Supplementary materials of

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

Gaosong Shi et al.

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

Contents

Contents I
S1 Supplementary prediction maps and uncertainty diagnostics 1
S1.1 SAND fraction (SAND) 1
S1.2 SILT fraction (SILT) 3
S1.3 CLAY fraction (CLAY) 5
S1.4 pH in water (PHH2O) 7
S1.5 Bulk density (BD) 9
S1.6 Organic carbon concentration (SOC) 11
S1.7 Total carbon (TC) 13
S1.8 Total nitrogen (TN) 15
S1.9. Effective cation exchange capacity (CEC) 17
S1.10 Cation exchange capacity at pH 7 (CECPH7) 19
S1.11 Volumetric coarse fragments (cfvo) 21
S1.12 Gravimetric coarse fragments (cfgr) 23
S1.13 Fine-earth bulk density at 33 kPa (bdfi33) 25
S1.14 Volumetric water content at 33 kPa (wv0033) 27
S1.15 Volumetric water content at 1500 kPa (wv1500) 29
S1.16 Volumetric water content at 10 kPa (wv0010) 31
S1.17 Gravimetric water content at 33 kPa (wg0033) 33
S1.18 Gravimetric water content at 1500 kPa (wg1500) 35
S1.19 Gravimetric water content at 10 kPa (wg0010) 37
S1.20 Extractable phosphorus, Bray I (PBR) 39
S2. Profile harmonization and quality control 41
S2.1 Data standardization and structural checks 41
S2.2 Physical-range and statistical-outlier screening 41
S2.3 Pedological consistency checks 42
S2.4 Duplicate and environmental-context screening 43
S2.5 Final retention 43
S2.6 Vertical harmonization of profile observations 43
S3. Regional-product harmonization 44
S3.1 Spatial harmonization 44
S3.2 Vertical harmonization of layer-average products 45

S3.3 Vertical harmonization of discrete-depth products 45
S3.4 Fusion eligibility and provenance assessment 46
Table S1. Inventory of soil-profile and soil-observation data sources used in the development of GSDEv2. 47
Table S2. Candidate environmental covariates considered in the GSDEv2 modelling framework before property-, depth-, and soil-domain-specific predictor selection. Source products and common harmonization procedures are identified by Source ID and Processing ID and documented in Tables S3a and S3b, thereby avoiding repeated text while preserving full provenance. Native grid spacing refers to the source or derived layer rather than the final nominal 3-arcsec GSDEv2 grid. 50
Table S3. Source datasets, versions, temporal support, references, and common covariate-harmonization protocols. 65
Table S3a. Source datasets and provenance. 65
Table S3b. Common preprocessing and resampling protocols. 67
Table S4. HWSD soil-type codes used to define the GSDEv2 soil-forming domains. 69
Table S5. Hyperparameters used for the considered soil properties at the 0–5 cm depth interval. See Table 1 for abbreviations and units of the soil properties considered. 70
Table S6. Descriptive statistics of the soil-profile observations for the 20 GSDEv2 properties at six standard depth intervals. Refer to Table 1 for the abbreviations and units of the soil properties considered. 71
Table S7. Accuracy evaluation of the soil properties prediction with GSDEv2 global baseline predictions before regional fusion, GSDEv1, SoilGrids 2.0, and HWSD 2.0, based on the randomly held-out soil profiles. The "N" column indicates the number of samples used during testing. See Table 1 for the abbreviations and units of the soil properties considered. 76
Table S8. Regional-fusion implementation, vertical harmonisation, and source-overlap assessment for the regional and national soil products used in GSDEv2. 80
Table S9. Configuration of the CoLM2024 soil-input sensitivity experiments driven by GSDEv2, GSDEv1, and SoilGrids 2.0. 83
References 84

S1 Supplementary prediction maps and uncertainty diagnostics

S1.1 SAND fraction (SAND)

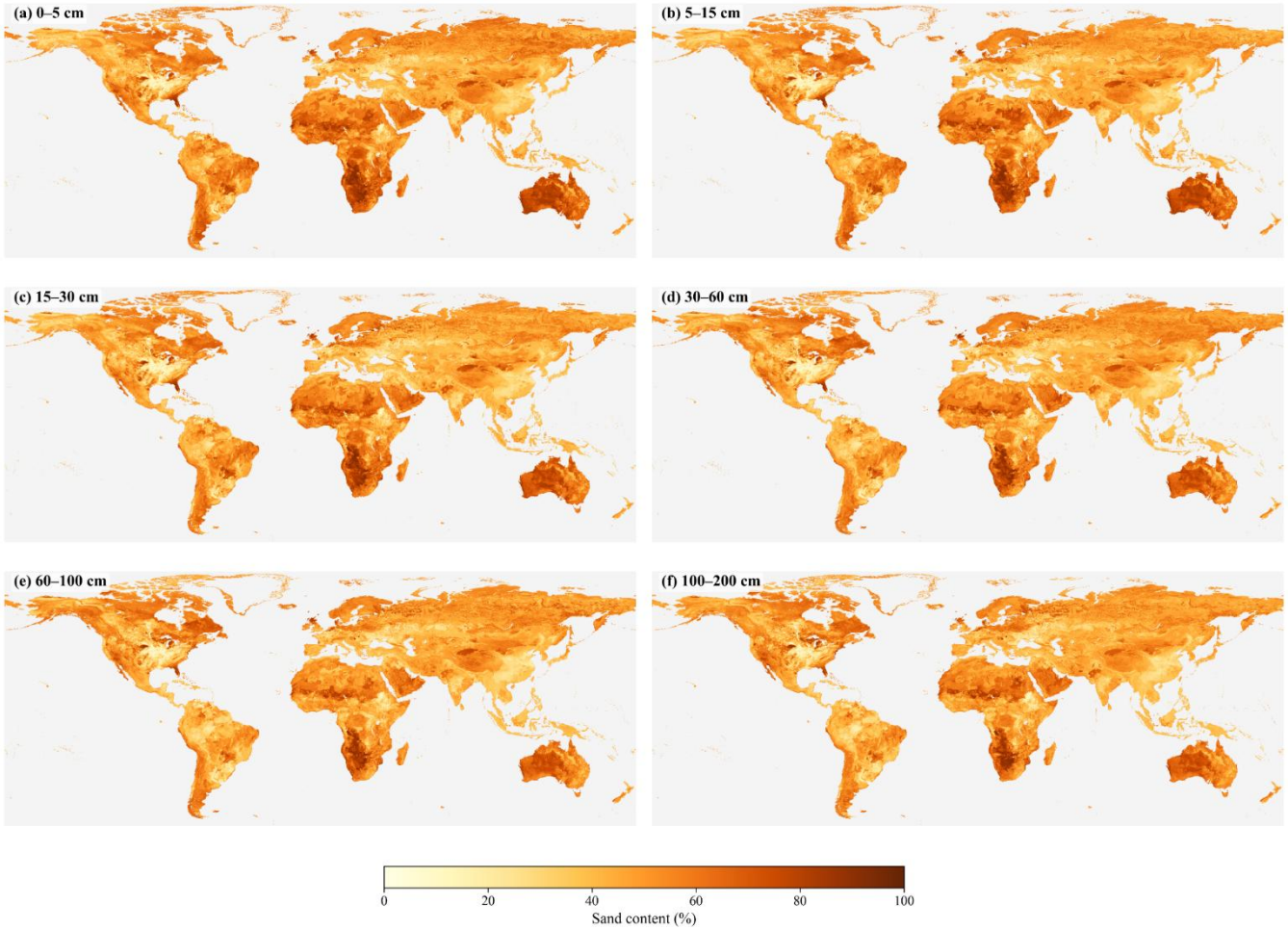


Figure S1. Final GSDEv2 mean predictions of SAND fraction (SAND) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

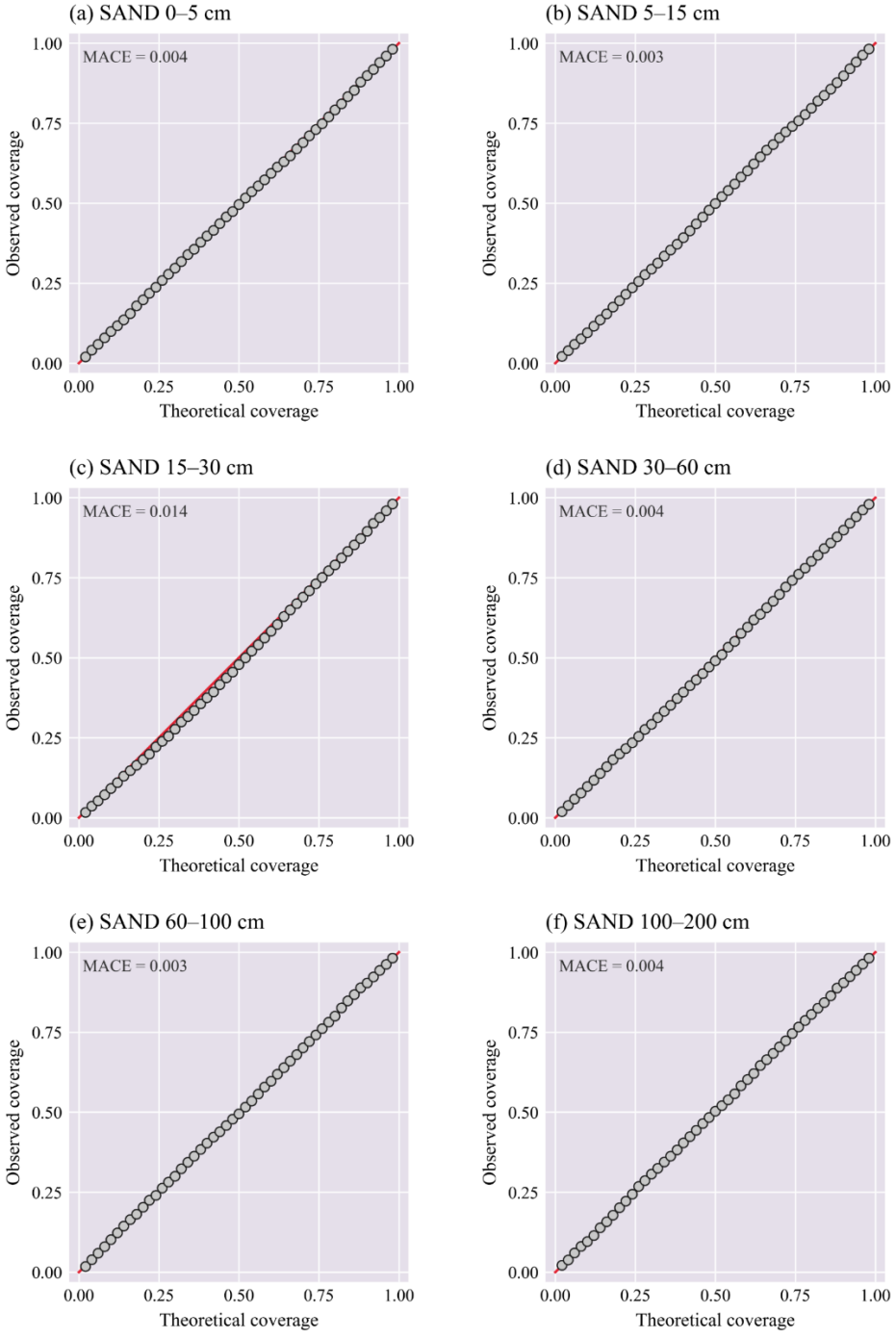


Figure S2. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for SAND fraction (SAND) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.2 SILT fraction (SILT)

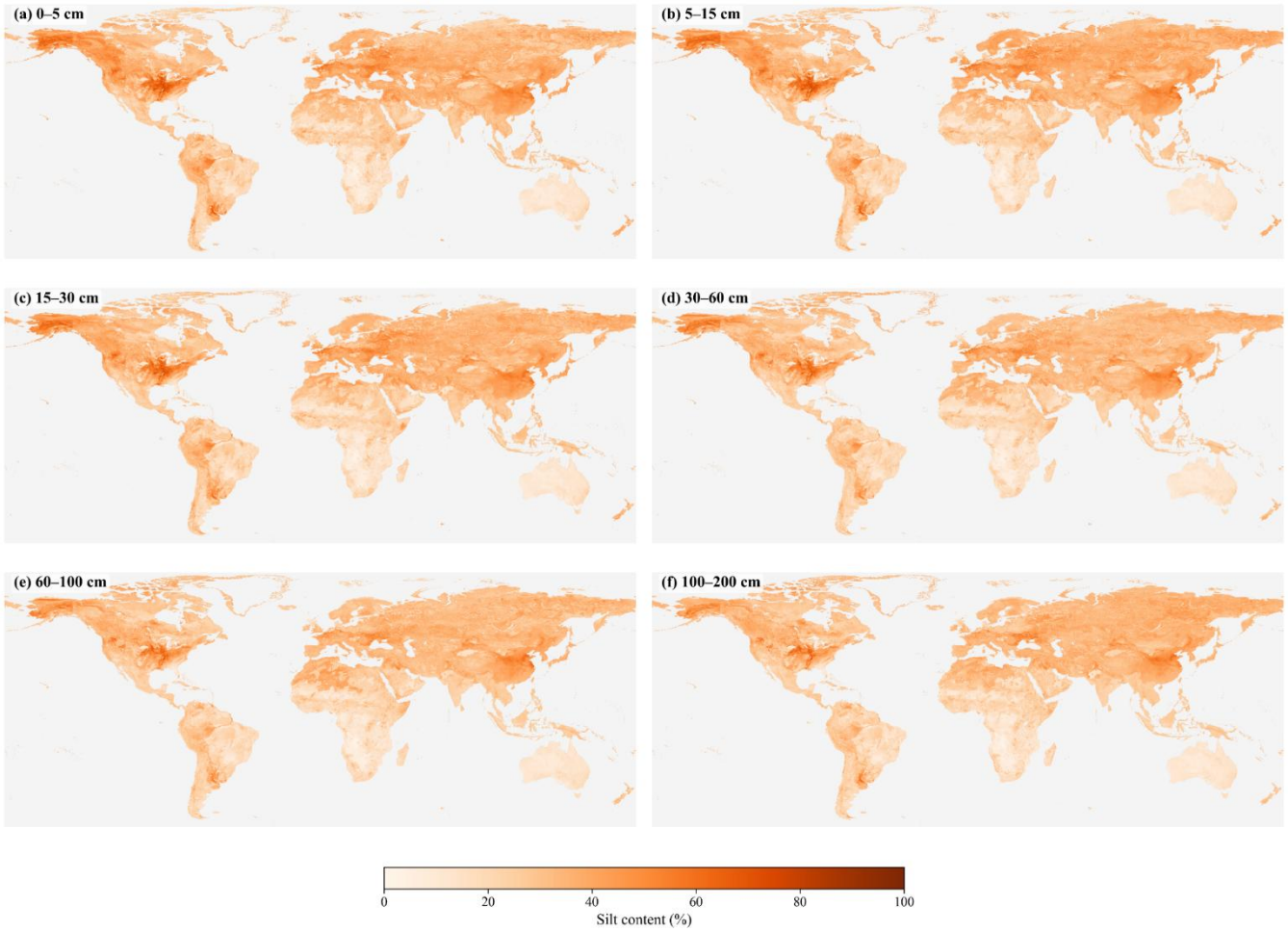
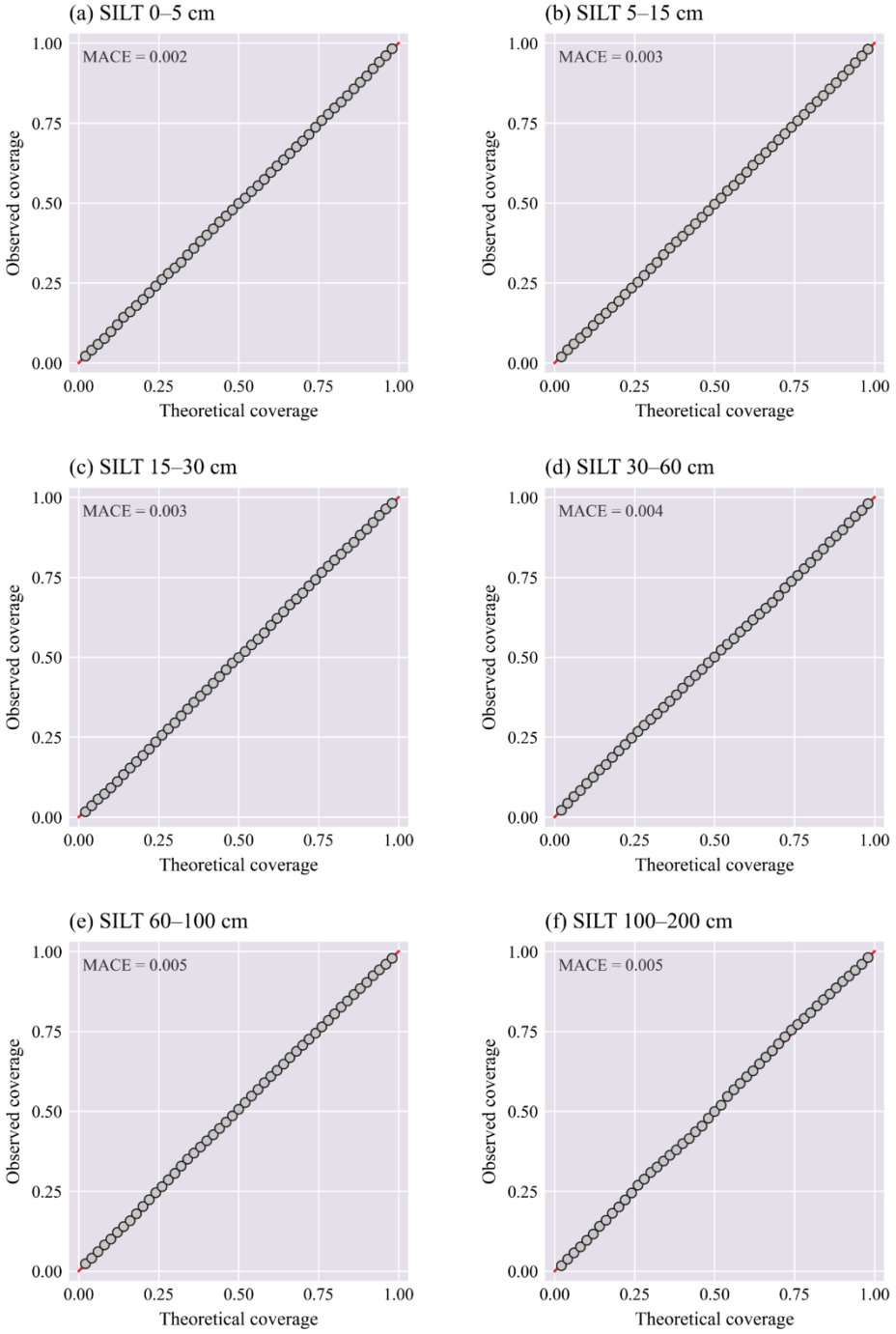


Figure S3. Final GSDEv2 mean predictions of SILT fraction (SILT) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.



