https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

A Global Soil Dataset for Earth System Modeling (Version 2, GSDEv2)

Gaosong Shi[1,2], Anne Gobin[2], Zhongwang Wei[1], Hua Yuan[1], Qingliang Li[3], Wei Shangguan[1,*], and Yongjiu Dai[1]

[1]School of Atmospheric Sciences, Sun Yat-sen University, Zhuhai 519082, China
[2]Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001 Leuven, Belgium
[3]School of Artificial Intelligence, Changchun Normal University, Changchun 130032, China

Correspondence to: Wei Shangguan (Email: shgwei@mail.sysu.edu.cn)

Abstract.

Soil properties are essential inputs to Earth system models. However, existing global soil datasets vary significantly in property coverage, spatial and vertical resolution, uncertainty characterization, and the integration of regional soil information. Here, we present the Global Soil Dataset for Earth System Modeling version 2 (GSDEv2), a model-oriented global dataset developed within a framework that synergizes observation-informed spatial prediction, soil-domain-specific modelling, and stacking-based incorporation of regional soil products with a globally consistent baseline. GSDEv2 provides predictions of 20 physical and chemical soil properties for six standard depth intervals on a nominal 90 m grid, together with spatially aggregated 1 km mean and conditional-quantile products. Compatible regional soil products for China, the United States, Africa, Canada, and Australia were incorporated as auxiliary predictors through stacking-based fusion. Global baseline performance was evaluated using a profile-level random hold-out design with 90 % of profiles used for training and 10 % for testing, characterizing interpolation performance within the sampled observational domain. The incremental value of regional fusion was evaluated for surface soil organic carbon in China using nested 60 km spatial-block validation. Among 54 property–depth combinations jointly supported by GSDEv2, GSDEv1, SoilGrids 2.0, and HWSD 2.0, GSDEv2 showed the highest mean modelling efficiency coefficient and the lowest mean root-mean-square error across the evaluated combinations. Median RMSE reductions relative to GSDEv1, SoilGrids 2.0, and HWSD 2.0 were 22.2 %, 5.8 %, and 24.6 %, respectively. In the spatially blocked China case study, regional stacking reduced RMSE by 31.8 % and the mean 90 % prediction-interval width by 38.7 % while maintaining approximately nominal coverage. In a downstream CoLM2024 sensitivity experiment, the GSDEv2-driven simulation reduced the annual-mean spatial RMSE of surface soil moisture from 0.109 to 0.089 $m^3 m^{-3}$ relative to the SoilGrids-driven experiment and increased the spatial correlation with adjusted ERA5-Land from 0.741 to 0.802. GSDEv2 thus provides a vertically harmonized and spatially detailed soil resource for regional-to-global Earth system applications. The GSDEv2 dataset is freely available at https://doi.org/10.11888/Terre.tpdc.303613 or https://www.scidb.cn/s/my2aea (Shi et al., 2026).

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

1. Introduction

Soil is a fundamental component of the Earth system, regulating exchanges of water, energy, carbon, nutrients, and solutes among the land surface, vegetation, groundwater, and atmosphere (Evangelista et al., 2024). Soil physical and chemical properties control key hydrological, thermal, biogeochemical, and ecological processes, and uncertainty in their spatial and vertical representation can propagate into simulations of soil moisture, runoff, evapotranspiration, vegetation dynamics, carbon storage, and land–atmosphere feedbacks (Dai et al., 2019a; Du et al., 2026; Fatichi et al., 2020; Lagacherie, 2025; Todd-Brown et al., 2013). Reliable and spatially consistent soil information is therefore essential for land-surface, Earth-system, hydrological, and biogeochemical modelling.

A range of global soil datasets have been developed to support large-scale modelling and environmental applications (FAO and IIASA, 2023; Francos et al., 2025; Padarian et al., 2022; Poggio et al., 2021; Shangguan et al., 2014; Turek et al., 2023). Legacy products such as the Harmonized World Soil Database integrated soil maps, profile observations, and expert-derived estimates to provide globally consistent soil information (FAO and IIASA, 2023). GSDEv1 extended this approach for Earth system modelling by integrating global, regional, and national soil information at 30 arc-second resolution for eight soil layers down to 2.3 m (Shangguan et al., 2014). More recent digital soil mapping products, including SoilGrids 2.0 and OpenLandMap-soildb, combine harmonized soil observations, environmental covariates, and machine-learning methods to generate spatially continuous predictions with explicit uncertainty information (Hengl et al., 2017, 2026; Poggio et al., 2021). These developments have substantially improved the spatial detail, accessibility, and uncertainty characterization of global soil information.

Despite these advances, existing global soil products differ in variable coverage, spatial and vertical support, modelling objectives, and intended applications. Fine nominal grid spacing does not necessarily correspond to uniform effective spatial resolution, and point-based predictive accuracy alone cannot fully characterize the spatial structure of gridded soil products (Rossiter and Poggio, 2025). For land-surface and Earth-system applications, an additional challenge is how to combine globally consistent soil predictions with complementary regional information. Regional and national products may benefit from denser observations, higher-resolution covariates, and locally optimized mapping, but direct substitution or mosaicking can introduce inconsistencies in property definitions, units, vertical support, grid specifications, provenance, and spatial transitions. GSDEv2 was developed to address this problem by combining harmonized multi-source soil profiles, standardized environmental covariates, soil-domain-specific Quantile Random Forest modelling, conditional uncertainty estimation, and stacking-based incorporation of compatible regional soil products. The dataset provides 20 physical and chemical soil properties for six standard depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm) on a nominal 90 m grid, together with spatially aggregated 1 km products. Regional and national gridded soil products from China, the United States, Africa, Canada, and Australia were incorporated as auxiliary predictors rather than directly replacing the global baseline.

Accordingly, this study aims to:

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

(1) compile and harmonize a global multi-source soil-profile database and develop property-, depth-, and soil-domain-specific QRF models;

(2) incorporate compatible regional soil products through stacking-based fusion while retaining a globally consistent baseline; and

(3) evaluate global baseline performance and uncertainty, quantify the incremental value of regional fusion, and illustrate the downstream sensitivity of CoLM simulations to alternative global soil inputs.

The overall GSDEv2 development, evaluation, and release workflow is summarized in Fig. 1.

2. Source data

GSDEv2 was developed using three main classes of input data: soil-profile observations, environmental covariates, and existing global, regional, and national gridded soil products. Soil-profile observations served as response data for model development, environmental covariates provided spatially continuous predictors, and existing soil products were used either for benchmarking or as auxiliary predictors in regional fusion. Benchmark datasets were not used to train the GSDEv2 global baseline models, whereas selected regional products were incorporated only within their corresponding fusion domains.

2.1 Soil profile observations

GSDEv2 was developed from a global compilation of soil-profile observations assembled from multiple global, regional, and national sources. Major sources included the World Soil Information Service (WoSIS; Batjes et al., 2024), the Second National Soil Survey of China (National Soil Survey Office, 1996), and additional regional and national datasets described by Shangguan et al. (2014), Chen et al. (2025), and Hengl et al. (2026). Source versions, geographical coverage, sampling periods, coordinate information, licences, and numbers of profiles or sites before and after screening are summarized in Table S1.

After harmonization, duplicate resolution, and quality control, the compiled database contained 234,785 unique soil profiles or observation sites with at least one valid GSDEv2 target-property measurement. Because property availability and sampling depths differ among sources, the number of observations used for individual property–depth models varies accordingly.

2.2 Soil properties and depth setting

GSDEv2 provides 20 physical and chemical soil properties covering soil texture, bulk density, coarse fragments, soil acidity, cation exchange capacity, carbon and nitrogen concentrations, Bray-I extractable phosphorus, and soil water retention. Variable definitions, acronyms, units, and raster conversion factors are summarized in Table 1.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Source observations were harmonized to consistent property definitions, analytical methods, units, and depth support. Properties derived using analytically distinct procedures were retained as separate target variables where methodological differences affected interpretation. Detailed compatibility rules and quality-control procedures are provided in Sect. S2. Bray-I phosphorus was retained as a method-specific property and should be interpreted cautiously in alkaline and calcareous soils (Ebeling et al., 2008).

Six standard depth intervals were adopted following the GlobalSoilMap framework: 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm (Arrouays et al., 2014). Observations reported for irregular horizons were harmonized to these intervals as described in Sect. 3.1.

2.3 Environmental covariates

A total of 158 environmental covariates were compiled following the SCORPAN framework (McBratney et al., 2003), representing climate, vegetation and land-surface conditions, terrain, parent material and geology, remote-sensing information, and existing spatial soil information. Major covariate groups and representative source products are summarized in Table 2, while the complete predictor inventory, source versions, spatial resolutions, and preprocessing protocols are provided in Tables S2, S3a, and S3b.

All covariates were transformed to WGS84 (EPSG:4326) and aligned to the common nominal 3-arcsec GSDEv2 modelling grid. Continuous and categorical variables were resampled using data-type-appropriate procedures, with detailed preprocessing specifications provided in Table S3b.

2.4 Benchmark and regional soil products

Existing gridded soil products were used for two distinct purposes. GSDEv1, SoilGrids 2.0, and HWSD 2.0 served exclusively as global benchmark products and were not used to train the GSDEv2 global baseline models. Because some underlying soil observations may overlap among these datasets, their comparison with GSDEv2 is interpreted as relative benchmarking rather than fully independent external validation.

Regional and national gridded soil products were incorporated as auxiliary predictors for China, the United States, Africa, Canada, and Australia. These products were selected based on data accessibility, compatibility with GSDEv2 target properties, spatial coverage, and availability of sufficiently detailed gridded information. Their native spatial resolutions, variables, vertical support, and references are summarized in Table 3.

Before fusion, regional products were harmonized to GSDEv2 property definitions, units, grid specifications, and target depth intervals. Fusion was restricted to property–depth combinations supported by the native vertical range of each regional product; unsupported combinations retained the global baseline prediction. Detailed product-specific harmonization and fusion eligibility are documented in Table S8 and Sect. 3.4.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

3. Methods

3.1 Profile harmonization and quality control

Soil observations from heterogeneous global, regional, and national sources were harmonized to consistent target-property definitions, analytical methods, units, coordinates, and depth descriptions following established global soil-data standards (Arrouays et al., 2014; Batjes et al., 2024; Shi et al., 2025). Quality control included structural and metadata checks, property-specific physical and pedological screening, duplicate resolution, and environmental-context inspection. Records with invalid coordinates or depths, ambiguous or non-convertible units, physically implausible values, or confirmed duplicates were removed. Equivalent records occurring in multiple sources were resolved to a single primary entry, whereas co-located observations representing distinct sampling information were retained. Statistical outliers were treated as diagnostic flags rather than automatically excluded, and disagreement with surrounding environmental conditions alone was not considered sufficient for rejection. Detailed quality-control, consistency, and duplicate-resolution procedures are provided in Sect. S2.

Following quality control, equal-area quadratic smoothing splines were fitted separately for each soil property and profile using at least two valid, non-overlapping measured horizons (Bishop et al., 1999; Arrouays et al., 2014; Ma et al., 2019). Interval means were obtained by integrating the fitted depth functions over the six GSDEv2 target intervals. A target interval was retained only when it was fully supported by the observed vertical range of the corresponding property; no extrapolation was permitted below the deepest measured horizon. Unsupported intervals were treated as missing during model training. Detailed spline-eligibility and vertical-support rules are provided in Sect. S2.6.

3.2 Soil-forming domain stratification

Valid soil-covered land cells were divided into three mutually exclusive modelling domains using the 30-arcsec HWSD-based soil-type layer: (1) Histosols, (2) Fluvisol–Arenosol–Gypsisol–dune soils, and (3) other mineral soils (FAO and IIASA, 2023). The same domain mask was used for grouping profile observations and assigning prediction cells. Detailed soil-type codes and domain definitions are provided in Table S4.

Domain stratification was intended to reduce excessive shrinkage of predictions in pedogenically distinct environments. For each property–depth–domain combination, a domain-specific model was fitted when at least 2,000 valid observations were available. This minimum sample size was adopted as a pragmatic operational threshold and was not optimized through sensitivity analysis. When fewer than 2,000 observations were available, the corresponding non-stratified property–depth model fitted to all available observations was used as a fallback.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

3.3 Global baseline modelling

The GSDEv2 global baseline was developed using Quantile Random Forests (QRF), which provide both central predictions and conditional response quantiles (Breiman, 2001; Meinshausen, 2006). The mean of individual tree predictions was used as the central estimate.

