil databases are increasingly used to infer spatial patterns and environmental
70 controls of SOM, but uncertainty associated with how observations are sampled, represented,
71 and modeled can propagate into both predictions and variable-importance inference. Because soil
72 observations are typically sparse, collected at discrete locations, and unevenly distributed across
73 large agricultural regions, direct measurements alone cannot characterize the spatial distribution
74 of SOM. Predictive models are therefore used to extend location-specific observations to
75 unsampled locations by relating measured SOM to soil, environmental, and other spatially
76 available predictors. Such predictions enable spatially explicit assessment of SOM at regional
77 and national scales and provide a basis for evaluating spatial patterns and potential management
78 needs. Advances in large-scale monitoring, soil information systems, and agricultural data
79 modeling have further expanded the capacity for spatial prediction of soil properties and crop
80 performance (Farina et al., 2021).

81 However, extending observations to unsampled locations through spatial prediction
82 introduces uncertainty, because these predictions depend on several sources of uncertainty,
83 including the composition of available observations and sampling design, the representation of
84 environmental covariates, and model specification and parameterization. Soil samples capture
85 only a part of spatial variation in SOM and are subject to sampling and measurement error,
86 whereas soil and environmental predictors are imperfect proxies for its underlying drivers (Piikki
87 et al., 2021; Lamichhane et al., 2019). Different model specifications may also produce divergent
88 predictions from the same data (Sylvain et al., 2021; Rohmer, 2025). These sources of



89 uncertainty are interconnected and can propagate through the modeling process, affecting the
90 reliability of SOM estimates across large agricultural regions.

91 This issue is particularly important for national-scale assessment, where soil information
92 systems provide large numbers of observations for spatial prediction, yet data volume alone does
93 not ensure reliable SOM estimates. A model may achieve high predictive accuracy while
94 remaining sensitive to sampling, response-variable perturbation, or model specification.
95 Quantifying the relative contributions of these uncertainty sources can identify whether
96 sampling, response-variable perturbation, or model specification most warrant improvement.
97 Assessing predictive uncertainty alongside prediction accuracy is therefore essential for
98 evaluating the reliability of large-scale SOM estimates and prioritizing improvements across
99 these components.

100 National and continental soil information systems now offer extensive observational
101 coverage for spatial prediction. However, applications of these systems have often emphasized
102 predictive accuracy, with comparatively less attention given to how predictive uncertainty
103 propagates through sampling, response-variable perturbation, and model specification (Piikki et
104 al., 2021; Sylvain et al., 2021). The Republic of Korea (hereafter South Korea) provides an
105 example of a national soil information system built on systematic soil testing and monitoring.
106 The Rural Development Administration (RDA) maintains a comprehensive soil test database and
107 associated soil attribute maps, compiling laboratory analyses of organic matter, pH,
108 exchangeable nutrients, and electrical conductivity from agricultural fields over several decades
109 (Lee, 2025). Similar large-scale soil information systems have been developed in other countries.
110 In Europe, the LUCAS topsoil database provides approximately 38,000 georeferenced samples
111 with harmonized observations of soil physical and chemical properties, together with derived
112 raster indicators for environmental modeling and policy analysis (Orgiazzi et al., 2018). In the
113 United States, the Soil Survey Geographic Database, maintained by the USDA Natural
114 Resources Conservation Service, provides detailed soil maps, including organic carbon (NRCS,
115 2004). In China, national-scale soil organic carbon datasets compiled from 7,852 soil profiles
116 collected between 2010 and 2024 have been standardized with geographic and environmental
117 covariates to support spatial prediction and modeling across diverse landscapes (Chen et al.,
118 2025).

119 Building on these datasets, this study develops nationwide models of SOM under rice
120 cultivation and evaluates three sources of predictive uncertainty. Uncertainty is considered across
121 the interconnected stages through which observed soil information is converted into a spatial
122 prediction, rather than being treated as a single property of the final model. We therefore quantify
123 uncertainty associated with sampling, response-variable perturbation, and model-specification
124 and compare their relative contributions to total predictive uncertainty. This decomposition can
125 help identify which components of these sources most strongly constrain the reliability of SOM
126 estimates, providing a basis for prioritizing improvements in monitoring and modeling strategies.

127 Paddy rice systems are well suited to examine this issue because their soils experience
128 environmental conditions that differ substantially from those of upland agricultural systems
129 (Geisseler et al., 2017). Periodic flooding alters soil oxygen availability and redox conditions,
130 thereby affecting microbial processes, organic matter decomposition, and carbon stabilization,
131 which in turn influence SOM dynamics that differ from those of upland systems (Conrad, 2020;
132 Chen et al., 2021). Paddy fields also occupy a large proportion of agricultural land in South
133 Korea and are distributed widely across the country. In 2025, the total paddy field area was
134 755,952 ha, accounting for approximately 50.4% of the national cropland, of which 673,930 ha



were cultivated with paddy rice (KOSIS, 2026). The specific objectives were therefore to (1) develop nationwide models relating SOM to soil chemical, physical, and categorical properties while incorporating environmental influences, (2) quantify predictive uncertainty stemming from sampling, response-variable perturbation, and model specification, and (3) identify influential predictors of SOM through variable-importance analysis.

2 Materials and Methods

2.1 Soil chemical data and sources

Soil chemical property data were obtained from the soil information system operated by the Rural Development Administration (RDA) of South Korea (NAS, 2025). This database compiles routine soil test results used to optimize manure and nitrogen fertilizer applications and to determine lime requirements. The database includes eight soil chemical properties: soil pH (1:5 H₂O), available phosphorus (AP; P₂O₅, mg kg⁻¹), available silicon dioxide (AS; SiO₂, mg kg⁻¹), organic matter (OM, g kg⁻¹), exchangeable magnesium (Mg, cmolc kg⁻¹), exchangeable potassium (K, cmolc kg⁻¹), exchangeable calcium (Ca, cmolc kg⁻¹), and electrical conductivity (EC, dS m⁻¹). All properties were determined using standardized laboratory protocols established by the National Institute of Agricultural Sciences (NAS, 2010). These soil observations were collected at multiple sampling locations within fields to characterize soil conditions across the field and support soil management and fertilizer recommendations for the 1,000 m² (10 a) reference-area scale in South Korea (NAS, 2019).

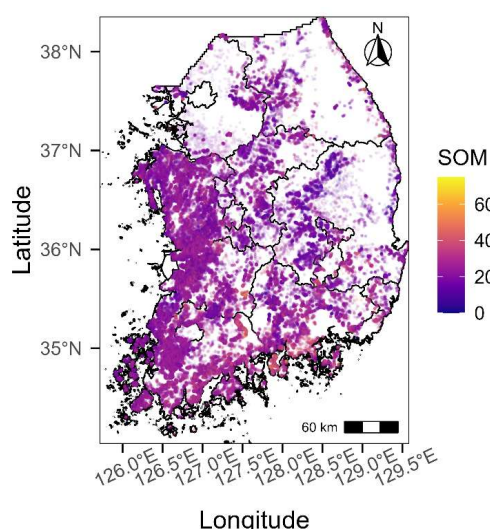
Soil chemical records were retrieved using 10-digit administrative division codes (BJD codes) corresponding to legally defined local administrative units in South Korea. A list of BJD codes was compiled from administrative division boundary data available through the National Spatial Data Infrastructure (MOLIT, 2026b, c), yielding 20,235 codes in the version used for this study. Soil records were then queried for each BJD code through an application programming interface, and units for which records were unavailable were excluded (Lee, 2025). For each administrative unit with available records, individual fields were geocoded based on their recorded road-name addresses and saved as separate unit-level datasets, resulting in 16,683 datasets. These datasets were compiled to include all soil records analyzed between 2022 and 2025, along with associated metadata such as postal address, agricultural land use, and sampling date. All data were accessed under publicly available terms of use.

For this study, the dataset was filtered to include only paddy soils and screened for quality control (Lee, 2025), resulting in a nationwide collection of 763,625 soil records (Figure 1). Each record represents a unique combination of field, sampling date, and laboratory replication on the same date. Individual fields were identified from field addresses recorded in the soil information system. The geographic coordinates of fields were spatially matched with cadastral parcel polygons to assign parcel-based identification numbers (PNU codes) (MOLIT, 2026a). The resulting dataset provided nationwide spatial coverage for statistical modeling of soil chemical variability.

Systematic error screening was applied to the compiled raw soil dataset to improve data reliability and ensure its suitability for subsequent analyses. The procedure followed the error-screening methods described by Lee (2025), addressing instrumental and human-induced errors commonly encountered in large soil monitoring databases. Instrumental anomalies were removed first, including records with values outside analytical detection limits or physically implausible



181 measurement ranges. Record-level quality control checks were then conducted to identify
182 human-induced errors associated with manual data entry and data management. These
183 procedures included the removal of duplicate records; records with all soil variables missing;
184 records in which all soil variables had identical values, including cases with only a single non-
185 missing variable; records exhibiting repeated identical soil test values across consecutive
186 sampling dates; and records lacking valid spatial (X–Y) coordinate information.
187
188



189
190
191 **Figure 1.** Study area and spatial distribution of the paddy-field sampling locations ($n = 283,618$).
192 Points represent location-level mean soil organic matter (SOM; g kg^{-1}) from soil samples
193 collected from 2022 to 2025.

196 2.2 Soil data integration with auxiliary data

197
198 The dataset used for SOM modeling included eight soil chemical variables and 32 auxiliary
199 attributes: 14 categorical soil thematic attributes, 10 topographic attributes, seven climate
200 variables, and one vegetation variable. Soil thematic attributes were obtained from the RDA soil
201 information system, which provides nationwide coverage of 27 physical and land suitability
202 properties at a 125 m spatial resolution (1:25,000 scale). All 27 attributes were initially retrieved
203 using the unique PNU codes, of which 14 were retained as predictors for SOM modeling. The
204 excluded attributes represented land-use suitability for rice paddies, upland crops, orchards,
205 pasture, and forest. These variables were excluded because they describe crop- production
206 suitability rather than inherent soil conditions. The retained categorical soil thematic variables
207 represented soil profile characteristics and soil classification. They were linked to the soil



chemical records via PNU identifiers and used as categorical predictors representing spatial variation in soil characteristics and soil-forming conditions.