15 **Figure S4.** Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for SILT fraction (SILT) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.3 CLAY fraction (CLAY)

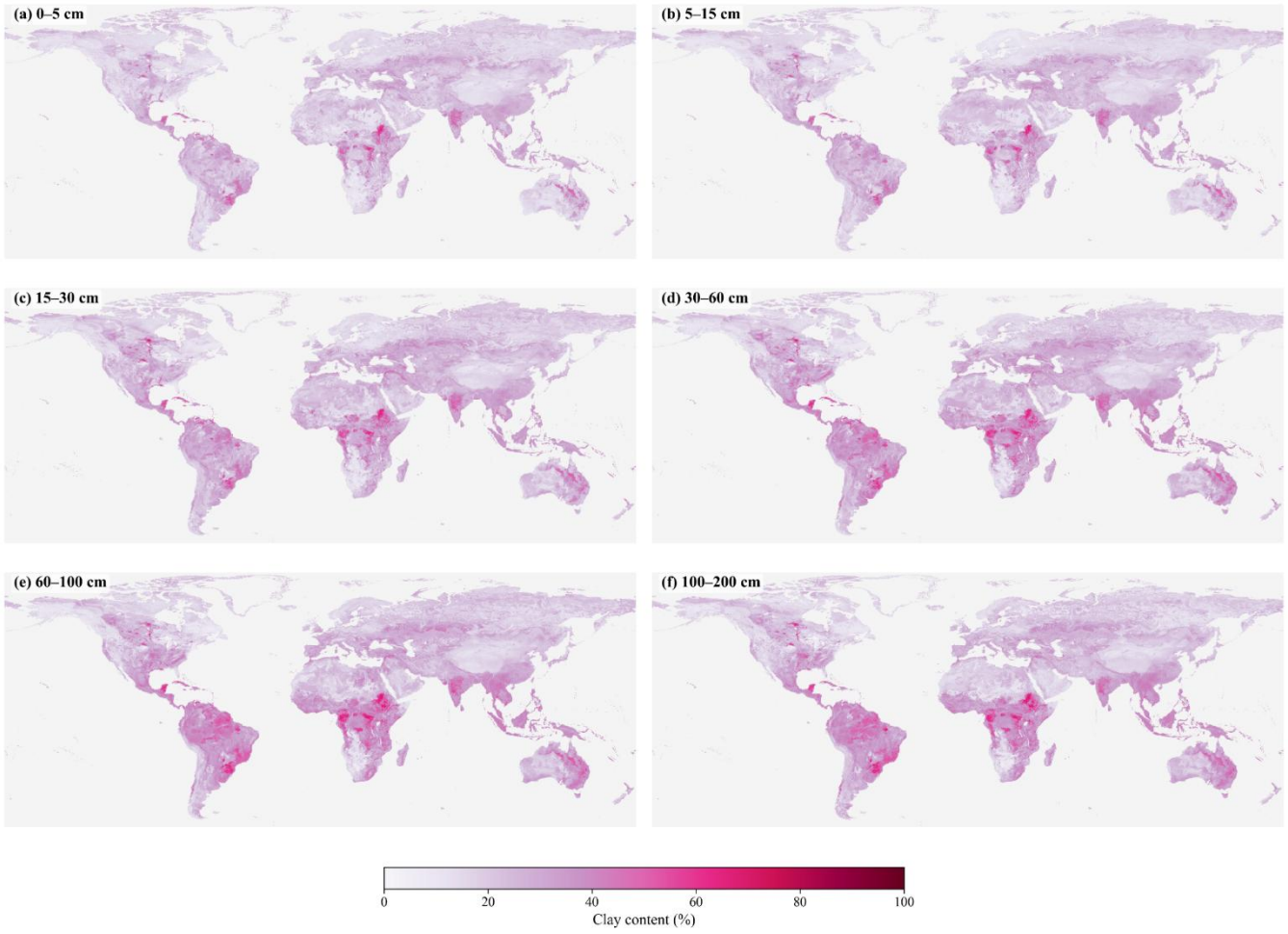


Figure S5. Final GSDEv2 mean predictions of CLAY fraction (CLAY) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

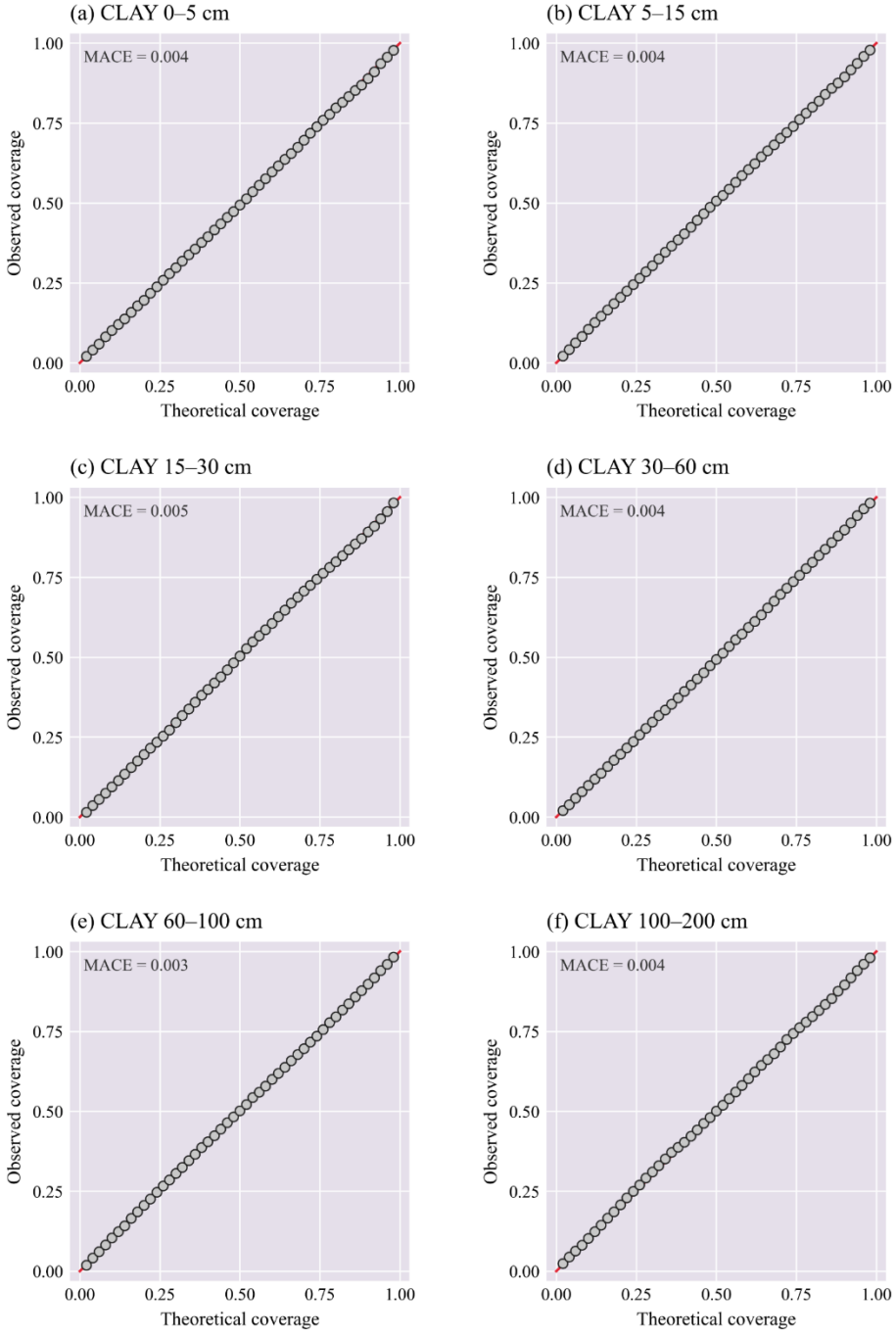


Figure S6. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for CLAY fraction (CLAY) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.4 pH in water (PHH2O)

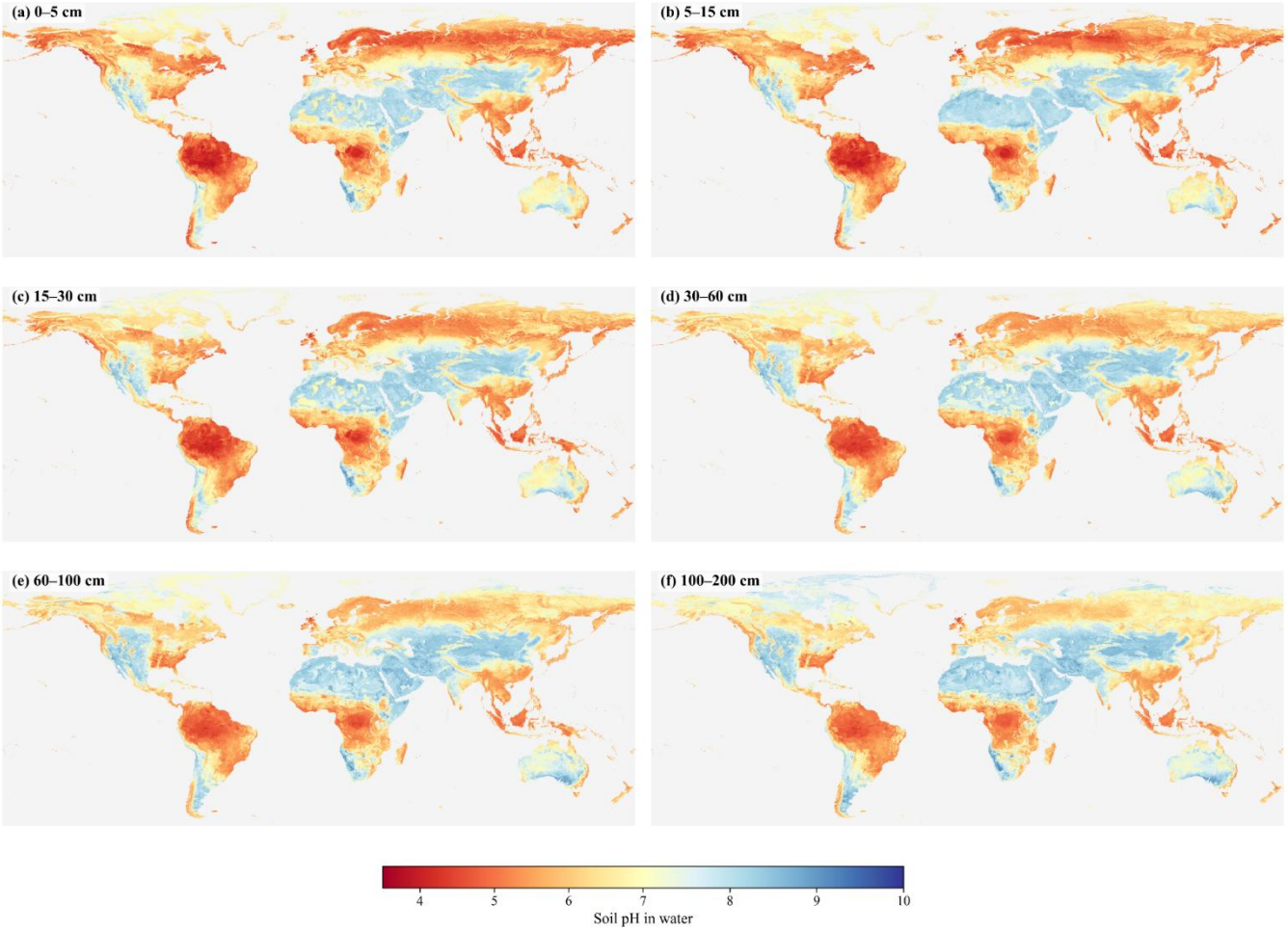


Figure S7. Final GSDEv2 mean predictions of pH in water (PHH2O) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

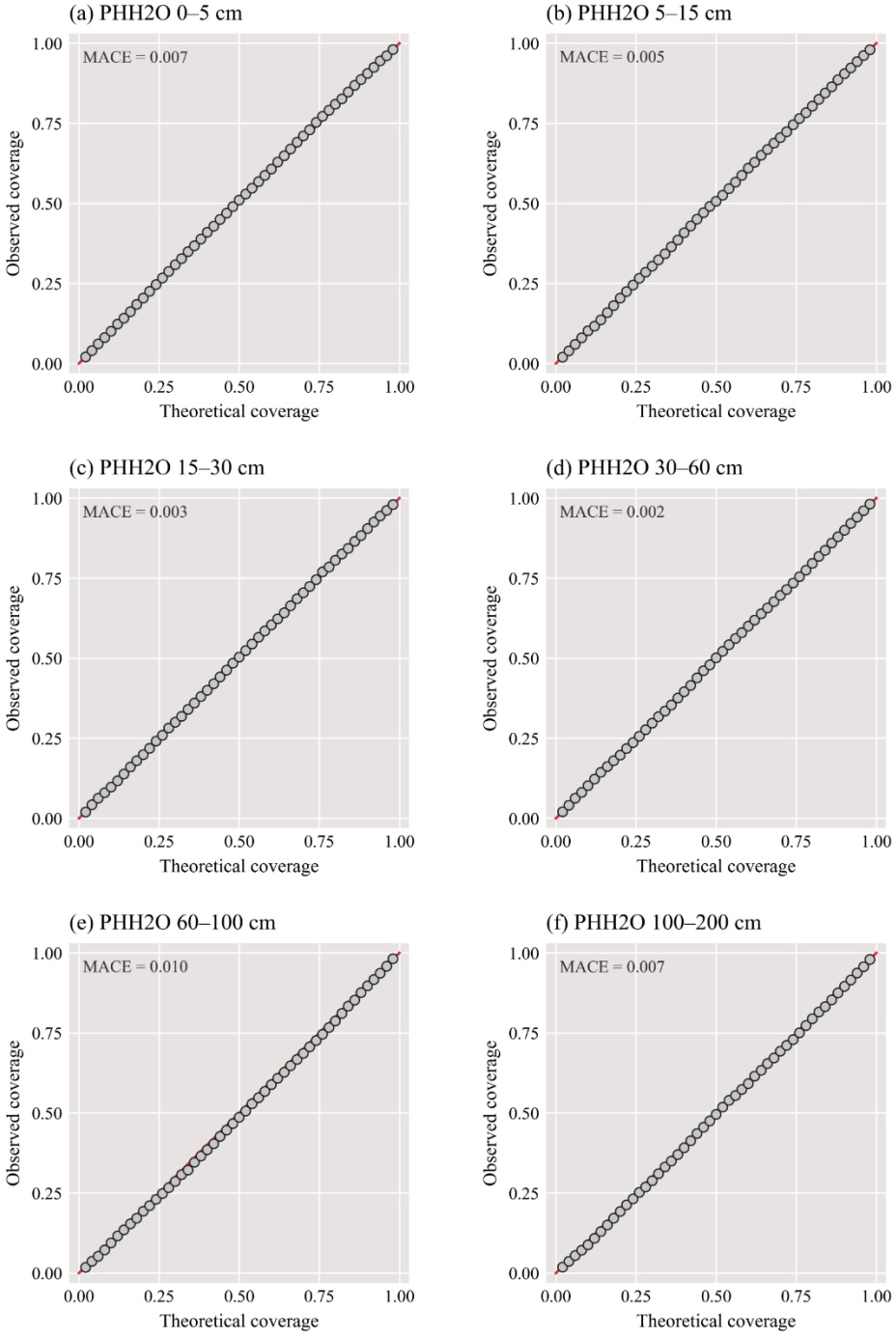


Figure S8. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for pH in water (PHH2O) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.5 Bulk density (BD)

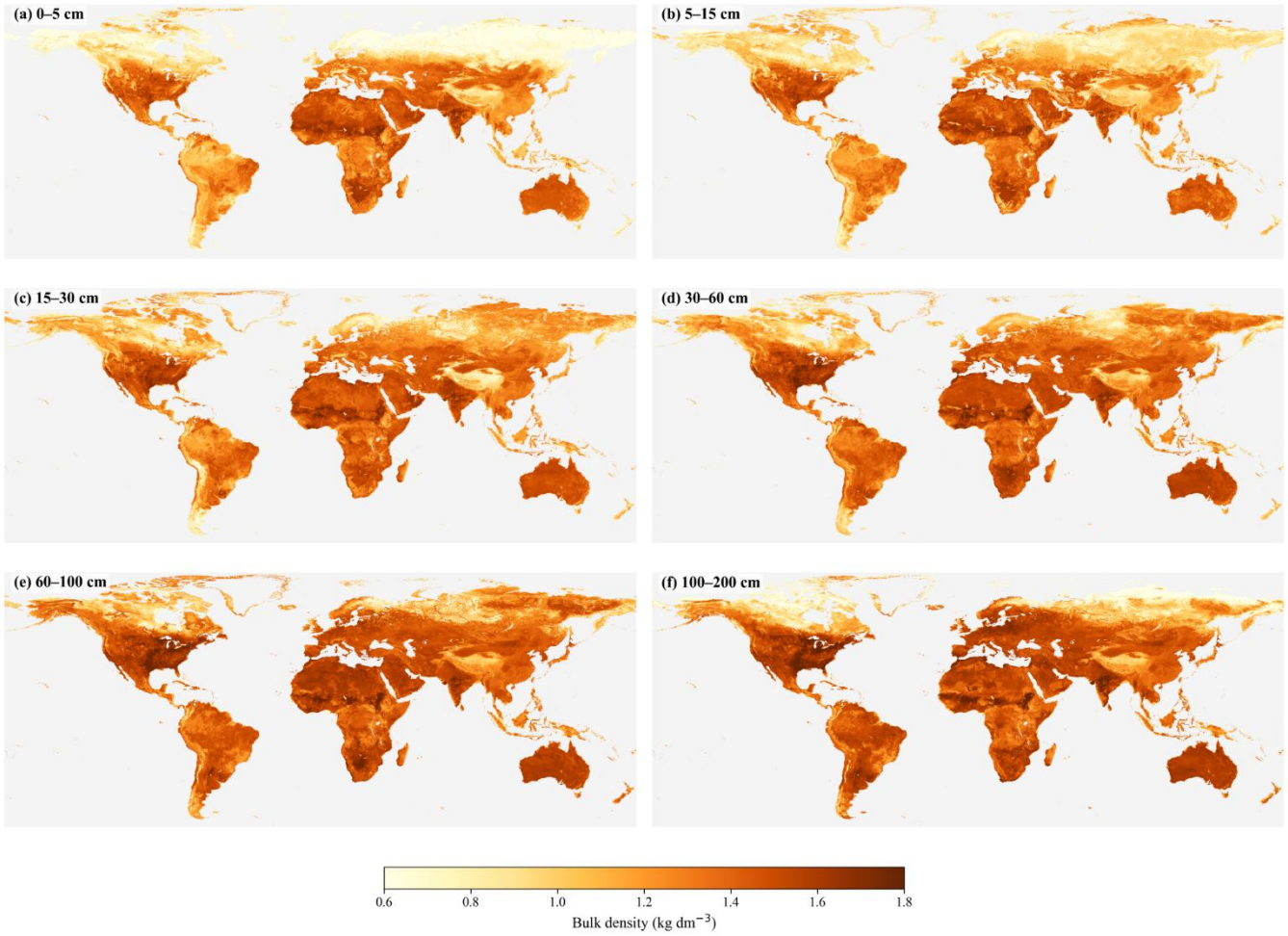


Figure S9. Final GSDEv2 mean predictions of Bulk density (BD) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

