Research Square

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

Mapping soil organic carbon with Soil Health Card observations and multisensor satellite data in Odisha, India

Pavan Kumar Bellam
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Murali Krishna Gumma
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Venkata Reddy Keesara
National Institute of Technology Warangal
Mukund Dhavaji Patil
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)
Narayanarao Bhogapurapu
bnarayanarao@nitw.ac.in
National Institute of Technology Warangal

Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11377365/v1

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

Additional Declarations: No competing interests reported.

Mapping soil organic carbon with Soil Health Card observations and multisensor satellite data in Odisha, India

Pavan Kumar Bellam[1,2], Murali Krishna Gumma[1], Venkata Reddy Keesara[2], Mukund Dhavaji Patil[1], Narayanarao Bhogapurapu[2*]

[1]Digital Agriculture, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India.
[2]Department of Civil Engineering, National Institute of Technology Warangal, Warangal, India.

*Corresponding author(s). E-mail(s): bnarayanarao@nitw.ac.in;

Abstract

Soil organic carbon contributes to soil fertility and terrestrial carbon storage, making its spatial distribution important for soil assessment. We used 13,284 organic carbon measurements from India's Soil Health Card programme and 17 optical, radar, terrain and climate covariates to map organic carbon in the upper 15 cm across Odisha, India, using a random forest on a nominal 30 m grid. A reserved 20% subset gave $R^2 = 0.501$ and a root mean square error of 0.137 percentage points. In separate five-fold assessments, R^2 declined from 0.468 with random folds to 0.346 when observations within 30 m were grouped, and to -0.103 with geographic-cluster folds. Coverage of nominal 90% prediction intervals declined from 90.14% under random validation to 81.96% under geographic-cluster validation. The area of applicability, defined using covariate dissimilarity, included 97.6% of mapped cropland and 68.1% of the full extent. Cropland within this domain covered 5.32 million ha and yielded upper-15-cm stock estimates of 58.83 and 77.66 Tg C under two Walkley–Black recovery assumptions. The results distinguish covariate representation from geographic predictive performance when interpreting soil organic carbon maps and stock estimates.

Keywords: Soil organic carbon, Soil Health Card, Digital soil mapping, Spatial validation, Remote sensing, Odisha

Introduction

Soil organic carbon (SOC) contributes to soil fertility and carbon storage, making its spatial distribution important for agricultural management and regional carbon assessment [1]. Digital soil mapping extends point measurements across a landscape by relating soil observations to environmental covariates [2, 3]. Satellite observations complement terrain and climate data by describing surface conditions at finer spatial scales [4]. Hengl et al. combined soil observations with Earth observation data to produce 30 m soil-property maps for Africa and evaluated them using spatial cross-validation [5]. These applications illustrate the value of predicting soil properties from existing measurements and environmental covariates, while sampling coverage and validation design remain important when interpreting the predictions [3, 6].

Existing measurements from agricultural soil-testing programmes can be reused for regional SOC mapping. The Soil Health Card (SHC) programme in India provides one such source. It was developed to inform farmers about the soil nutrient status of their holdings and provide recommendations on fertilizers and soil amendments. Organic carbon (OC) is among its measured properties [7]. These OC measurements can be used to fit predictive models with satellite-derived, terrain and climate covariates and estimate OC at unsampled locations. A national assessment using SHC observations has examined OC and pH across Indian agricultural soils, illustrating the wider value of the archive [8]. When predictions are extended to other land covers, their covariate combinations need to be compared with those represented by the agricultural observations.

In Odisha, cultivated areas span coastal plains, central tablelands and the Eastern Ghats. The state includes coastal alluvial soils, red and lateritic uplands and smaller areas of black soils [9, 10]. Previous work in eastern Odisha evaluated multi-frequency synthetic aperture radar (SAR) backscatter for SOC prediction [11, 12]. At the statewide scale, the wider range of land uses and environmental conditions makes both covariate coverage and geographic evaluation important for interpreting maps based on agricultural observations.

Geographic evaluation needs to account for the distribution of those observations. In densely sampled areas, nearby observations can fall within the same 30 m predictor cell; even samples in neighbouring cells may receive similar information from terrain and climate layers with coarser native resolutions. Random partitioning can consequently place nearby sites with similar covariates in both the fitting and validation data, so the resulting performance statistics may describe prediction near sampled fields better than prediction in a geographically separate area [13, 14]. Keeping observations within 30 m in the same fold examines the influence of immediate proximity, whereas assigning whole geographic groups to validation folds tests prediction in held-out geographic groups [15, 16]. Jafari et al. examined spatial predictors and validation designs in SOC modelling, illustrating why spatial dependence needs to be considered when interpreting model performance [17].

Geographic validation can be complemented by assessing similarity in predictor space between prediction locations and the model-development observations. Geographically distant locations may share similar covariate combinations, while nearby locations may differ. The dissimilarity index (DI) compares the combination of predictors at

each mapped cell with those at the model-development observations, and a threshold delineates the area of applicability (AOA) [18]. This describes representation in predictor space rather than the accuracy of an individual prediction. Prediction intervals provide a further assessment through their empirical coverage at validation observations [19]. Together, these assessments distinguish representation in predictor space from predictive performance and interval coverage measured at validation locations.

In this study, we used SHC OC measurements and multisensor satellite, terrain and climate covariates to develop a random-forest model and map OC in the upper 15 cm across Odisha on a nominal 30 m grid. The mapping model was evaluated on a reserved subset, and separate five-fold assessments compared random allocation, proximity grouping and geographic-cluster allocation. We assessed covariate dissimilarity across the mapped area and prediction-interval coverage at observation locations. We then estimated regional carbon stocks from the predicted OC surface, reporting cropland inside the AOA separately from the full mapped extent.

Methods

Study area

Odisha is located on the eastern coast of India and covers approximately 156,000 km^2 between about 17.8° and 22.6° N and 81.4° and 87.5° E. The state extends from the coastal and deltaic plains associated with the Mahanadi, Brahmani and Baitarani river systems to central tablelands, western rolling uplands and the Eastern Ghats. The climate is tropical monsoonal, with most rainfall received from June to September. Agricultural land is dominated by rice-based systems together with pulses, oil seeds, and other seasonal crops. Regional soil descriptions were based on the ICAR–NBSS&LUP survey of Odisha [9]. The combination of marked topographic relief, soil differences, rainfall gradients, and contrasting land use produces a broad range of environments relevant to SOC accumulation and redistribution.

Soil observations

Soil observations were obtained from India's SHC programme for 2020 and represented the surface 0–15 cm layer. OC was determined using the Walkley–Black wet-oxidation method [20]. Coordinate records were screened for completeness and alignment with the covariate grid, and observations lacking complete covariate data were excluded. After the additional covariate screening described below, 13,284 observations remained for modelling (Fig. 1). Measured OC ranged from 0.01 to 2.49% by mass, with a mean of 0.540% and a standard deviation of 0.189%.

The observations were collected primarily from agricultural land and were not spatially balanced across Odisha. Their measured OC distribution was therefore not interpreted as an area-weighted summary of the state. We retained observations at their reported coordinates without spatial averaging and used the 30 m grouped validation to examine the influence of closely located samples.