Vegetation was represented by annual net primary production (NPP, $\text{g C m}^{-2} \text{ yr}^{-1}$) for 2019–2021, derived from MODIS satellite observations (Terra and Aqua) (NEO, 2023). These data were obtained at a spatial resolution of 0.1° ($\sim 11 \text{ km}$) in GeoTIFF format. Climate variables were obtained from the Modified Korean-Parameter-elevation Regressions on Independent Slopes Model (MK-PRISM, version 1.2) provided by the Korea Meteorological Administration (KMA, 2023). Climate predictors included mean annual air temperature ($^\circ\text{C}$), annual maximum and minimum air temperature ($^\circ\text{C}$), annual precipitation (mm), solar irradiation ($\text{MJ m}^{-2} \text{ d}^{-1}$), relative humidity (%), and wind speed (m s^{-1}), all provided at a spatial resolution of approximately 1 km. All climate variables were aggregated to long-term means over the available observation period, while precipitation was summed annually and then averaged. These vegetation and climate layers were used as environmental covariates in the modeling framework.

A digital elevation model (DEM) for South Korea was obtained from the National Spatial Data Infrastructure at 90 m spatial resolution and provided in the GRS80 / Korea 2000 Central Belt coordinate reference system (EPSG:5186). The DEM was cropped and masked to the national boundary. Topographic attributes, including slope, aspect, topographic position index (TPI), terrain ruggedness index (TRI), surface roughness, and flow direction, were derived from the projected DEM using the *terra* package (Hijmans, 2025). Hillshade was calculated from slope and aspect using a solar elevation angle of 45° and an azimuth of 315° . Sinks in the DEM were filled before calculating the upslope contributing area and topographic wetness index (TWI) using the TOPMODEL approach implemented in the *topmodel* package (Buytaert, 2018). The upslope contributing area was log-transformed, whereas TWI was retained on its original scale. The resulting elevation and terrain layers were subsequently reprojected to geographic coordinates (WGS84; EPSG:4326) for integration with other spatial datasets. These variables were used as topographic covariates in subsequent analyses.

Soil chemical data were spatially integrated with auxiliary datasets based on field coordinates. Following data validation and identification of complete cases across the variables used for modeling, the dataset consisted of 283,618 soil records (Table 1). Given the large sample size, no imputation was applied; only records that passed all quality control procedures were retained for modeling. Continuous variables were summarized using the mean, median, standard deviation (SD), interquartile range (IQR), and range (maximum–minimum). The coefficient of variation (CV) was calculated as the SD divided by the mean and expressed as a percentage. CV was not reported for variables with means close to zero or non-positive means, for which the metric is not informative, or for DEM-derived variables, which had zero-centered distributions.



Table 1. Summary statistics of numerical variables used in the dataset after quality control ($n = 283,618$).

Variable	Mean	Median	SD	IQR	CV (%)	Range
SOM (g kg ⁻¹)	23.7	23.0	9.0	10.0	38.0	1–300
pH (1:5 H ₂ O)	6.2	6.1	0.6	0.7	9.3	3.5–9.5
AP (mg kg ⁻¹)	148.7	99.3	183.0	112.3	123.1	1–9999
AS (mg kg ⁻¹)	245.6	198.0	192.5	171.1	78.4	5–7032
Mg (cmolc kg ⁻¹)	1.6	1.3	1.2	1.1	71.4	0.1–25
K (cmolc kg ⁻¹)	0.5	0.3	0.5	0.3	102.9	0–15
Ca (cmolc kg ⁻¹)	5.9	5.5	2.6	2.6	43.9	0.1–35
EC (dS m ⁻¹)	0.9	0.5	1.4	0.6	153.0	0–30
DEM (m)	59.8	26.0	79.4	76.0	132.8	0–769
Slope (rad)	0.0	0.0	0.1	0.1	—	0–0.5
Aspect (rad)	3.3	3.4	1.5	2.6	—	0–6.3
TPI (m)	−1.1	−0.6	1.7	1.6	—	−19.4–18
TRI (m)	3.0	2.0	3.2	3.6	—	0–39.7
Roughness (m)	10.0	6.6	10.5	12.0	—	0–118.6
Flow direction (°)	25.7	14.4	29.4	33.9	—	0.5–128
Hillshade	0.7	0.7	0.0	0.0	—	0.3–0.9
Log upslope area	10.6	10.7	0.7	1.1	—	9–13.1
TWI	12.1	11.9	1.9	2.6	—	6.3–25.9
TA (°C)	12.8	13.0	1.1	1.3	8.4	6.9–15.2
TAMAX (°C)	18.5	18.6	0.9	1.1	5.0	12.2–20.6
TAMIN (°C)	7.9	8.1	1.5	1.8	18.9	0.9–11.7
RN (mm)	1,297.9	1,250.7	170.3	216.2	13.1	871.1–2636.7
RHM (%)	70.9	71.6	3.1	4.8	4.4	38.6–77.6
SI (MJ m ⁻² d ⁻¹)	13.9	13.9	0.6	0.9	4.5	9.7–15.9
WS (m s ⁻¹)	1.8	1.7	0.5	0.7	25.2	0.2–4.1
NPP (g C m ⁻² y ⁻¹)	183.1	180.5	14.6	13.8	8.0	151.5–255

SOM, soil organic matter; pH, soil pH; AP, available phosphorus; AS, available SiO₂; Mg, exchangeable magnesium; K, exchangeable potassium; Ca, exchangeable calcium; EC, electrical conductivity; DEM, digital elevation model; TPI, topographic position index; TRI, terrain ruggedness index; TWI, topographic wetness index; TA, mean annual temperature; TAMAX, mean annual maximum temperature; TAMIN, mean annual minimum temperature; RHM, mean annual relative humidity; RN, annual precipitation; SI, solar irradiation; WS, wind speed; NPP, net primary productivity. —, dimensionless.



2.3 Ensemble modeling, validation, and predictive uncertainty

Predictive variability associated with sampling, response-variable perturbation uncertainty, and model specification was evaluated using an ensemble random forest (RF) modeling framework (Figure 2). Soil observations collected at identical spatial coordinates during the sampling period were aggregated by calculating the arithmetic mean of each soil chemical property. The resulting values therefore represent location-level mean soil conditions over the sampling period rather than measurements from a single sampling occasion. The resulting modeling dataset contained 283,618 location-level observations with complete predictor and response data.

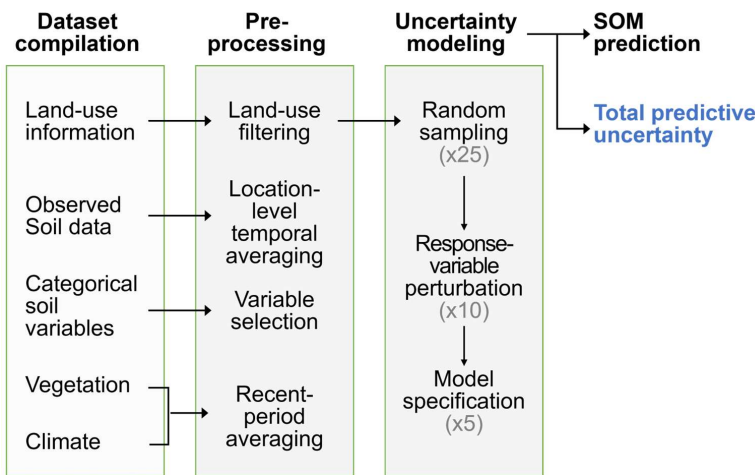


Figure 2. Uncertainty modeling framework for quantifying uncertainty by source at the national scale, including sampling, response-variable perturbation, and model-specification uncertainty.

Sampling variability was represented by repeatedly partitioning the dataset into calibration and validation subsets. This component quantified prediction variability arising from alternative calibration samples drawn from the available observations. Because the nationwide soil-testing data were collected through operational or farmer requests rather than according to a predetermined spatial sampling strategy, the available observations may not provide spatially uniform representation of soil conditions. For each of 25 sampling iterations, the dataset was randomly divided into calibration (60%) and validation (40%) subsets using stratified random sampling after dividing the response variable into six quantile-based bins. Each iteration was initialized using a different random seed. The corresponding validation subset was retained for independent prediction and was not used for model fitting.

Response-variable perturbation uncertainty was represented by an assumed 10% relative perturbation of observed SOM values to approximate uncertainty associated with the response measurements and representation. For each sampling iteration, 10 independent perturbed



calibration datasets were generated. Multiplicative Gaussian noise was applied to the original SOM values as:

$$y_i^* = \max(0, y_i(1 + \varepsilon_i))$$

where y_i is the observed SOM value at location i , y_i^* is the perturbed SOM value after introducing measurement error, and ε_i is a random error term drawn from $N(0, \sigma_y^2)$. The standard deviation was set to $\sigma_y = 0.10$, representing an assumed 10% relative uncertainty in the SOM response variable (Goidts et al., 2009; Samsonova et al., 2023). Perturbed values were constrained to non-negative values using the maximum function and subsequently transformed using the shifted natural logarithm, $\ln(1 + y_i^*)$, prior to model fitting. Validation observations were transformed using the same shifted natural logarithm but were not subjected to measurement-error perturbation.