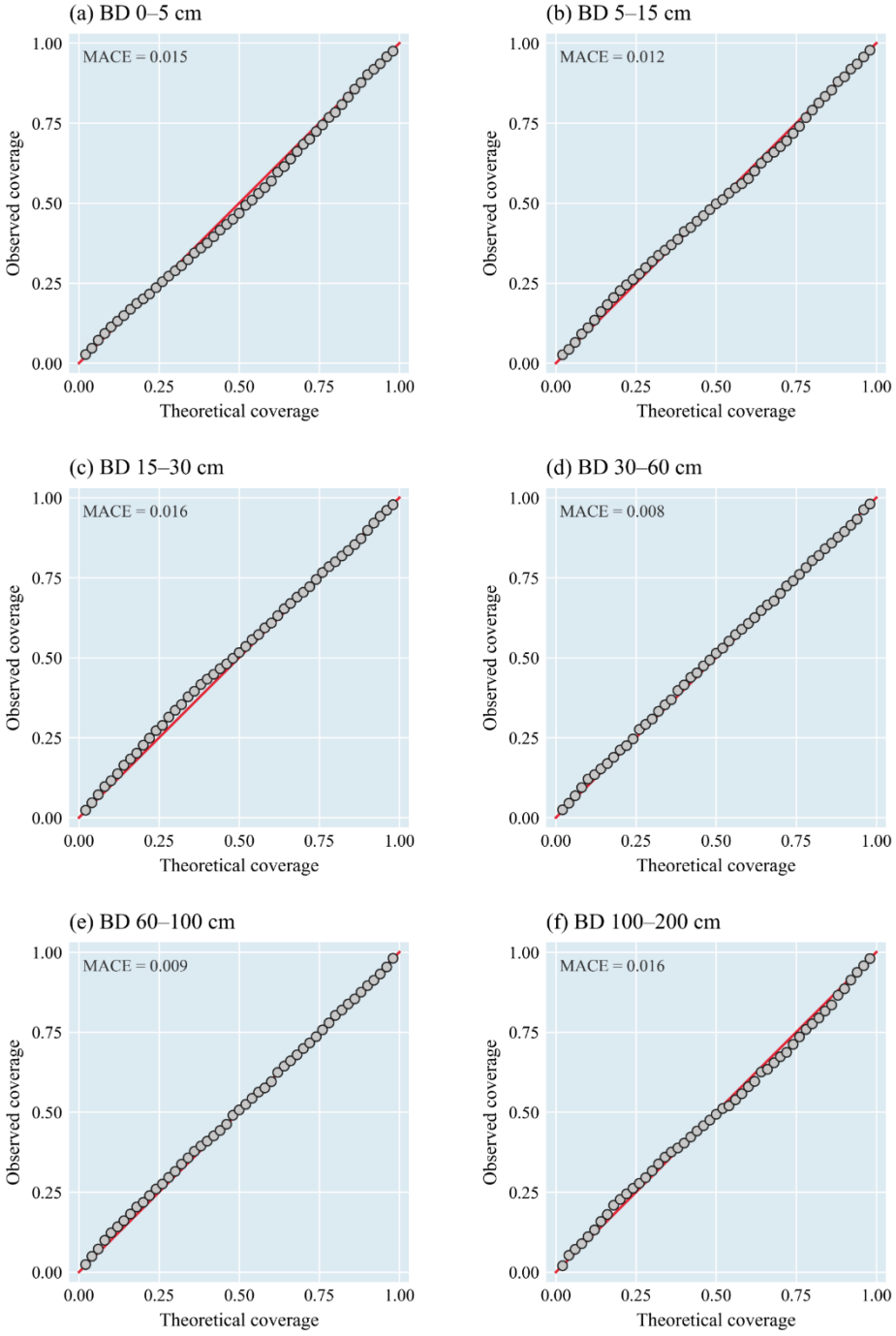


Figure S10. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Bulk density (BD) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.6 Organic carbon concentration (SOC)

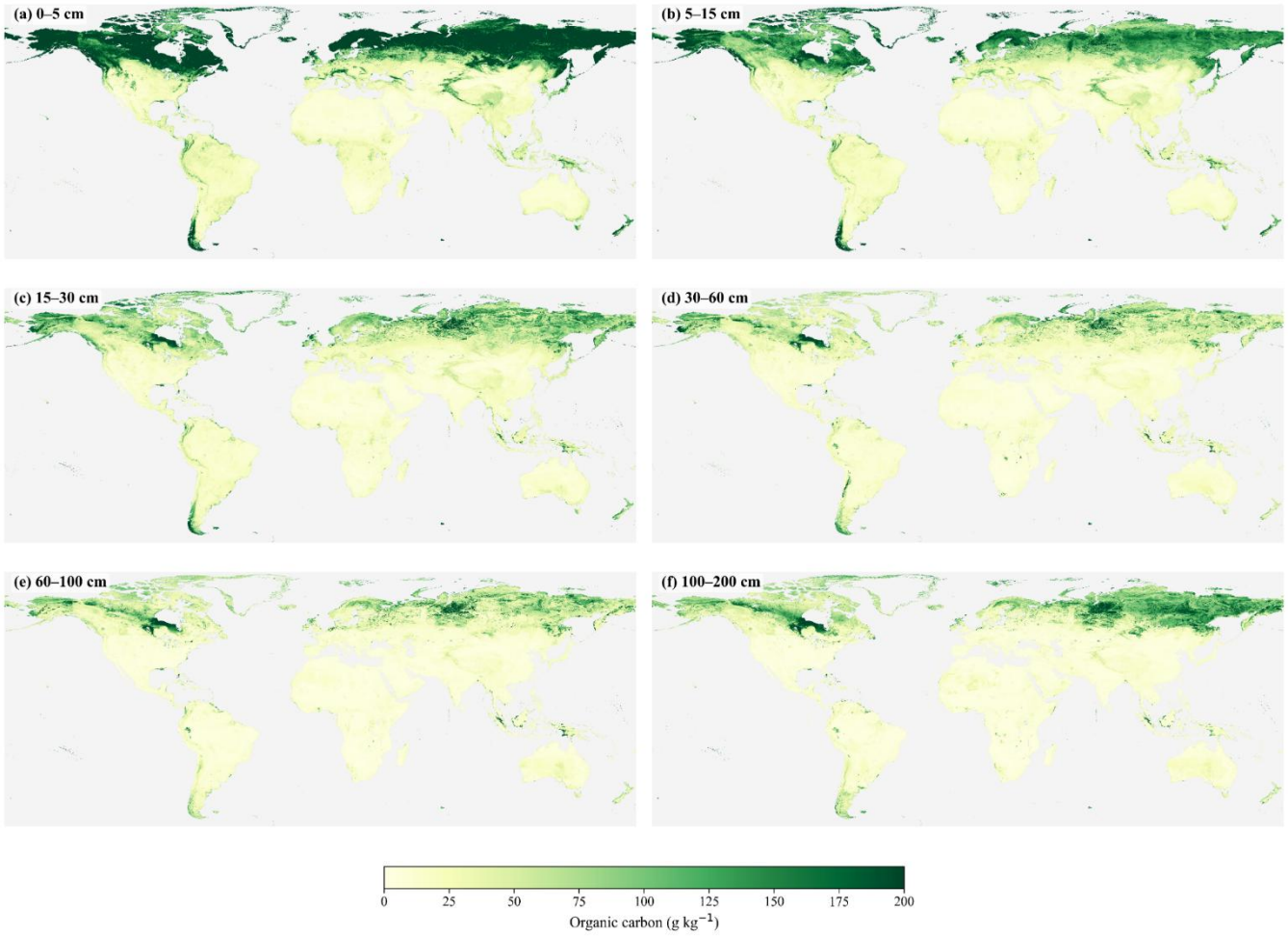


Figure S11. Final GSDEv2 mean predictions of Organic carbon concentration (SOC) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

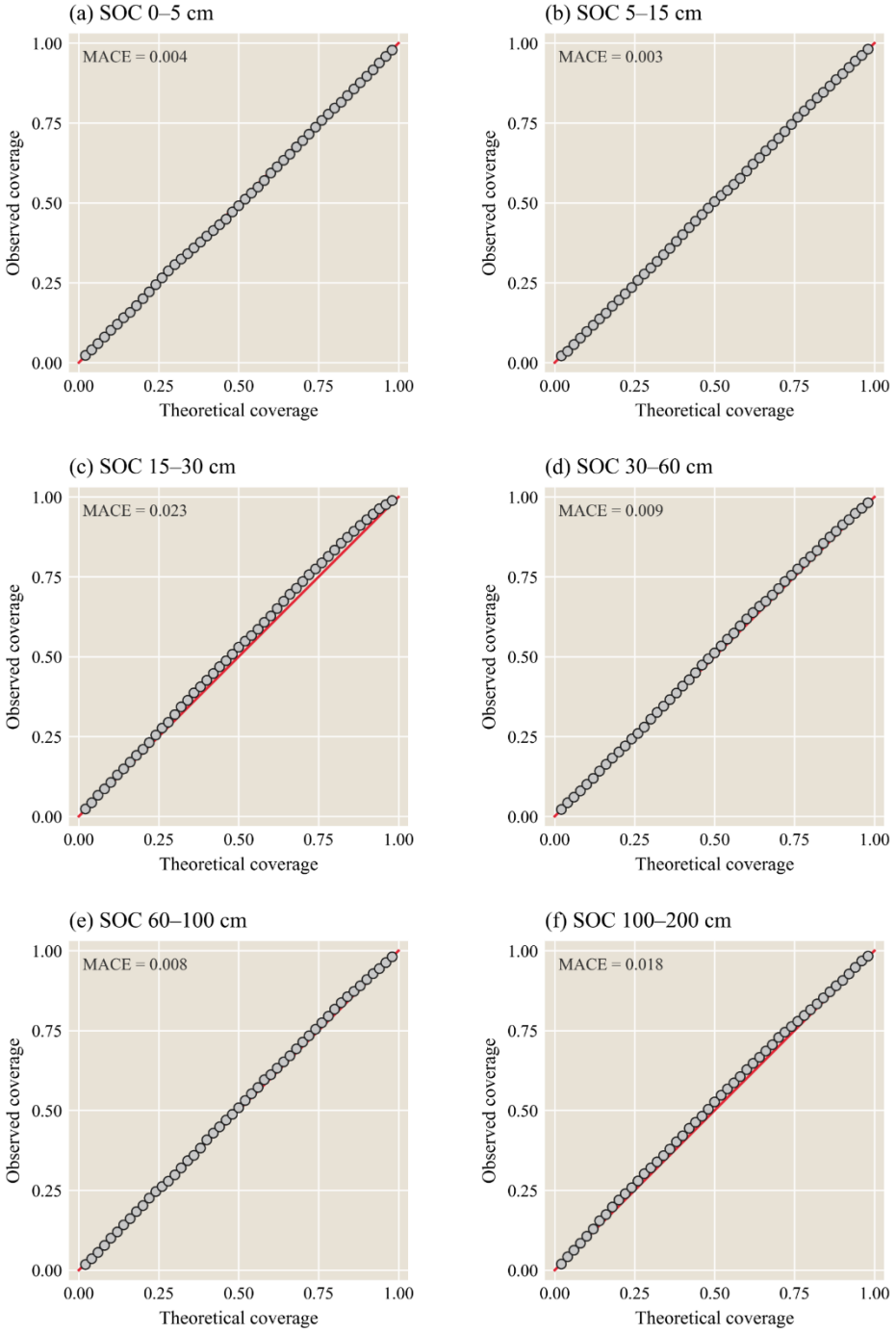


Figure S12. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Organic carbon concentration (SOC) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.7 Total carbon (TC)

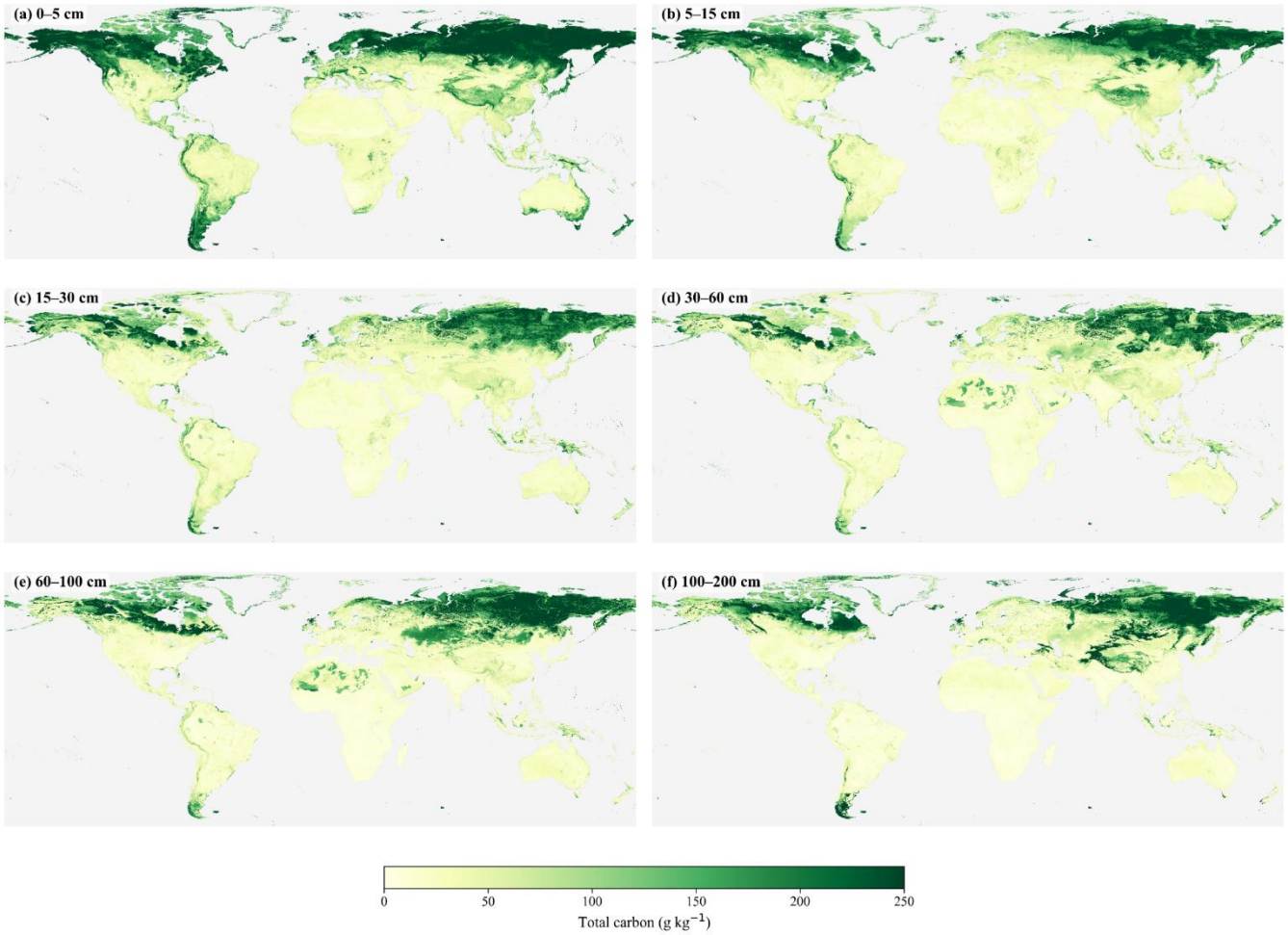
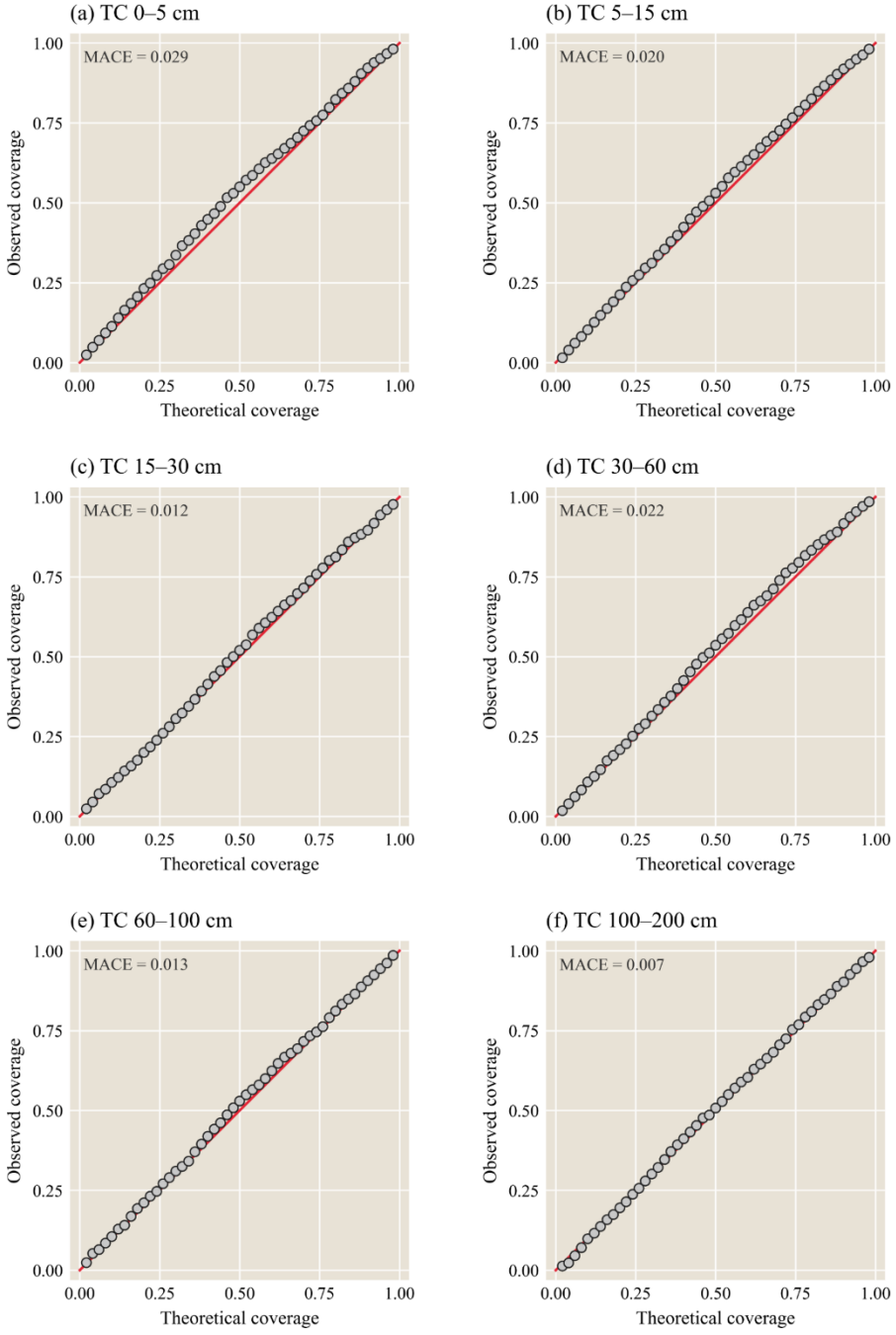


Figure S13. Final GSDEv2 mean predictions of Total carbon (TC) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.