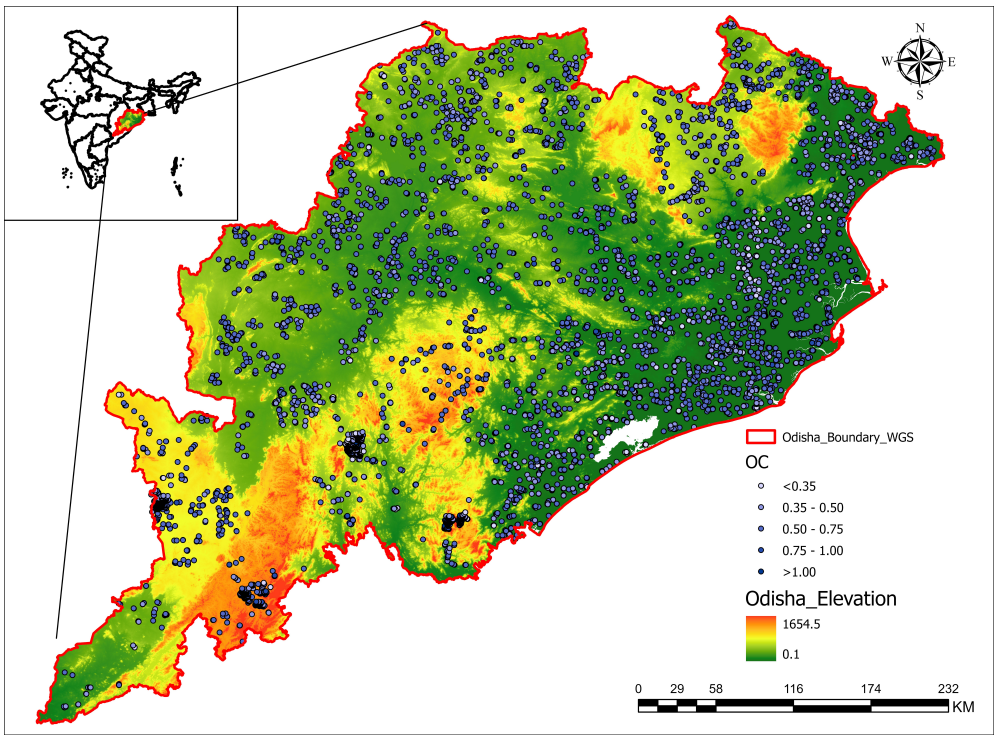


Fig. 1 Study area, elevation and spatial distribution of the soil observations in Odisha, India. Point colours indicate measured OC classes; the inset locates Odisha within India.

Environmental covariates and preprocessing

Environmental covariates were selected to represent terrain, climate, optical reflectance and radar backscatter. The final predictor set contained 17 variables (Supplementary Table S1): blue, near-infrared (NIR) and short-wave infrared (SWIR1) surface reflectance and the normalised difference tillage index (NDTI) from Harmonized Landsat and Sentinel-2 (HLS) data [21]; Sentinel-1 VV and VH backscatter; ALOS-2/PALSAR-2 HH and HV backscatter; elevation, slope, aspect, topographic wetness index (TWI), topographic position index (TPI), local elevation variability and local slope variability derived from MERIT DEM [22]; and mean annual precipitation (BIO12) and precipitation seasonality (BIO15) from WorldClim v2.1 [23].

The optical predictors were derived from cloud-screened HLSL30 and HLSS30 observations acquired during May–June in 2018, 2019 and 2020. This three-year seasonal window provided a larger pool of seasonally comparable observations and reduced dependence on individual acquisitions. Per-pixel medians were calculated across the pooled valid observations for six common spectral bands (blue, green, red, NIR, SWIR1 and SWIR2) and aligned to the 30 m grid. The enhanced vegetation index (EVI) and NDTI were initially derived as candidate optical indices. NDTI was calculated as $(\mathrm{SWIR1}-\mathrm{SWIR2})/(\mathrm{SWIR1}+\mathrm{SWIR2})$; no broader set of spectral indices was evaluated.

Radar predictors comprised incidence-angle-normalised Sentinel-1 C-band VV and VH [24] and ALOS-2/PALSAR-2 L-band HH and HV backscatter [25] from 2020. Temporal medians were calculated for both radar datasets, and the resulting composites were retained in linear units and aligned to the 30 m grid.

Before covariate extraction, the optical and radar predictors were decomposed with a single-level two-dimensional Haar wavelet transform. The threshold was estimated from the median absolute deviation of the detail coefficients, which were soft-thresholded before inverse reconstruction [26]. Elevation was Gaussian-smoothed with a standard deviation of three pixels before extraction. The contribution of wavelet denoising to predictive performance was not evaluated separately because an otherwise identical model using unfiltered predictors was not fitted.

After records with missing covariates were removed, observations outside the 0.1st–99.9th percentile range of any optical or radar candidate predictor were excluded as potential sensor extremes. This reduced the complete dataset from 13,562 to 13,284 observations. Iterative variance-inflation-factor screening, using a cutoff of 10, subsequently removed EVI and the green, red and SWIR2 bands as separate predictors. The retained optical variables were blue, NIR and SWIR1 reflectance and NDTI. NDTI was retained after redundancy screening. Both covariate-tail and variance-inflation screening were response-blind and were completed before the validation partitions were generated.

All covariates were aligned to the common prediction grid in geographic coordinates (EPSG:4326), with nominal 30 m spacing. Continuous terrain and climate layers were resampled bilinearly where required. Optical and radar mosaics entered the workflow on 30 m grids, whereas the effective source resolutions of MERIT DEM and WorldClim were approximately 90 m and 1 km, respectively. Resampling these layers did not increase their spatial information content. Covariate values were extracted from the corresponding grid cell at each soil-observation coordinate. Metric distances used for grouped validation were calculated separately in UTM zone 45N (EPSG:32645).

Random-forest model

Measured OC was related to the 17 covariates using a random forest (RF). The model used 400 trees, a maximum depth of 16 and a minimum of four observations per leaf; $\sqrt{p}$ predictors were considered at each split, where p is the number of predictors [27]. Predictors were centred by their median and scaled by their interquartile range. The response was transformed as $\log(1+OC)$ for fitting and back-transformed for prediction. Hyperparameters were fixed before validation. The statewide mapping model was fitted to the 80% model-development subset, excluding the stratified 20% validation subset from RF fitting. The three five-fold analyses described below were separate resampling assessments using all observations.

To examine which covariates the fitted model relied on, permutation importance was calculated on the 20% hold-out subset using ten permutations per predictor. Importance was reported as the decrease in hold-out R^2 after permutation. It was used to aid interpretation of the mapped patterns and was treated as model dependence rather than as a causal ranking of SOC controls.

Model validation and spatial transfer

Model validation was conducted in two stages: a stratified random hold-out assessment of the mapping model and separate five-fold assessments under random and spatially grouped partitions. First, the observations were divided into an 80% model-development subset and a 20% hold-out validation subset ($n = 2{,}657$). The split was stratified by deciles of observed OC so that both subsets retained a similar response distribution. The validation subset was excluded from model fitting and used once to calculate the conventional hold-out validation statistics.

The second stage used all 13,284 observations in three five-fold validation designs. In each design, the model was fitted five times using four folds for fitting and the remaining fold for validation, so that every observation was evaluated once. The random design allocated individual observations among folds without considering their locations and therefore assessed prediction among spatially interspersed samples.

For the 30 m grouped design, coordinates were projected to UTM zone 45N (EPSG:32645), and observations separated by 30 m or less were assigned to the same fold. This kept nearly coincident observations together, including samples that could share the same or adjacent predictor cells and receive identical or similar covariate values.

For the geographic-cluster design, the observation locations were divided into 20 geographically compact clusters using k-means clustering, and the clusters were allocated among five folds. In each fold, all observations from the selected clusters were excluded from model fitting and used only for validation. This design tested whether a model fitted in other parts of Odisha could reproduce OC variation in geographically separate locations.

Validation statistics were calculated from the pooled predictions so that each observation contributed once to each five-fold result. All three designs used the same model specification and random seed. Geographic-cluster validation was interpreted as a test of spatial transfer within the SHC archive, not as a design-unbiased estimate of statewide map accuracy [28]. To determine whether the result depended on dividing the locations into exactly 20 clusters, the analysis was repeated using 10, 30, and 40 clusters.