RF model-specification variability was represented by varying the number of predictor variables randomly selected at each tree split (*mtry*). For each combination of sampling iteration and response-variable perturbation, RF models were fitted using the ranger implementation (Breiman, 2001; Wright and Ziegler, 2017). Five *mtry* values (5, 10, 15, 20, and 35) were evaluated for each calibration dataset and response-variable perturbation, spanning a range of subset sizes relative to the available predictors. Tail weighting was applied during model fitting to increase the influence of high-SOM observations and reduce their potential underprediction, with the 90th percentile used as the tail threshold and a tail weight of 5. Models were fitted without an additional internal cross-validation procedure because *mtry* was evaluated across the ensemble.

The combination of sampling, response-variable perturbation, and model specification resulted in 1,250 RF ensemble model realizations ($S = 25$ sampling iterations $\times D = 10$ response-variable perturbations $\times M = 5$ *mtry* values). Each realization generated predictions for its corresponding 40% validation subset. Consequently, each location received approximately 500 predictions on average across the ensemble, depending on the sampling iterations in which it was included in the validation subset. These predictions were used both to evaluate model performance across *mtry* values and to quantify predictive uncertainty.

For each location, empirical predictive uncertainty was characterized from the distribution of available ensemble predictions. The 2.5th and 97.5th percentiles of ensemble SOM predictions were calculated on the log1p-transformed scale, and the resulting limits were back-transformed to the original SOM scale using the inverse transformation, $\exp(x) - 1$. These limits defined a 95% empirical predictive interval. Total predictive variance was additionally decomposed into sampling, response-variable perturbation, and model-specification components based on the nested ensemble structure. Sampling variance was estimated from predictions averaged across response-variable perturbations and *mtry* values and then quantified across sampling iterations. Response-variable variance was estimated from predictions averaged across *mtry* values and then quantified across response-variable perturbations. Model-specification variance was estimated from variation among *mtry* values and then averaged across sampling iterations and response-variable perturbations. The three components were summed to obtain total predictive variance.

The *mtry* value was selected based on pooled validation performance across the ensemble. For each combination of sampling iteration and response-variable perturbation,



validation predictions were generated separately for each of the five *mtry* values. Squared validation errors and the corresponding numbers of validation predictions were accumulated separately for each *mtry* value across all sampling iterations and response-variable perturbations. The *mtry*-specific root mean squared error (RMSE) was then calculated as:

$$SS_m = \sum_{s=1}^S \sum_{d=1}^D \sum_{i \in V_s} (\ln(1 + y_i) - \hat{y}_{i,s,d,m})^2$$

$$RMSE_m = \sqrt{\frac{SS_m}{N_m}}$$

where SS_m is the pooled sum of squared validation errors, y_i is the observed SOM value in the validation subset, $\hat{y}_{i,s,d,m}$ is the corresponding prediction from the model fitted using the perturbed calibration SOM, N_m is the total number of validation predictions contributed to the RMSE for *mtry* value m , and V_s is the validation subset for sampling iteration s . RMSE was calculated on the transformed scale. The *mtry* value with the lowest aggregated RMSE was selected for the final model.

A final RF model was subsequently fitted using the selected *mtry*. A new 60:40 stratified calibration-validation split was generated using the same sampling procedure and a fixed random seed (999). SOM values in both subsets were transformed using $\ln(1 + y_i)$, and no response-variable perturbation was applied to the final calibration data. The final model used the same tail-weighting configuration as the ensemble models and was fitted without internal cross-validation because *mtry* had already been selected through the ensemble evaluation. Final predictive performance was assessed using the independent validation subset and summarized using RMSE, mean absolute error (MAE), and the coefficient of determination (R^2), calculated on the original SOM scale after back-transformation of predictions and observations using the inverse log1p transformation (`expm1`). Bias was calculated as the mean difference between predicted and observed SOM values on the original scale.

2.4 Decomposition of predictive uncertainty

Predictive uncertainty was decomposed into sampling, response-variable perturbation, and model-specification components using predictions from the ensemble validation framework. These components represent variability induced by the corresponding dimensions of the ensemble and should not be interpreted as mutually independent physical error sources. For each validation observation, predictions were indexed across three nested dimensions: sampling iteration (s), response-variable perturbation (d), and model-specification configuration (m), represented by five *mtry* values.

Model-specification variance was first calculated within each sampling iteration and response-variable perturbation as the variance of predictions across the five *mtry* configurations:

$$V_{model,i,s,d} = \frac{1}{M-1} \sum_{m=1}^M (\hat{y}_{i,s,d,m} - \bar{\hat{y}}_{i,s,d})^2$$

where



$$\bar{\hat{y}}_{i,s,d} = \frac{1}{M} \sum_{m=1}^M \hat{y}_{i,s,d,m} .$$

The resulting variances were then averaged across the ($S = 25$) sampling iterations and ($D = 10$) response-variable perturbations for each validation observation:

$$V_{model,i} = \frac{1}{SD} \sum_{s=1}^S \sum_{d=1}^D V_{model,i,s,d}$$

Response-variable perturbation variance was estimated after averaging predictions across the five *mtry* configurations within each sampling iteration and perturbation. For each sampling iteration (s), the variance of these prediction means was calculated across the $D = 10$ response-variable perturbations:

$$V_{measurement,i,s} = \frac{1}{D-1} \sum_{d=1}^D (\bar{\hat{y}}_{i,s,d} - \bar{\hat{y}}_{i,s})^2$$

where

$$\bar{\hat{y}}_{i,s} = \frac{1}{D} \sum_{d=1}^D \bar{\hat{y}}_{i,s,d} .$$

The resulting variances were then averaged across the $S = 25$ sampling iterations:

$$V_{measurement,i} = \frac{1}{S} \sum_{s=1}^S V_{measurement,i,s} .$$

Sampling variance was estimated after averaging predictions across both response-variable perturbations and model-specification configurations within each sampling iteration:

$$\bar{\hat{y}}_{i,s} = \frac{1}{DM} \sum_{d=1}^D \sum_{m=1}^M \hat{y}_{i,s,d,m}$$

The sample variance of these prediction means was then calculated across the $S = 25$ sampling iterations:

$$V_{sampling,i} = \frac{1}{S-1} \sum_{s=1}^S (\bar{\hat{y}}_{i,s} - \bar{\hat{y}}_i)^2$$

where

$$\bar{\hat{y}}_i = \frac{1}{S} \sum_{s=1}^S \bar{\hat{y}}_{i,s} .$$

Total predictive variance was calculated as the sum of the three uncertainty components:

$$V_{total,i} = V_{sampling,i} + V_{measurement,i} + V_{model,i}$$

The total predictive standard deviation was obtained as the square root of the total variance.

$$SD_{total,i} = \sqrt{V_{total,i}}$$

The relative contribution of each uncertainty source was calculated as its proportion of total predictive variance:

$$f_{source,i} = \frac{V_{source,i}}{V_{total,i}}$$



where the *source* denotes sampling, response-variable perturbation, or model-specification uncertainty and *i indexes* the validation location. The dominant uncertainty source at each validation location was identified as the component with the largest relative contribution.

Variable importance scores were extracted from each of the 1,250 ensemble model realizations and summarized across the ensemble. For each predictor, the mean, standard deviation, and coefficient of variation (CV) of permutation importance were calculated across all realizations, encompassing the 25 sampling iterations, 10 response-variable perturbations, and five *mtry* configurations. These statistics characterized both the average contribution of each predictor and the stability of its importance across the ensemble.

2.5 Analysis software

All analyses were conducted in the R statistical computing environment (R Core Team, 2025). Data preprocessing and analysis workflows were implemented using the *data.table* package for data manipulation (Barrett et al., 2026), the *ggplot2* package for visualization (Wickham et al., 2025), and the *ranger* package for RF modeling (Wright et al., 2026). Random seeds were specified for all stochastic procedures to ensure reproducibility of the ensemble simulations.

3 Results and Discussion

3.1 Large-scale dataset for SOM modeling

The nationwide soil dataset comprised 283,618 quality-controlled observations distributed across the national agricultural domain (Table 1), corresponding to approximately 37 observations km⁻² of paddy-rice area. This large sample size provided extensive spatial coverage and substantially greater representation of local soil conditions than would be feasible from conventional soil monitoring networks alone. Such dense observations are particularly valuable for national-scale spatial modeling because they improve representation of spatial variability and major environmental gradients, although high sample density does not necessarily eliminate uncertainty associated with uneven environmental representation (Dalle Vaglie et al., 2026).

SOM had a mean of 23.7 g kg⁻¹, a median of 23.0 g kg⁻¹, and a CV of 38%, with values ranging from 1 to 300 g kg⁻¹ (Table 1). The distribution was strongly right-skewed: 62.3% of observations were below 25 g kg⁻¹, 99.9% were below 75 g kg⁻¹, and only a small fraction occurred in the upper tail. Although the dataset encompassed the recommended SOM range for paddy soils (20–30 g kg⁻¹; NAS (2019)), the sparse representation of high-SOM conditions provides substantially less empirical support for predictions at the upper end of the distribution. Such imbalance can limit model performance for extreme values and increase the sensitivity of tail predictions to sampling and model specification. The spatial distribution of observed SOM illustrates substantial heterogeneity across the country (Figure 1), consistent with differences in soil, land use, management, and environmental conditions.

The predictor dataset included seven non-response soil chemical variables and 32 auxiliary attributes representing soil thematic, topographic, climatic, and vegetation conditions (Table 1). These variables covered broad environmental gradients, including elevation (0–769 m), mean annual temperature (6.9–15.2°C), annual precipitation (871–2,637 mm), and NPP (152–255 g C m⁻² y⁻¹). The combination of extensive sampling and environmental coverage across these gradients therefore provided a strong basis for representing the heterogeneous



environmental gradients associated with SOM variation across the country. At the same time, this heterogeneity introduces complex and potentially interacting controls on SOM, making predictor relationships spatially variable and potentially contributing to uncertainty in variable importance and predictions.