Regression matrices were constructed separately for each property, depth interval, and modelling domain. Feature selection proceeded in three stages. First, among continuous predictor pairs with $|r| > 0.8$, the variable less strongly correlated with the response was removed. Second, predictors were iteratively removed until all variance inflation factors were ≤5 (Dormann et al., 2013; Nussbaum et al., 2018). Third, RFECV with 10-fold cross-validation and MEC as the optimization criterion was used to select the final predictor subset, with a minimum of 25 predictors retained.

QRF hyperparameters were optimized within the training data using grid search and 10-fold cross-validation, maximizing mean MEC. Tuned parameters included the number of trees, candidate predictors per split, split criterion, minimum leaf size, sampling fraction, and bootstrap strategy. Surface-soil parameter settings are summarized in Table S5, with complete property–depth–domain configurations and selected predictors provided in the accompanying machine-readable model inventory.

Up to 360 property–depth–domain model combinations were defined for 20 properties, six depths, and three soil domains. Combinations with fewer than 2,000 observations used the corresponding non-stratified property–depth model as a fallback. Domain-specific predictions were then mosaicked to produce the global baseline layers.

3.4 Regional stacking fusion

Regional stacking was implemented for China, the United States, Africa, Canada, and Australia using the regional products summarized in Table 3. Fusion was restricted to property–depth combinations supported by the native vertical range and compatible definitions of each regional product; unsupported combinations retained the global baseline. Regional rasters were harmonized to the GSDEv2 units and modelling grid before fusion. Target-depth values were derived using overlap-thickness-weighted averaging for products reported over depth intervals and equal-area spline integration for products reported at discrete depth points; extrapolation beyond the native vertical support of a regional product was not permitted. Detailed regional-product harmonization, vertical-support, and eligibility rules are provided in Sect. S3 and Table S8.

For each supported property–depth combination, a second-level QRF meta-model was fitted using harmonized regional soil products, spatially out-of-fold GSDEv2 baseline predictions, and the five highest-ranked environmental covariates from the corresponding baseline model (Wolpert, 1992). Regional products therefore entered as auxiliary predictors rather than direct replacements for the global baseline. Potential overlap between source observations underlying the regional products and the GSDEv2 profile database was assessed from available provenance information and is summarized in Table S8. After validation, the regional meta-model for each supported property–depth combination was refitted using all eligible

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

observations. Spatially out-of-fold baseline predictions were retained as the baseline predictor during meta-model fitting, whereas predictions from the fully refitted global baseline model were used during gridded regional prediction.

To reduce abrupt transitions between regional stacking predictions and the surrounding global baseline, an outward distance-weighted buffer was applied during final map assembly. Regional stacking predictions were retained unchanged within each supported fusion region. Outside the regional boundary, their weight decreased linearly from 1 to 0 over empirically selected transition widths of 10 km for China, 20 km for the United States, 50 km for Africa, 20 km for Canada, and 50 km for Australia, based on visual inspection of boundary transitions in preliminary maps.

Let d_i denote the outward distance of grid cell i from the boundary of fusion region r, and let D_r denote the corresponding transition width. The regional stacking weight was defined as

$$w_i = \begin{cases} 1, & i \in r, \\ 1 - \frac{d_i}{D_r}, & 0 < d_i < D_r, \\ 0, & d_i \geq D_r. \end{cases} \quad (1)$$

The final prediction was then calculated as

$$\hat{y}_{i,final} = w_i \hat{y}_{i,stack} + (1 - w_i)\hat{y}_{i,global} \quad (2)$$

Thus, stacking predictions were retained within each fusion region, blended linearly with the global baseline within the outward transition buffer, and replaced entirely by the global baseline beyond the buffer.

3.5 Global validation, product benchmarking, and evaluation of uncertainty estimates

The predictive performance of the global baseline models was evaluated using profile-level random hold-out validation. Following quality control and depth harmonization, the available profiles were randomly divided into a 90 % training subset and a 10 % held-out test subset using a fixed random seed of 42. To avoid within-profile and duplicate-site leakage, all observations associated with the same deduplicated profile or observation-site group were assigned to the same partition.

Held-out test profiles were excluded from all response-dependent model-development procedures. The fitted QRF models were then applied to the corresponding covariate stacks to generate out-of-sample predictions. This random hold-out design evaluates interpolation performance within the sampled observational domain. Because no global spatial-block or leave-region-out validation was conducted, the resulting metrics should not be interpreted as measures of geographical transferability to sparsely sampled or environmentally distinct regions.

GSDEv2 was benchmarked against GSDEv1, SoilGrids 2.0, and HWSD 2.0 using the same held-out test profiles. For each property–depth combination, performance metrics were calculated using the same set of observations for which predictions were available from all evaluated products. Because the benchmark products may share some underlying source observations with GSDEv2 and differ in native spatial support, these comparisons were interpreted as relative benchmarking rather than fully independent external validation.

Regional-fusion performance was evaluated using nested spatial-block validation with a 60 km block size. Profile coordinates were projected to EPSG:6933, and populated blocks were assigned to five outer folds using a greedy balancing

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



procedure with a fixed random seed of 42. Within each outer training subset, five-fold internal spatial-block partitioning was used to generate out-of-fold GSDEv2 baseline predictions for meta-model training. Entire outer test blocks were excluded from fitting both the baseline and stacking models. The 60 km block size represented a pragmatic compromise between spatial separation and retaining sufficient observations per fold rather than an optimized universal separation distance (Roberts et al., 2017; Valavi et al., 2019).

This nested design avoided leakage from baseline predictions fitted to the same response observations but could not remove information already embedded in regional products developed using overlapping source observations. Regional-fusion performance was therefore interpreted as a spatially blocked comparative assessment rather than fully source-independent external validation.

For both global baseline evaluation and regional-fusion assessment, predictive performance was summarized using the modelling efficiency coefficient (MEC), root-mean-square error (RMSE), and mean error (ME):

$$MEC = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad , \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}, \tag{4}$$

and

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i), \tag{5}$$

where y_i and $\hat{y}_i$ are the observed and predicted values for held-out observation i, respectively; $\bar{y}$ is the mean of the held-out observations; and n is the number of evaluated observations. Higher MEC and lower RMSE indicate better overall predictive performance. Positive ME indicates overall overprediction, whereas negative ME indicates underprediction.

For the regional-fusion assessment, the relative RMSE change for each spatial block was calculated as:

$$RI_{RMSE} = 100\frac{RMSE_{Base} - RMSE_{Stacking}}{RMSE_{Base}} \tag{6}$$

Positive values indicate reduced prediction error following regional stacking fusion. The point-level absolute error change was calculated as:

$$\Delta AE_i = \left|\hat{y}_{Stacking,i} - y_i\right| - \left|\hat{y}_{Base,i} - y_i\right|. \tag{7}$$

Negative values of ΔAE_i indicate reduced absolute error after regional fusion.

Model-based prediction uncertainty was quantified using the conditional response distributions generated by the QRF models. For global baseline models, prediction-interval calibration and sharpness were evaluated using the 10 % random hold-out test set. For the regional-fusion case study, the corresponding diagnostics were calculated from the outer-fold spatially out-of-fold predictions generated by the nested spatial-block validation. Specifically, the 5th ($q_{0.05}$), 50th ($q_{0.50}$),

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



and 95th ($q_{0.95}$) conditional quantiles were extracted. The $q_{0.05}$ and $q_{0.95}$ quantiles defined the lower and upper bounds of the nominal 90 % conditional prediction interval, whereas $q_{0.50}$ represented the conditional median. For each validation framework, prediction-interval performance was evaluated using the corresponding held-out predictions, following the principle that informative probabilistic predictions require joint optimization of statistical calibration and predictive sharpness (Gneiting et al., 2007; Schmidinger and Heuvelink, 2023). For a nominal coverage level c_k, the prediction-interval coverage probability was calculated as:

$$PICP_k = \frac{1}{n}\sum_{i=1}^{n} I[q_{(1-c_k)/2,i} \le y_i \le q_{(1+c_k)/2,i}] \quad (8)$$

, where $q_{(1-c_k)/2,i}$ and $q_{(1+c_k)/2,i}$ are the lower and upper QRF conditional quantiles for observation i, respectively, and $I(\cdot)$ is the indicator function. Overall calibration across the evaluated nominal coverage levels was summarized using the mean absolute calibration error (MACE):

$$MACE = \frac{1}{K}\sum_{k=1}^{K}|PICP_k - c_k| \quad (9)$$

, where K is the number of evaluated nominal coverage levels. Lower MACE indicates closer agreement between empirical and nominal coverage. For the nominal 90 % prediction interval, sharpness was additionally summarized using the mean interval width:

$$MIW_{90} = \frac{1}{n}\sum_{i=1}^{n}(q_{0.95,i} - q_{0.05,i}) \quad (10)$$

Calibration and sharpness were interpreted jointly, as narrower intervals are informative only when empirical coverage remains close to the nominal level. All probabilistic diagnostics were computed from 90 m native-resolution conditional quantiles rather than the aggregated 1 km quantile products.

The global random hold-out and regional spatial-block assessments therefore address complementary questions: interpolation performance within the sampled global domain and incremental fusion performance under spatial separation, respectively.

3.6 Global prediction, post-processing and data organization

Following validation, the property–depth–domain QRF models were refitted using all available quality-controlled observations. Final predictions were generated for global land areas between 60° S and 90° N on the common 3-arcsec geographic grid; the nominal 90 m designation therefore does not imply uniform ground resolution.

Domain-specific baseline predictions were first mosaicked into global layers, after which regional stacking predictions were incorporated for supported property–depth combinations using the boundary-transition procedure described in Sect. 3.4. The mean QRF prediction was retained as the central estimate, while $q_{0.05}$, $q_{0.50}$, and $q_{0.95}$ surfaces were generated for uncertainty processing. Final 90 m layers underwent land-mask application, property-specific physical-range screening, cross-property consistency checks, and alignment to common grid and tile boundaries. A 1 km cell was retained when at least 50 % of its underlying 90 m terrestrial cells contained valid predictions. For supported fusion regions and their outward

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



transition buffers, baseline and stacking conditional-quantile surfaces were combined using the same transition weights as the corresponding mean predictions. For $\alpha \in \{0.05, 0.50, 0.95\}$, the blended surfaces were calculated as:

$$\tilde{q}_a(s) = w_r(s) q_{a,stack\,(s)} + [1 - w_r(s)] q_{a,base\,(s)}, \quad (11)$$

where $w_r(s)$ is the region-specific transition weight defined in Sect. 3.4.

The 1 km conditional quantile products were derived from the corresponding 90 m conditional-quantile surfaces using the same property-specific valid-area-weighted spatial aggregation procedures applied to the mean-prediction layers. The resulting 1 km quantile layers represent spatially aggregated native-grid conditional-quantile surfaces and should not be interpreted as formally calibrated predictive quantiles of the true 1 km block-mean soil property. For a 1 km grid cell J, the aggregated quantile value was computed as:

$$q_{a,J}^{1km} = A_p\left(\{\tilde{q}_{a,i}^{90m}\}_{i \in J}\right), a \in \{0.05, 0.50, 0.95\}, \quad (12)$$

where A_p denotes the aggregation operator used for soil property p. The aggregated quantile layers were checked to ensure that $q_{0.05} \le q_{0.50} \le q_{0.95}$ and that both bounds remained within the property-specific physically plausible ranges.

Distributed rasters were stored using the property-specific integer scale factors listed in Table 1; product organization is described in Sect. 4.6.

3.7 CoLM soil-input sensitivity experiment

To illustrate how differences among global soil datasets propagate in downstream land-surface simulations, three controlled sensitivity experiments with different input global soil property datasets (i.e., GSDEv1, GSDEv2 and SoilGrids 2.0) were conducted using the Common Land Model version CoLM2024 (CoLM2024; Dai et al., 2003; https://github.com/CoLM-SYSU/CoLM). Global land surface simulations were performed at 0.25° resolution from 2013 to 2018 using a time step of 1800 s. The 2013–2017 period was used for model spin-up, and the 2018 outputs were used for evaluation. All three experiments used the same ERA5 atmospheric forcing, initial soil state, and other model settings (Table S9). The only manipulated variable was the soil-input dataset.