Covariate dissimilarity and applicability

Covariate dissimilarity was assessed using an adaptation of the DI and AOA framework of Meyer and Pebesma [18]. The calculation used the 80% model-development subset, matching the observations used to fit the statewide model. For this analysis, each of the 17 predictors was centred by its mean and scaled by its standard deviation in that subset. Predictors were not weighted by model importance and contributed equally to the distance calculation.

To make the raster calculation computationally feasible and reproducible, a seeded random sample of 3,000 model-development observations was used as the reference set. For each prediction location, DI was defined as its Euclidean distance in standardised predictor space to the nearest reference observation, divided by a reference mean

pairwise distance:

$$DI(\mathbf{x}) = \frac{\min_i \|\mathbf{z}(\mathbf{x}) - \mathbf{z}(\mathbf{x}_i)\|_2}{\bar{d}}. \quad (1)$$

Here, $\mathbf{z}(\mathbf{x})$ is the standardised predictor vector at location $\mathbf{x}$, $\mathbf{z}(\mathbf{x}_i)$ is the corresponding vector for reference observation i, and $\bar{d}$ is the mean pairwise distance within a seeded subsample of 2,000 reference observations. DI was also calculated for every model-development observation relative to the fixed reference set; when an observation occurred in that set, its distance to itself was excluded. The AOA cutoff was the upper Tukey fence of these observation-level DI values, calculated as the third quartile plus 1.5 times the interquartile range. This produced an absolute DI cutoff of 0.529. For presentation, pixel-level DI was divided by this cutoff, so relative DI values at or below 1 were inside the AOA and values above 1 were outside. Because this implementation used an unweighted reference sample rather than fold-specific distances, the resulting AOA was interpreted as a covariate-support domain and not as an area with guaranteed cross-validation accuracy.

Map generation

The RF fitted to the 80% model-development subset was applied across Odisha on the nominal 30 m grid. A 30 m cropland mask representing 2020, prepared using the Landsat time-series mapping approach described by Gumma et al. [29], delineated cropland. The mask was aligned to the prediction grid using mode resampling, and pixels classified as cropland defined the domain for the AOA overlay and cropland stock summaries. We mapped the AOA alongside OC to identify prediction locations whose covariate combinations fell within the specified DI threshold relative to the reference observations. Predictions outside the AOA were retained in the statewide map, while cropland inside the AOA was reported separately as the domain most consistent with the agricultural sampling base. For the joint display of predicted OC and relative DI, 100,000 jointly valid mapped pixels were sampled uniformly without replacement using a fixed random seed (42), retaining their cropland and AOA classifications.

Observation-level prediction intervals

Prediction intervals for individual observations were estimated using nested split conformal prediction, in which residuals from a calibration subset determine the interval width [30]. Conformal methods have also been applied to uncertainty estimation in soil modelling [31]. Within each outer validation fold, the remaining observations were divided into model-training and calibration subsets. Absolute calibration residuals defined symmetric prediction intervals at 80%, 90% and 95% nominal coverage:

$$I_{1-\alpha}(x) = [\hat{y}(x) - q_{1-\alpha},\ \hat{y}(x) + q_{1-\alpha}], \quad (2)$$

where $q_{1-\alpha}$ is the ranked absolute calibration residual used for the target coverage. Within each outer fold, 20% of the available model-fitting observations were reserved for calibration. The random design used a random inner split, whereas complete 30 m groups or geographic clusters were kept intact in the grouped designs. The conformal radius was the finite-sample higher order statistic at rank $\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil$ of the

absolute calibration residuals. Lower interval limits below zero were truncated at zero. The usual marginal coverage guarantee assumes exchangeability between calibration and evaluation observations. Spatial dependence or differences in the distributions of covariates and OC between these observations can violate this assumption, so empirical coverage was assessed under each design. Intervals were evaluated only for observations excluded from fitting in each outer fold and were not propagated to mapped pixels or regional stock totals.

Scenario-based carbon stock estimation

To estimate regional stocks, mapped OC was converted to upper-15-cm soil carbon density as

$$SOC_{\text{density}}\ (\text{t C ha}^{-1}) = OC\ (\%) \times BD\ (\text{g cm}^{-3}) \times d\ (\text{cm}) \times f, \tag{3}$$

where BD is bulk density, $d = 15$ cm and f is the Walkley–Black recovery factor. Pixel densities were multiplied by latitude-corrected pixel area and summed to obtain totals in teragrams of carbon (Tg C). The recovery factor was applied only during this calculation.

Results were calculated for $f = 1.00$ and $f = 1.32$. The latter is commonly used to account for incomplete Walkley–Black oxidation, but recovery depends on soil and procedure [20, 32]. The two factors therefore define alternative analytical-recovery scenarios; they are not lower and upper statistical uncertainty bounds.

Bulk density for 0–15 cm was calculated as the thickness-weighted mean of the SoilGrids 2.0 mean predictions for 0–5 and 5–15 cm. Values were converted from cg cm^{-3} to g cm^{-3} by dividing by 100 and aligned to the prediction grid using nearest-neighbour resampling [4]; its effective spatial resolution remained approximately 250 m. Coarse-fragment content was unavailable and set to zero, so the calculation treated the entire soil volume as fine earth. Totals were calculated for the full mapped extent and for cropland within the AOA. Joint uncertainty from the OC model, bulk density, coarse fragments and recovery factor was not propagated.

Results

Spatial distribution of soil observations

The observations were locally clustered (Fig. 1). The median nearest-neighbour distance was 31.2 m; 48.9% of observations had another observation within 30 m and 27.8% within 5 m. The 90th and 99th percentiles were 0.973 and 5.15 km, respectively. The distribution therefore combined dense local sampling with gaps extending over several kilometres, providing the spatial context for the validation comparisons.

Conventional hold-out validation

The reserved 20% subset provided the hold-out assessment of the mapped model. For these 2,657 observations, the RF gave $R^2 = 0.501$, root mean square error (RMSE) = 0.137 percentage points, mean absolute error (MAE) = 0.092 percentage points,

bias = −0.0048 percentage points and a concordance correlation coefficient of 0.635. Predictions ranged from 0.105 to 1.325% OC, compared with 0.030–2.370% in the observations. The model reproduced part of the observed variation but drew the most extreme OC values towards the centre of the distribution.

Five-fold validation under different spatial partitions

Predictive performance was lower under proximity grouping and geographic-cluster allocation than under random folds (Table 1; Fig. 2). The largest decline occurred when whole geographic clusters were assigned to validation folds.

Table 1 Predictive performance under the three five-fold validation designs (n = 13,284). OC errors and bias are expressed in percentage points. SD denotes standard deviation.

Statistic	Random folds	30 m grouped folds	Geographic-cluster folds
R^2	0.468	0.346	−0.103
RMSE	0.138	0.153	0.199
MAE	0.091	0.102	0.138
Bias	−0.006	−0.011	+0.017
Pearson r	0.692	0.603	0.145
Concordance correlation coefficient	0.609	0.470	0.113
Slope of observed on predicted	1.157	1.251	0.298
Predicted/observed SD ratio	0.598	0.482	0.485
Predicted range (% OC)	0.095–1.467	0.221–1.536	0.317–0.925

Random five-fold validation gave $R^2 = 0.468$, compared with 0.501 for the separate 20% validation subset. Predictions again occupied a narrower range than the observations, with a predicted/observed SD ratio of 0.598 (Fig. 2a). Keeping observations within 30 m in the same fold reduced R^2 to 0.346 and the SD ratio to 0.482 (Fig. 2b). The positive relationship remained, but high-OC observations were more strongly underpredicted. This decline is consistent with immediate proximity contributing to the stronger random-validation performance, although useful predictive information remained after nearby observations were grouped.