Overall, the dataset's large sample size and extensive environmental coverage were major strengths for national-scale SOM modeling, but they did not imply uniformly strong information across the response distribution or environmental space. The highly skewed SOM distribution, sparse upper tail, and heterogeneous environmental gradients create complementary challenges: the model is well supported for common SOM conditions but has less empirical support for extreme SOM values and must represent complex relationships across diverse environments. These characteristics provide important context for interpreting the predictive performance and uncertainty decomposition presented below.

3.2 Total predictive uncertainty and its components

The modeling framework (Figure 2) enabled the quantification of total predictive uncertainty, its decomposition into sampling, response-variable perturbation, and model-specification components, and the assessment of their spatial and SOM-level patterns at the national scale. Figure 3 shows the median predicted SOM and empirical predictive interval width relative to predicted SOM, while Figure 4 shows the spatial distribution of the dominant uncertainty sources. Table 2 summarizes the median contributions of sampling, response-variable perturbation, and model-specification uncertainty across five SOM quantile-based groups.

Table 2. Median contributions of sampling, response-variable perturbation, and model-specification uncertainty across five empirical soil organic matter (SOM) quantile groups.

<i>n</i>	Median SOM (g kg ⁻¹)	95% interval width (g kg ⁻¹)	Relative interval width (%)	Sampling (%)	Response (%)	Model (%)
59,176	14.0 (1–17)	2.90	16.1	30.5	16.0	48.7
54,288	19.5 (>17–21.2)	2.77	13.2	27.2	21.5	45.6
63,205	23.1 (>21.2–25)	2.92	12.8	24.2	25.3	41.4
50,561	27.0 (>25–29)	3.32	13.4	24.3	26.7	39.2
56,388	34.0 (>29–300)	4.60	16.1	28.7	24.9	37.5

As shown by the national-scale pattern of observed SOM (Figure 1), the model reproduced a generally similar spatial pattern of SOM (Figure 3). The predicted SOM distribution had a mean of 23.2 g kg⁻¹ and a median of 22.8 g kg⁻¹, with a CV of 24.5% and an IQR of 5.8 g kg⁻¹. However, the model underpredicted the observed range, with predicted SOM ranging from 1.4 to 81.1 g kg⁻¹. The narrower predicted range suggests that the model captured the dominant spatial pattern of SOM but had limited ability to reproduce the most extreme observed values. Such range compression is common in predictive models when observations at the extremes are relatively sparse or when predictor–response relationships do not fully represent the processes generating local extremes (Sylvain et al., 2021).



The first three SOM quantiles, which encompassed 62.2% of the observations, spanned approximately 1–25 g kg⁻¹ and showed relatively widespread spatial distributions (Table 2). In contrast, SOM values above approximately 29 g kg⁻¹ were more spatially localized, occurring as distinct patches, and were less well predicted by the model. Although total predictive uncertainty increased toward the upper end of the SOM distribution, the contribution of model-specification uncertainty declined with increasing SOM. The weaker representation of these localized high-SOM areas likely reflects the greater difficulty of predicting conditions that are less prevalent in the calibration data and/or are associated with environmental combinations that are less well represented across the prediction domain. High SOM may also depend on management history, organic amendments, residue inputs, water management, soil depth, microtopography, drainage and redox conditions, or other local processes that are not fully represented by the available covariates. This interpretation is consistent with previous studies showing that SOM is influenced by interacting environmental and management factors, including climate, soil properties, vegetation inputs, and hydrological conditions (Viscarra Rossel et al., 2019; Li et al., 2023; Li et al., 2022). The pattern is also consistent with the greater predictive uncertainty observed toward the upper end of the SOM distribution (Table 2), suggesting that uncertainty was partly related to the distribution and representativeness of the observations used for model development (Sylvain et al., 2021).

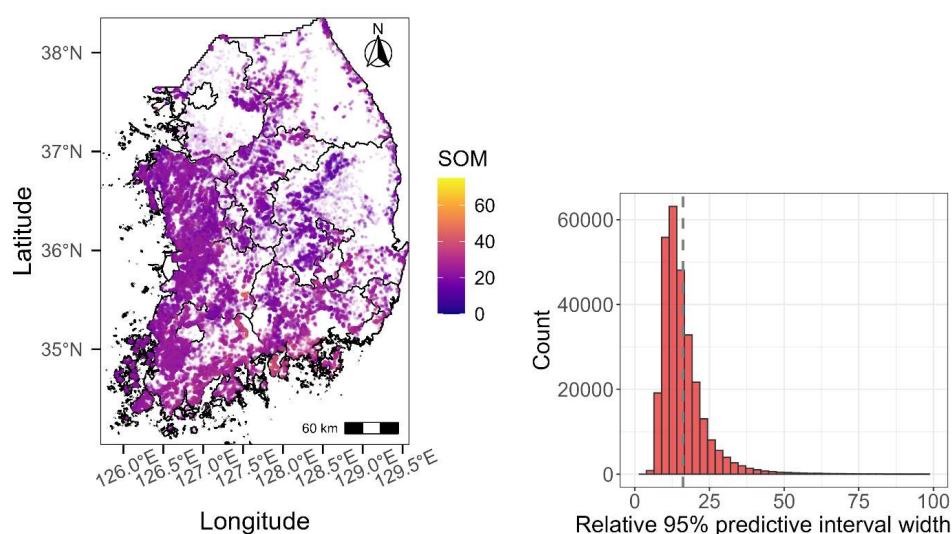


Figure 3. Median predicted soil organic matter (SOM; g kg⁻¹) and distribution of the empirical 95% predictive interval width relative to predicted SOM (%). The x-axis is truncated at 100% for visualization; values above 100% were retained in the analysis but are not shown. The dashed vertical line indicates the median relative predictive interval width.

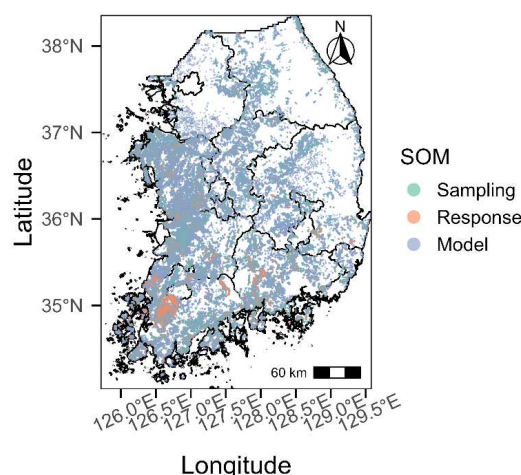


Figure 4. Spatial distribution of the dominant sources of predictive uncertainty in nationwide soil organic matter (SOM) prediction, classified as sampling, response-variable perturbation, or model-specification uncertainty.

The relative 95% predictive interval width showed considerable spatial variation in SOM prediction uncertainty, with a distinctly right-skewed distribution (Table 2; Figure 3). The strongly right-skewed distribution indicates that predictive uncertainty was generally modest across the prediction domain but increased substantially in a relatively small number of locations. This spatial concentration of uncertainty is important because a single summary measure of uncertainty would obscure substantial local differences in prediction reliability. Specifically, relatively narrow predictive intervals characterized most of the study area, with the highest density of predictions concentrated at interval widths between 10% and 15%. The median relative predictive interval width was approximately 16%. Conversely, a protracted right tail extending beyond 40% highlighted localized areas of high uncertainty. Thus, although predictive uncertainty was relatively low across most of the study area, substantially greater uncertainty occurred in localized areas, highlighting the spatially uneven reliability of the SOM predictions. Spatially heterogeneous uncertainty is expected in digital soil mapping because prediction reliability depends not only on model performance but also on the local availability and representativeness of observations and environmental covariates (Rossiter et al., 2022).

Total predictive uncertainty varied substantially across the national paddy-rice domain, with a median predictive SD of 3.18 g kg^{-1} (IQR: 1.86 g kg^{-1}). This spatial variation indicates that prediction reliability differed considerably across the target domain, despite the use of a common modeling framework. Model-specification uncertainty accounted for 41.9% of total predictive variance, followed by sampling uncertainty (29.3%) and response-variable perturbation uncertainty (28.8%). Model-specification uncertainty was the dominant source at 55.8% of locations ($n = 158,379$), whereas sampling and response-variable uncertainty



dominated at 23.6% and 20.5%, respectively. The relatively large contribution of model-specification uncertainty indicates that variation among the RF specifications examined in this study was an important source of predictive variability, exceeding the contributions from alternative calibration samples and response-variable perturbations. However, the model-specification component quantified here represents only a subset of possible sources of model uncertainty (Sylvain et al., 2021; Rohmer, 2025). The sampling component represents variability among alternative observations and should not be interpreted as a complete characterization of uncertainty arising from the nationwide sampling process. The spatial distribution of these dominant sources (Figure 4) showed that model-specification uncertainty was widespread across the prediction domain, whereas sampling and response-variable uncertainty were more spatially localized and interspersed. Thus, although model specification accounted for the largest share of predictive variance overall within the ensemble framework examined here, the relative contributions of uncertainty sources varied substantially among locations. This spatial heterogeneity indicates that different components of the modeling process contributed to prediction reliability in different parts of the study area.

3.3 Predictive performance

Although uncertainty was quantified using the ensemble, the final tail-weighted RF model showed similar predictive performance between calibration and independent validation subsets (Table 3). The coefficient of determination (R^2) was 0.47 for calibration and 0.50 for independent validation, indicating that the model explained approximately half of the variation in independent SOM observations. Prediction errors were also similar between the two subsets, with RMSE values of 6.61 and 6.32 g kg⁻¹ and MAE values of 4.11 and 4.06 g kg⁻¹ for calibration and validation, respectively. The small differences between calibration and validation metrics indicate similar predictive ability between the two subsets, providing limited evidence of overfitting under the random validation design. Overall, the model demonstrated moderate predictive performance, with an R^2 of 0.50 for the independent validation subset, which is close to the median R^2 reported for large-scale soil organic carbon mapping studies (>10,000 km²) in a global review (Chen et al., 2022). The model exhibited a small negative bias of -0.38 g kg⁻¹ in calibration and -0.36 g kg⁻¹ in validation, indicating a slight tendency toward underprediction that remained consistent between the two subsets. Thus, the model captured large-scale spatial variation in SOM but did not fully reproduce local or extreme variability. This moderate predictive performance likely reflects the substantial spatial heterogeneity of SOM, which arises from interactions among soil, climate, vegetation, land use, and landscape properties (Viscarra Rossel et al., 2019; Doetterl et al., 2015). Such interactions can produce substantial variation in SOM across landscapes and regions, making complete representation of its spatial variability challenging at broad spatial scales.