55 **Figure S14.** Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Total carbon (TC) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.8 Total nitrogen (TN)

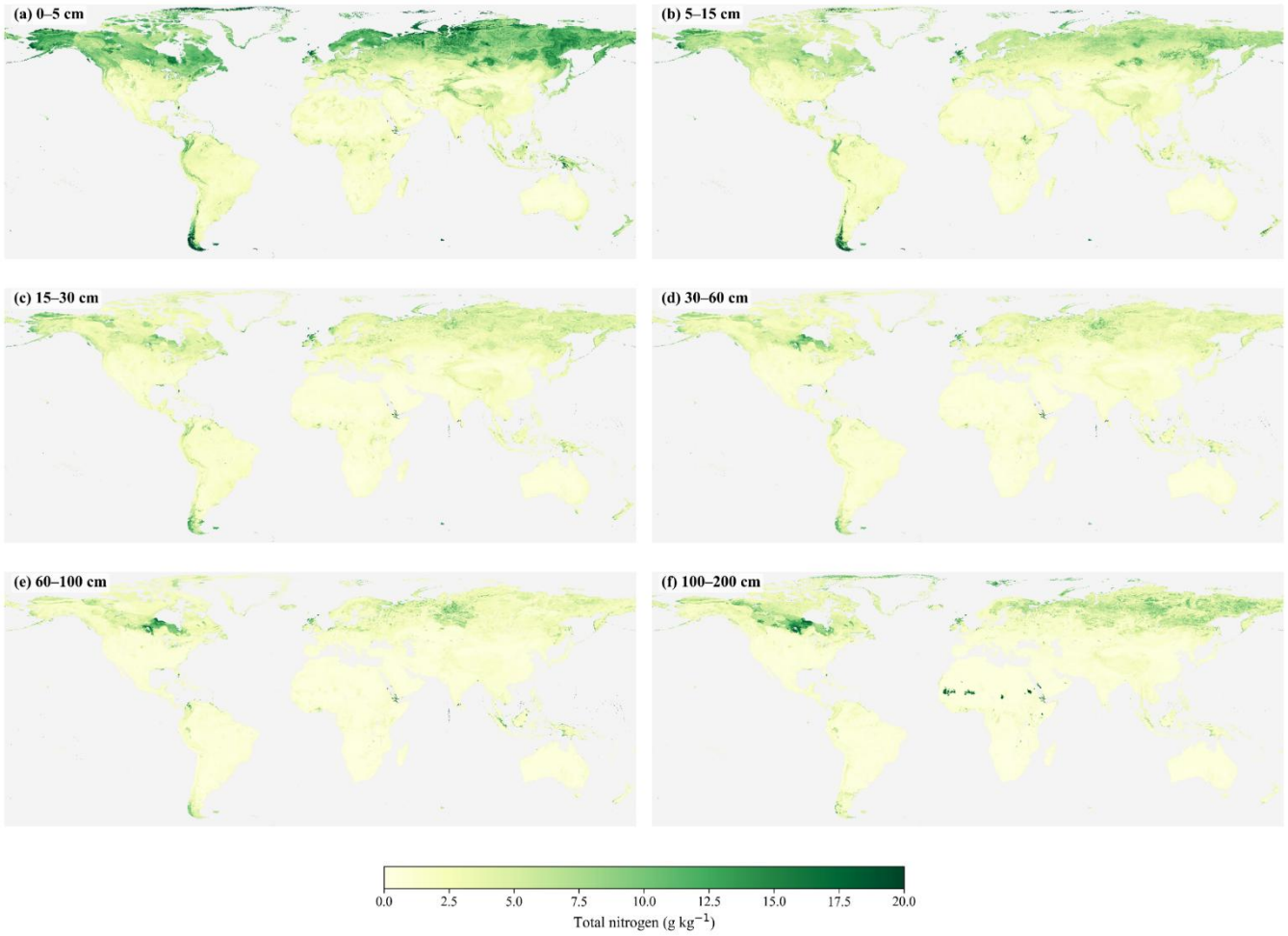


Figure S15. Final GSDEv2 mean predictions of Total nitrogen (TN) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

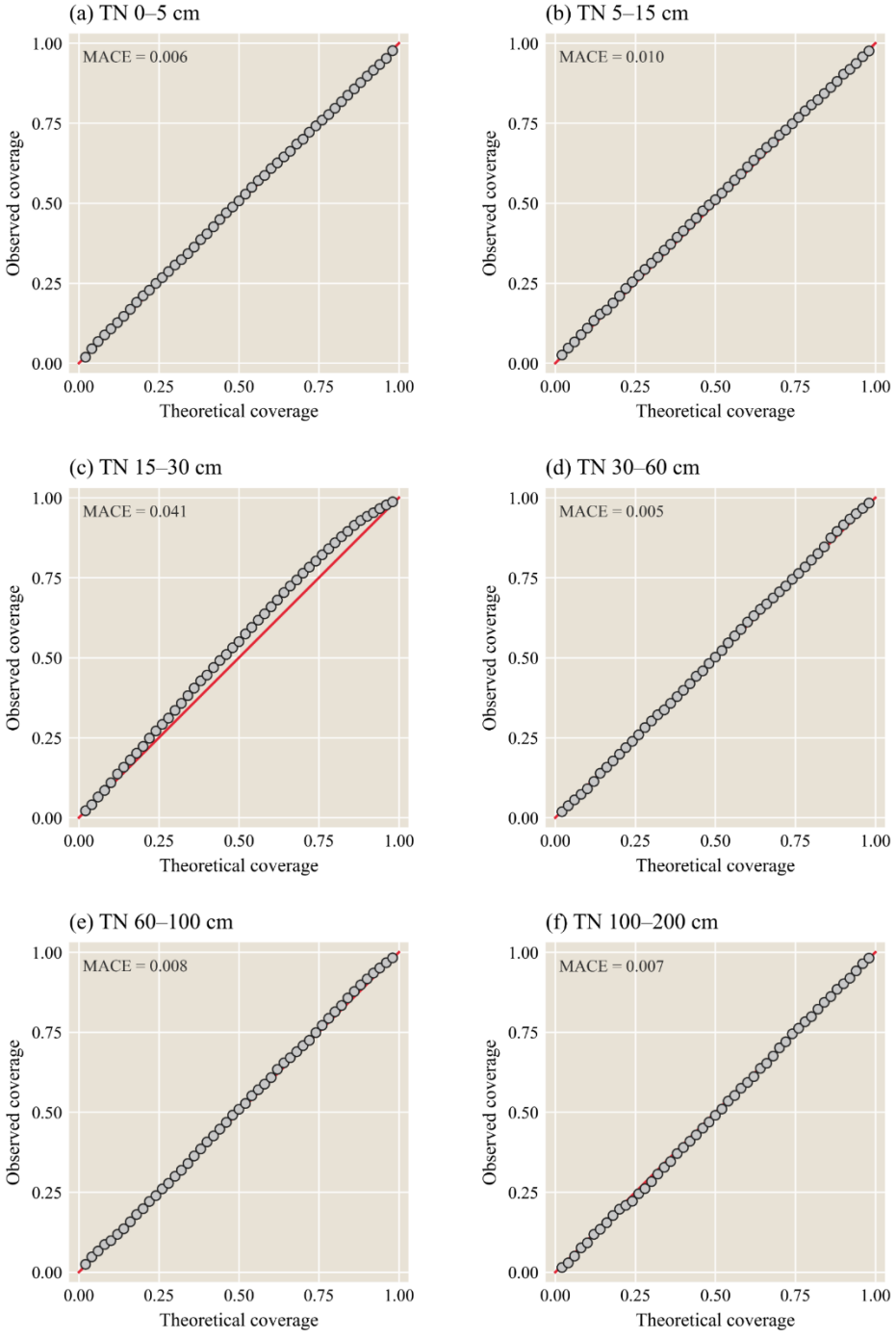


Figure S16. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Total nitrogen (TN) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.9. Effective cation exchange capacity (CEC)

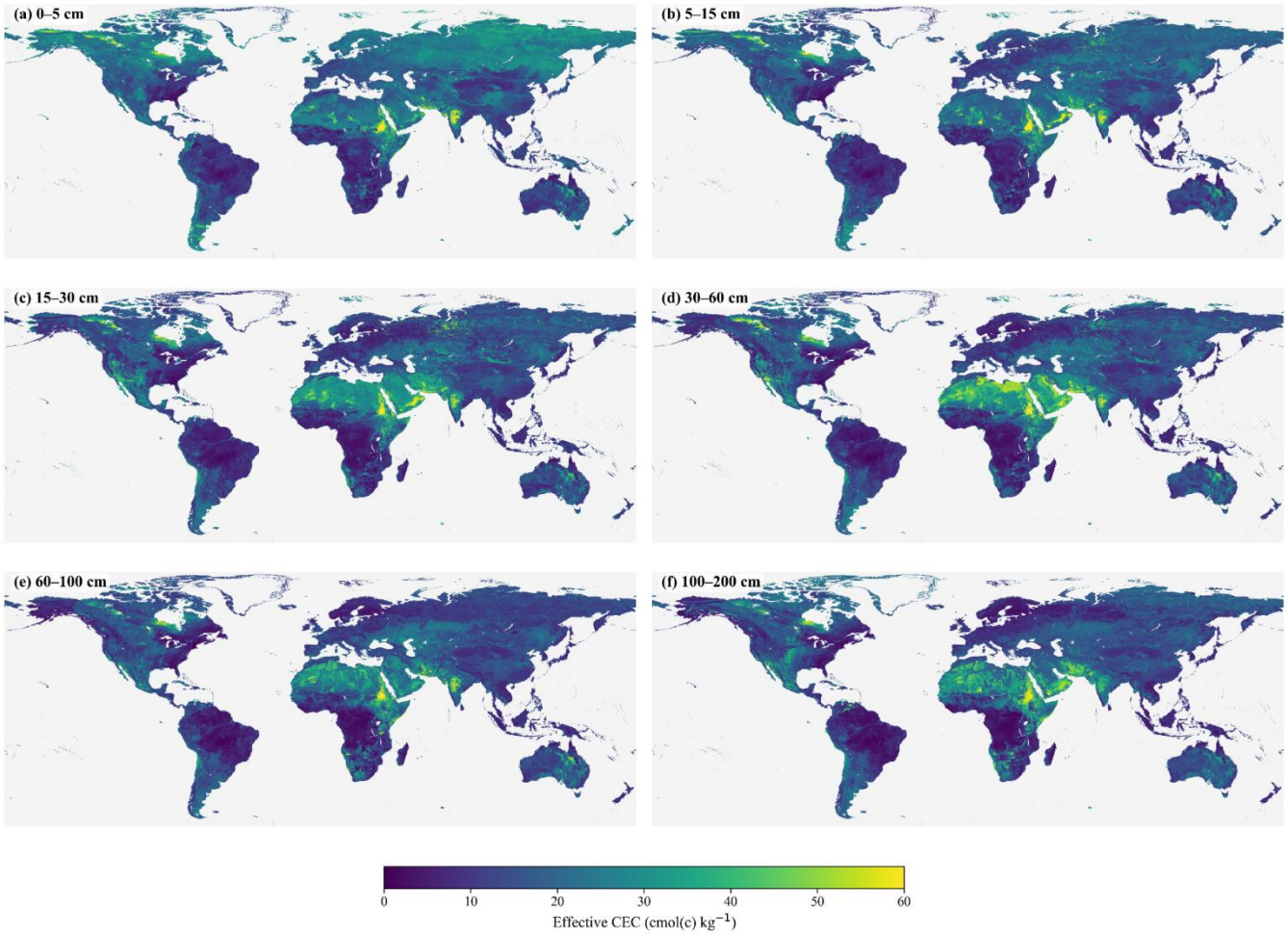


Figure S17. Final GSDEv2 mean predictions of Effective cation exchange capacity (CEC) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

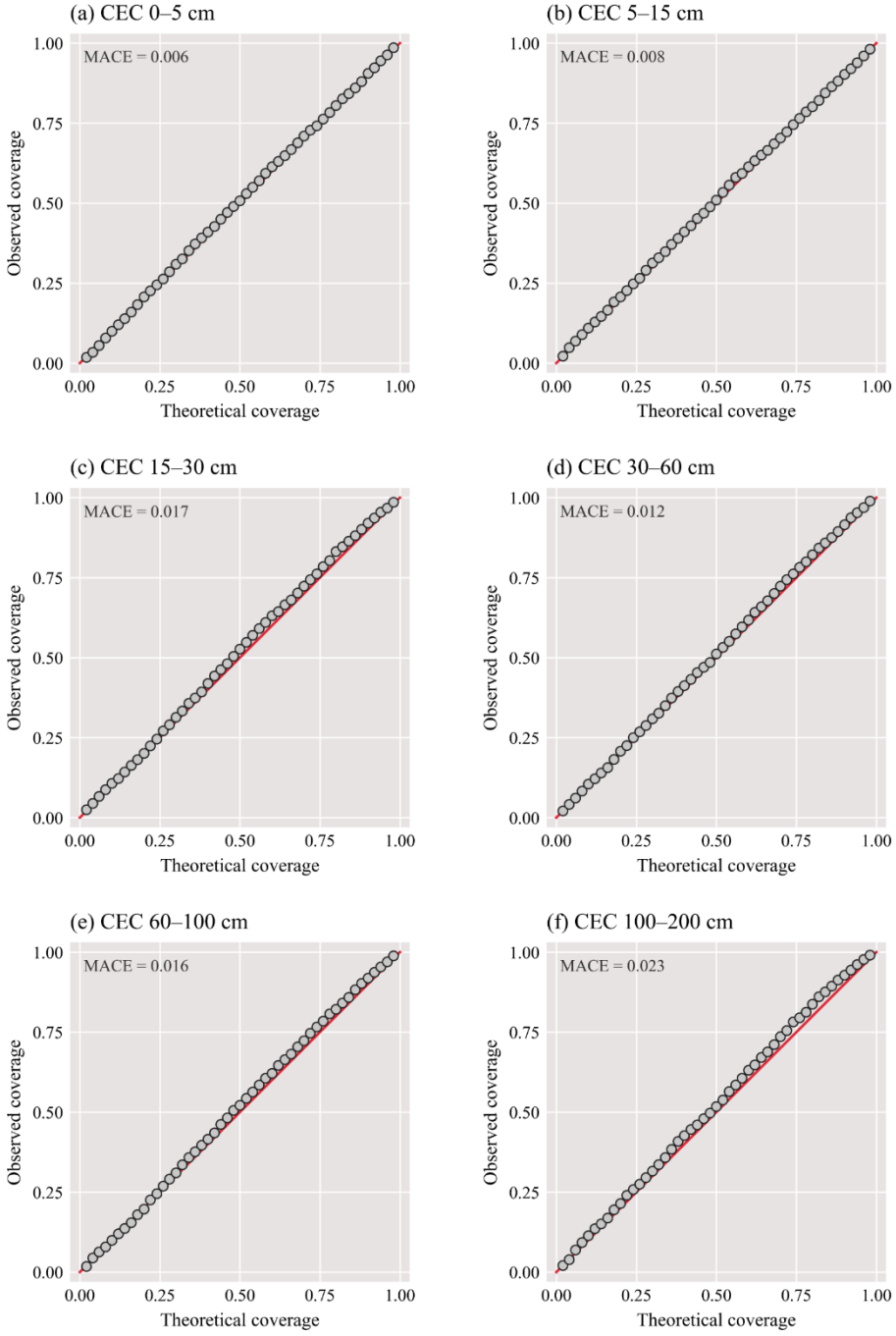


Figure S18. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Effective cation exchange capacity (CEC) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.10 Cation exchange capacity at pH 7 (CECPH7)

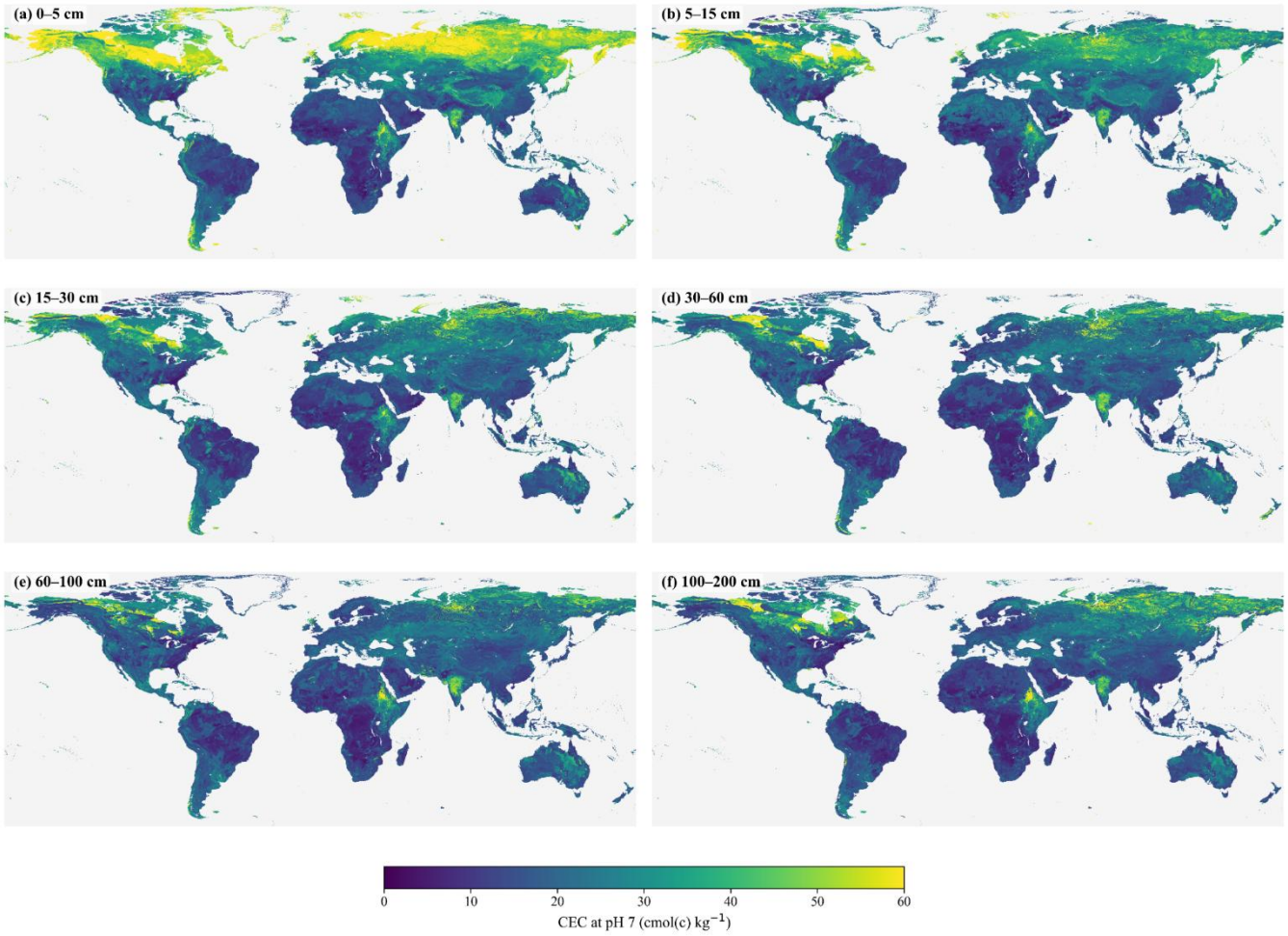


Figure S19. Final GSDEv2 mean predictions of Cation exchange capacity at pH 7 (CECPH7) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