Geographic-cluster validation produced a larger decline. Predictions occupied a narrow portion of the observed OC range and showed little ability to distinguish low- from high-OC observations (Pearson $r = 0.145$; Fig. 2c). Across observed-OC deciles, the contrast between overprediction at low OC and underprediction at high OC was more pronounced under the grouped designs, particularly geographic-cluster validation (Supplementary Fig. S2).

The pooled R^2 of −0.103 indicates greater squared error than the benchmark defined by the mean OC of all observations in that pooled evaluation. This benchmark differs from a mean-only model fitted separately to each training fold. RMSE increased to 0.199 percentage points, while the small bias (+0.017 percentage points) indicates that the main loss was in reproducing variation rather than a large overall shift in

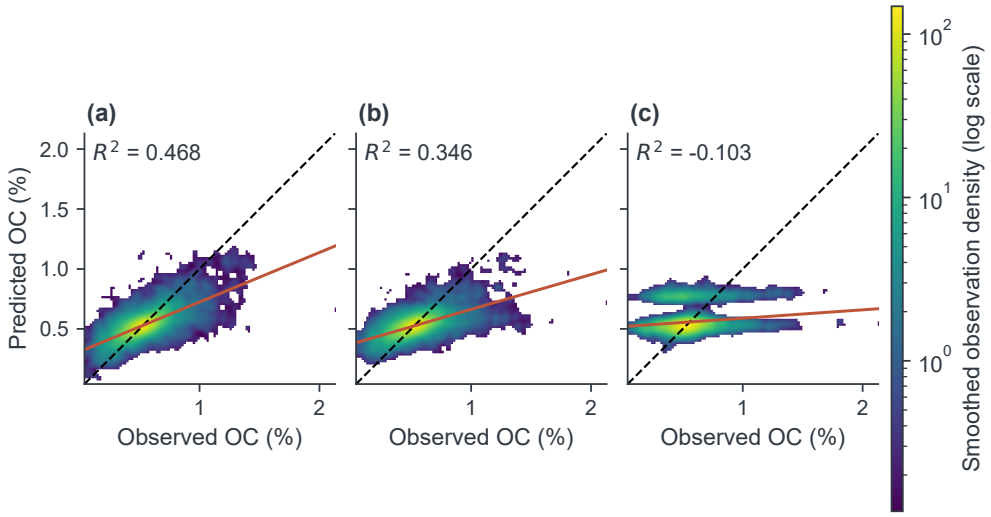


Fig. 2 Observed and predicted OC under (a) random, (b) 30 m grouped and (c) geographic-cluster five-fold validation. Colour indicates smoothed observation density. Dashed lines show 1:1 agreement; solid red lines show fitted linear relationships.

predicted OC. Cross-validated residuals formed coherent areas of underprediction and overprediction (Supplementary Fig. S1).

The weak geographic-transfer result did not depend only on the choice of 20 clusters. Pooled R^2 values were −0.009, −0.103, −0.012 and −0.005 when the observation locations were divided into 10, 20, 30 and 40 clusters, respectively; RMSE ranged from 0.190 to 0.199 percentage points. Each cluster number produced different geographic groupings, so these values do not form a sequence of increasingly strict spatial partitions. Their consistently near-zero or negative values nevertheless show that limited transfer was common across the tested spatial partitions.

Predictor importance

Permutation importance identified the covariates on which predictions for the reserved validation subset depended most strongly. Elevation had the highest importance (mean decrease in R^2 = 0.185), followed by ALOS-2/PALSAR-2 HH (0.137), HV (0.113), annual precipitation (0.105) and precipitation seasonality (0.098). HLS SWIR1 and NIR each had an importance of 0.046. Figure 3 shows the 12 highest-ranked covariates from the model trained using all 17 predictors listed in Supplementary Table S1. These values describe dependence of the fitted model and were used to interpret the broad mapped pattern.

Predicted surface and applicability

The statewide map showed broad regional variation in predicted topsoil organic carbon. Across the mapped extent, predicted topsoil OC averaged 0.551%, with a standard deviation of 0.055 percentage points and a 5th–95th percentile range of 0.465–0.635%

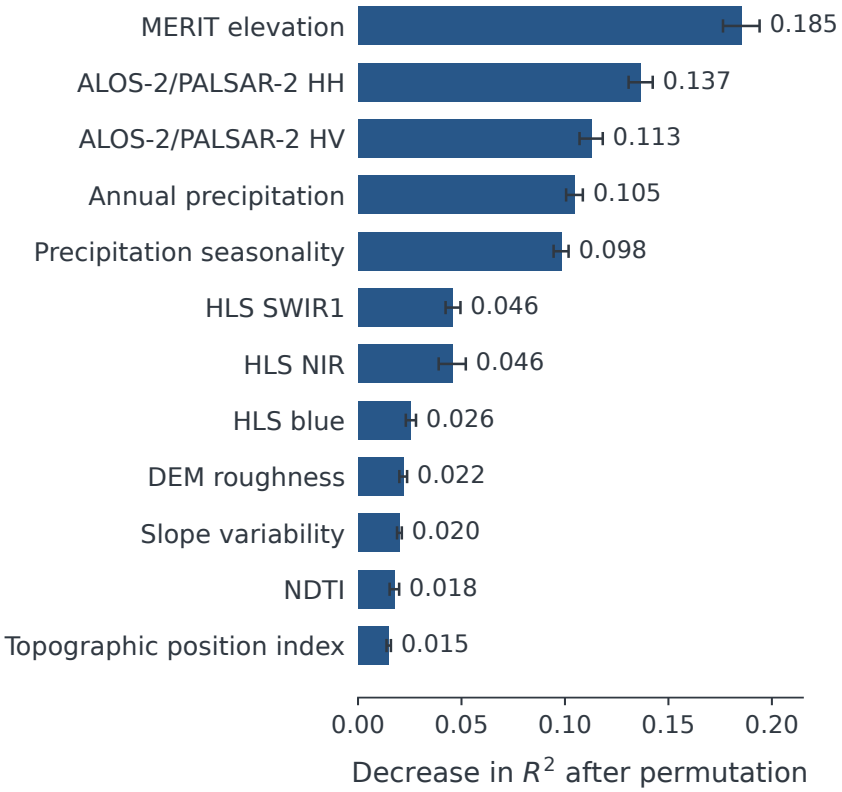


Fig. 3 The 12 highest-ranked covariates by permutation importance among the 17 predictors used in the random-forest model. Importance was evaluated on the reserved 20% validation subset. Bars show the mean decrease in R^2 across ten permutations per predictor; error bars show one standard deviation across those permutations.

(Fig. 4). Lower values formed an extensive zone across the eastern coastal plain and northeastern Odisha. Higher values occurred more frequently in the western and southern uplands, with smaller high-OC areas in the north. The mapped range was narrower than the observed range, consistent with the contraction toward intermediate OC values observed during model evaluation.

Covariate dissimilarity differed between the full mapped extent and cropland. Relative DI was below 1 across most cultivated areas, indicating that their combinations of terrain, climate and remotely sensed covariates resembled those at the model-development observations (Supplementary Fig. S3). Areas with relative DI above 1 were concentrated mainly in central, western and northern Odisha. Using an absolute DI cutoff of 0.529, 68.1% of the full mapped extent was classified as inside the AOA.