Observed–predicted comparisons were also similar between the calibration and validation subsets, although the upper tail of the SOM distribution was less well captured (Figure 5). This weaker performance at higher SOM levels was consistent with the uncertainty analysis, which showed greater predictive uncertainty toward the upper end of the SOM distribution (Table 2). Together, these results suggest that predictive reliability was not uniform across the observed SOM range, with greater caution warranted when interpreting predictions toward the extremes. Because the validation data were sampled from the same environmental domain as the calibration data, these results primarily evaluate the model's ability to predict independent



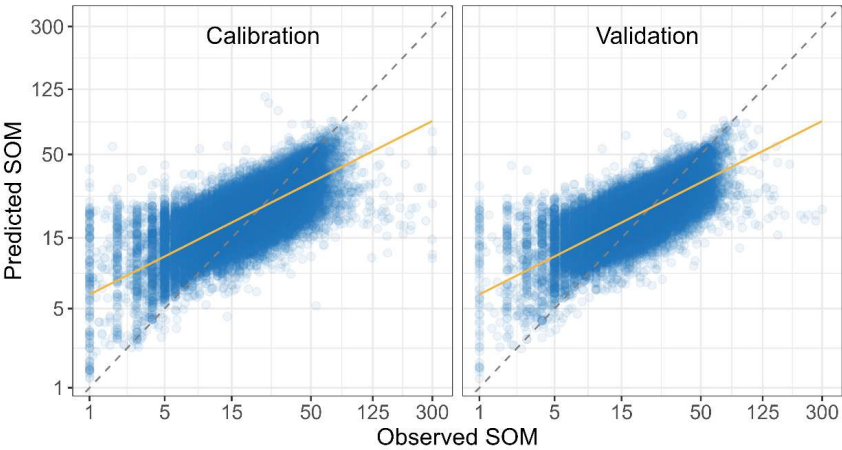
587 observations within the environmental space represented by the dataset, rather than its
588 performance under extrapolation to substantially different conditions (Rossiter et al., 2022). The
589 stratified validation design ensured that the validation subset retained observations across the
590 SOM distribution, providing a more balanced assessment of predictive performance than a
591 purely unstratified random split would have provided (Rendleman et al., 2022). Thus, the
592 validation results provide evidence of generalization within the represented data space, but do
593 not constitute an independent test of model transferability to environmental conditions outside
594 that space.

595
596
597 **Table 3.** Calibration and independent validation performance of the tail-weighted Random Forest
598 model for SOM prediction.

Subset	<i>n</i>	R^2	RMSE	MAE	Bias
Calibration	170,174	0.47	6.61	4.11	−0.38
Validation	113,444	0.50	6.32	4.06	−0.36

600 Performance metrics included the coefficient of determination (R^2), root mean square error
601 (RMSE), mean absolute error (MAE), and prediction bias, all calculated on the original SOM
602 scale after back-transformation of predictions and observations from the log1p scale using
603 $\exp(x) - 1$. Bias was calculated as the mean difference between predicted and observed SOM
604 values (predicted – observed) on the original scale.

605
606
607



608
609 **Figure 5.** Observed versus predicted soil organic matter (SOM) values for the calibration and
610 independent validation subsets using the final tail-weighted Random Forest model. Model
611 performance is reported in Table 3.

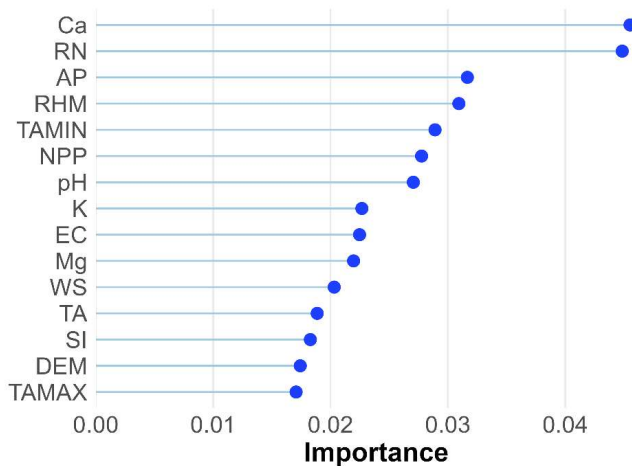
612
613



614 **3.4 Variable importance and SOM–environment relationships**

615
616 The variable importance analysis identified Ca (0.0455) and RN (0.0449) as the most influential
617 predictors of SOM, followed by AP (0.0317), RHM (0.0309), TAMIN (0.0289), NPP (0.0278),
618 and pH (0.0271) (Figure 6). Other relatively influential predictors included K (0.0227), EC
619 (0.0225), Mg (0.0220), WS (0.0203), TA (0.0189), SI (0.0183), DEM (0.0174), and TAMAX
620 (0.0171). Thus, the predictors contributing most to SOM prediction comprised a combination of
621 soil chemical, climatic, vegetation, and terrain variables, rather than a single predictor group. The
622 relatively high importance of Ca, AP, and pH indicates that soil chemical conditions were
623 important predictors of SOM and may reflect their associations with nutrient availability and
624 SOM dynamics (Rasmussen et al., 2018; Vermeiren et al., 2022). The importance of RN, RHM,
625 TAMIN, TAMAX, and NPP indicates that climatic and vegetation-related variables contributed
626 substantially to SOM prediction, potentially reflecting their associations with soil moisture,
627 productivity, and SOM turnover (Min et al., 2026; Lu et al., 2026). In agricultural soils, long-
628 term applications of amendments and fertilizers may also contribute to spatial variation in soil
629 chemical properties and, consequently, to their importance in SOM prediction (Hou et al., 2021).

630
631



632
633 **Figure 6.** Variable importance ranking of predictors in the final tail-weighted Random Forest
634 model. Predictor abbreviations are defined in Table 1.

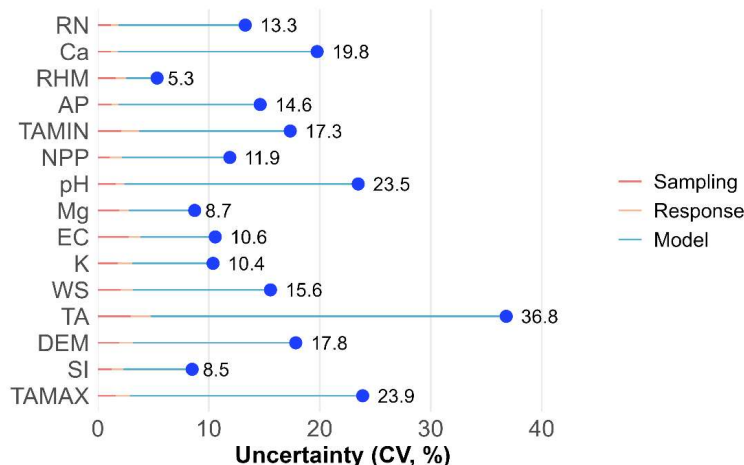
635
636
637 The importance rankings showed noticeable variation across model configurations
638 (Figure 7). Model specification contributed the largest component of variable-importance
639 variability, with contributions ranging from 2.8% to 32.0%, whereas sampling and response-
640 variable perturbation uncertainty contributed 1.1–3.0% and 0.6–1.8%, respectively. The
641 relatively large model-specification contribution indicates that the estimated importance of some
642 predictors was sensitive to model configuration. Ca and RN consistently ranked among the more
643 influential predictors across model configurations, although their exact relative rankings varied.



644 Thus, individual importance rankings should not be interpreted as a fixed ranking of factors
645 influencing SOM. Rather, they represent model-dependent estimates whose relative magnitudes
646 varied with model specification and other sources of uncertainty. The results therefore suggest
647 that SOM prediction was supported by multiple predictors, with their estimated contributions not
648 fully independent of the modeling framework.

649 In contrast, several terrain and categorical soil variables showed comparatively low
650 importance. AS had an importance of 0.0079, while TRI (0.0062), roughness (0.0061), TPI
651 (0.0039), TWI (0.0024), and upslope area (0.0023) contributed relatively little to the model.
652 Categorical soil variables, including soil color (0.0021), great group (0.0018), flow direction
653 (0.0017), aspect (0.0015), and suborder and soil type (both 0.0013), also showed low importance.
654 The lower importance of these variables may indicate that other continuous soil, climatic, and
655 vegetation predictors captured more of the spatial variation in SOM represented in the model,
656 although correlated predictors may also have reduced the apparent importance of individual
657 variables.