The analysis focused on the volumetric soil water content of the first CoLM soil layer. Daily values for 2018 were used to calculate annual-mean surface soil moisture. The evaluation domain was restricted to land areas between 60°S and 90°N. Surface soil moisture was selected because of its high sensitivity to soil properties, and it is a robust indicator for diagnosing soil-input-induced discrepancies in land-surface model simulations (Huang et al., 2022; Xu et al., 2023; Zhang et al., 2023). Adjusted ERA5-Land soil moisture was used as a pseudo-reference for model evaluation. This product combines the spatially continuous coverage of ERA5-Land (Muñoz-Sabater et al., 2021) with a mean–variance adjustment informed by SMAP L4 soil moisture, thereby reducing systematic biases in the original ERA5-Land fields while retaining daily global coverage (Wang et al., 2026). Physically implausible values outside 0–0.70 $m^3\ m^{-3}$ were excluded. Agreement in the annual-mean spatial patterns was quantified using the area-weighted RMSE, ME, and spatial correlation coefficient, with grid-cell

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

weights proportional to the cosine of latitude. Latitude-zonal means were also calculated to characterize latitudinal distribution discrepancies.

4. Results

4.1 Profile coverage and vertical variation

The 234,785 soil profiles retained after harmonization and quality control covered broad climatic, topographic, vegetation, parent-material, and land-use gradients, but their spatial distribution remained heterogeneous (Fig. 2). Observation densities were generally higher in regions with long-established soil-survey programmes and lower in deserts, high mountains, tropical forests, wetlands, and high-latitude environments. Sample availability also varied among properties and target depths, ranging from 1,084 to 135,170 observations (Table S6).

The harmonized observations showed systematic vertical variation (Fig. 3; Table S6). Mean SOC decreased from 49.60 g kg^{-1} at 0–5 cm to 8.50 g kg^{-1} at 100–200 cm, whereas mean bulk density increased from 1.23 to 1.54 kg dm^{-3}. Mean clay content increased from 20.86 % at 0–5 cm to 28.80 % at 60–100 cm and declined slightly to 27.57 % at 100–200 cm, while mean pH increased from 6.18 to 6.59.

Depth effects were significant for all four properties evaluated on complete six-depth profiles (Friedman tests, $p < 0.001$; Fig. 3). Kendall's W was 0.742 for SOC, 0.245 for bulk density, 0.083 for pH, and 0.132 for clay content. These analyses characterize vertical patterns in the harmonized database rather than independently validating the depth-harmonization procedure.

4.2 Predictive accuracy

The GSDEv2 global baseline models were evaluated before regional stacking and final map assembly (Table 4; Table S7). Across the 120 property–depth combinations, MEC ranged from 0.29 to 0.71. Lower values occurred for volumetric coarse fragments (0.29–0.37), Bray-I phosphorus (0.30–0.34), and gravimetric coarse fragments (0.31–0.38), whereas higher values were obtained for pH (0.64–0.67), silt (0.54–0.71), SOC (0.56–0.64), and bulk density (0.55–0.62).

Among the 54 property–depth combinations jointly supported by GSDEv2, GSDEv1, SoilGrids 2.0, and HWSD 2.0, GSDEv2 had the highest mean MEC and the lowest mean RMSE. Median RMSE reductions relative to GSDEv1, SoilGrids 2.0, and HWSD 2.0 were 22.2 %, 5.8 %, and 24.6 %, respectively. Relative to SoilGrids 2.0, RMSE reductions ranged from 0.4 % to 21.9 %.

Predictive performance generally decreased with depth for several properties; for example, MEC declined from 0.58 to 0.45 for sand, from 0.64 to 0.56 for SOC, and from 0.54 to 0.43 for total nitrogen between 0–5 and 100–200 cm. In contrast, pH showed little depth-related change (0.64–0.67). GSDEv2 had the smallest absolute ME in 31 of the 54 common comparisons.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

4.3 Calibration of conditional prediction intervals for representative properties

Conditional prediction intervals for SOC and bulk density generally showed empirical coverage close to the corresponding nominal levels across the six target depths (Fig. 4). For SOC, MACE ranged from 0.003 to 0.023, with a depth-averaged value of 0.011; for bulk density, MACE ranged from 0.008 to 0.016, with a depth-averaged value of 0.013. Calibration diagnostics for the remaining properties are provided in the Supplementary Material.

The mean width of the nominal 90 % interval ranged from approximately 33 to 108 g kg^{-1} for SOC, with a depth-averaged value of 62.39 g kg^{-1}, and from approximately 0.65 to 0.74 kg dm^{-3} for bulk density, with a depth-averaged value of 0.675 kg dm^{-3} (Fig. 4m, n). Thus, interval calibration was generally close to nominal levels for the two representative properties, while sharpness varied with property and depth.

4.4 Global spatial patterns and comparison with existing products

The final GSDEv2 maps reproduced broad soil–environment patterns expected from major climatic, geological, and pedological gradients (Fig. 5). For example, high sand contents occurred across major arid regions, lower bulk density was associated with organic-rich and humid soils, and higher pH values occurred predominantly in arid and semi-arid environments. Complete property-specific maps are provided in Figs. S1–S40.

Spatial texture was compared qualitatively using surface SOC (0–5 cm) from GSDEv2, GSDEv1, SoilGrids 2.0, and HWSD 2.0 across three contrasting regions: the Colorado Rockies–Great Plains transition, Southwest Sichuan, and the Great Dividing Range–Murray Basin system (Fig. 6). GSDEv2 showed comparatively continuous regional gradients and finer within-region spatial variation, SoilGrids 2.0 showed broadly similar patterns with locally smoother transitions, whereas GSDEv1 and HWSD 2.0 retained more polygon-like spatial structures associated with mapping units.

These visual comparisons illustrate differences in spatial representation among products but do not constitute independent validation of map accuracy, because terrain and vegetation covariates contributing to the GSDEv2 patterns were also used as model predictors.

4.5 Evaluation of regional fusion for surface SOC in China

Regional stacking substantially improved spatially blocked SOC prediction at 0–5 cm in China (Fig. 7). Across 10,047 spatially out-of-fold observations, MEC increased from 0.305 for the nested global baseline to 0.676 for the stacking model, while RMSE decreased from 42.76 to 29.18 g kg^{-1} (31.8 %) and the absolute ME decreased from 2.35 to 0.50 g kg^{-1}.

The observed–predicted relationship moved closer to the 1:1 line after stacking (Fig. 7a, b), particularly at low and intermediate SOC concentrations, although underprediction remained at the upper end of the observed distribution. Among 393 valid 60 km blocks, RMSE decreased in 332 blocks (84.5 %) and increased in 61 (15.5 %) (Fig. 7c, d). The median block-level RMSE reduction was 28.9 % (95 % bootstrap CI: 26.2 %–31.3 %), and paired block-level improvements were significantly greater than zero (one-sided Wilcoxon signed-rank test, $p < 0.001$).

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Empirical coverage of the nominal 90 % interval remained approximately 0.90 for both baseline and stacking predictions, while the mean interval width decreased from 90.6 to 55.5 g kg^{-1} (38.7 %) (Fig. 7e, f). Interval-width changes were spatially heterogeneous, with both narrowing and widening among individual observations.

This evaluation quantifies predictive performance at profile locations but does not directly assess continuity across fusion-region boundaries. In addition, unresolved source overlap between regional products and the GSDEv2 profile database means that the results represent a spatially blocked comparative assessment rather than fully source-independent external validation.

4.6 Dataset records and organization

The final GSDEv2 release contains 120 property–depth combinations. Mean predictions are distributed at nominal 90 m resolution, whereas the 1 km release includes the spatially aggregated mean, $q_{0.05}$, $q_{0.50}$, and $q_{0.95}$ layers. Prediction-interval width can be calculated as $q_{0.95} - q_{0.05}$. The 90 m products are organized by property, depth interval, regional subdivision, and 5° × 5° tile, whereas the 1 km products are provided as global rasters. File naming, units, scale factors, NoData values, coordinate reference system, aggregation rules, and provenance are documented in the accompanying data documentation.

4.7 Sensitivity of simulated surface soil moisture to soil inputs

All three CoLM experiments reproduced similar large-scale patterns of annual-mean surface soil moisture, but regional differences emerged among the soil-input datasets (Fig. 8a–e). Relative to CoLM-GSDEv1, CoLM-GSDEv2 generally produced wetter conditions over parts of northern North America and Eurasia, whereas CoLM-SoilGrids 2.0 produced substantially wetter conditions than CoLM-GSDEv2 across extensive northern high-latitude regions. The largest differences occurred mainly at middle and high northern latitudes.

CoLM-GSDEv2 showed the closest overall spatial agreement with adjusted ERA5-Land among the three experiments (Fig. 8f). Area-weighted RMSE was 0.089 m^3 m^{-3} for CoLM-GSDEv2, compared with 0.092 m^3 m^{-3} for CoLM-GSDEv1 and 0.109 m^3 m^{-3} for CoLM-SoilGrids 2.0. Spatial correlations were 0.802, 0.786, and 0.741, respectively. CoLM-GSDEv2 also had a smaller magnitude of negative ME than CoLM-GSDEv1, whereas CoLM-SoilGrids 2.0 had a ME closer to zero.

These results show that alternative soil inputs affected simulated surface-soil-moisture patterns under the evaluated configuration. Because the experiment covered one year at 0.25° resolution and only the first model soil layer, it evaluates aggregate soil-input sensitivity rather than the contribution of nominal 90 m spatial detail.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



5. Discussion

5.1 Advances over GSDEv1 and complementarity with contemporary global products

GSDEv2 represents a methodological transition from the soil taxonomic-linkage and map-unit framework of GSDEv1 toward observation-driven digital soil mapping. Whereas GSDEv1 relied primarily on soil classes, taxotransfer relationships, and discrete mapping units (Shangguan et al., 2014), GSDEv2 combines harmonized soil-profile observations, environmental covariates, soil-domain-specific modelling, conditional uncertainty estimation, and stacking-based incorporation of regional soil products. GSDEv2 nevertheless retains some dependence on legacy soil geographic information, because HWSD-derived soil classes contribute to the definition of modelling domains and to the predictor set (FAO and IIASA, 2023).

Benchmarking indicated improved average interpolative performance relative to GSDEv1, SoilGrids 2.0, and HWSD 2.0, although the improvement over SoilGrids 2.0 was comparatively modest. This is consistent with the fact that both GSDEv2 and SoilGrids 2.0 use harmonized soil observations, environmental covariates, and machine-learning approaches, whereas GSDEv1 and HWSD 2.0 rely more strongly on map-unit-based representations (Shangguan et al., 2014; Poggio et al., 2021; FAO and IIASA, 2023). However, the observed performance differences reflect combined effects of training data, covariates, vertical harmonization, spatial support, domain stratification, and model structure and therefore cannot be attributed to any single methodological component.

These results should be interpreted as evidence of improved interpolation within the sampled observational domain rather than universal superiority. Random hold-out validation does not directly quantify geographical transferability, particularly where test observations remain environmentally or spatially similar to training data (Roberts et al., 2017; Meyer and Pebesma, 2021). GSDEv2 is therefore best viewed as a complementary global soil resource within an increasingly diverse product ecosystem that includes SoilGrids 2.0 and OpenLandMap-soildb (Poggio et al., 2021; Hengl et al., 2026). Its principal distinction lies in the combination of broad property coverage, standardized depths to 2 m, conditional uncertainty information, and stacking-based use of regional gridded soil products within a globally consistent framework.

5.2 Contribution of quality control and domain stratification

Pedologically informed quality control is important because observational errors and inconsistencies can be propagated by machine-learning models into spatial predictions. The GSDEv2 workflow therefore combined structural, metadata, physical, and cross-property checks while retaining statistically unusual observations when they remained physically plausible. This approach was intended to reduce identifiable data errors without systematically removing valid extremes in global soil-property distributions (Minasny et al., 2024). Nevertheless, residual uncertainty remains because analytical methods, coordinate precision, sampling periods, and metadata quality vary among source datasets.

Domain stratification was introduced to reduce excessive shrinkage of predictions in pedogenically distinct environments, particularly organic and strongly contrasting mineral-soil settings. Such stratification may help preserve

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.


property distributions that could otherwise be diluted within a single global model, but its effectiveness depends on sample size, geographical coverage, and environmental representativeness within each domain. Misclassification of domain membership and discrete domain boundaries may also introduce prediction discontinuities or inappropriate model assignment. Moreover, the present study did not conduct a formal stratified-versus-unstratified ablation analysis; the independent contribution of domain stratification to predictive improvement therefore remains unresolved.

5.3 Benefits and limitations of regional fusion

Regional stacking provides a way to incorporate national and continental gridded soil products as auxiliary information without directly replacing the globally consistent baseline. In contrast to simple mosaicking, the meta-model learns relationships among regional products, the GSDEv2 baseline, environmental covariates, and in-situ observations, allowing regional information to contribute according to its predictive association with the target variable.