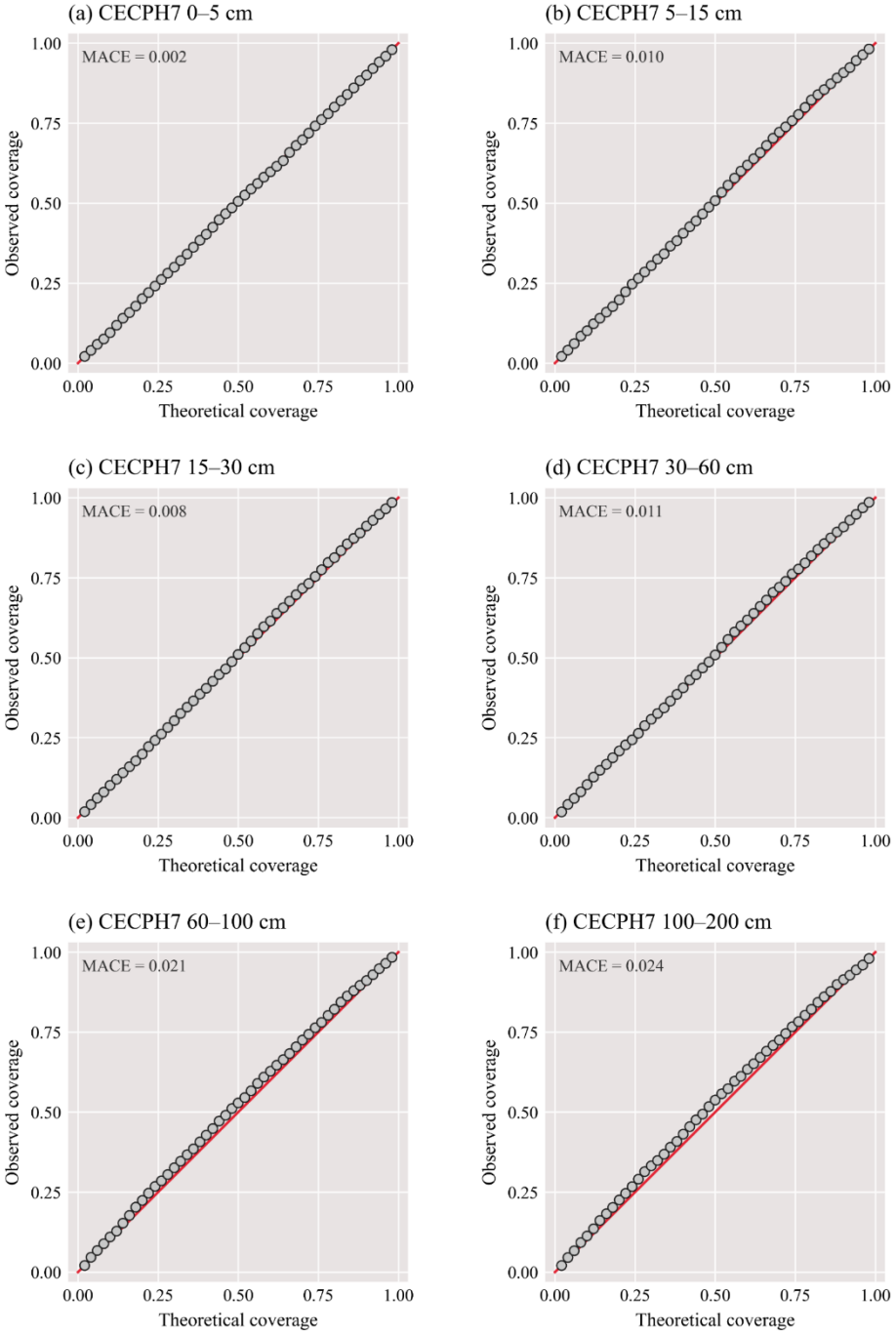


Figure S20. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Cation exchange capacity at pH 7 (CECPH7) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.11 Volumetric coarse fragments (cfvo)

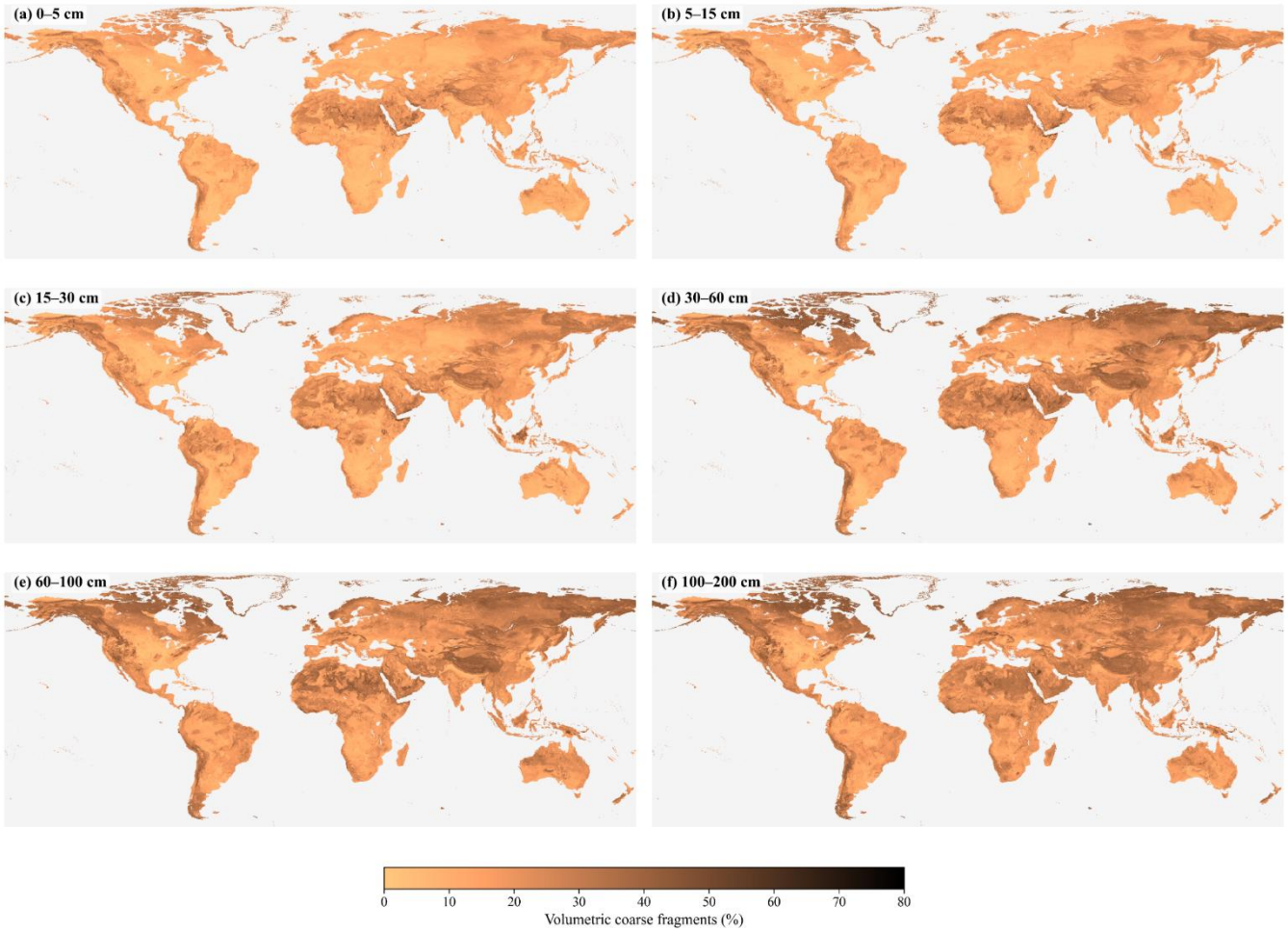


Figure S21. Final GSDEv2 mean predictions of Volumetric coarse fragments (cfvo) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

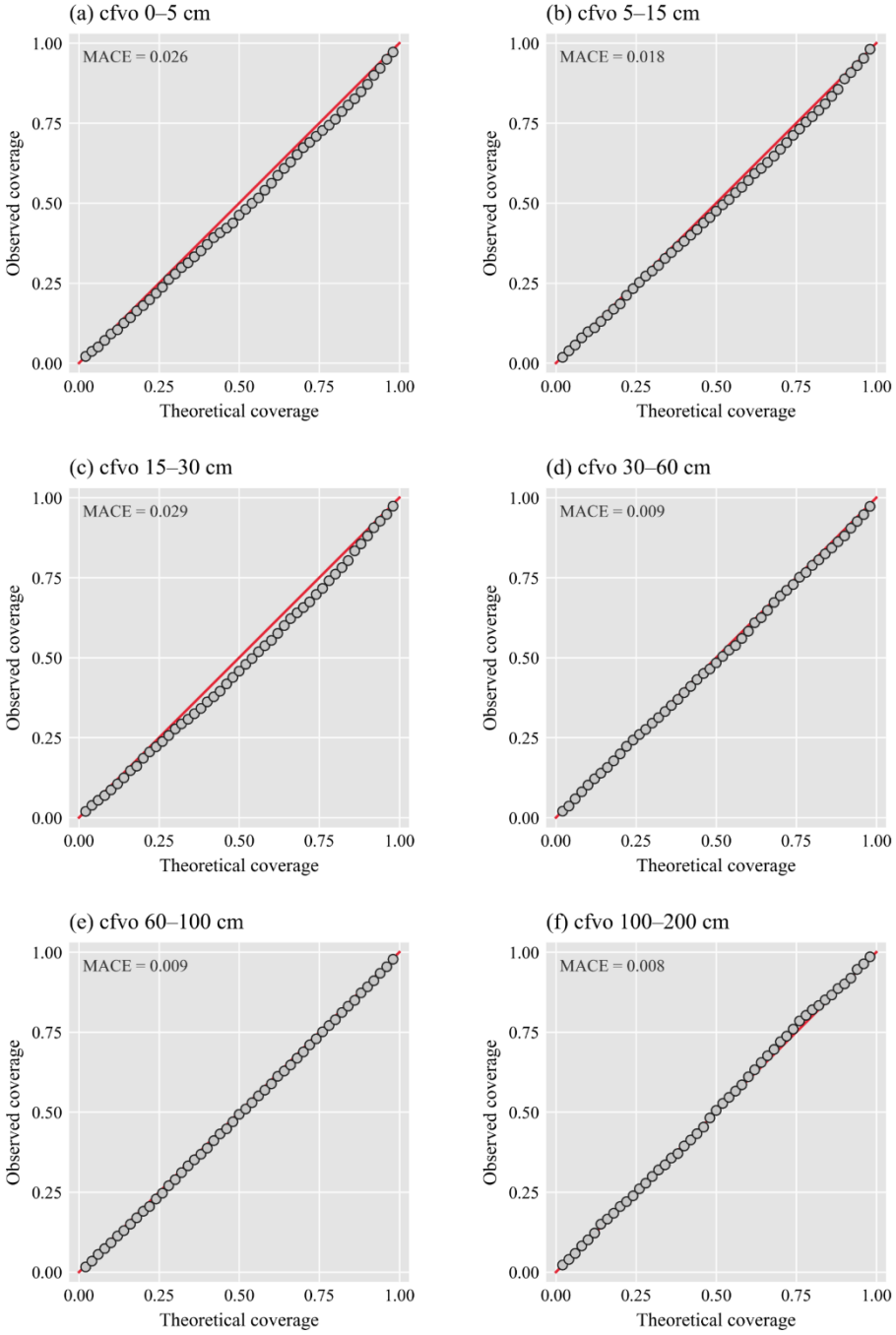


Figure S22. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Volumetric coarse fragments (cfvo) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.12 Gravimetric coarse fragments (cfgr)

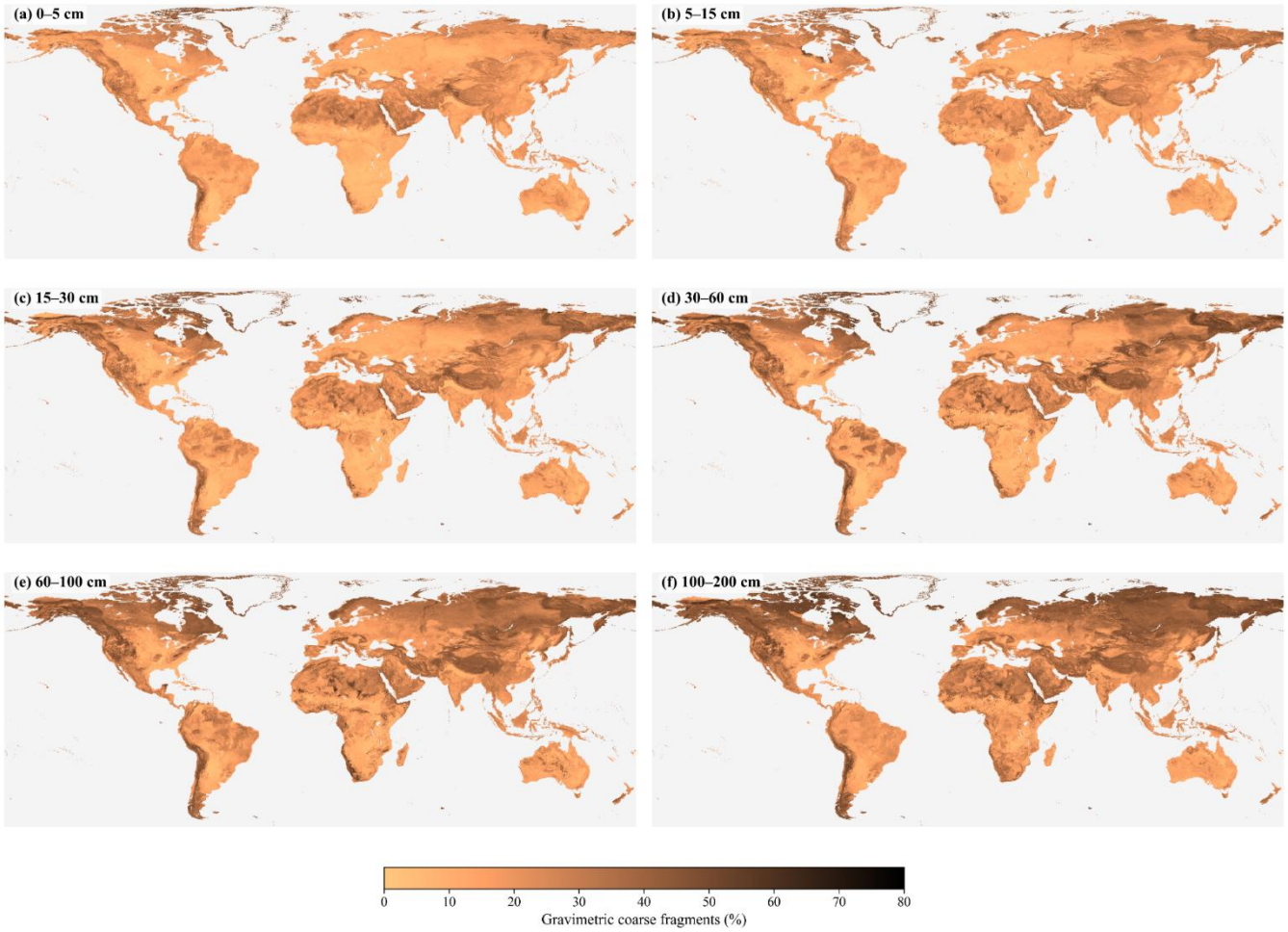


Figure S23. Final GSDEv2 mean predictions of Gravimetric coarse fragments (cfgr) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