Overlaying the AOA with the cropland mask showed that 97.6% of mapped cropland, equivalent to 5.32 million hectares, was inside the AOA (Fig. 5). Cropland outside

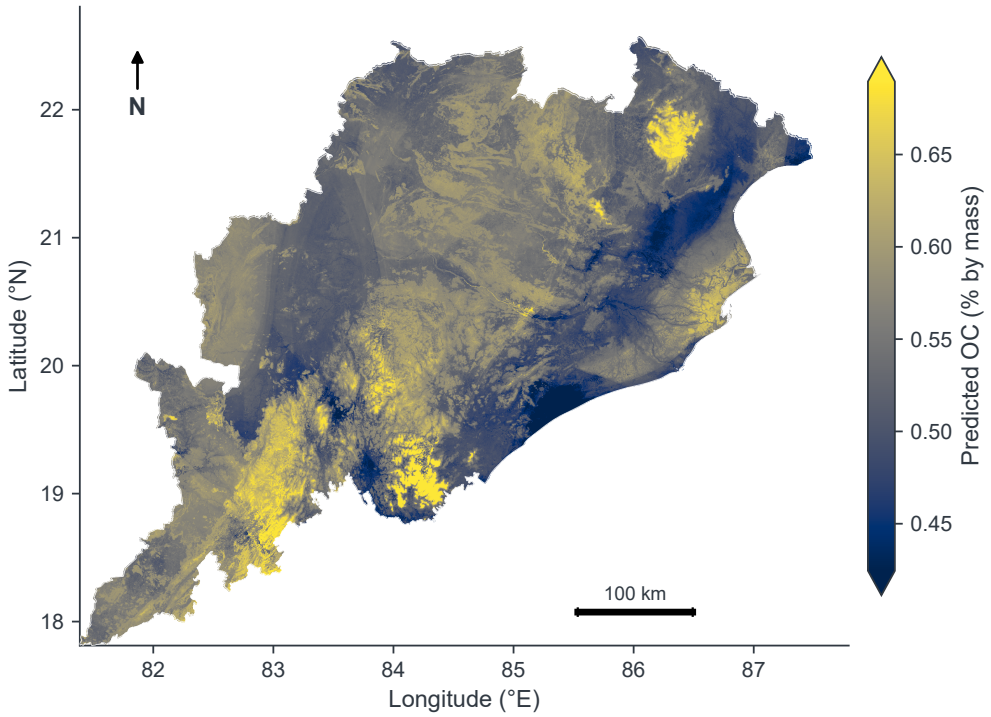


Fig. 4 Predicted topsoil OC (0–15 cm) across Odisha on a 30 m grid. The display colour scale spans the 1st–99th percentiles; values outside these limits remain in the prediction raster.

the AOA occurred in small, dispersed patches, whereas most pixels outside the AOA belonged to non-cropland environments.

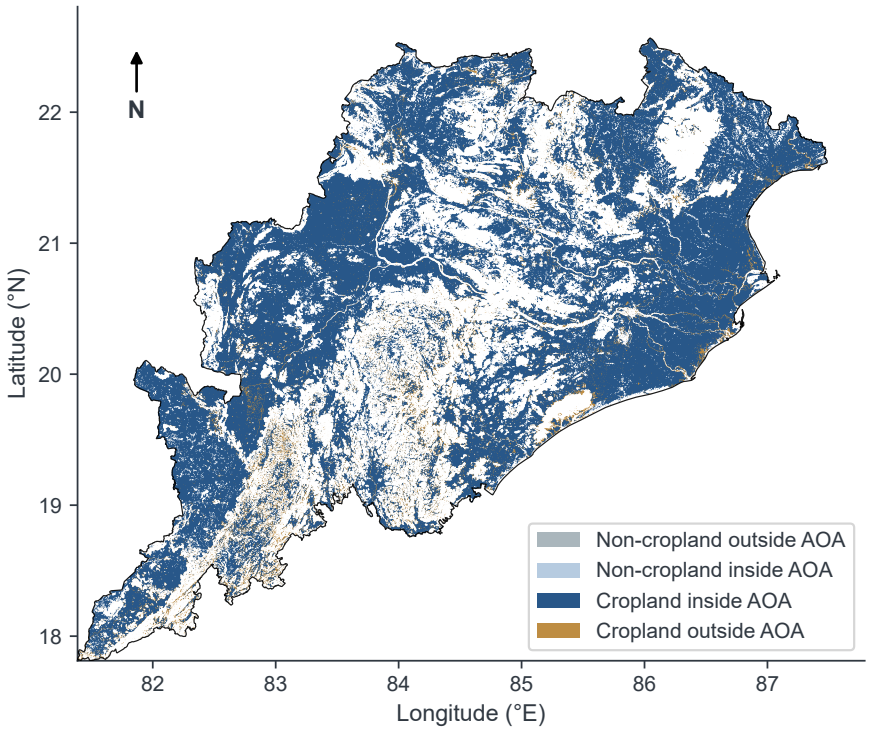


Fig. 5 Cropland and non-cropland within and outside the area of applicability (AOA). Cropland inside the AOA is the reporting domain most consistent with the agricultural origin of the soil observations.

This contrast was also evident in the relative-DI distributions (Supplementary Fig. S4). Cropland values were concentrated below the AOA boundary, while the distribution for the full mapped extent had a longer tail above relative DI = 1. Within cropland inside the AOA, predicted OC averaged 0.527%, with a median of 0.525% and a 5th–95th percentile range of 0.455–0.595%. The corresponding distributions for the full extent, all cropland and cropland within the AOA are shown in Supplementary Fig. S5.

The spatial contrast in the AOA map (Fig. 5) was reflected in the joint distribution of relative DI and predicted OC (Fig. 6). In the 100,000-pixel sample, median relative DI was 0.569 for cropland inside the AOA and 1.531 for non-cropland outside it. Most cropland outside the AOA lay near the boundary, with a median relative DI of 1.130 and a 95th percentile of 1.705. Non-cropland outside the AOA had a higher 95th percentile of 3.034 and accounted for most of the long DI tail.

Predicted OC distributions nevertheless overlapped across the four classes. In the sample, cropland inside the AOA had a median predicted OC of 0.527% and a 5th–95th percentile range of 0.454–0.593%, compared with 0.586% and 0.504–0.671% for

non-cropland outside it. Similar OC predictions therefore occurred on both sides of the AOA boundary, even where their covariate combinations differed.

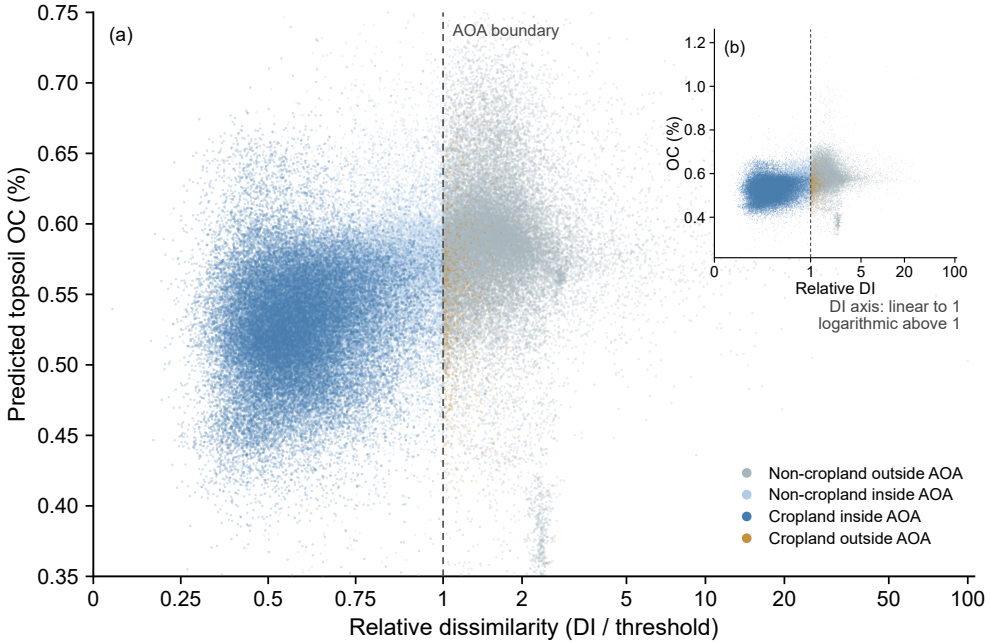


Fig. 6 Predicted topsoil OC and relative covariate dissimilarity for a uniform sample of 100,000 mapped pixels. (a) The main panel focuses on 0.35–0.75% OC, containing 99.5% of sampled predictions. (b) The inset shows the full OC range and all sampled points. Colours correspond to the four classes in Fig. 5. The dashed line marks the AOA boundary at relative DI = 1. Both horizontal axes are linear from 0 to 1 and logarithmic above 1.