The China surface-SOC case study demonstrated substantial improvements under nested spatial-block validation, including lower RMSE and narrower prediction intervals while maintaining approximately nominal coverage. The improvement occurred across most evaluated spatial blocks, although localized error increases remained, indicating that fusion performance was spatially heterogeneous.

The generalizability of this result remains limited by regional-product coverage, vertical support, training provenance, and inherited uncertainties. Nested spatial-block validation avoids leakage from baseline predictions fitted to the same response observations but cannot remove information already embedded in external regional products that may share source observations with the GSDEv2 database. Differences in native spatial resolution, depth definitions, and data provenance introduce additional harmonization uncertainty. The China SOC analysis should therefore be interpreted as a validated regional case rather than evidence that stacking necessarily improves every property, depth, or fusion region.

The outward boundary-transition procedure was designed to reduce abrupt changes between regional stacking predictions and the surrounding global baseline while preserving stacking predictions within the supported fusion regions. Its effectiveness in improving spatial continuity, however, was not quantitatively evaluated and should therefore be regarded as a map-assembly procedure rather than an independently validated performance improvement.

5.4 Remaining data limitations

GSDEv2 integrates soil observations collected over different periods with multi-temporal environmental covariates and should therefore be interpreted as a long-term soil baseline rather than a time-specific representation. It is not designed to detect recent soil change without additional time-resolved observations.

Vertical data support also decreases with depth because target intervals were retained only when fully supported by observed horizons. Consequently, deeper predictions are based on fewer observations and are more sensitive to their geographical and environmental representativeness. GSDEv2 additionally does not explicitly represent spatially varying soil

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.


thickness or bedrock depth; predictions for deep intervals should therefore be interpreted cautiously in shallow-soil and shallow-bedrock environments.

5.5 Conditional model uncertainty and fitness for use

The distributed conditional-quantile layers characterize uncertainty associated with the fitted QRF relationships, available observations, predictors, and model structure. They do not fully propagate uncertainty arising from measurement and analytical errors, coordinates, depth harmonization, covariates, spatial support, spatial aggregation, or externally derived regional products (Schmidinger and Heuvelink, 2023; Shi et al., 2024). Conditional prediction intervals may therefore underestimate total uncertainty in sparsely sampled or environmentally novel regions (Meyer and Pebesma, 2021).

GSDEv2 predictions should consequently be interpreted together with the available uncertainty information, observation support, regional-fusion provenance, and environmental applicability of the training data. The nominal 90 m grid provides a common prediction support but should not be interpreted as uniform effective spatial resolution or site-scale accuracy (Hengl, 2006). The 1 km quantile layers are spatial aggregates of native-grid conditional quantiles rather than formally calibrated predictive distributions for 1 km block means. Local evaluation and downstream uncertainty propagation remain advisable for applications sensitive to absolute soil-property values.

5.6 Implications for Earth system modelling

A key motivation for GSDEv2 is to provide globally consistent soil information that can be remapped to the diverse horizontal and vertical discretizations used by land-surface and Earth-system models. The nominal 90 m product provides spatially detailed information for regional analyses and aggregation to application-specific model grids, whereas the spatially aggregated 1 km product offers a more computationally and storage-efficient option for continental and global simulations.

The physical and chemical properties included in GSDEv2 support a range of land-surface and Earth-system model applications. Soil texture, bulk density, coarse fragments, and water-retention properties provide information commonly used to derive hydraulic and thermal parameters controlling infiltration, soil-water storage, hydraulic conductivity, runoff generation, heat transfer, and land–atmosphere exchanges (Dai et al., 2019a, b; Fatichi et al., 2020; Montzka et al., 2017; Turek et al., 2023). Carbon- and nutrient-related variables, including SOC, TC, TN, pH, CEC, CECPH7, and Bray-I extractable phosphorus, can further support biogeochemical and nutrient-cycle applications. However, the water-retention values provided at 10, 33, and 1500 kPa do not constitute a complete hydraulic function, and additional pedotransfer relationships or calibration may be required to derive model-specific hydraulic parameters.

GSDEv2 variables may therefore be used directly where model formulations are compatible or transformed into model-specific parameters through appropriate pedotransfer functions and parameterization schemes. Care is required because land-surface and Earth-system models differ in soil-layer structures, units, hydraulic formulations, and treatments of coarse fragments and organic soils. Any vertical remapping, unit conversion, pedotransfer procedure, or spatial aggregation applied before model implementation should therefore be documented explicitly. The CoLM2024 sensitivity experiment presented in

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.


Sect. 4.7 further illustrates that differences among global soil-input datasets can propagate into measurable changes in simulated surface soil moisture, although the experiment does not isolate the contribution of individual soil properties or establish general performance across land-surface processes.

The fine nominal grid spacing of GSDEv2 also provides opportunities to investigate spatial soil heterogeneity in regional hydrological, ecological, and land–atmosphere applications. Nevertheless, as emphasized in Sect. 5.5, the nominal 90 m grid should not be interpreted as uniform site-scale accuracy, and uncertainty in the soil predictions can propagate into downstream simulations. GSDEv2 should also be regarded as a static long-term soil baseline rather than a monitoring product and therefore should not be used to infer recent decadal soil changes without independent time-series observations. Application-specific sensitivity analysis, local validation against in-situ measurements, and downstream uncertainty propagation remain important for robust interpretation of model simulations using GSDEv2.

6. Code and data availability

A Python implementation of the core GSDEv2 processing workflow, including soil-profile quality control and depth harmonization, feature preparation, QRF model training, global prediction, regional stacking fusion, and validation, is archived at https://doi.org/10.5281/zenodo.22859059 (Shi, 2026).

The GSDEv2 dataset is freely available from the National Tibetan Plateau Data Center (TPDC; Shi et al., 2026; https://doi.org/10.11888/Terre.tpdc.303613), with an additional record available from Science Data Bank (https://www.scidb.cn/s/my2aea).

7. Conclusions

This study presents GSDEv2, a global soil dataset providing spatially continuous predictions of 20 physical and chemical soil properties across six standardized depth intervals to 2 m. Mean predictions are provided on a nominal 90 m grid, with spatially aggregated mean and conditional-quantile products additionally available at 1 km. GSDEv2 combines observation-driven global baseline modelling with stacking-based incorporation of compatible regional soil products, while maintaining a globally consistent prediction framework.

Random hold-out validation showed improved average interpolative performance relative to GSDEv1, SoilGrids 2.0, and HWSD 2.0 across the 54 commonly supported property–depth combinations. In the spatially blocked China surface-SOC case study, regional stacking reduced RMSE by 31.8 % and the mean width of the nominal 90 % prediction interval by 38.7 % while maintaining approximately nominal coverage. The CoLM2024 sensitivity experiment further showed that alternative global soil inputs can propagate into measurable differences in simulated surface soil moisture, with the GSDEv2-driven experiment showing lower spatial RMSE and higher spatial correlation with adjusted ERA5-Land than the SoilGrids-driven experiment under the evaluated configuration.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.


These results should be interpreted within the scope of the adopted validation and data support. The global random hold-out assessment characterizes interpolation within the sampled observational domain rather than geographical transferability, and the regional-fusion evaluation represents a validated China SOC case rather than evidence of universal improvement across all properties and regions. The nominal 90 m grid should not be interpreted as uniform site-scale accuracy, and the 1 km conditional-quantile layers are spatial aggregates of native-grid conditional quantiles rather than formally calibrated predictive distributions for true 1 km block means. GSDEv2 should therefore be used together with information on observation support, conditional uncertainty, regional-fusion provenance, and application-specific validation. Within these limitations, GSDEv2 provides a model-oriented global soil resource for regional-to-global land-surface and Earth-system applications.

8. Author contributions

WS and YD conceptualized the study and provided overall scientific guidance. GS developed the methodological and computational workflow, compiled and harmonized the data, implemented the modelling and data-fusion framework, developed the code, performed the experiments and analyses, produced the visualizations, and wrote the original draft. AG contributed to the development of the regional stacking concept, supervised the work during the research stay at KU Leuven, contributed to scientific discussion and interpretation of the results, and reviewed and edited the manuscript. ZW contributed to scientific discussion, interpretation of the results, and review and editing of the manuscript. HY contributed to scientific discussion and review of the manuscript. QL reviewed the manuscript. WS supervised the study, contributed to methodological development and interpretation of the results, reviewed and edited the manuscript, and acquired funding. YD supervised the study, contributed to methodological development and scientific discussion, provided the CoLM model and computational resources for the land-surface modelling experiment, and reviewed and edited the manuscript. All authors reviewed and approved the final manuscript.

9. Competing interests

The authors declare that they have no conflict of interest.

10. Acknowledgements

The authors gratefully acknowledge the organizations, soil survey programmes, and data providers that made the soil-profile observations and gridded soil products used in this study available. In particular, we thank the developers and maintainers of the World Soil Information Service (WoSIS) and the providers of the global, national, and regional soil-information products used for model development, regional fusion, and benchmarking. We also acknowledge the computing facilities and

© Author(s) 2026. CC BY 4.0 License.



institutional support that enabled the large-scale data processing, model training, and global prediction performed in this study.

11. Financial support

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM902), National Natural Science Foundation of China (42375144, 4227515), Guangdong Basic and Applied Basic Research Foundation (2021B0301030007) and China Meteorological Administration Youth Innovation Team (CMA2024QN01).

References

Arrouays, D., Grundy, M. G., Hartemink, A. E., Hempel, J. W., Heuvelink, G. B. M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A. B., McKenzie, N. J., Mendonca-Santos, M. d.L., Minasny, B., Montanarella, L., Odeh, I. O. A., Sanchez, P. A., Thompson, J. A., and Zhang, G.-L.: GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties, in: Advances in Agronomy, vol. 125, Elsevier, 93–134, https://doi.org/10.1016/B978-0-12-800137-0.00003-0, 2014.

Batjes, N. H., Calisto, L., and De Sousa, L. M.: Providing quality-assessed and standardised soil data to support global mapping and modelling (WoSIS snapshot 2023), Earth Syst. Sci. Data, 16, 4735–4765, https://doi.org/10.5194/essd-16-4735-2024, 2024.

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

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

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, Sci Data, 12, 1480, https://doi.org/10.1038/s41597-025-05863-3, 2025.

Dai, Y., Zeng, X., Dickinson, R. E., Baker, I., Bonan, G. B., Bosilovich, M. G., Denning, A. S., Dirmeyer, P. A., Houser, P. R., Niu, G., Oleson, K. W., Schlosser, C. A., and Yang, Z.-L.: The Common Land Model, Bulletin of the American Meteorological Society, 84, 1013–1024, https://doi.org/10.1175/BAMS-84-8-1013, 2003.

Dai, Y., Shangguan, W., Wei, N., Xin, Q., Yuan, H., Zhang, S., Liu, S., Lu, X., Wang, D., and Yan, F.: A review of the global soil property maps for Earth system models, SOIL, 5, 137–158, https://doi.org/10.5194/soil-5-137-2019, 2019a.

Dai, Y., Xin, Q., Wei, N., Zhang, Y., Shangguan, W., Yuan, H., Zhang, S., Liu, S., and Lu, X.: A Global High-Resolution Data Set of Soil Hydraulic and Thermal Properties for Land Surface Modeling, J Adv Model Earth Syst, 11, 2996–3023, https://doi.org/10.1029/2019MS001784, 2019b.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36, 27–46, https://doi.org/10.1111/j.1600-0587.2012.07348.x, 2013.

Du, Y., Minasny, B., Ng, W., Zhang, L., Saby, N.P., Zhou, Y., van Wesemael, B., Ma, Y., Wang, Z., Luo, Z. and Shi, Z. Spatiotemporal modelling of soil organic carbon: integrating process-based and machine learning approaches. Geoderma, 473, p.117967. https://doi.org/10.1016/j.geoderma.2026.117967, 2026.

Ebeling, A. M., Bundy, L. G., Kittell, A. W., and Ebeling, D. D.: Evaluating the Bray P1 Test on Alkaline, Calcareous Soils, Soil Sci. Soc. Am. J., 72, 985–991, https://doi.org/10.2136/sssaj2006.0347, 2008.

Evangelista, S.J., Field, D.J., McBratney, A.B., Minasny, B., Ng, W., Padarian, J., Dobarco, M.R. and Wadoux, A.M.C. Soil security—Strategizing a sustainable future for soil. Advances in Agronomy, 183, p.1-70. https://doi.org/10.1016/bs.agron.2023.10.001, 2024.

FAO and IIASA: Harmonized World Soil Database version 2.0, FAO, Rome, Italy, and IIASA, Laxenburg, Austria, https://doi.org/10.4060/cc3823en, 2023.