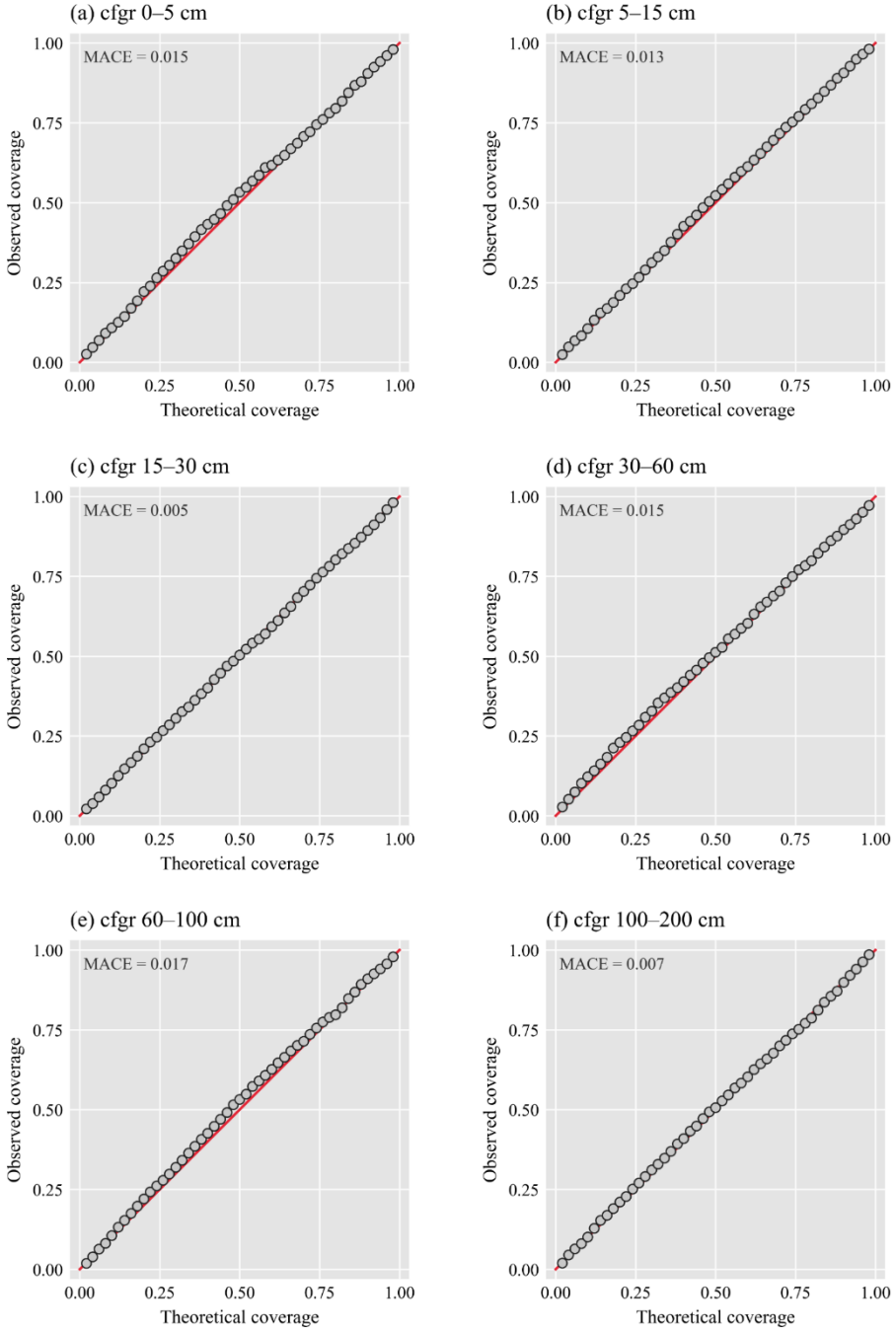


Figure S24. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Gravimetric coarse fragments (cfgr) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.13 Fine-earth bulk density at 33 kPa (bdfi33)

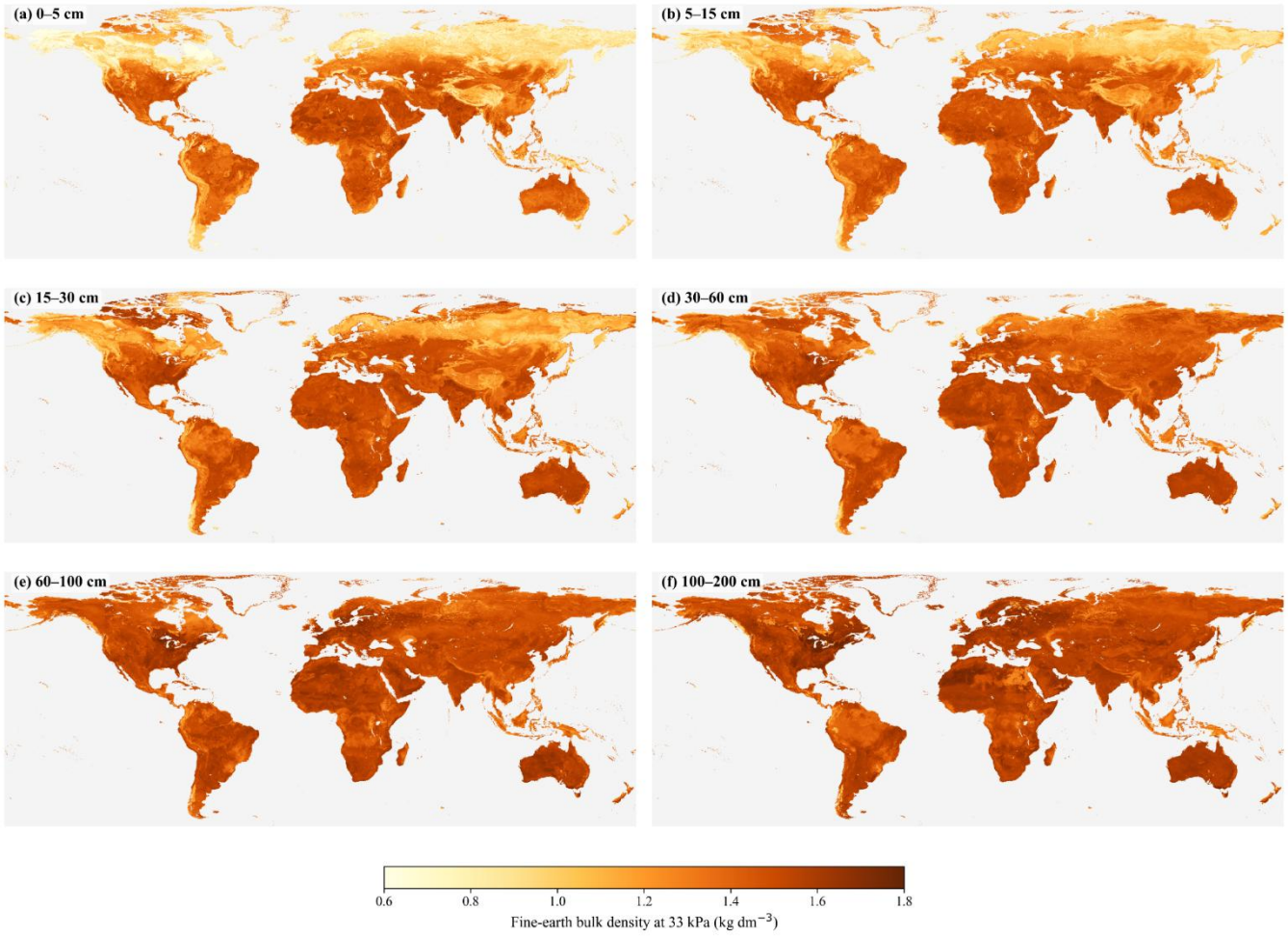


Figure S25. Final GSDEv2 mean predictions of Fine-earth bulk density at 33 kPa (bdfi33) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

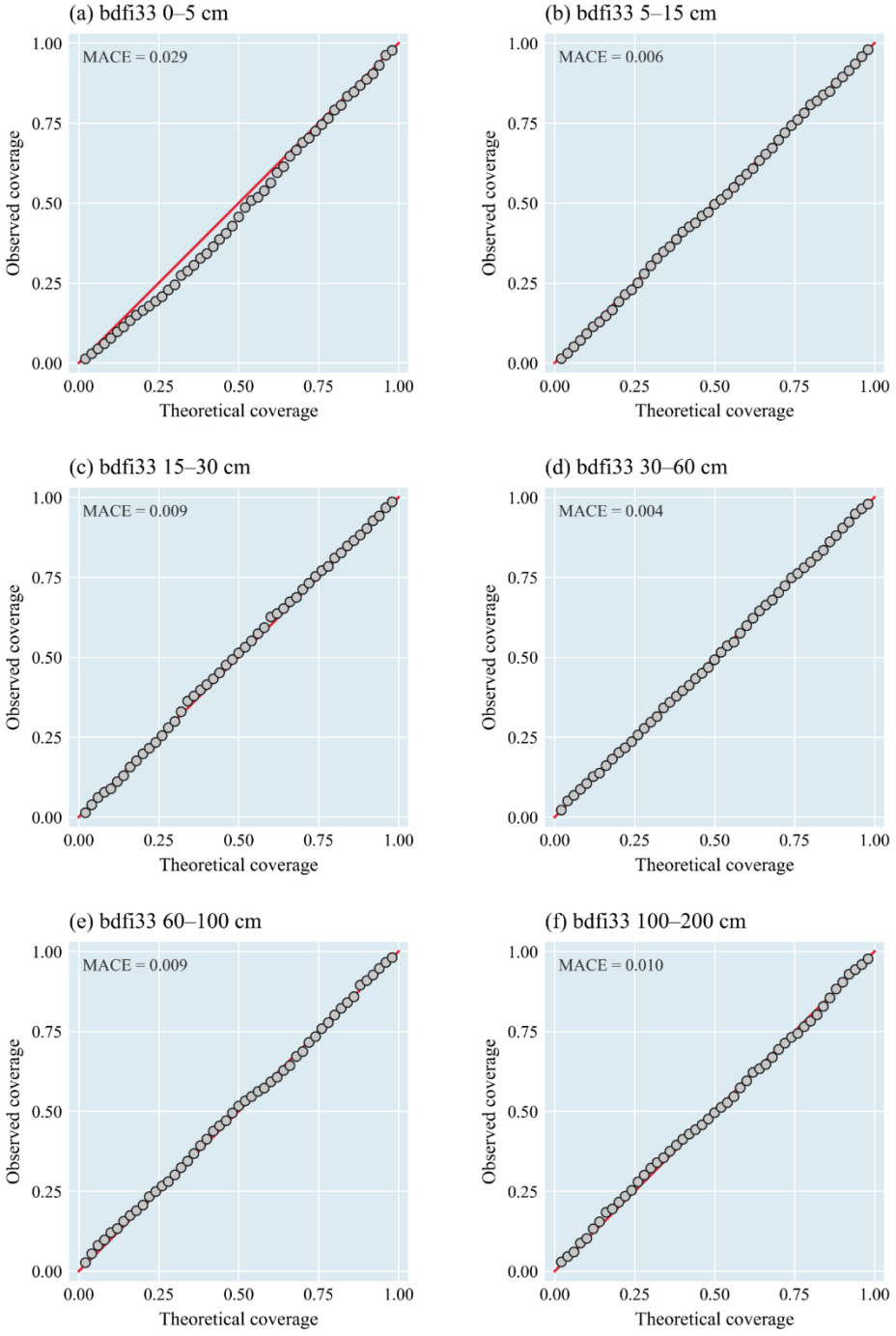


Figure S26. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Fine-earth bulk density at 33 kPa (bdfi33) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.14 Volumetric water content at 33 kPa (wv0033)

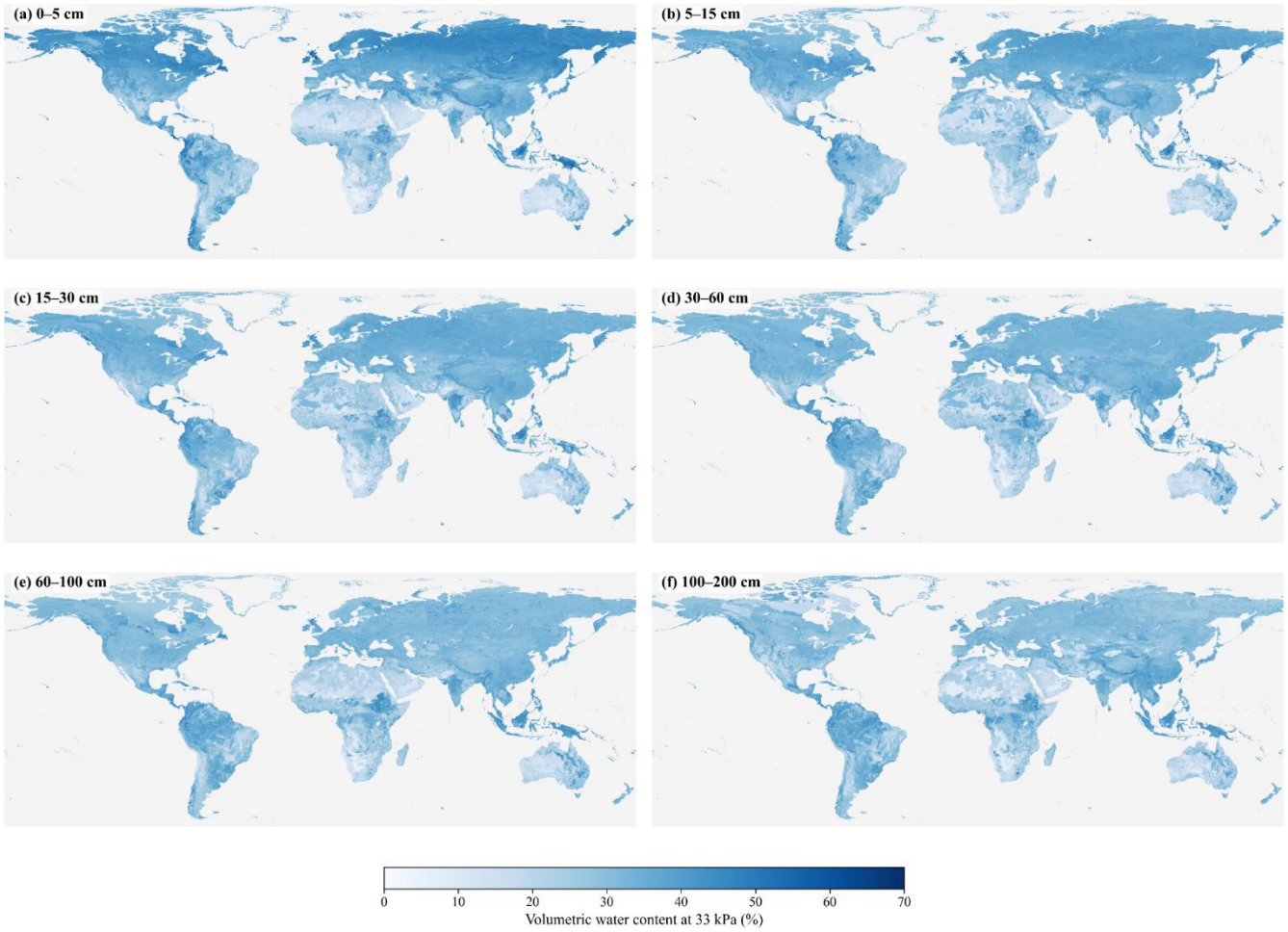


Figure S27. Final GSDEv2 mean predictions of Volumetric water content at 33 kPa (wv0033) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

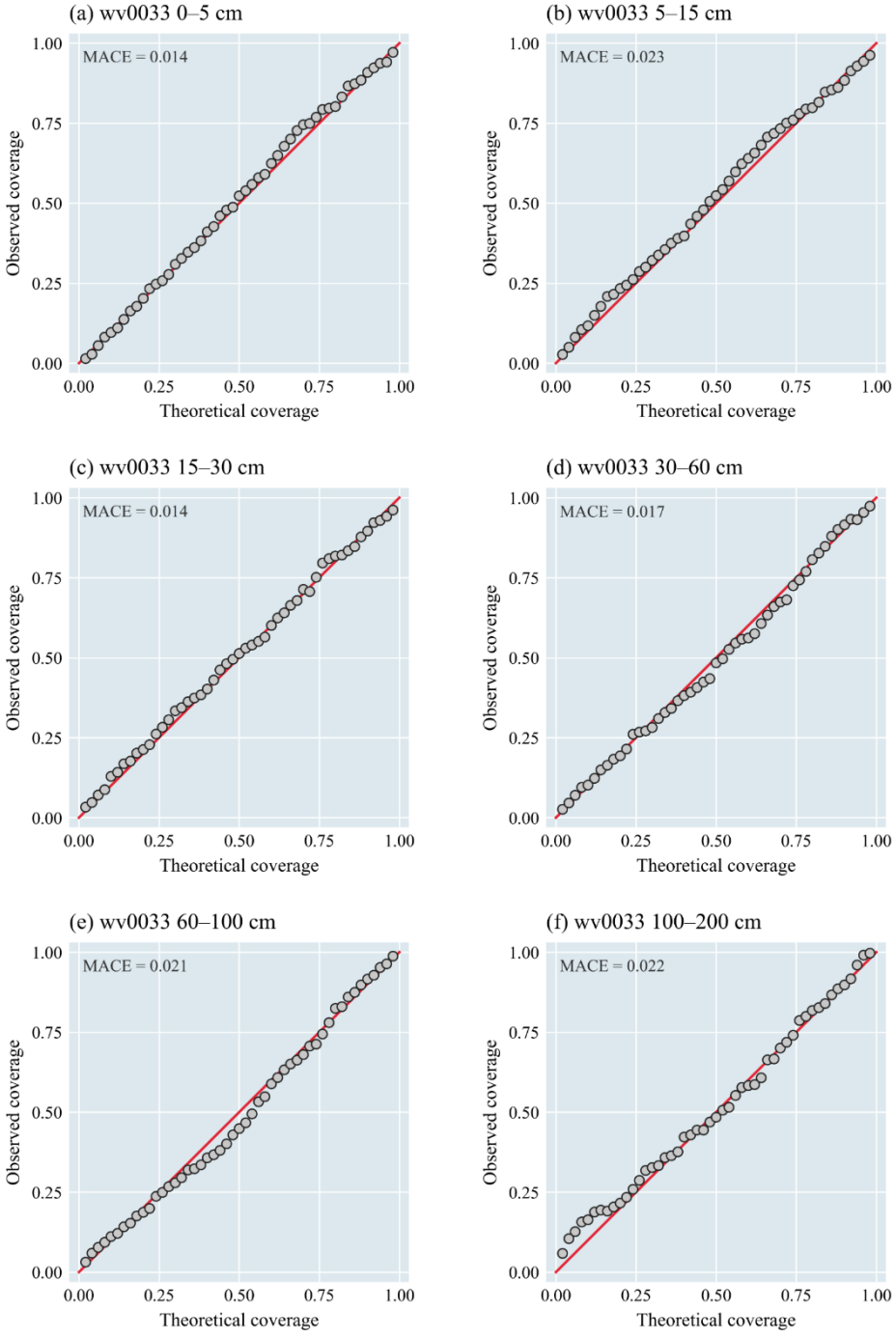
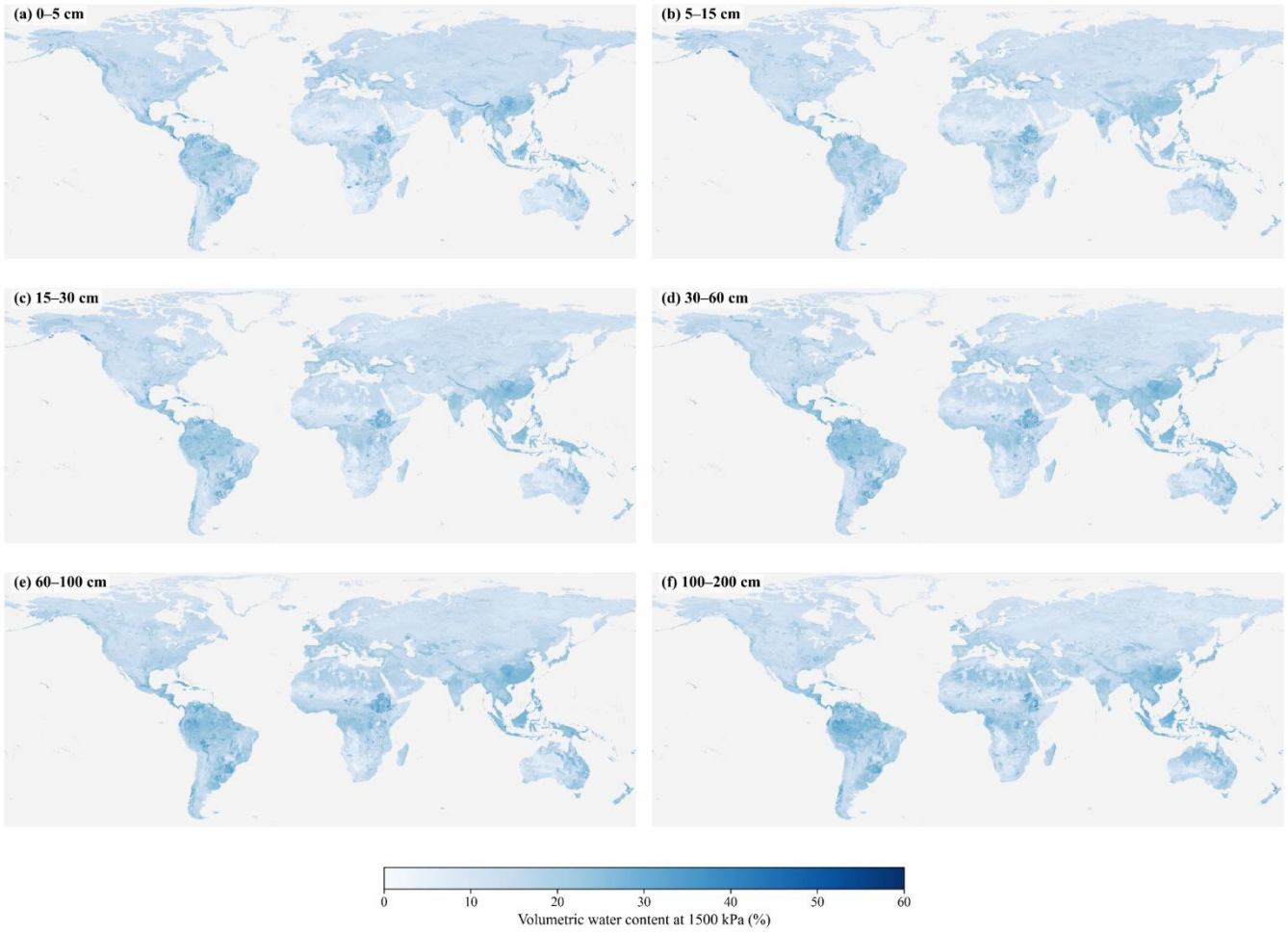