Validation of prediction intervals

Prediction intervals were validated using empirical coverage and interval width under the same three five-fold designs. Mean interval widths are reported below, while Fig. 7b shows the median width in each fold. Under random validation, empirical coverage closely followed the nominal 80%, 90% and 95% levels (Fig. 7a). Coverage of the nominal 90% interval was 90.14%, with a mean width of 0.424 percentage points. The 30 m grouped design gave similar 90% coverage (89.95%), although mean width increased to 0.456 percentage points (Fig. 7b).

Under geographic-cluster validation, coverage fell below the nominal level. Observed coverage was 72.06%, 81.96% and 87.65% for the nominal 80%, 90% and 95% intervals, respectively. The mean width of the 90% intervals increased to 0.538 percentage points and varied more among folds. Thus, wider intervals did not fully compensate for the additional errors encountered in geographically separate locations. The example intervals show the same contrast: most randomly validated observations were enclosed,

whereas more observations from the geographic-cluster folds fell outside their intervals (Fig. 7c,d).

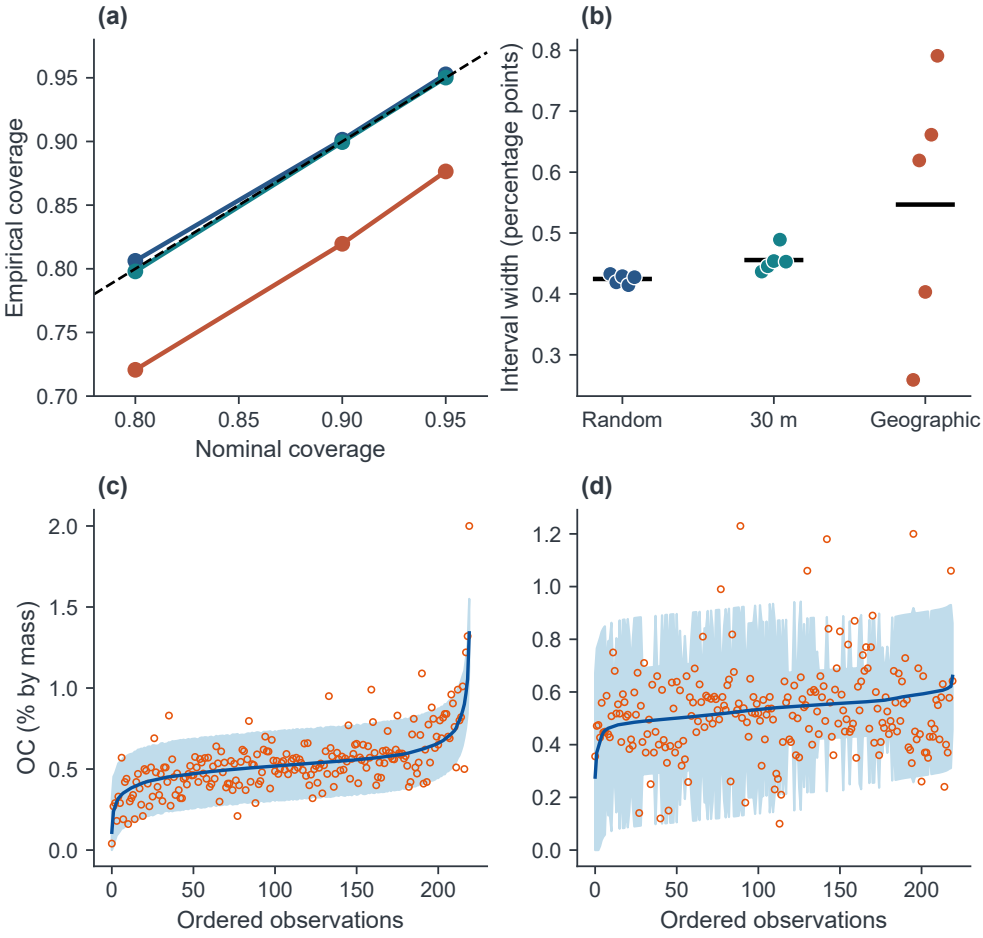


Fig. 7 Validation of observation-level conformal prediction intervals. (a) Empirical versus nominal coverage; the dashed line indicates equal coverage. (b) Median width of the 90% interval in each fold; black bars show the mean of the five fold medians. Blue, teal and orange denote random, 30 m grouped and geographic-cluster validation, respectively. Panels (c) and (d) show 220 illustrative observations ordered by predicted OC under random and geographic-cluster validation. Shading shows the 90% intervals, dark lines the predictions and open orange circles the observations.

Scenario-based upper-15-cm soil carbon stocks

Carbon stocks were compared between the full mapped extent and cropland within the AOA under two Walkley–Black recovery assumptions (Supplementary Table S2). These are analytical scenarios rather than uncertainty bounds. Without a recovery correction, mean carbon density was 11.581 t C ha^{-1} over the full mapped extent and 11.064

t C ha^{-1} for cropland within the AOA. These values corresponded to upper-15-cm soil carbon stocks of 180.74 and 58.83 Tg C, respectively. Applying the 1.32 recovery factor increased mean carbon density to 15.286 and 14.604 t C ha^{-1} and increased the corresponding totals to 238.57 and 77.66 Tg C. Corrected pixel-level stocks ranged from 5.858 to 38.221 t C ha^{-1} across the full extent and from 5.858 to 34.113 t C ha^{-1} within cropland inside the AOA.

Cropland within the AOA represented 34.1% of the mapped area and 32.5% of the uncorrected full-extent stock. Its slightly smaller stock share reflected a mean carbon density about 4.5% below that of the full mapped extent.

The full-extent total included the 31.92% of mapped pixels outside the AOA, whereas the cropland total was restricted to cropland pixels inside the DI-based AOA. Neither total has a formal confidence interval because uncertainty from the OC model, bulk density, coarse fragments and recovery factor was not propagated jointly.

Discussion

Regional OC patterns and spatial validation

The statewide map extends the use of SHC observations from sampled agricultural fields to predicted regional OC patterns. Higher predictions occurred more frequently in the western and southern uplands, while lower values extended across much of the coastal plain and northeastern Odisha. Elevation, L-band HH and HV backscatter, and precipitation were the predictors on which the model depended most strongly. These associations suggest that the model used regional contrasts in terrain, moisture regime and surface condition, but they do not establish independent controls on OC. The narrower predicted range also shows that these covariates did not reproduce the full observed variation. Farm management, including residue return, irrigation, manure use and crop rotation, was unavailable and may account for part of the remaining variation.