Fatichi, S., Or, D., Walko, R., Vereecken, H., Young, M. H., Ghezzehei, T. A., Hengl, T., Kollet, S., Agam, N., and Avissar, R.: Soil structure is an important omission in Earth System Models, Nat Commun, 11, 522, https://doi.org/10.1038/s41467-020-14411-z, 2020.

Francos, N., Sharififar, A., Flynn, T., Styc, Q., Evangelista, S.J., Padarian, J., Ng, W., Field, D.J., Minasny, B. and McBratney, A.B. The global pedogenon map: Combining and spatialising the factors of soil formation. Geoderma, 460, p.117462. https://doi.org/10.1016/j.geoderma.2025.117462, 2025.

Geng, X., He, J., Grima, V., Jiang, Y., Tetreau, M., Crittenden, S., Kiley, S., VandenBygaart, A. J., and Vanrobaeys, J.: 100 m soil landscape grids of Canada, Sci Data, 12, 1178, https://doi.org/10.1038/s41597-025-05460-4, 2025.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness, Journal of the Royal Statistical Society Series B: Statistical Methodology, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Hengl, T.: Finding the right pixel size, Computers & Geosciences, 32, 1283–1298, https://doi.org/10.1016/j.cageo.2005.11.008, 2006.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., Yemefack, M., Wendt, J., MacMillan, R. A., Wheeler, I., and Crouch, J.: African soil

© Author(s) 2026. CC BY 4.0 License.



properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning, Sci Rep, 11, 6130, https://doi.org/10.1038/s41598-021-85639-y, 2021.

Hengl, T., Consoli, D., Tian, X., Nauman, T. W., Nussbaum, M., Isik, M. S., Parente, L., Ho, Y.-F., Simoes, R., Gupta, S., Samuel-Rosa, A., Zborowski Horst, T., Safanelli, J. L., and Harris, N.: OpenLandMap-soildb: global soil information at 30 m spatial resolution for 2000–2022+ based on spatiotemporal Machine Learning and harmonized legacy soil samples and observations, Earth Syst. Sci. Data, 18, 989–1036, https://doi.org/10.5194/essd-18-989-2026, 2026.

Huang, S., Eisner, S., Haddeland, I., and Tadege Mengistu, Z.: Evaluation of two new-generation global soil databases for macro-scale hydrological modelling in Norway, Journal of Hydrology, 610, 127895, https://doi.org/10.1016/j.jhydrol.2022.127895, 2022.

Lagacherie, P.: Operational Digital Soil Mapping: Achievements, Challenges and Future Strategies to Go Beyond, European J Soil Science, 76, e70139, https://doi.org/10.1111/ejss.70139, 2025.

Liu, F., Zhang, G.-L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., and Yang, F.: High-resolution and three-dimensional mapping of soil texture of China, Geoderma, 361, 114061, https://doi.org/10.1016/j.geoderma.2019.114061, 2020.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., and Zhang, G.-L.: Mapping high resolution National Soil Information Grids of China, Science Bulletin, 67, 328–340, https://doi.org/10.1016/j.scib.2021.10.013, 2022.

Ma, Y., Minasny, B., Malone, B.P. and McBratney, A.B. Pedology and digital soil mapping (DSM). European Journal of Soil Science, 70(2), p.216-235. https://doi.org/10.1111/ejss.12790, 2019.

Malone, B. P., Searle, R., Stenson, M., McJannet, D., Zund, P., Román Dobarco, M., Wadoux, A. M. J.-C., Minasny, B., McBratney, A., and Grundy, M.: Update and expansion of the soil and landscape grid of Australia, Geoderma, 455, 117226, https://doi.org/10.1016/j.geoderma.2025.117226, 2025.

McBratney, A., Mendonça Santos, M., and Minasny, B.: On Digital Soil Mapping, Geoderma, 117, 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, 2006.

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

Minasny, B., Bandai, T., Ghezzehei, T.A., Huang, Y.C., Ma, Y., McBratney, A.B., Ng, W., Norouzi, S., Padarian, J., Sharififar, A. and Styc, Q. Soil science-informed machine learning. Geoderma, 452, p.117094. https://doi.org/10.1016/j.geoderma.2024.117094, 2024.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Montzka, C., Herbst, M., Weihermüller, L., Verhoef, A., and Vereecken, H.: A global data set of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves, Earth Syst. Sci. Data, 9, 529–543, https://doi.org/10.5194/essd-9-529-2017, 2017.

National Soil Survey Office: Chinese Soil Genus Records. vol. 6 China Agriculture Press, Beijing (in Chinese), 1996.

Nauman, T. W., Kienast-Brown, S., Roecker, S. M., Brungard, C., White, D., Philippe, J., and Thompson, J. A.: Soil landscapes of the United States (SOLUS): Developing predictive soil property maps of the conterminous United States using hybrid training sets, Soil Science Soc of Amer J, 88, 2046–2065, https://doi.org/10.1002/saj2.20769, 2024.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental covariates, SOIL, 4, 1–22, https://doi.org/10.5194/soil-4-1-2018, 2018.

Padarian, J., Stockmann, U., Minasny, B. and McBratney, A.B. Monitoring changes in global soil organic carbon stocks from space. Remote Sensing of Environment, 281, p.113260. https://doi.org/10.1016/j.rse.2022.113260, 2022.

Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881, 2017.

Rossiter, D. G. and Poggio, L.: Representing soil landscapes from digital soil mapping products – helping the map to speak for itself, SOIL, 11, 849–881, https://doi.org/10.5194/soil-11-849-2025, 2025.

Schmidinger, J. and Heuvelink, G. B. M.: Validation of uncertainty predictions in digital soil mapping, Geoderma, 437, 116585, https://doi.org/10.1016/j.geoderma.2023.116585, 2023.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil data set for earth system modeling, J. Adv. Model. Earth Syst., 6, 249–263, https://doi.org/10.1002/2013MS000293, 2014.

Shi, G.: GSDEv2 Code Release v1.0.0, Zenodo [code], https://doi.org/10.5281/zenodo.22859059, 2026.

Shi, G. Shangguan, W.: A global soil dataset for Earth system modeling (Version 2, GSDEv2), National Tibetan Plateau / Third Pole Environment Data Center [data set], https://doi.org/10.11888/Terre.tpdc.303613, 2026.

Shi, G., Shangguan, W., Zhang, Y., Li, Q., Wang, C., and Li, L.: Reducing location error of legacy soil profiles leads to improvement in digital soil mapping, Geoderma, 447, 116912, https://doi.org/10.1016/j.geoderma.2024.116912, 2024.

Shi, G., Sun, W., Shangguan, W., Wei, Z., Yuan, H., Li, L., Sun, X., Zhang, Y., Liang, H., Li, D., Huang, F., Li, Q., and Dai, Y.: A China dataset of soil properties for land surface modelling (version 2, CSDLv2), Earth Syst. Sci. Data, 17, 517–543, https://doi.org/10.5194/essd-17-517-2025, 2025.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., and Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, Biogeosciences, 10, 1717–1736, https://doi.org/10.5194/bg-10-1717-2013, 2013.

Turek, M.E., Poggio, L., Batjes, N.H., Armindo, R.A., van Lier, Q.D.J., de Sousa, L. and Heuvelink, G.B. Global mapping of volumetric water retention at 100, 330 and 15 000 cm suction using the WoSIS database. International Soil and Water Conservation Research, 11(2), p.225-239. https://doi.org/10.1016/j.iswcr.2022.08.001, 2023.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G.: BLOCK CV : An R package for generating spatially or environmentally separated folds for *k* -fold cross-validation of species distribution models, Methods Ecol Evol, 10, 225–232, https://doi.org/10.1111/2041-210X.13107, 2019.

Wang, W., Feng, S., Zhang, Y., Wei, Z., Dong, J., Weihermüller, L., Liu, C.-Q., and Vereecken, H.: Fusing ERA5-Land and SMAP L4 for an improved global soil moisture product (1950–2025), Earth Syst. Sci. Data, 18, 1061–1088, https://doi.org/10.5194/essd-18-1061-2026, 2026.

Wolpert, D. H.: Stacked generalization, Neural Networks, 5, 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1, 1992.

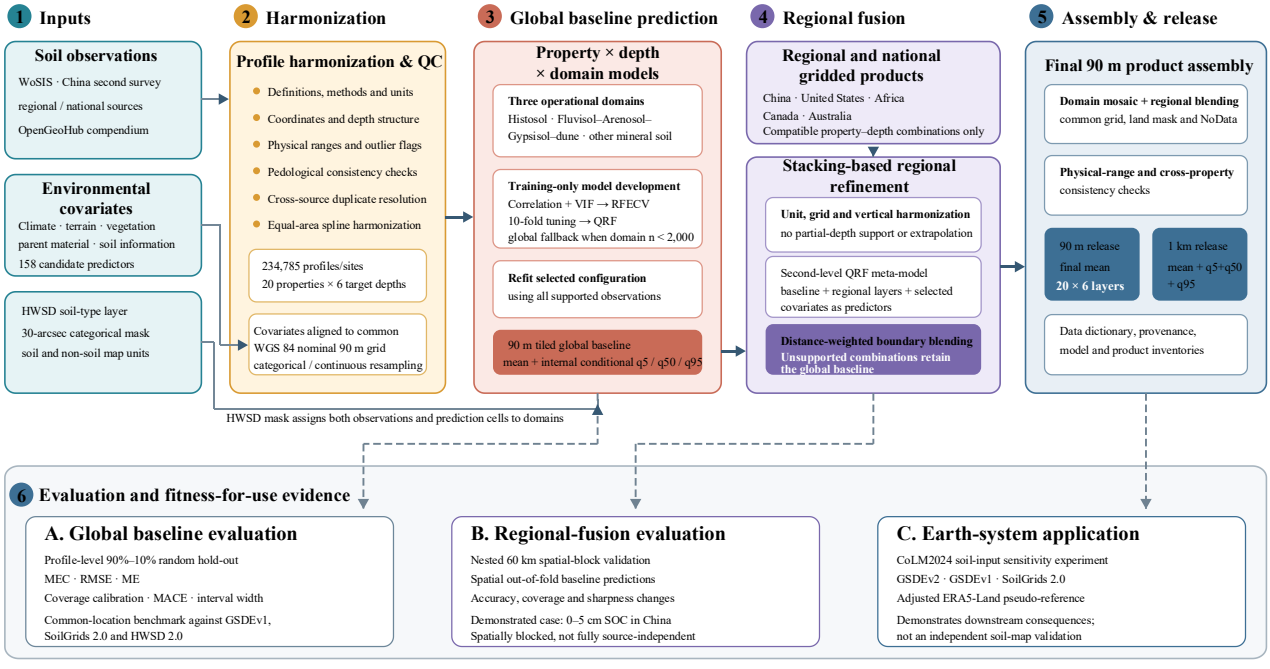
Xu, C., Torres-Rojas, L., Vergopolan, N., and Chaney, N. W.: The Benefits of Using State-Of-The-Art Digital Soil Properties Maps to Improve the Modeling of Soil Moisture in Land Surface Models, Water Resources Research, 59, e2022WR032336, https://doi.org/10.1029/2022WR032336, 2023.

Zhang, Z., Laux, P., Baade, J., Arnault, J., Wei, J., Wang, X., Liu, Y., Schmullius, C., and Kunstmann, H.: Impact of alternative soil data sources on the uncertainties in simulated land-atmosphere interactions, Agricultural and Forest Meteorology, 339, 109565, https://doi.org/10.1016/j.agrformet.2023.109565, 2023.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

Figure 1. GSDEv2 development, regional fusion, evaluation, and release workflow. Harmonized soil profiles and environmental covariates were used to generate property-, depth-, and soil-domain-specific QRF global baseline predictions. Compatible regional products were incorporated through stacking, followed by outward boundary blending and physical-consistency checks. Dashed arrows indicate evaluation and downstream application tracks.