Figure S28. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Volumetric water content at 33 kPa (wv0033) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.15 Volumetric water content at 1500 kPa (wv1500)



115

Figure S29. Final GSDEv2 mean predictions of Volumetric water content at 1500 kPa (wv1500) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

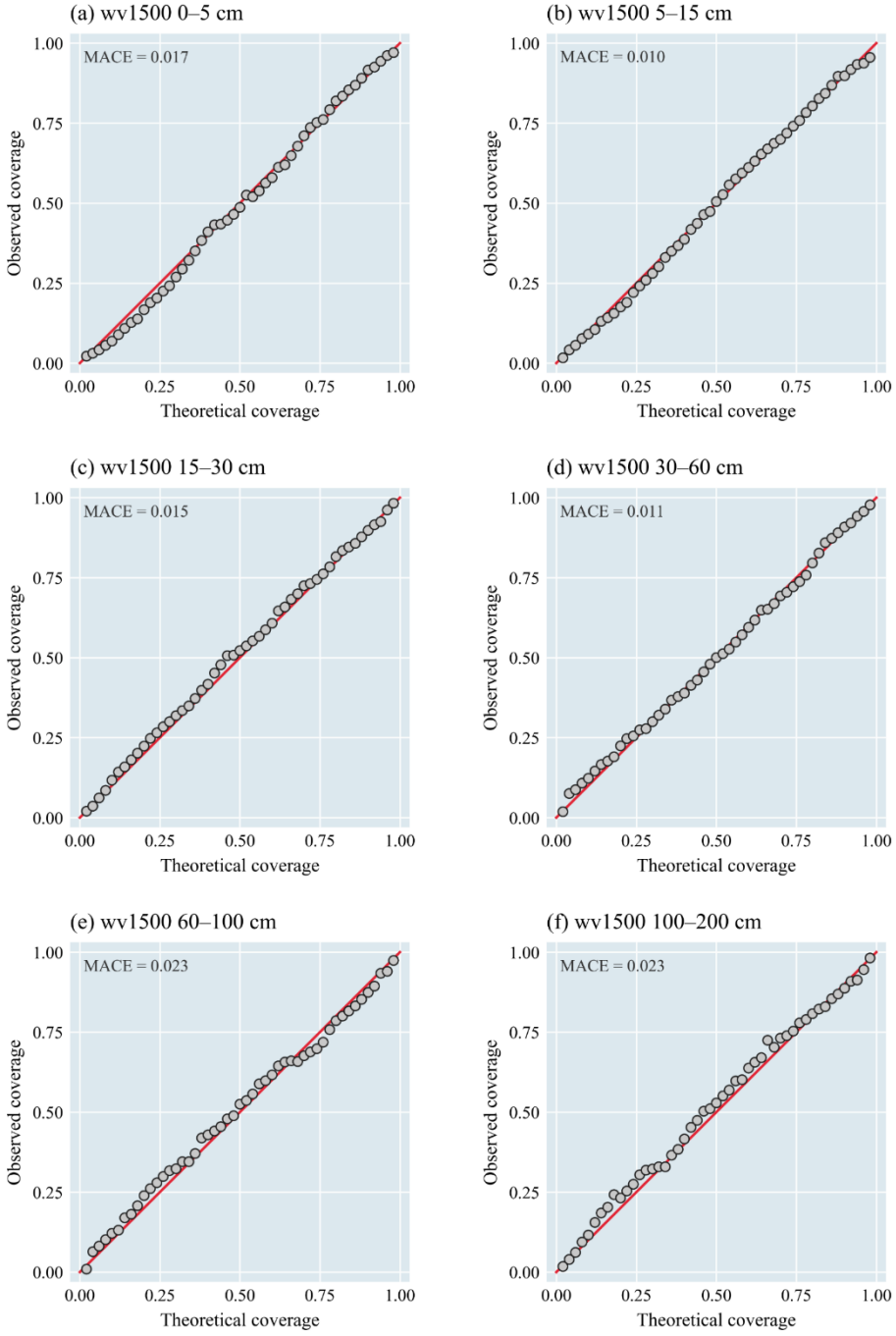


Figure S30. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Volumetric water content at 1500 kPa (wv1500) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.16 Volumetric water content at 10 kPa (wv0010)

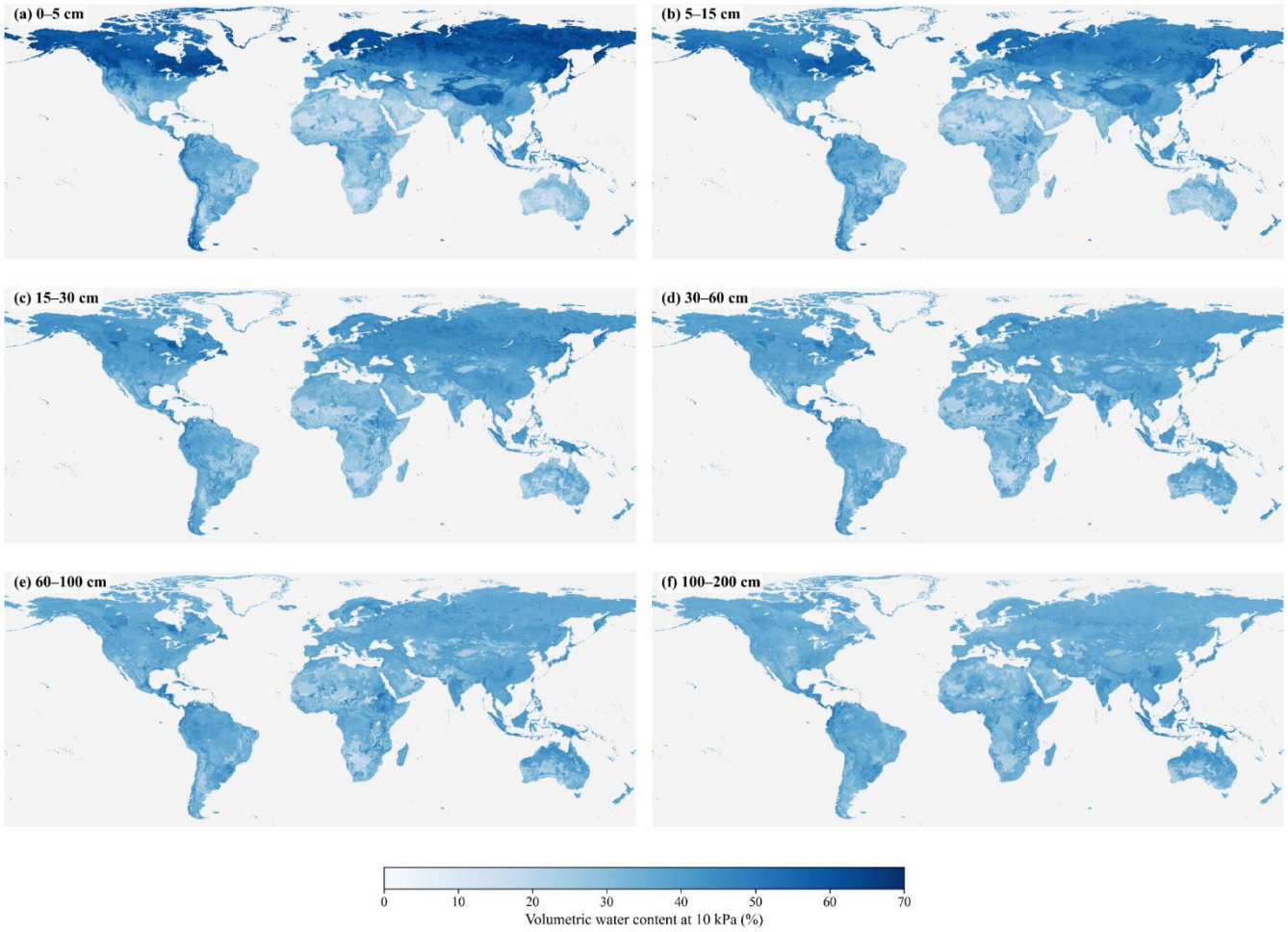


Figure S31. Final GSDEv2 mean predictions of Volumetric water content at 10 kPa (wv0010) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

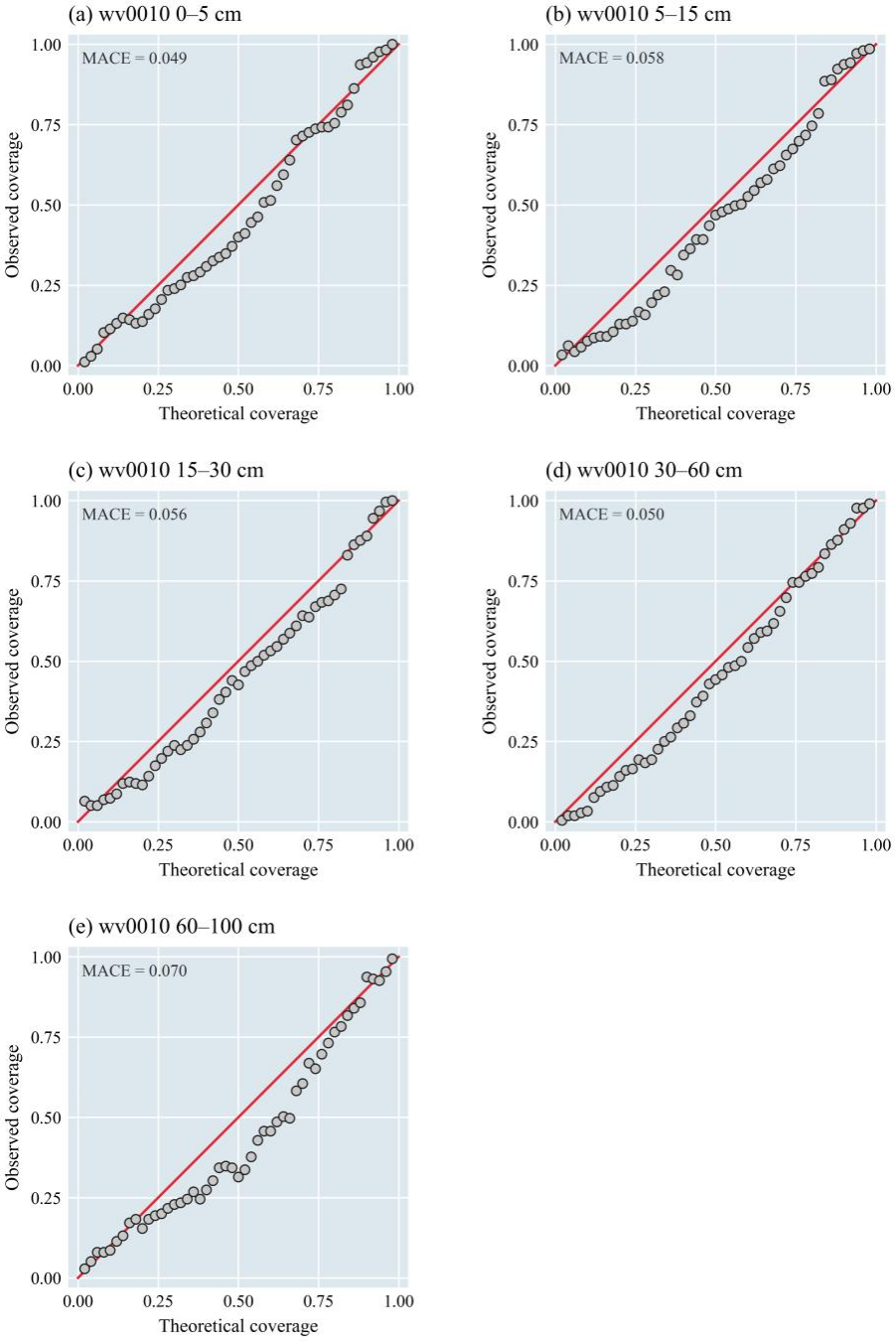


Figure S32. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Volumetric water content at 10 kPa (wv0010) at five soil depths. Panels (a–e) correspond to 0–5, 5–15, 15–30, 30–60, and 60–100 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.17 Gravimetric water content at 33 kPa (wg0033)

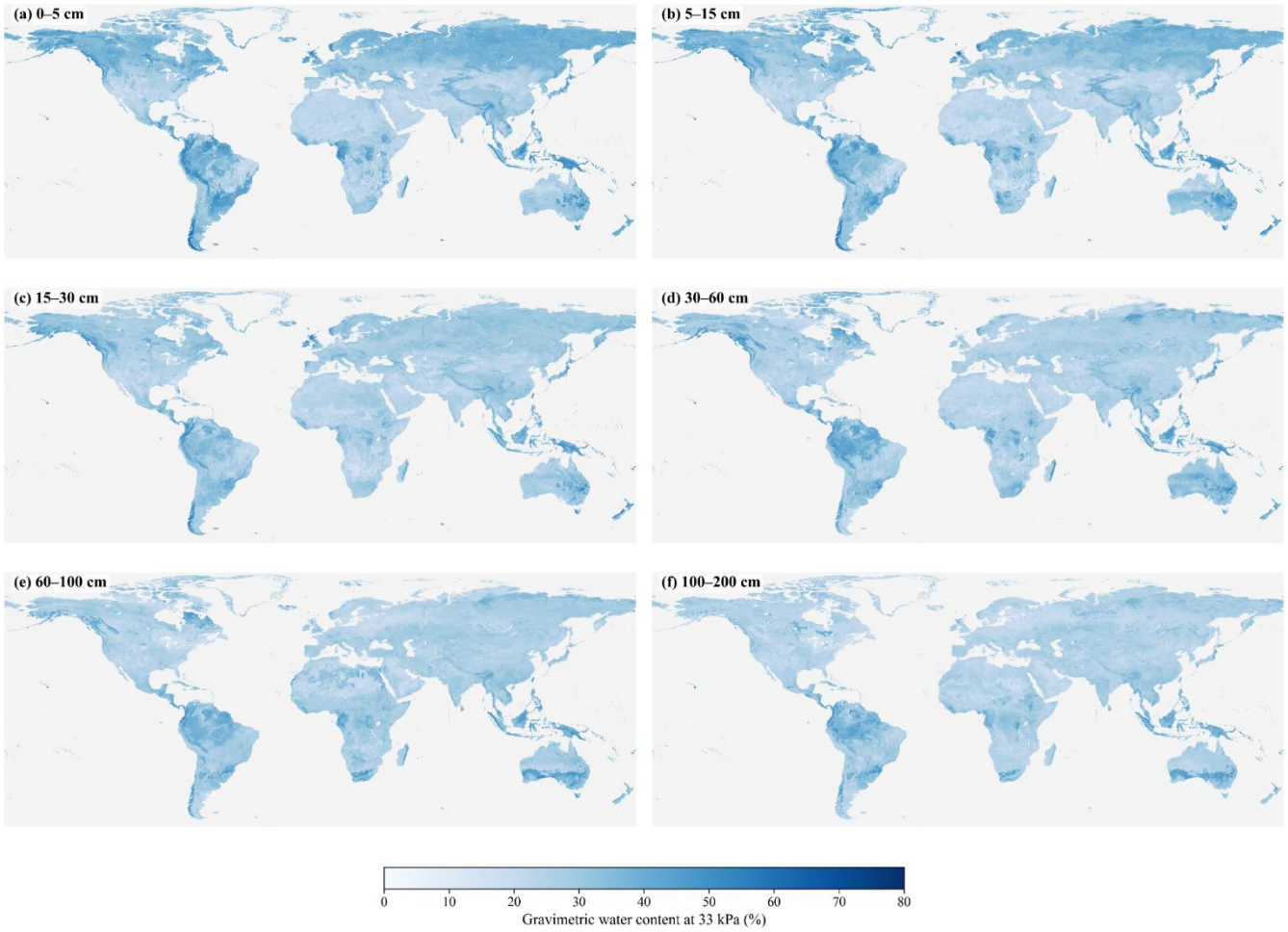
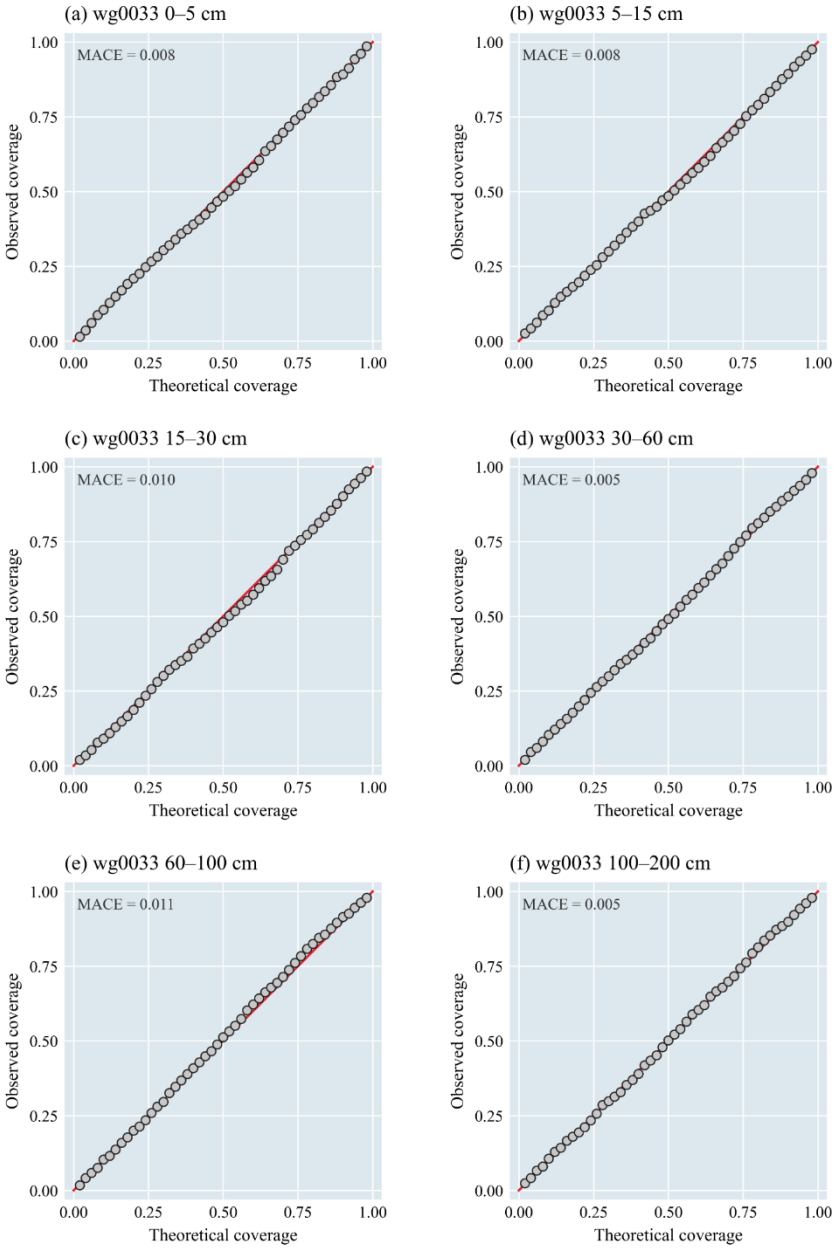