The validation comparisons help distinguish performance among spatially interspersed observations from transfer to held-out geographic clusters. Hold-out and random five-fold validation gave similar performance (R^2 = 0.501 and 0.468, respectively), and grouping observations within 30 m retained a positive relationship while reducing R^2 to 0.346. Geographic-cluster validation produced a much larger loss: R^2 became negative, correlation weakened and predictions occupied only a narrow part of the observed range. Together with the spatially coherent residuals, these results show that prediction among spatially interspersed observations was more successful than transfer to geographically separate parts of Odisha. The nominal 30 m grid describes the output spacing; accuracy at that scale was not established for individual fields.

This contrast is consistent with the different questions posed by random and geographic validation. Random folds allow fitting and evaluation observations to share local environmental context, whereas geographic-cluster folds test predictions for groups of locations excluded from model fitting [13, 14, 33, 34]. Jafari et al. compared spatial predictors and validation strategies in SOC modelling [17]; our analysis kept the model specification fixed to examine how the evaluation changed with spatial partitioning. The near-zero or negative results across 10–40 clusters show that weak transfer was common to the tested partitions. They describe transfer within the SHC archive, while

design-based assessment of statewide map accuracy would require an independent probability sample [28].

Covariate applicability and prediction intervals

The AOA identifies the portion of the mapped extent with covariate combinations represented within the specified DI threshold. It included 97.6% of mapped cropland and 68.1% of the full extent, consistent with the agricultural origin of the observations. Most pixels outside the AOA belonged to non-cropland environments, where combinations of measured covariates were less well represented. DI measures similarity in the selected predictors, whereas geographic validation evaluates errors at locations excluded from fitting [18]. The high cropland AOA coverage therefore identifies where predictor combinations fall within the DI-based applicability domain; it does not establish accurate OC predictions or transfer of the fitted relationship between regions.

The interval-coverage results provide further evidence of the difference between random and geographic validation. Coverage remained close to the nominal 90% level under random and 30 m grouped validation, but fell to 81.96% under geographic-cluster validation despite wider intervals. Calibration residuals therefore did not fully represent the errors encountered in geographically separate locations. As emphasized by Lilburne et al. [19], uncertainty information needs evaluation for its intended prediction domain. Our intervals were evaluated at observation locations; they do not provide mapped pixel uncertainty or confidence intervals for regional stocks.

Implications for carbon stock estimation

The contrast in covariate coverage also affects how the regional stock estimates are reported. Cropland within the AOA covered 5.32 million hectares and yielded upper-15-cm stock estimates of 58.83 and 77.66 Tg C under the two recovery assumptions. The full-extent estimates were 180.74 and 238.57 Tg C, but almost one-third of that area was outside the AOA. The full-extent estimates therefore include contributions from locations whose covariate combinations are dissimilar to those in the reference observations.

Within each reporting domain, the difference between the two estimates reflects the analytical-recovery assumption. Actual Walkley–Black recovery varies among soils and laboratory procedures [32], and the estimates also inherit error from predicted OC, the coarser bulk-density layer and the assumption of zero coarse fragments. Khanal et al. mapped forest SOC stocks in Nepal using quantile random forest and examined model uncertainty and sensitivity to the spatial distribution of sample plots [35]. Their stock assessment illustrates the need to distinguish a predicted regional total from the uncertainty attached to it. Here, observation-level intervals were not propagated through the stock calculation, so the totals remain conditional regional estimates for carbon assessment and planning.

Relation to previous SOC mapping studies

The measured OC mean of 0.540% and the predicted mean of 0.527% for cropland within the AOA fell within the 0.39–1.06% range of agro-ecological-region means

reported by Kandadai and Dadhwal from more than 39 million SHC observations [8]. This provides national context for the magnitude of the Odisha values, rather than independent validation of their spatial distribution. Earlier work in eastern Odisha also evaluated multi-frequency SAR backscatter for SOC prediction [12]. In the present model, L-band HH and HV were among the predictors with the highest permutation importance. Direct comparisons of accuracy statistics remain limited by differences in study extent, observations, covariates and validation design.

Hengl et al. demonstrated how compiled soil observations and Earth observation data can produce detailed regional soil information, with spatial cross-validation used to evaluate the predictions [5]. The present study uses SHC measurements for this regional mapping approach and derives stock estimates from the predicted OC surface, with cropland within the AOA reported separately under two analytical-recovery assumptions. The high AOA coverage of cropland and the weak geographic validation show that broad representation of predictor conditions does not establish transfer of the fitted OC relationship. This distinction is useful when existing agricultural soil records are extended beyond their sampled locations.

Limitations

The observation network was concentrated in agricultural land and was not designed to estimate statewide map accuracy. All evaluation observations came from the same programme, including those used for geographic-cluster validation. An independent probability-based survey covering cropland, non-cropland and the AOA boundary would therefore provide stronger evidence of map accuracy across Odisha.

Several processing choices remain untested. The effect of wavelet denoising cannot be isolated without an otherwise identical model fitted to unfiltered predictors. The AOA depends on equal predictor weights, a sampled reference set and the chosen cutoff; sensitivity to these choices was not evaluated. Covariate screening was response-blind but used the full archive before partitioning, so validation observations were excluded from RF fitting rather than from every preprocessing step. A strictly inductive evaluation would repeat data-dependent screening within the fitting partitions. The cropland classification was not independently validated for Odisha, and individual soil sampling dates could not be matched to satellite observations.

Stock estimation adds assumptions about bulk density, coarse fragments and analytical recovery to those affecting the OC map. A stronger assessment would combine independent soil sampling with measured bulk density and coarse fragments and propagate analytical, model and spatial uncertainty. Evaluation using SHC observations from other states would test the geographic transfer of this mapping approach.

Conclusions

Using 13,284 SHC observations and 17 satellite, terrain and climate covariates, we mapped OC in the upper 15 cm across Odisha on a nominal 30 m grid. The mapping model achieved $R^2 = 0.501$ and an RMSE of 0.137 percentage points on the reserved 20% subset. Reusing these agricultural measurements provided a statewide view of predicted OC beyond the sampled fields, although performance declined when nearby

observations were grouped and was weak for geographically separate clusters. These results make spatial evaluation integral to interpreting maps derived from agricultural soil archives.

Cropland inside the AOA covered 5.32 million ha, or 97.6% of mapped cropland. Stocks for this domain in the upper 15 cm were estimated at 58.83 and 77.66 Tg C under the two Walkley–Black recovery assumptions. These conditional estimates can inform regional carbon assessment and guide further measurements of OC, bulk density and coarse fragments. Quantifying uncertainty in these stock totals would strengthen their use in carbon assessment. The same mapping and evaluation approach could be tested for other soil properties in the SHC archive, with their sampling coverage and geographic transfer assessed separately.

Acknowledgements

The authors thank the Department of Agriculture and Farmers' Empowerment, Government of Odisha, for funding the soil sampling programme from which the observations were obtained, and the Charles Renard Analytical Laboratory (CRAL) at ICRISAT for the soil analyses. We acknowledge the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) and the National Institute of Technology Warangal for research facilities and technical expertise.

Data and code availability

Point-level Soil Health Card observations cannot be redistributed under the terms of access. Derived rasters and analysis code are available from the corresponding author, subject to data-provider permission and institutional requirements.

AI-assisted manuscript preparation

The authors used OpenAI's ChatGPT to assist with language editing, organization, consistency checks, and code drafting and review for selected preprocessing and visualization scripts. All scientific content, analyses, citations, figures, code, and final text were reviewed, revised, tested where applicable, and approved by the authors.