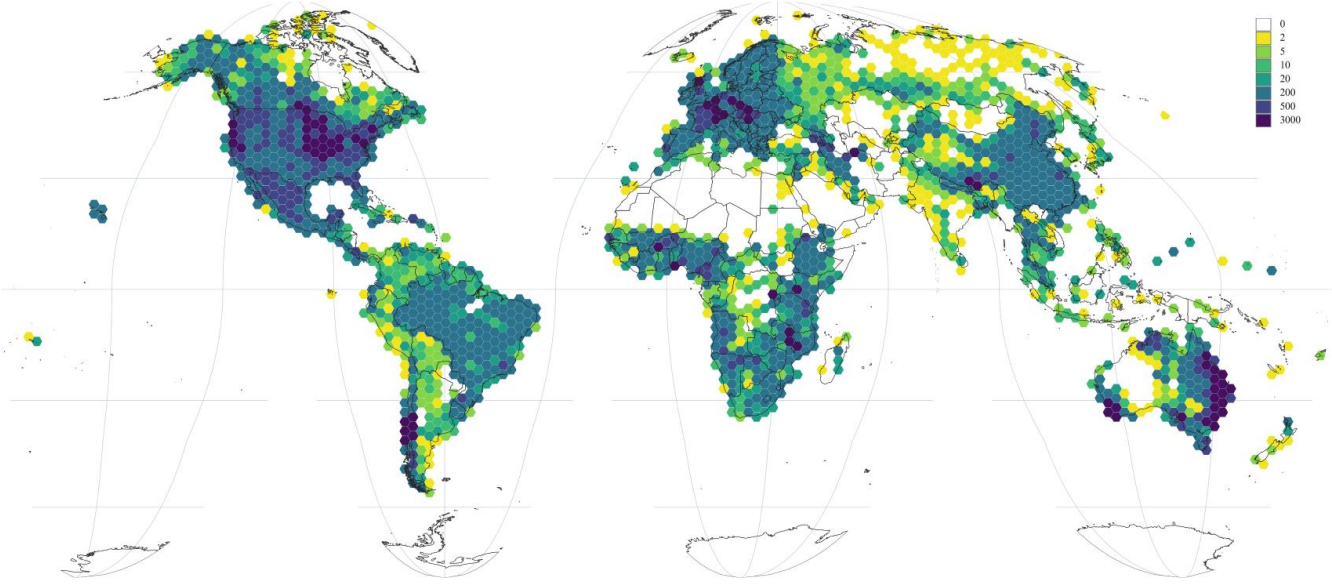


https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



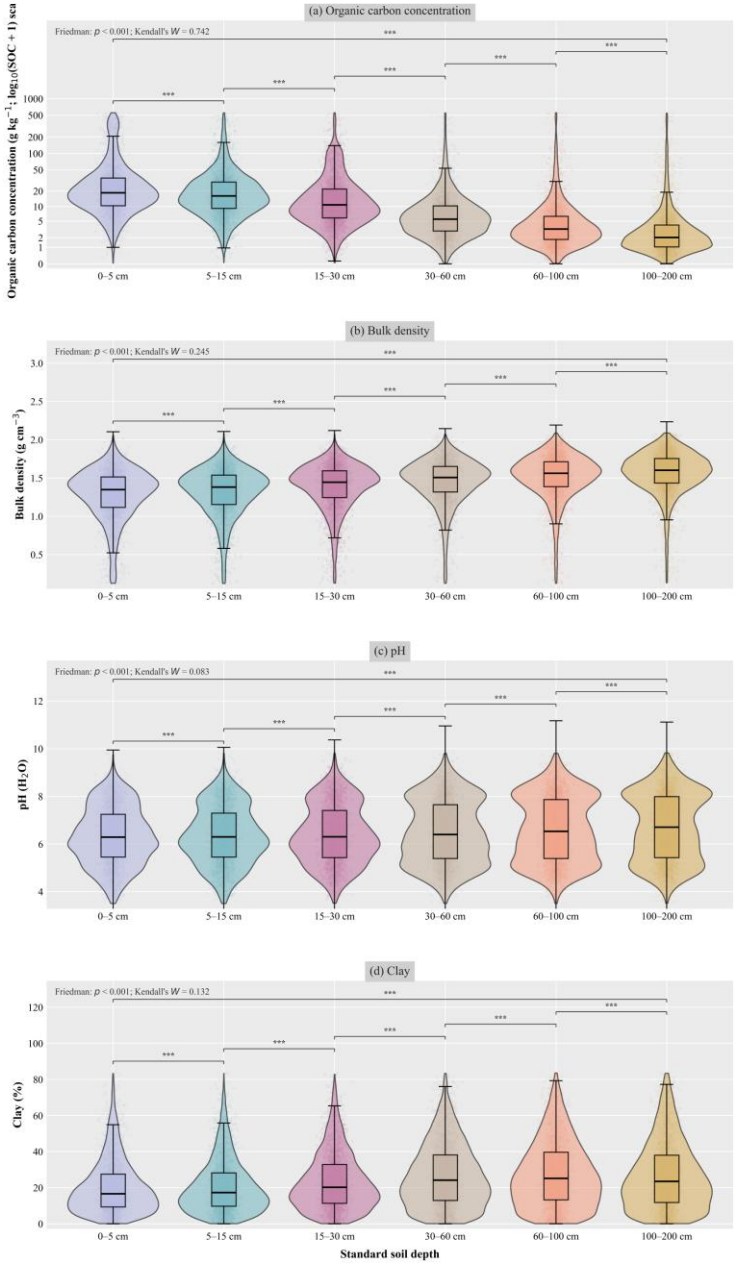
Open Access Earth System Science Data Discussions

Figure 2. Spatial distribution and sampling density of the 234,785 soil profiles retained after harmonization and quality control. Soil profile locations were aggregated into hexagonal bins, and colours indicate the number of unique profile locations per bin.



https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

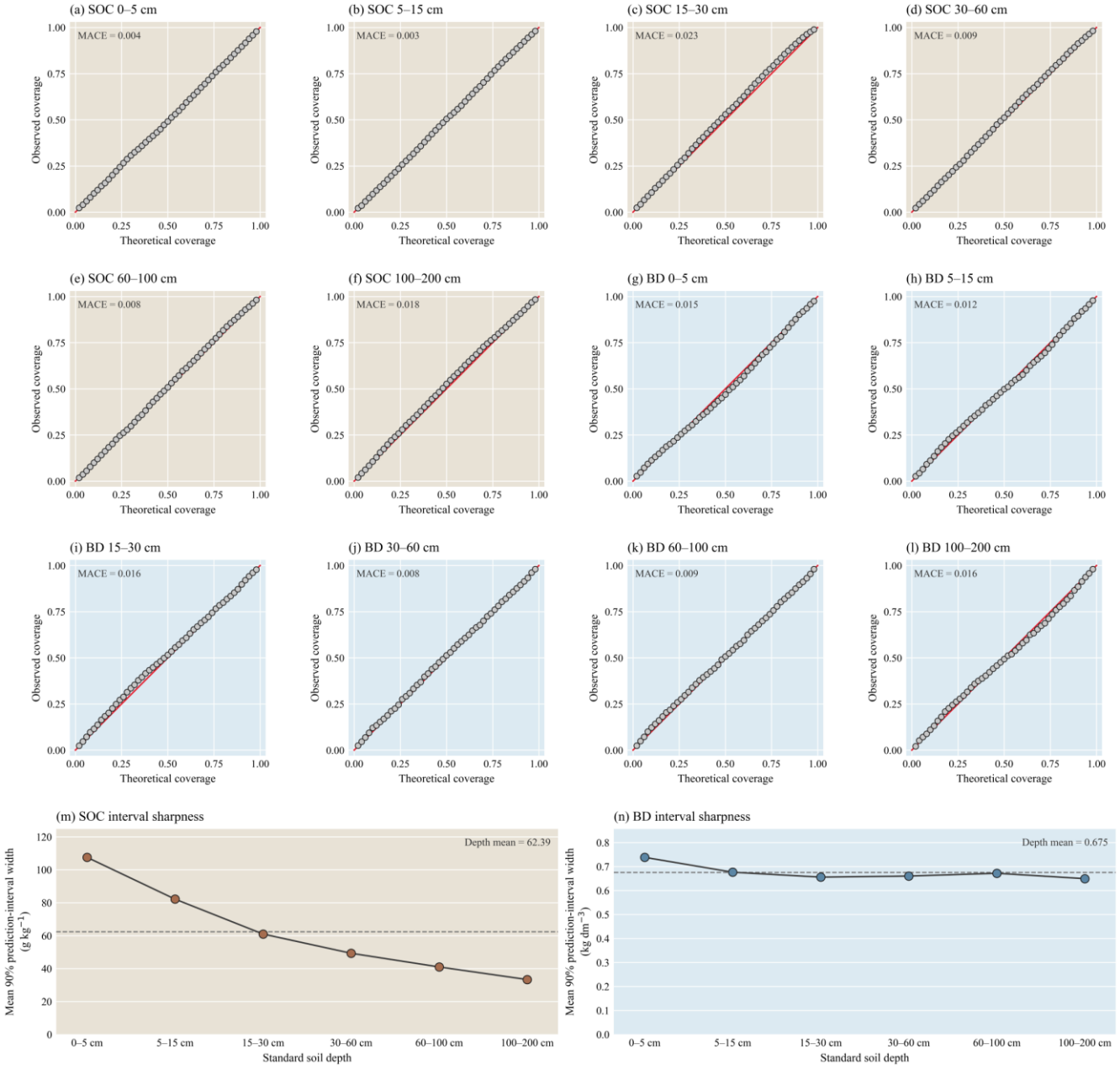
Figure 3. Distributions of (a) organic carbon concentration, (b) bulk density, (c) pH in water, and (d) clay content across the six depth-standardized intervals in the GSDEv2 soil-profile dataset. SOC was $\log_{10}$(SOC + 1)-transformed for visualization, whereas statistical analyses were performed using the original values. Overall depth effects were evaluated using Friedman tests, with Kendall's W as the effect size. Horizontal brackets denote selected pairwise Wilcoxon signed-rank tests with Holm correction, including comparisons between adjacent depths and between 0–5 and 100–200 cm. * indicates adjusted p < 0.001.**



https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

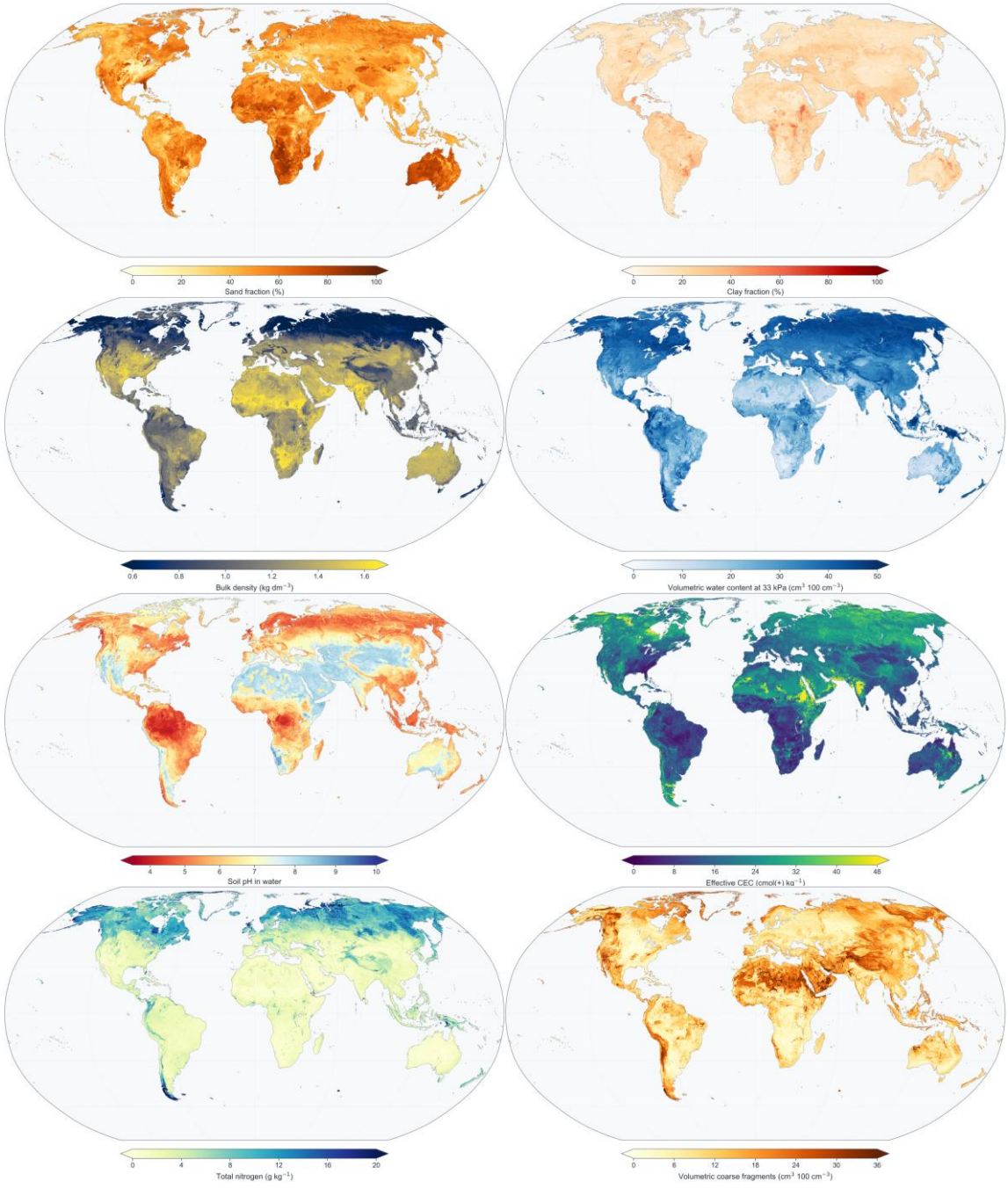
Open Access Earth System Science Data Discussions

Figure 4. Calibration and sharpness of conditional prediction intervals for organic carbon concentration (SOC; a–f, m) and bulk density (BD; g–l, n) across the six standard depth intervals. Grey points in panels (a–l) show observed coverage calculated from randomly held-out test profiles versus nominal coverage. The red diagonal line represents perfect calibration. MACE denotes the mean absolute calibration error across the evaluated nominal coverage levels. Panels (m) and (n) show the mean width of the nominal 90 % prediction interval across the six depth intervals for SOC and BD, respectively. Prediction intervals were derived from Quantile Random Forest conditional quantiles.



https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 5. Global spatial patterns of selected topsoil (0 – 5 cm) properties from the final 90 m GSDEv2 mean prediction dataset: sand content (a), clay content (b), bulk density (c), volumetric water content at 33 kPa (d), soil pH in water (e), effective cation exchange capacity (f), total nitrogen content (g), and volumetric coarse fragment content (h).