658
659



660
661 **Figure 7.** Relative variability in variable importance attributable to sampling, response-variable
662 perturbation, and model-specification uncertainty. Predictor abbreviations are defined in Table 1.
663

664
665 Overall, the variable-importance patterns indicate that nationwide paddy-soil SOM
666 prediction was supported by multiple environmental dimensions, including soil chemical
667 conditions, climate, vegetation, and topography. The relatively high importance of Ca and AP
668 points to the potential contribution of soil chemical interactions and nutrient retention, while the
669 contributions of RN, RHM, TAMIN, TAMAX, and NPP highlight the relevance of climatic and
670 vegetation-related conditions. The importance of Ca and AP may partly reflect their sensitivity to
671 agricultural management, including fertilizer application and organic amendments (Jindo et al.,
672 2023; Liu et al., 2024). Widespread applications of agricultural lime and Si fertilizers can affect
673 soil Ca and pH, while long-term phosphate fertilization can contribute to residual P accumulation



in agricultural soils (Makaza et al., 2026; Antonangelo et al., 2022; Khan et al., 2018). The importance of these variables may therefore reflect both their associations with SOM-related soil chemical conditions and the influence of historical management on those conditions. Climatic variables such as RN and RHM may similarly capture spatial and seasonal variation associated with the East Asian monsoon, including precipitation, soil moisture, and redox conditions in paddy soils (Kögel-Knabner et al., 2010).

These results are consistent with previous studies showing that SOM at large spatial scales is associated with multiple interacting environmental factors (Doetterl et al., 2025). Although the permutation importance of individual predictors was relatively small, their collective contributions indicate that SOM variability was associated with multiple environmental dimensions rather than a single dominant control. However, variable importance reflects the contribution of predictors to model predictions and does not establish causal relationships or specific mechanisms. The relatively high uncertainty associated with model specification further indicates that individual importance rankings should be interpreted cautiously. This is particularly relevant given the comparatively limited predictive performance in the lower and upper tails of the SOM distribution (Figure 5), suggesting that predictors that were important overall did not fully capture SOM variability at the extremes of the observed range.

3.5 Implications for large-scale SOM modeling

The results have several implications for large-scale SOM modeling. First, the substantial contribution of model specification to predictive and variable-importance uncertainty indicates that model uncertainty should not be overlooked. In this study, different model specifications produced meaningful variation in predictions and variable-importance rankings. Reliance on a single optimized model may therefore underestimate uncertainty, particularly when alternative model configurations provide similarly plausible representations of the data. Using multiple model specifications or model ensembles can provide a more robust assessment of predictive uncertainty and reduce dependence on a single modeling configuration.

Second, the relatively poor prediction of the upper SOM tail highlights the importance of retaining and adequately representing high-SOM observations. In this study, high-SOM observations were retained rather than excluded solely because they were difficult to predict, allowing the model to represent a wider range of SOM variability. During sampling-density optimization, however, excluding poorly predicted upper-tail observations could artificially improve apparent model performance and should therefore be avoided unless supported by independent evidence of measurement or other data-quality problems. The difficulty in predicting high SOM also suggests that additional observations and predictors that better represent the processes associated with high SOM may be needed. Complementary approaches that explicitly account for the upper tail, such as quantile regression, may also be useful (Lombardo et al., 2018).

Third, the contribution of sampling uncertainty indicates that sample size and representativeness remain important considerations even when very large soil databases are available. There is no universally accepted statistical standard for determining the sampling density required for national-scale SOM modeling (Saurette et al., 2025; Xia et al., 2022), and the appropriate density likely depends on the target spatial scale and environmental heterogeneity. Sampling requirements can instead be evaluated empirically by progressively



reducing the available sample size and assessing whether model performance remains stable under repeated sampling. Loiseau et al. (2021), for example, identified a performance threshold of approximately one soil profile per 2 km² in a region of approximately 5,000 km², whereas Guo et al. (2026) found that 434 samples were sufficient to maintain stable SOM prediction performance across approximately 6,174 km². These findings suggest that sampling density should be evaluated in relation to the intended modeling context, with particular attention to whether observations adequately represent the range of SOM and environmental conditions within the prediction domain.

Fourth, differences in the spatial support and resolution of predictor datasets represent an additional source of uncertainty that was not quantified in the present study. The predictors used here ranged from approximately 90 m for DEM-derived variables and 125 m for thematic soil variables to approximately 1 km for climate variables and 11 km for NPP. These differences mean that environmental conditions were represented at substantially different spatial scales, potentially introducing spatial representation uncertainty into SOM predictions (Chen et al., 2022; Piedallu et al., 2022; Han et al., 2022). Such differences may influence both variable importance and predicted spatial patterns, because fine-resolution soil and terrain variables can capture local heterogeneity that is not represented by coarser-resolution climate or vegetation variables. Future large-scale SOM modeling should therefore consider the spatial support of predictor datasets when interpreting model results and, where feasible, incorporate uncertainty associated with environmental representation.

Overall, these results indicate that improving large-scale SOM predictions requires attention not only to model fitting but also to the sources of uncertainty surrounding the data and predictors. Multiple model specifications can help characterize model uncertainty, adequate and representative sampling can reduce sampling uncertainty, and greater representation of the upper SOM tail can improve prediction of extreme conditions. Accounting for differences in predictor spatial support would further provide a more comprehensive assessment of uncertainty in large-scale SOM mapping.

4 Conclusions

Large national soil datasets enable large-scale prediction of SOM, but predictive reliability depends not only on model accuracy but also on how observations are sampled, responses are represented, and models are specified. We demonstrate that national-scale SOM predictions can remain substantially uncertain despite very large datasets, with uncertainty arising differently across sampling, response-variable perturbation, and RF model specification. Model specification was the largest contributor to predictive variance overall, although the dominant source varied spatially. Consequently, national SOM predictions should be interpreted together with predictive uncertainty and the variability in variable importance across model configurations. Ca and RN consistently ranked among the more influential predictors across model configurations, although their exact relative rankings varied. Thus, some predictors showed robust importance across model specifications, whereas their precise rankings remained model-dependent rather than representing fixed drivers of SOM. Predictive uncertainty was also spatially uneven and concentrated toward the upper end of the SOM distribution, indicating that high-SOM conditions remain less reliably predicted and warrant caution in interpretation. For paddy soils in South Korea, the relatively limited importance of topographic variables suggests that influential topographic controls on SOM may not have been fully captured by the predictors



766 considered here, whereas soil inputs and management-related factors appeared to be more
767 prominent.

768 These findings show that improving national-scale SOM prediction requires attention not
769 only to model fitting but also to the quality and representativeness of the underlying observations
770 and predictors. Exploring multiple model specifications can help characterize model uncertainty,
771 while adequate and representative sampling can reduce sampling uncertainty. Uncertainty
772 affected not only predicted SOM values, but also our inference about which predictors are
773 important. Increasing representation of high-SOM conditions may further improve prediction in
774 the upper tail of the SOM distribution. For soil management, spatially explicit uncertainty
775 information can help distinguish areas where predicted SOM is sufficiently reliable for broad-
776 scale assessment from those where field soil testing or additional monitoring may be appropriate.
777 This is particularly important for high-SOM areas, where greater predictive uncertainty may
778 limit confidence in management decisions based solely on spatial predictions. Future work
779 should further quantify uncertainty arising from differences in the spatial support and spatial
780 resolution of predictor datasets, which can be particularly important when integrating
781 heterogeneous environmental information at national scales. More generally, this framework can
782 be applied to other national and continental soil information systems to identify where
783 improvements in soil testing, monitoring, data integration, and modeling are most likely to
784 enhance the reliability of large-scale soil assessments.

785
786



787 **Data availability.** The quality-controlled dataset on soil, climate, and environmental variables is
788 openly available at Zenodo: <https://doi.org/10.5281/zenodo.23226302>.

789
790 **Code availability.** The R scripts used for data integration, quality control, and modeling are
791 openly available at Zenodo: <https://doi.org/10.5281/zenodo.23226302>.

792
793 **Author contributions.** JL formulated the research questions and designed the simulations with
794 WL. JL performed the simulation and the data analysis and wrote the manuscript with
795 contributions from SH and WL.

796
797 **Competing interests.** The authors declare that they have no conflict of interests.

798
799 **Acknowledgements.** The authors gratefully acknowledge the Rural Development Administration
800 (RDA) for making the soil test data available for this research.

801
802 **Financial support.** This study was funded by the National Research Foundation of Korea (NRF),
803 RS-2026-25536232.

804
805



References

- Amundson, R., Berhe, A. A., Hopmans, J. W., Olson, C., Sztein, A. E., and Sparks, D. L.: Soil and human security in the 21st century, *Science*, 348, 1261071, <https://doi.org/10.1126/science.1261071>, 2015.
- Antonangelo, J. A., Neto, J. F., Crusciol, C. A. C., Zhang, H., and Alleoni, L. R. F.: Lime and calcium-magnesium silicate cause chemical attributes stratification in no-till fields, *Soil and Tillage Research*, 224, 105522, <https://doi.org/10.1016/j.still.2022.105522>, 2022.
- Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B., and Krylov, I.: Extension of 'data.frame', R package version 1.18.2.1 [code], available at: <https://CRAN.R-project.org/package=data.table> (last access: 21 August 2026), 2026.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5-32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Buytaert, W.: topmodel: Implementation of the Hydrological Model TOPMODEL in R, R package version 0.7.5 [code], available at: <https://CRAN.R-project.org/package=topmodel> (last access: 21 August 2026), 2018.
- Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., and Walter, C.: Digital mapping of GlobalSoilMap soil properties at a broad scale: A review, *Geoderma*, 409, 115567, <https://doi.org/10.1016/j.geoderma.2021.115567>, 2022.
- Chen, X., Hu, Y., Xia, Y., Zheng, S., Ma, C., Rui, Y., He, H., Huang, D., Zhang, Z., Ge, T., Wu, J., Guggenberger, G., Kuzyakov, Y., and Su, Y.: Contrasting pathways of carbon sequestration in paddy and upland soils, *Global Change Biology*, 27, 2478-2490, <https://doi.org/10.1111/gcb.15595>, 2021.
- Chen, Z., Chen, L., Lu, R., Lou, Z., Zhou, F., Jin, Y., Xue, J., Guo, H., Wang, Z., Wang, Y., Liu, F., Song, X., Zhang, G., Su, Y., Ye, S., Shi, Z., and Chen, S.: A national soil organic carbon density dataset (2010–2024) in China, *Scientific Data*, 12, 1480, <https://doi.org/10.1038/s41597-025-05863-3>, 2025.
- Conrad, R.: Methane production in soil environments—anaerobic biogeochemistry and microbial life between flooding and desiccation, *Microorganisms*, 8, 881, <https://doi.org/10.3390/microorganisms8060881>, 2020.
- Cotrufo, M. F. and Lavelle, J. M.: Soil organic matter formation, persistence, and functioning: A synthesis of current understanding to inform its conservation and regeneration, in: *Advances in agronomy*, edited by: Sparks, D. L., Academic Press, 1-66, <https://doi.org/10.1016/bs.agron.2021.11.002>, 2022.
- Dalle Vaglie, M., Francini, S., Chirici, G., and Martellozzo, F.: A large-scale framework for estimating soil carbon, nitrogen, pH, and salinity dynamics for 1985–2023, *Proceedings of the National Academy of Sciences*, 123, e2534913123, <https://doi.org/10.1073/pnas.2534913123>, 2026.
- Doetterl, S., Berhe, A. A., Heckman, K., Lawrence, C., Schnecker, J., Vargas, R., Vogel, C., and Wagai, R.: A landscape-scale view of soil organic matter dynamics, *Nature Reviews Earth & Environment*, 6, 67-81, <https://doi.org/10.1038/s43017-024-00621-2>, 2025.
- Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., Casanova-Katny, A., Muñoz, C., Boudin, M., Zagal Venegas, E., and Boeckx, P.: Soil carbon storage controlled by interactions between geochemistry and climate, *Nature Geoscience*, 8, 780-783, <https://doi.org/10.1038/ngeo2516>, 2015.