Author Contributions

Pavan Kumar Bellam: Conceptualization, Methodology, Software, Formal analysis, Validation, Investigation, Data curation, Visualization, Writing – original draft, and Writing – review and editing. Murali Krishna Gumma: Conceptualization, Resources, Supervision, and Writing – review and editing. Venkata Reddy Keesara: Conceptualization, Methodology, Supervision, and Writing – review and editing. Mukund Dhavaji Patil: Investigation, and Data curation. Narayanarao Bhogapurapu: Conceptualization, Methodology, Supervision, and Writing – review and editing.

Competing interests

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.

References

[1] Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **304**, 1623–1627 (2004). https://doi.org/10.1126/science.1097396.

[2] McBratney, A.B., Mendonça Santos, M.L. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52 (2003). https://doi.org/10.1016/S0016-7061(03)00223-4.

[3] Wadoux, A.M.J.-C., Minasny, B. & McBratney, A.B. Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth Sci. Rev.* **210**, 103359 (2020). https://doi.org/10.1016/j.earscirev.2020.103359.

[4] Poggio, L. *et al.*. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL* **7**, 217–240 (2021). https://doi.org/10.5194/soil-7-217-2021.

[5] Hengl, T. *et al.* African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Sci. Rep.* **11**, 6130 (2021). https://doi.org/10.1038/s41598-021-85639-y.

[6] Safaee, S. *et al.*. Influence of sample size, model selection, and land use on prediction accuracy of soil properties. *Geoderma Reg.* **36**, e00766 (2024). https://doi.org/10.1016/j.geodrs.2024.e00766.

[7] Government of India. Soil Health Card Scheme: Frequently Asked Questions. Department of Agriculture & Cooperation, Ministry of Agriculture. https://soilhealth.dac.gov.in/files/FAQ_Final_English.pdf (accessed 28 September 2026) (n.d.).

[8] Kandadai, S. & Dadhwal, V.K. Large-scale spatial assessment of soil organic carbon, pH and their interrelation in Indian agricultural soils using Soil Health Card big data. *Environ. Monit. Assess.* **198**, 336 (2026). https://doi.org/10.1007/s10661-026-15184-6.

[9] Sarkar, D., Thampi, C.J., Sehgal, J. & Velayutham, M. *Soils of Orissa for Optimum Land Use*. ICAR–National Bureau of Soil Survey and Land Use Planning, Nagpur, India (1998).

[10] Mishra, A., Dash, P.K., Mohanty, S. & Mishra, K.N. Soils of Odisha: Types, Status and Management. In: *Land Restoration, Desertification & Drought Resilience*. State EIACP Hub, Centre for Environmental Studies, Government of Odisha, Bhubaneswar, pp. 7–19. https://www.cesodisha.org/admin/publication/pdf_73.

pdf (2024).

[11] Bellam, P.K., Gumma, M.K., Bhogapurapu, N. & Keesara, V.R. Soil organic carbon mapping using Sentinel-1 and ALOS-2 SAR data. In: *2024 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, pp. 1–4. https://doi.org/10.1109/InGARSS61818.2024.11105067 (2024).

[12] Bellam, P.K., Gumma, M.K., Bhogapurapu, N. & Keesara, V.R. Soil organic carbon mapping using multi-frequency SAR data and machine learning algorithms. *Land* **14**, 2105 (2025). https://doi.org/10.3390/land14112105.

[13] Roberts, D.R. *et al.*. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017). https://doi.org/10.1111/ecog.02881.

[14] Ploton, P. *et al.*. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **11**, 4540 (2020). https://doi.org/10.1038/s41467-020-18321-y.

[15] Pohjankukka, J., Pahikkala, T., Nevalainen, P. & Heikkonen, J. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int. J. Geogr. Inf. Sci.* **31**, 2001–2019 (2017). https://doi.org/10.1080/13658816.2017.1346255.

[16] Fourel, V., Wiesmeier, M., Kakhani, N. & Scholten, T. Compact CNN-Transformers match larger models for spatially-robust soil organic carbon mapping at regional scale. *Geoderma* **472**, 117933 (2026). https://doi.org/10.1016/j.geoderma.2026.117933.

[17] Jafari, A., Sarmadian, F., Heidari, A. & Rasaei, Z. Enhancing digital mapping of soil organic carbon through spatial modeling and validation. *Sci. Rep.* **16**, 8810 (2026). https://doi.org/10.1038/s41598-026-39496-2.

[18] Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.* **12**, 1620–1633 (2021). https://doi.org/10.1111/2041-210X.13650.

[19] Lilburne, L., Helfenstein, A., Heuvelink, G.B.M. & Eger, A. Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids. *Geoderma* **450**, 117052 (2024). https://doi.org/10.1016/j.geoderma.2024.117052.

[20] Walkley, A. & Black, I.A. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil Sci.* **37**, 29–38 (1934). https://doi.org/10.1097/00010694-193401000-00003.

[21] Claverie, M. *et al.*. The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sens. Environ.* **219**, 145–161 (2018). https://doi.org/10.1016/j.rse.2018.09.002.

[22] Yamazaki, D. *et al.*. A high-accuracy map of global terrain elevations. *Geophys. Res. Lett.* **44**, 5844–5853 (2017). https://doi.org/10.1002/2017GL072874.

[23] Fick, S.E. & Hijmans, R.J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017). https://doi.org/10.1002/joc.5086.

[24] Torres, R. *et al.* GMES Sentinel-1 mission. *Remote Sens. Environ.* **120**, 9–24 (2012). https://doi.org/10.1016/j.rse.2011.05.028.

[25] Kankaku, Y., Suzuki, S. & Osawa, Y. ALOS-2 mission and development status. In: *2013 IEEE International Geoscience and Remote Sensing Symposium–IGARSS*, 2396–2399 (2013). https://doi.org/10.1109/IGARSS.2013.6723302.

[26] Donoho, D.L. & Johnstone, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455 (1994). https://doi.org/10.1093/biomet/81.3.425.

[27] Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

[28] Wadoux, A.M.J.-C., Heuvelink, G.B.M., de Bruin, S. & Brus, D.J. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecol. Model.* **457**, 109692 (2021). https://doi.org/10.1016/j.ecolmodel.2021.109692.

[29] Gumma, M.K. *et al.*. Agricultural cropland extent and areas of South Asia derived using Landsat satellite 30-m time-series big-data using random forest machine learning algorithms on the Google Earth Engine cloud. *GISci. Remote Sens.* **57**, 302–322 (2020). https://doi.org/10.1080/15481603.2019.1690780.

[30] Shafer, G. & Vovk, V. A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9**, 371–421 (2008). https://jmlr.org/papers/v9/shafer08a.html.

[31] Huang, Y.-C., Padarian, J., Minasny, B. & McBratney, A.B. Using Monte Carlo conformal prediction to evaluate the uncertainty of deep-learning soil spectral models. *SOIL* **11**, 553–563 (2025). https://doi.org/10.5194/soil-11-553-2025.

[32] De Vos, B., Lettens, S., Muys, B. & Deckers, J.A. Walkley–Black analysis of forest soil organic carbon: recovery, limitations and uncertainty. *Soil Use Manag.* **23**, 221–229 (2007). https://doi.org/10.1111/j.1475-2743.2007.00084.x.

[33] Hateffard, F., Steinbuch, L. & Heuvelink, G.B.M. Evaluating the extrapolation potential of random forest digital soil mapping. *Geoderma* **441**, 116740 (2024). https://doi.org/10.1016/j.geoderma.2023.116740.

[34] Meyer, H. & Pebesma, E. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.* **13**, 2208 (2022). https://doi.org/10.1038/s41467-022-29838-9.

[35] Khanal, S., Nolan, R. H., Medlyn, B. E. & Boer, M. M. Mapping soil organic carbon stocks in Nepal's forests. *Sci. Rep.* **13**, 8090 (2023). https://doi.org/10.1038/s41598-023-34247-z.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SOCOdishaScientificreportssupplementary.pdf