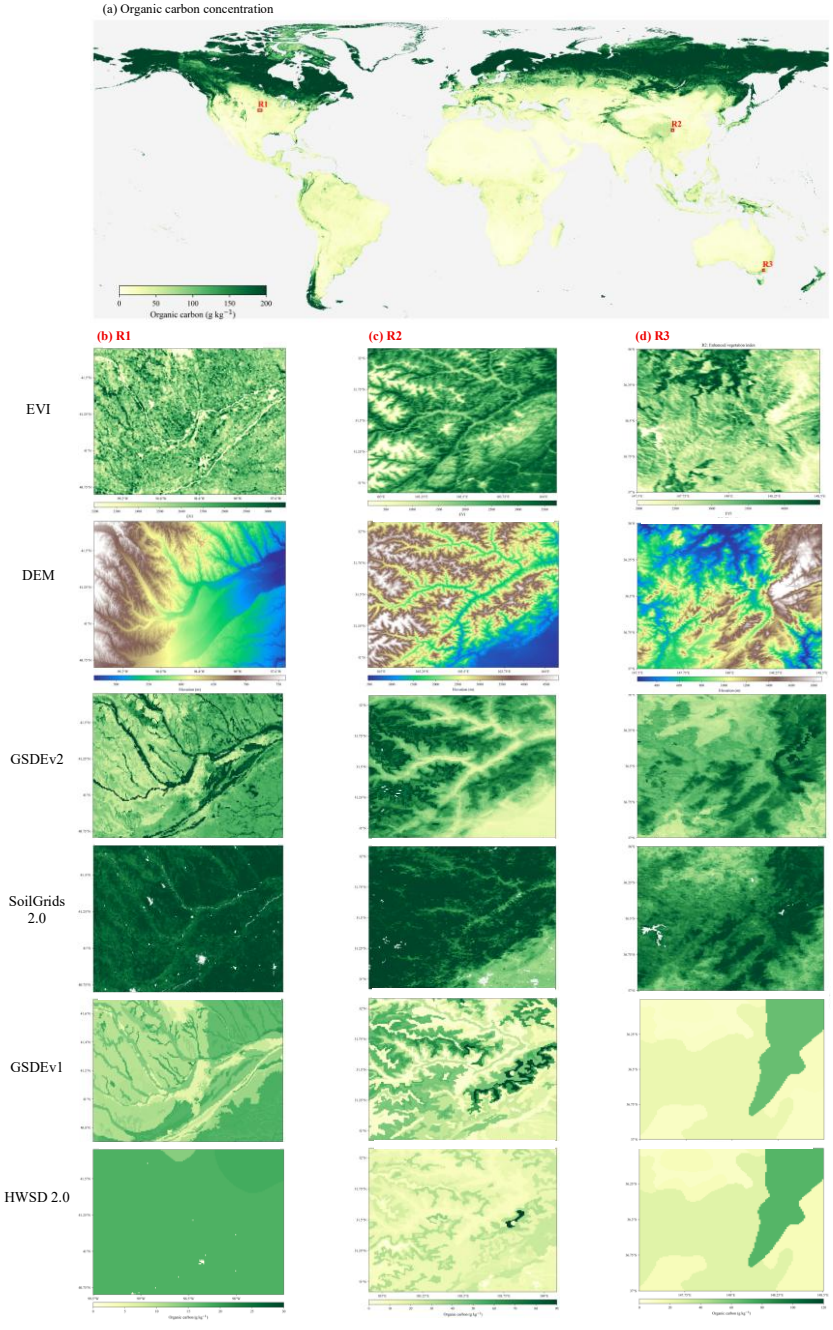


https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Figure 6. Global and regional comparison of surface SOC (0–5 cm) among GSDEv2, SoilGrids 2.0, GSDEv1, and HWSD 2.0. (a) Global GSDEv2 SOC and locations of R1–R3; (b–d) enhanced vegetation index (EVI), digital elevation model (DEM), and SOC products for the Colorado Rockies–Great Plains transition (R1), Southwest Sichuan (R2), and the Great Dividing Range–Murray Basin (R3). The four SOC products use a common colour scale within each region.

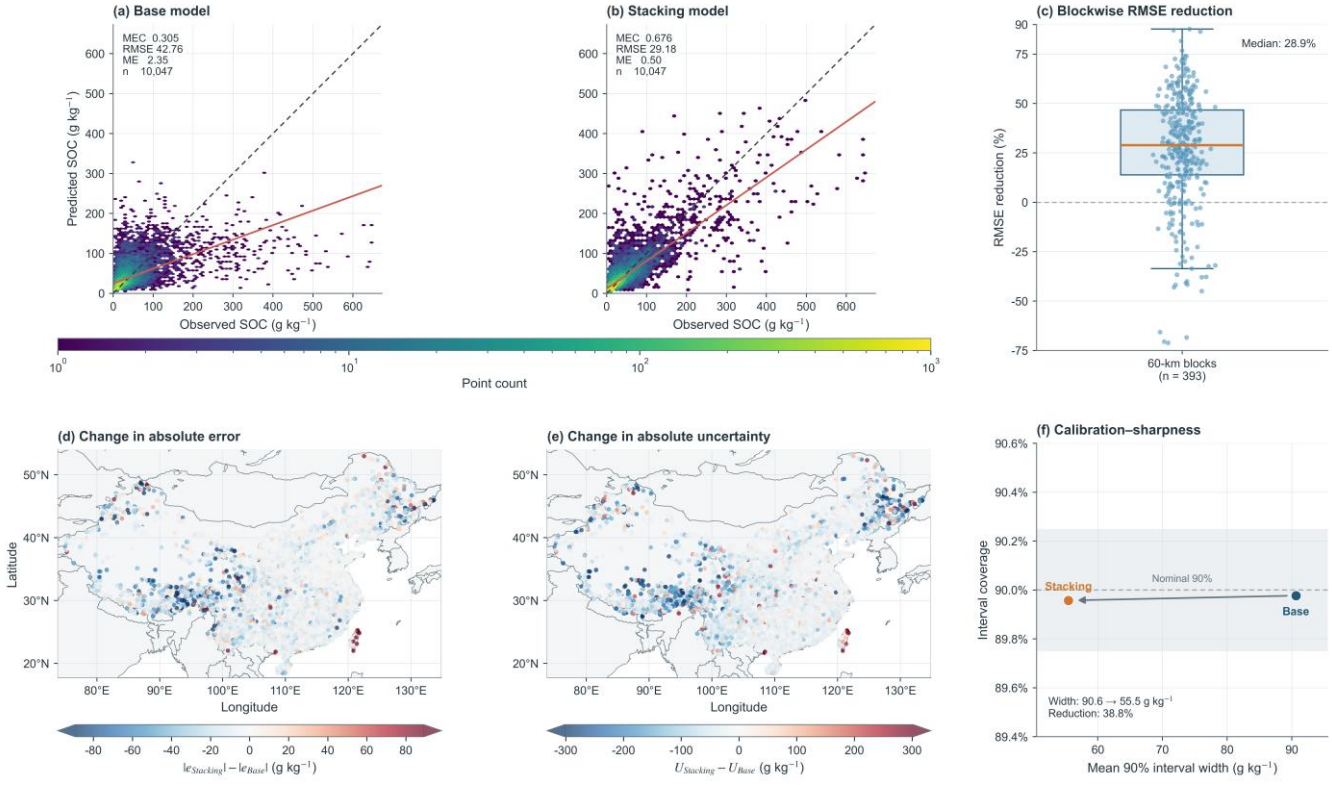


https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

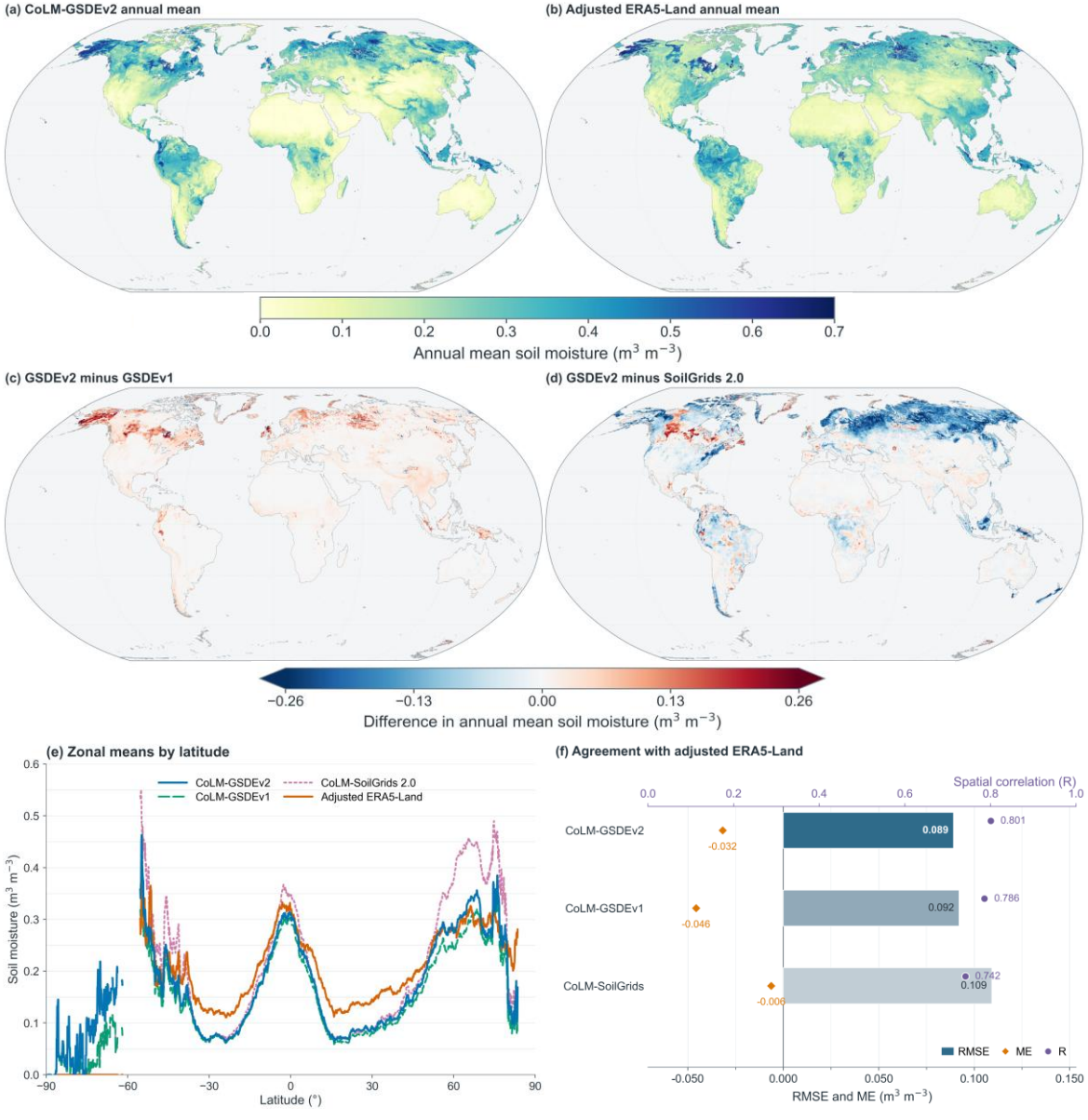
Figure 7. Spatial-block evaluation of stacking-based regional fusion for soil organic carbon prediction at 0–5 cm depth in China. Observed–predicted relationships for (a) the GSDEv2 baseline model and (b) GSDEv2 stacking model, (c) blockwise RMSE reductions within each 60 km block, spatial stacking-minus-baseline changes in (d) absolute prediction error and (e) conditional 90 % prediction-interval width, and (f) prediction-interval coverage versus mean interval width. MEC is the modelling efficiency coefficient and ME is mean error.



https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Open Access Earth System Science Data Discussions

740 **Figure 8. Sensitivity of CoLM2024-simulated annual-mean first-layer soil moisture to alternative global soil datasets in 2018. (a) CoLM simulation driven by GSDEv2 inputs; (b) adjusted ERA5-Land annual-mean surface soil moisture; (c) difference between CoLM-GSDEv2 and CoLM-GSDEv1; (d) difference between CoLM-GSDEv2 and CoLM-SoilGrids 2.0; (e) latitude-zonal means for the three CoLM experiments and adjusted ERA5-Land; and (f) area-weighted spatial root-mean-square error (RMSE), mean error (ME), and spatial correlation coefficient (R) relative to**
745 **adjusted ERA5-Land.**



https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.