- 852 Farina, R., Sándor, R., Abdalla, M., Álvaro-Fuentes, J., Bechini, L., Bolinder, M. A., Brilli, L.,
853 Chenu, C., Clivot, H., De Antoni Migliorati, M., Di Bene, C., Dorich, C. D., Ehrhardt, F.,
854 Ferchaud, F., Fitton, N., Francaviglia, R., Franko, U., Giltrap, D. L., Grant, B. B., Guenet, B.,
855 Harrison, M. T., Kirschbaum, M. U. F., Kuka, K., Kulmala, L., Liski, J., McGrath, M. J., Meier,
856 E., Menichetti, L., Moyano, F., Nendel, C., Recous, S., Reibold, N., Shepherd, A., Smith, W. N.,
857 Smith, P., Soussana, J.-F., Stella, T., Taghizadeh-Toosi, A., Tsutsikh, E., and Bellocchi, G.:
858 Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-
859 fallow soils, *Global Change Biology*, 27, 904-928, <https://doi.org/10.1111/gcb.15441>, 2021.
860 Geisseler, D., Linquist, B. A., and Lazicki, P. A.: Effect of fertilization on soil microorganisms in
861 paddy rice systems – A meta-analysis, *Soil Biology and Biochemistry*, 115, 452-460,
862 <https://doi.org/10.1016/j.soilbio.2017.09.018>, 2017.
863 Goidts, E., Van Wesemael, B., and Crucifix, M.: Magnitude and sources of uncertainties in soil
864 organic carbon (SOC) stock assessments at various scales, *European Journal of Soil Science*, 60,
865 723-739, <https://doi.org/10.1111/j.1365-2389.2009.01157.x>, 2009.
866 Guo, J., Zhang, Y., Kong, D., Zhao, X., Li, X., and Luo, C.: Optimization of sampling density in
867 regional-scale soil organic matter mapping and its impact on prediction accuracy, *Remote*
868 *Sensing*, 18, 1485, <https://doi.org/10.3390/rs18101485>, 2026.
869 Han, S. Y., Filippi, P., Singh, K., Whelan, B. M., and Bishop, T. F. A.: Assessment of global,
870 national and regional-level digital soil mapping products at different spatial supports, *European*
871 *Journal of Soil Science*, 73, e13300, <https://doi.org/10.1111/ejss.13300>, 2022.
872 Hartmann, M. and Six, J.: Soil structure and microbiome functions in agroecosystems, *Nature*
873 *Reviews Earth & Environment*, 4, 4-18, <https://doi.org/10.1038/s43017-022-00366-w>, 2023.
874 Hijmans, R. J.: terra: Spatial Data Analysis. , R package version 1.8-80 [code], available at:
875 <https://CRAN.R-project.org/package=terra> (last access: 21 August 2026), 2025.
876 Hou, L., Liu, Z., Zhao, J., Ma, P., and Xu, X.: Comprehensive assessment of fertilization, spatial
877 variability of soil chemical properties, and relationships among nutrients, apple yield and orchard
878 age: A case study in Luochuan County, China, *Ecological Indicators*, 122, 107285,
879 <https://doi.org/10.1016/j.ecolind.2020.107285>, 2021.
880 Jindo, K., Audette, Y., Olivares, F. L., Canellas, L. P., Smith, D. S., and Paul Voroney, R.: Biotic
881 and abiotic effects of soil organic matter on the phytoavailable phosphorus in soils: a review,
882 *Chemical and Biological Technologies in Agriculture*, 10, 29, [https://doi.org/10.1186/s40538-](https://doi.org/10.1186/s40538-023-00401-y)
883 [023-00401-y](https://doi.org/10.1186/s40538-023-00401-y), 2023.
884 Khan, A., Lu, G., Ayaz, M., Zhang, H., Wang, R., Lv, F., Yang, X., Sun, B., and Zhang, S.:
885 Phosphorus efficiency, soil phosphorus dynamics and critical phosphorus level under long-term
886 fertilization for single and double cropping systems, *Agriculture, Ecosystems & Environment*,
887 256, 1-11, <https://doi.org/10.1016/j.agee.2018.01.006>, 2018.
888 KMA: Climate Change Scenarios [dataset], 2023, accessed 1 October 2026,
889 http://www.climate.go.kr/home/CCS/contents_2021/35_download.php.
890 Kögel-Knabner, I., Amelung, W., Cao, Z., Fiedler, S., Frenzel, P., Jahn, R., Kalbitz, K., Kölbl, A.,
891 and Schloter, M.: Biogeochemistry of paddy soils, *Geoderma*, 157, 1-14,
892 <https://doi.org/10.1016/j.geoderma.2010.03.009>, 2010.
893 KOSIS: Cultivated Area of Food Crops, Korean Statistical Information Service [dataset], 2026,
894 accessed 25 August 2026, <http://kosis.kr/eng/>.
895 Lamichhane, S., Kumar, L., and Wilson, B.: Digital soil mapping algorithms and covariates for
896 soil organic carbon mapping and their implications: A review, *Geoderma*, 352, 395-413,
897 <https://doi.org/10.1016/j.geoderma.2019.05.031>, 2019.



- 898 Lee, J.: Data errors in soil tests compromise large-scale assessment of soil organic matter in
899 Korea, *Applied Biological Chemistry*, 68, 83, <https://doi.org/10.1186/s13765-025-01055-0>, 2025.
- 900 Li, Q., Wang, L., Fu, Y., Lin, D., Hou, M., Li, X., Hu, D., and Wang, Z.: Transformation of soil
901 organic matter subjected to environmental disturbance and preservation of organic matter bound
902 to soil minerals: a review, *Journal of Soils and Sediments*, 23, 1485-1500,
903 <https://doi.org/10.1007/s11368-022-03381-y>, 2023.
- 904 Li, Z., Li, X., Zhou, S., Yang, X., Fu, Y., Miao, C., Wang, S., Zhang, G., Wu, X., Yang, C., and
905 Deng, Y.: A comprehensive review on coupled processes and mechanisms of soil-vegetation-
906 hydrology, and recent research advances, *Science China Earth Sciences*, 65, 2083-2114,
907 <https://doi.org/10.1007/s11430-021-9990-5>, 2022.
- 908 Liu, X., Zhang, Y., Wang, Z., and Chen, Z.: The contribution of organic and chemical fertilizers
909 on the pools and availability of phosphorus in agricultural soils based on a meta-analysis,
910 *European Journal of Agronomy*, 156, 127144, <https://doi.org/10.1016/j.eja.2024.127144>, 2024.
- 911 Loiseau, T., Arrouays, D., Richer-de-Forges, A. C., Lagacherie, P., Ducommun, C., and Minasny,
912 B.: Density of soil observations in digital soil mapping: A study in the Mayenne region, France,
913 *Geoderma Regional*, 24, e00358, <https://doi.org/10.1016/j.geodrs.2021.e00358>, 2021.
- 914 Lombardo, L., Saia, S., Schillaci, C., Mai, P. M., and Huser, R.: Modeling soil organic carbon
915 with Quantile Regression: Dissecting predictors' effects on carbon stocks, *Geoderma*, 318, 148-
916 159, <https://doi.org/10.1016/j.geoderma.2017.12.011>, 2018.
- 917 Lu, Y., Yang, L., and Zhou, C.: The Dual Role of NPP in Mediating and Moderating Climate-Soil
918 Carbon Pathways Under Warming and Drought Across European Ecosystems, *Global Change*
919 *Biology*, 32, e70766, <https://doi.org/10.1111/gcb.70766>, 2026.
- 920 Lundy, M. E., Pittelkow, C. M., Linquist, B. A., Liang, X. Q., van Groenigen, K. J., Lee, J., Six,
921 J., Venterea, R. T., and van Kessel, C.: Nitrogen fertilization reduces yield declines following no-
922 till adoption, *Field Crops Research*, 183, 204-210, <https://doi.org/10.1016/j.fcr.2015.07.023>,
923 2015.
- 924 Makaza, W., Khiari, L., and El Achaby, M.: The meta-analysis study on the effects of the quality
925 of lime materials on the soil physicochemical properties and crop yields in acid soils, *Frontiers in*
926 *Soil Science*, 6, <https://doi.org/10.3389/fsoil.2026.1725559>, 2026.
- 927 Min, K., Yang, Y., Wahab, L., Woo, S., Oh, M., Ghezzehei, T. A., and Asefaw Berhe, A.: Soil
928 organic matter dynamics under changing precipitation regimes, *New Phytologist*, 249, 2179-
929 2195, <https://doi.org/10.1111/nph.70804>, 2026.
- 930 MOLIT: Nationwide Continuous Cadastral Map, Ministry of Land, Infrastructure and Transport
931 [dataset], 2026a, accessed 19 March 2026,
932 https://www.vworld.kr/dtmk/dtmk_ntads_s002.do?dsId=30563.
- 933 MOLIT: Administrative divisions of South Korea at the submunicipal level: RI (village), version
934 2026-04-15, Ministry of Land, Infrastructure and Transport [dataset], 2026b, accessed 30 April
935 2026, https://www.vworld.kr/dtmk/dtmk_ntads_s002.do?dsId=30602.
- 936 MOLIT: Administrative divisions of South Korea at the submunicipal level: Eup (towns), Myeon
937 (townships), and Dong (neighborhoods), version 2026-04-15, Ministry of Land, Infrastructure
938 and Transport [dataset], 2026c, accessed 30 April 2026,
939 www.vworld.kr/dtmk/dtmk_ntads_s002.do?dsId=30603.
- 940 NAS: Methods of Soil Chemical Analysis, National Institute of Agricultural Sciences, Rural
941 Development Administration Suwon, Korea, 319 pp., ISBN 978-89-480-0913-2, 2010.
- 942 NAS: Fertilizer Recommendation Guide for Crops, National Institute of Agricultural Sciences,
943 Rural Development Administration, Jeonju, Korea, 420 pp., ISBN 978-89-480-6048-5, 2019.