Table 1. Soil properties included in GSDEv2, with their acronyms, units, raster conversion factors, and descriptions.

No	Soil property	Acronym	Units	Conversion factor*	Description
1	Sand fraction	SAND	%	10	Mass percentage of sand-sized (2-0.05 mm) particles in the fine-earth fraction.
2	Silt fraction	SILT	%	10	Mass percentage of silt-sized (0.05-0.002 mm) particles in the fine-earth fraction.
3	Clay fraction	CLAY	%	10	Mass percentage of clay-sized (< 0.002 mm) particles in the fine-earth fraction.
4	pH (water)	PHH2O	unitless	100	Soil acidity or alkalinity measured in a soil–water suspension.
5	Bulk density	BD	kg dm^{-3}	100	Bulk density of the fine earth fraction oven dry.
6	Organic carbon concentration	SOC	g kg^{-1}	10	Mass concentration of organic carbon in the fine-earth fraction.
7	Total carbon	TC	g kg^{-1}	10	Mass concentration of total carbon, including organic and inorganic carbon forms.
8	Total nitrogen	TN	g kg^{-1}	10	Mass concentration of total nitrogen in soil, including organic and inorganic nitrogen forms.
9	Effective cation exchange capacity	CEC	cmol(+) kg^{-1}	10	Capacity of the fine earth fraction to hold exchangeable cations.
10	Cation exchange capacity at pH 7	CECPH7	cmol(+) kg^{-1}	10	Cation exchange capacity determined after buffering soil to pH 7.
11	Volumetric coarse fragments	cfvo	cm^3 100 cm^{-3}	100	Volume percentage of soil fragments larger than 2 mm in the whole soil.
12	Gravimetric coarse fragments	cfgr	g 100 g^{-1}	100	Mass percentage of soil fragments larger than 2 mm in the whole soil.
13	Fine-earth bulk density at 33 kPa	bdfi33	kg dm^{-3}	100	Bulk density of the fine-earth fraction after equilibration or desorption at 33 kPa.
14	Volumetric water content at 33 kPa	wv0033	cm^3 100 cm^{-3}	100	Volumetric soil water content retained at a matric potential of 33 kPa.
15	Volumetric water content at 1500 kPa	wv1500	cm^3 100 cm^{-3}	100	Volumetric soil water content retained at a matric potential of 1500 kPa.
16	Volumetric water content at 10 kPa	wv0010	cm^3 100 cm^{-3}	100	Volumetric soil water content retained at a matric potential of 10 kPa.
17	Gravimetric water content at 33 kPa	wg0033	g 100 g^{-1}	100	Gravimetric soil water content retained at a matric potential of 33 kPa.
18	Gravimetric water	wg1500	g 100 g^{-1}	100	Gravimetric soil water content retained at a matric potential of

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

No	Soil property	Acronym	Units	Conversion factor*	Description
	content at 1500 kPa				1500 kPa.
19	Gravimetric water content at 10 kPa	wg0010	g 100 g^{-1}	100	Gravimetric soil water content retained at a matric potential of 10 kPa.
20	Extractable phosphorus, Bray I	PBR	mg kg^{-1}	10	Extractable phosphorus determined using the Bray I extraction procedure, mainly applicable to acidic soils.

Note. The conversion factor indicates the scale factor used for the distributed raster files. Users should divide the stored raster value by the conversion factor to obtain values in the listed units. For example, a stored value of 735 for PHH2O corresponds to PHH2O = 7.35 when the conversion factor is 100.

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 2. Environmental covariate groups used to characterize soil-forming conditions in GSDEv2. Representative predictors and principal source datasets are shown. The complete covariate inventory is provided in Table S2, and the corresponding source datasets and preprocessing protocols are documented in Tables S3a and S3b.

Category	Representative covariates	Main products	Soil-forming processes represented
Soil information	Soil-group proportions and soil-domain indicators	HWSD-derived layers	Broad soil-geographical patterns and differentiation among desert, organic-soil, and other mineral-soil domains
Climate	Long-term surface temperature, precipitation, potential solar radiation, cloud cover, and snow or ice indicators	WorldClim, CHELSA, MODIS, and WorldGrids	Hydrothermal controls on weathering, leaching, soil moisture, organic-matter turnover, and carbonate accumulation
Vegetation and land surface	NDVI, EVI, LAI, NPP, surface reflectance, land cover, land use, and water or snow indices	Landsat, Sentinel-2, MODIS, and ESA land-cover products	Vegetation productivity, organic-matter inputs, surface cover, land management, and land-surface conditions
Terrain	Elevation, slope, aspect, curvature, topographic position index, wetness index, and drainage-related attributes	MERIT DEM and derived terrain layers	Redistribution of water and materials, erosion and deposition, drainage, and local microclimatic variation
Parent material and geology	Lithology, geological classes, and sediment or regolith thickness	GLiM and global sediment-thickness datasets	Mineralogical composition, soil texture, coarse fragments, nutrient status, pH, and carbonate occurrence

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

Table 3. Regional and national soil products used in the GSDEv2 stacking-based data-fusion framework, with their native grid spacing, variables, vertical support and references.

Fusion region	Regional product	Native resolution	Variables used in fusion	Native vertical support	Reference
China	CSDLv2	90 m	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC, cfvo	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Shi et al., 2025
China	National Soil Information Grids of China	90 m	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC, cfvo	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Liu et al., 2020, 2022
United States	SOLUS100m	100 m	SAND, SILT, CLAY, PHH2O, BD, SOC, CEC, CECPH7, cfvo	0, 5, 15, 30, 60, 100, and 150 cm	Nauman et al., 2024
Africa	iSDAsoil	30 m	SAND, SILT, CLAY, PHH2O, BD, SOC, TC, TN, CEC, cfgr	0-20, and 20-50 cm	Hengl et al., 2021
Canada	Canadian soil grids	100 m	SAND, SILT, CLAY, BD, SOC, CEC	0-5, 5-15, 15-30, 30-60, and 60-100 cm	Geng et al., 2025
Australia	SLGA	90 m	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Malone et al., 2025

https://doi.org/10.5194/essd-2026-770
Preprint. Discussion started: 12 November 2026
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data Discussions

760 **Table 4. Predictive performance of selected soil properties in GSDEv2 global baseline predictions before regional fusion and three existing global soil products, evaluated for selected property–depth combinations using common randomly held-out test profiles. The "Number" is the evaluation sample size. Complete results are provided in Table S7; property abbreviations and units are defined in Table 1.**

Soil property	Depth interval (cm)	Number	GSDEv2			GSDEv1			SoilGrids 2.0			HWSD 2.0		
			MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME	MEC	RMSE	ME
SAND	0-5	8668	0.58	18.03	0.09	0.20	24.88	-2.20	0.54	18.87	-1.64	0.21	24.73	1.35
	5-15	9579	0.56	17.09	-0.31	0.20	23.04	-2.21	0.52	17.85	-1.65	0.20	23.04	1.37
	15-30	9482	0.55	17.60	-0.20	0.19	23.61	-2.34	0.50	18.55	-1.62	0.21	23.32	1.26
	30-60	9090	0.54	18.65	0.09	0.17	25.05	-2.34	0.47	20.02	-1.65	0.18	24.90	1.34
	60-100	8036	0.48	19.68	0.10	0.14	25.31	-1.62	0.42	20.78	-1.86	0.17	24.86	1.21
	100-200	5936	0.45	21.10	0.07	0.11	26.84	-1.31	0.38	22.40	-1.93	0.13	26.54	1.23
CLAY	0-5	10698	0.46	10.99	0.77	0.11	14.11	0.22	0.41	11.48	-0.48	0.12	14.13	0.23
	5-15	12185	0.46	10.97	0.78	0.08	14.32	0.03	0.40	11.56	-0.49	0.09	14.24	0.01
	15-30	11962	0.45	12.03	-0.16	0.10	15.39	0.51	0.40	12.56	-0.50	0.11	15.30	0.48
	30-60	11188	0.44	13.34	-0.34	0.09	17.01	0.12	0.37	14.15	-0.51	0.08	17.10	0.10
	60-100	9943	0.43	13.95	-0.17	0.07	17.82	-0.21	0.36	14.78	-0.51	0.09	17.63	-0.22
	100-200	7300	0.44	14.05	-0.46	0.08	18.01	0.22	0.37	14.90	-0.50	0.08	18.01	0.21
PHH2O	0-5	10504	0.67	0.73	0.04	0.49	0.91	-0.02	0.60	0.80	-0.12	0.39	0.99	-0.38
	5-15	11216	0.64	0.73	-0.01	0.50	0.86	-0.03	0.57	0.80	-0.15	0.35	0.98	-0.12
	15-30	10861	0.66	0.69	-0.02	0.39	0.92	-0.35	0.55	0.79	-0.23	0.38	0.93	-0.12
	30-60	10722	0.65	0.77	0.00	0.42	0.99	-0.05	0.54	0.88	-0.25	0.37	1.03	-0.17
	60-100	9590	0.65	0.82	0.01	0.45	1.03	0.04	0.54	0.94	-0.27	0.39	1.08	-0.17
	100-200	8640	0.64	0.87	-0.01	0.51	1.02	-0.03	0.51	1.02	-0.35	0.51	1.02	-0.09
BD	0-5	3240	0.62	0.21	0.01	0.12	0.32	0.00	0.53	0.23	0.01	0.03	0.34	0.14
	5-15	3054	0.60	0.23	0.01	0.13	0.34	0.01	0.57	0.24	0.01	0.01	0.36	0.17
	15-30	3011	0.59	0.27	-0.01	0.12	0.40	0.00	0.53	0.29	0.01	0.01	0.42	0.11
	30-60	2969	0.58	0.16	0.01	0.11	0.23	0.01	0.52	0.17	-0.00	0.02	0.24	0.21
	60-100	2879	0.56	0.22	0.00	0.09	0.32	0.02	0.50	0.23	-0.01	0.02	0.33	0.25
	100-200	2663	0.55	0.24	0.01	0.07	0.35	0.04	0.42	0.27	-0.01	0.01	0.36	0.31
SOC	0-5	8953	0.64	49.18	-1.43	0.21	72.85	-1.89	0.41	62.96	0.53	0.09	78.19	-0.48
	5-15	9803	0.61	36.19	-1.87	0.17	52.80	-1.75	0.48	41.79	0.05	0.07	55.89	-1.10
	15-30	10410	0.59	45.71	-1.51	0.13	66.59	-1.23	0.43	53.90	0.04	0.09	68.10	-0.47
	30-60	8471	0.58	32.25	-0.95	0.11	46.95	-1.51	0.38	39.18	0.01	0.05	48.50	-0.45
	60-100	7049	0.58	36.62	-1.32	0.09	53.90	-1.23	0.35	45.56	0.63	0.04	55.36	-0.25
	100-200	6988	0.56	21.85	-0.73	0.08	31.60	-1.11	0.31	27.36	0.51	0.04	32.27	-0.27