- 944 NAS: Chemical Data for Soil Test, National Institute of Agricultural Sciences, Rural
 945 Development Administration [dataset], 2025, accessed 25 August 2026,
 946 www.data.go.kr/data/15073569/openapi.do.
 947 NEO: Net Primary Productivity (1 year - TERRA/MODIS), NASA Earth Observations [dataset],
 948 2023, accessed 20 February 2023, <https://neo.gsfc.nasa.gov>.
 949 NRCS: Soil Survey Geographic (SSURGO) Database, U.S. Department of Agriculture, Natural
 950 Resources Conservation Service [dataset], 2004, accessed 19 March 2026,
 951 <https://sdmdataaccess.nrcs.usda.gov/>.
 952 Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., and Fernández-Ugalde, O.: LUCAS Soil, the
 953 largest expandable soil dataset for Europe: a review, *European Journal of Soil Science*, 69, 140-
 954 153, <https://doi.org/10.1111/ejss.12499>, 2018.
 955 Piedallu, C., Pedersoli, E., Chaste, E., Morneau, F., Seynave, I., and Gégout, J.-C.: Optimal
 956 resolution of soil properties maps varies according to their geographical extent and location,
 957 *Geoderma*, 412, 115723, <https://doi.org/10.1016/j.geoderma.2022.115723>, 2022.
 958 Piikki, K., Wetterlind, J., Söderström, M., and Stenberg, B.: Perspectives on validation in digital
 959 soil mapping of continuous attributes—A review, *Soil Use and Management*, 37, 7-21,
 960 <https://doi.org/10.1111/sum.12694>, 2021.
 961 Pittelkow, C. M., Liang, X. Q., Linquist, B. A., van Groenigen, K. J., Lee, J., Lundy, M. E., van
 962 Gestel, N., Six, J., Venterea, R. T., and van Kessel, C.: Productivity limits and potentials of the
 963 principles of conservation agriculture, *Nature*, 517, 365-U482,
 964 <https://doi.org/10.1038/nature13809>, 2015a.
 965 Pittelkow, C. M., Linquist, B. A., Lundy, M. E., Liang, X. Q., van Groenigen, K. J., Lee, J., van
 966 Gestel, N., Six, J., Venterea, R. T., and van Kessel, C.: When does no-till yield more? A global
 967 meta-analysis, *Field Crops Research*, 183, 156-168, <https://doi.org/10.1016/j.fcr.2015.07.020>,
 968 2015b.
 969 R Core Team: R: A Language and Environment for Statistical Computing, 4.4.3 [code], available
 970 at: <https://www.R-project.org> (last access: 2026-09-29), 2025.
 971 Rasmussen, C., Heckman, K., Wieder, W. R., Keiluweit, M., Lawrence, C. R., Berhe, A. A.,
 972 Blankinship, J. C., Crow, S. E., Druhan, J. L., Hicks Pries, C. E., Marin-Spiotta, E., Plante, A. F.,
 973 Schädell, C., Schimel, J. P., Sierra, C. A., Thompson, A., and Wagai, R.: Beyond clay: towards an
 974 improved set of variables for predicting soil organic matter content, *Biogeochemistry*, 137, 297-
 975 306, [10.1007/s10533-018-0424-3](https://doi.org/10.1007/s10533-018-0424-3), 2018.
 976 Reeves, D. W.: The role of soil organic matter in maintaining soil quality in continuous cropping
 977 systems, *Soil and Tillage Research*, 43, 131-167, [https://doi.org/10.1016/S0167-1987\(97\)00038-](https://doi.org/10.1016/S0167-1987(97)00038-X)
 978 [X](https://doi.org/10.1016/S0167-1987(97)00038-X), 1997.
 979 Rendleman, M. C., Smith, B. J., Canahuat, G., Braun, T. A., Buatti, J. M., and Casavant, T. L.:
 980 Representative random sampling: an empirical evaluation of a novel bin stratification method for
 981 model performance estimation, *Statistics and Computing*, 32, 101,
 982 <https://doi.org/10.1007/s11222-022-10138-7>, 2022.
 983 Rohmer, J.: Importance ranking of data and model uncertainties in quantile regression forest-
 984 based spatial predictions when data are sparse, imprecise and clustered, *Ecological Informatics*,
 985 92, 103459, <https://doi.org/10.1016/j.ecoinf.2025.103459>, 2025.
 986 Rossiter, D. G., Poggio, L., Beaudette, D., and Libohova, Z.: How well does digital soil mapping
 987 represent soil geography? An investigation from the USA, *Soil*, 8, 559-586,
 988 <https://doi.org/10.5194/soil-8-559-2022>, 2022.



- 989 Samsonova, V. P., Meshalkina, J. L., Dobrovolskaya, V. A., Kondrashkina, M. I., Dyadkina, S.
990 E., Filippova, O. I., Krotov, D. G., Morozova, T. M., and Krasilnikov, P. V.: Investigation of
991 uncertainty in organic carbon stock estimates on a field scale, *Eurasian Soil Science*, 56, 1765-
992 1775, <https://doi.org/10.1134/S106422932360183X>, 2023.
- 993 Saurette, D. D., Heck, R. J., Gillespie, A. W., Berg, A. A., and Biswas, A.: Post-hoc evaluation of
994 sample size in a regional digital soil mapping project, *Land*, 14, 545,
995 <https://doi.org/10.3390/land14030545>, 2025.
- 996 Six, J., Conant, R. T., Paul, E. A., and Paustian, K.: Stabilization mechanisms of soil organic
997 matter: Implications for C-saturation of soils, *Plant and Soil*, 241, 155-176,
998 <https://doi.org/10.1023/A:1016125726789>, 2002.
- 999 Sylvain, J.-D., Anctil, F., and Thiffault, É.: Using bias correction and ensemble modelling for
1000 predictive mapping and related uncertainty: A case study in digital soil mapping, *Geoderma*, 403,
1001 115153, <https://doi.org/10.1016/j.geoderma.2021.115153>, 2021.
- 1002 Vermeiren, C., Kerckhof, P., Reheul, D., and Smolders, E.: Increasing soil organic carbon content
1003 can enhance the long-term availability of phosphorus in agricultural soils, *European Journal of*
1004 *Soil Science*, 73, e13191, <https://doi.org/10.1111/ejss.13191>, 2022.
- 1005 Viscarra Rossel, R. A., Lee, J., Behrens, T., Luo, Z., Baldock, J., and Richards, A.: Continental-
1006 scale soil carbon composition and vulnerability modulated by regional environmental controls,
1007 *Nature Geoscience*, 12, 547-552, <https://doi.org/10.1038/s41561-019-0373-z>, 2019.
- 1008 Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani,
1009 H., Dunnington, D., van den Brand, T., and Posit, P.: Create Elegant Data Visualisations Using
1010 the Grammar of Graphics, R package version 4.0.1 [code], available at:
1011 <https://ggplot2.tidyverse.org> (last access: 21 August 2026), 2025.
- 1012 Wright, M., Wager, S., and Probst, P.: A Fast Implementation of Random Forests, R package
1013 version 0.18.0 [code], available at: <https://github.com/imbs-hl/ranger> (last access: 21 August
1014 2026), 2026.
- 1015 Wright, M. N. and Ziegler, A.: ranger: A fast implementation of Random Forests for high
1016 dimensional data in C++ and R, *Journal of Statistical Software*, 77, 1-17,
1017 <https://doi.org/10.18637/jss.v077.i01>, 2017.
- 1018 Xia, Y., McSweeney, K., and Wander, M. M.: Digital mapping of agricultural soil organic carbon
1019 using soil forming factors: A review of current efforts at the regional and national Scales,
1020 *Frontiers in Soil Science*, 2, <https://doi.org/10.3389/fsoil.2022.890437>, 2022.
- 1021 Young, M. D., Ros, G. H., and de Vries, W.: Impacts of agronomic measures on crop, soil, and
1022 environmental indicators: A review and synthesis of meta-analysis, *Agriculture, Ecosystems &*
1023 *Environment*, 319, 107551, <https://doi.org/10.1016/j.agee.2021.107551>, 2021.

1024