Figure S33. Final GSDEv2 mean predictions of Gravimetric water content at 33 kPa (wg0033) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.



135 **Figure S34.** Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Gravimetric water content at 33 kPa (wg0033) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.18 Gravimetric water content at 1500 kPa (wg1500)

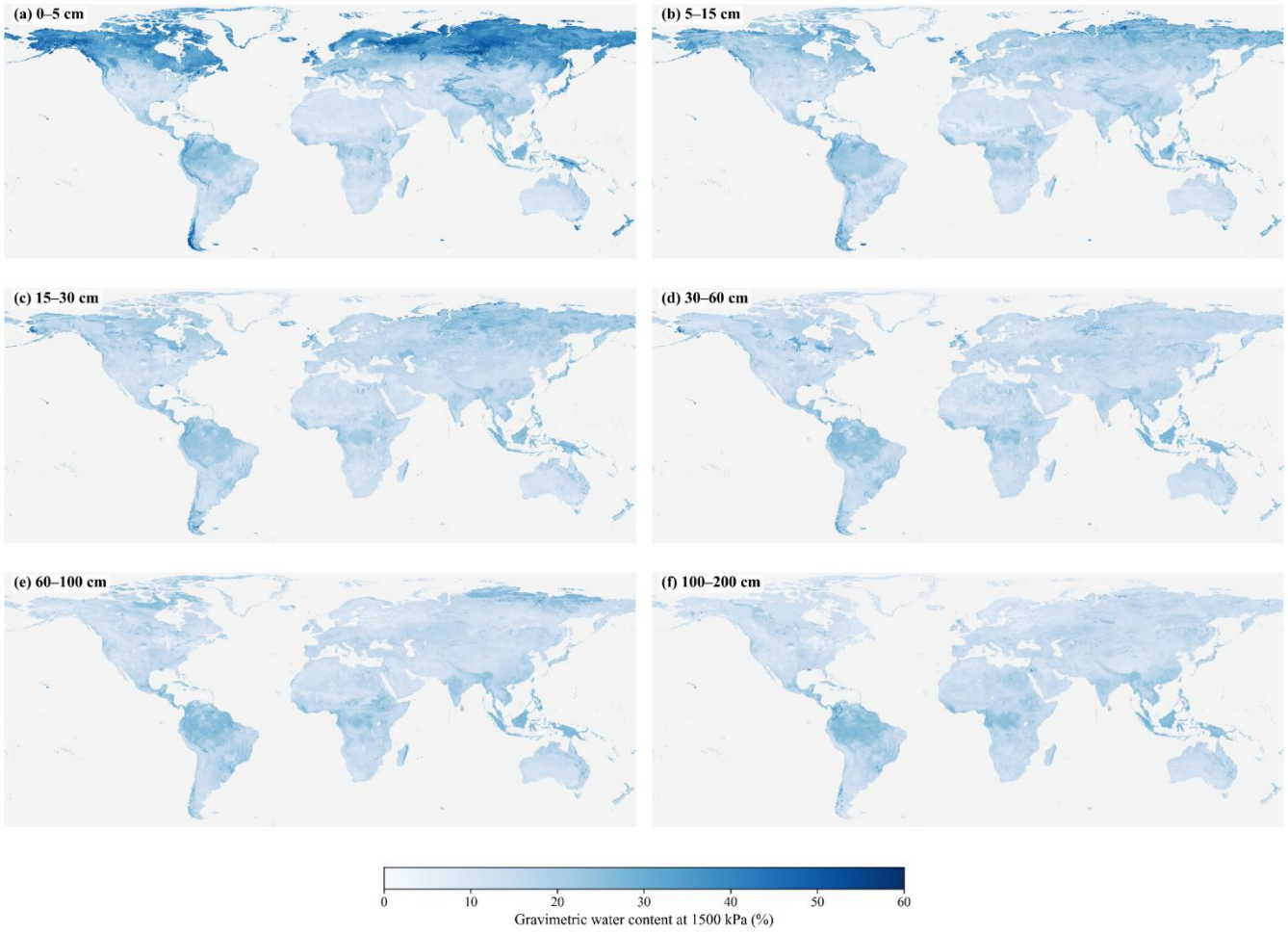


Figure S35. Final GSDEv2 mean predictions of Gravimetric water content at 1500 kPa (wg1500) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

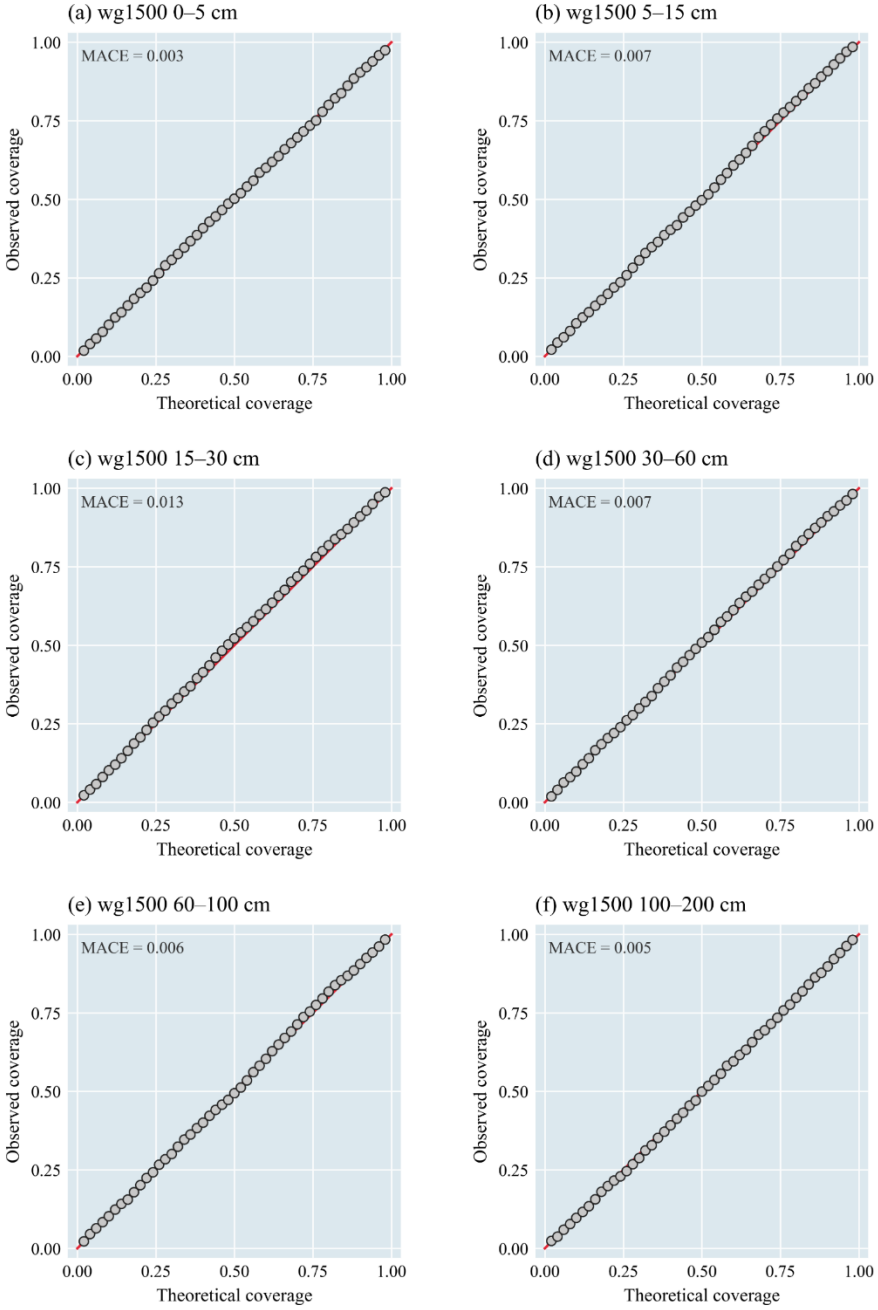


Figure S36. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Gravimetric water content at 1500 kPa (wg1500) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.19 Gravimetric water content at 10 kPa (wg0010)

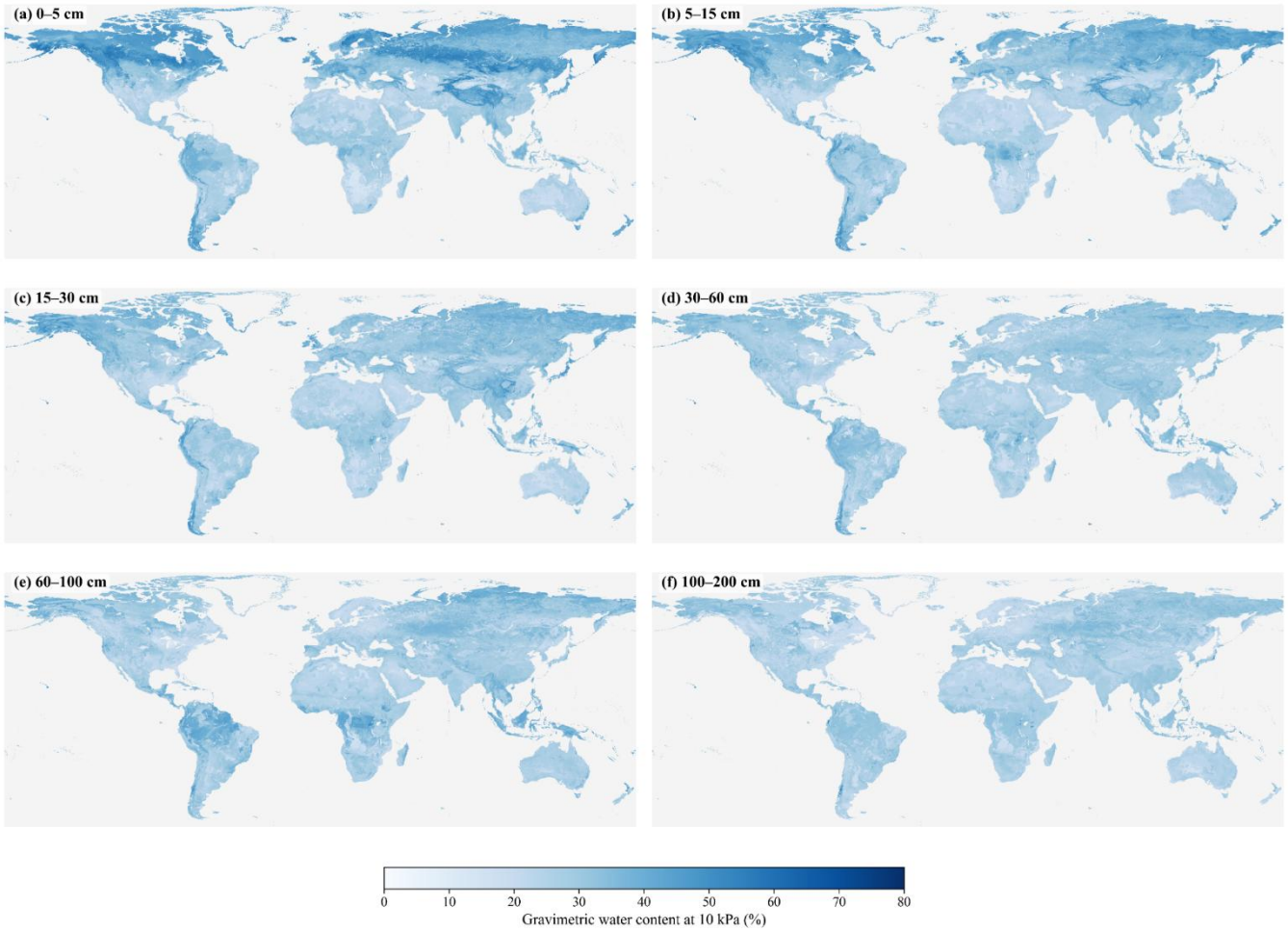


Figure S37. Final GSDEv2 mean predictions of Gravimetric water content at 10 kPa (wg0010) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

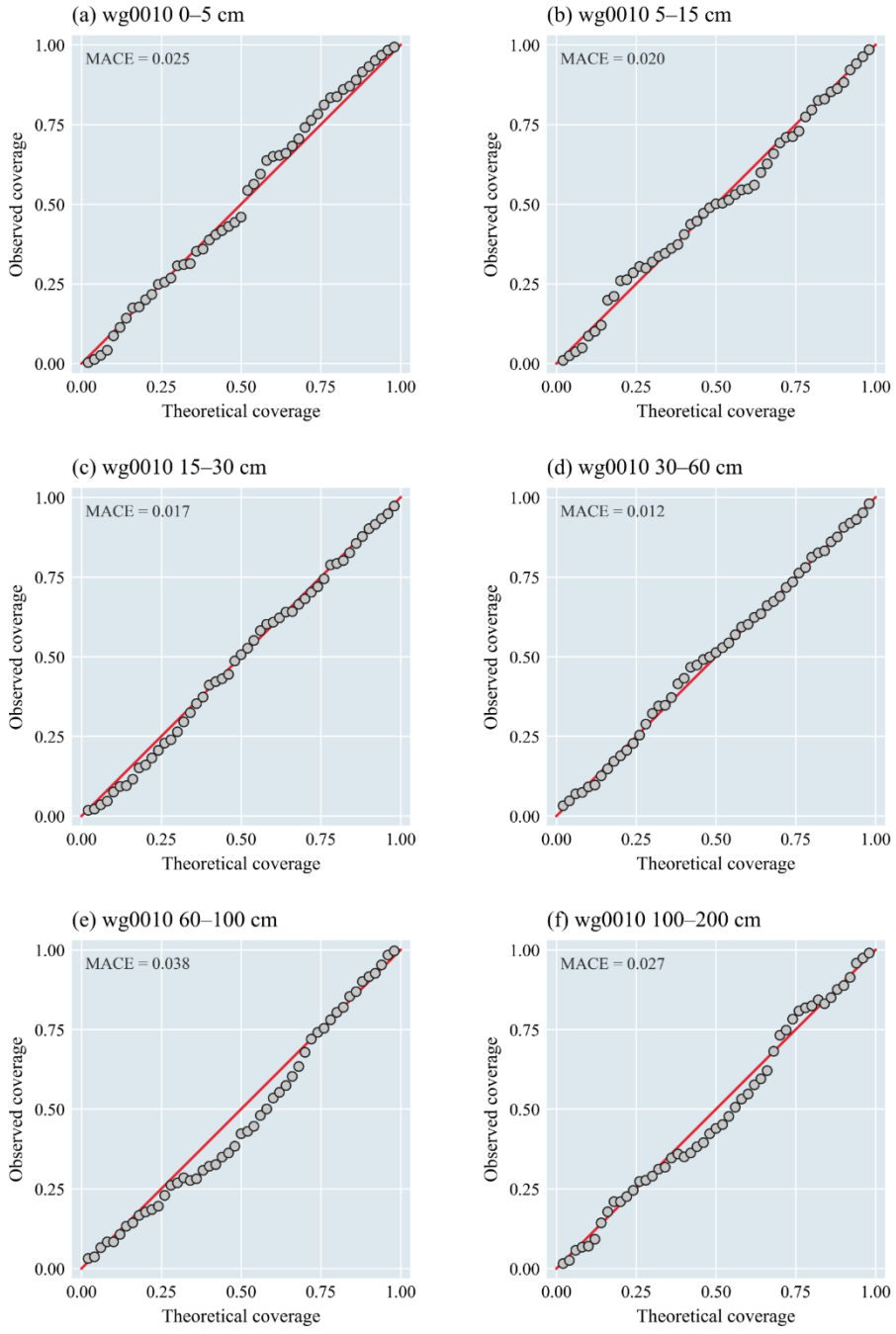


Figure S38. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Gravimetric water content at 10 kPa (wg0010) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S1.20 Extractable phosphorus, Bray I (PBR)

Usage note. Bray-I extractable phosphorus is a method-specific indicator principally intended for acidic soils. The PBR layers should therefore be interpreted with caution in alkaline and calcareous environments.

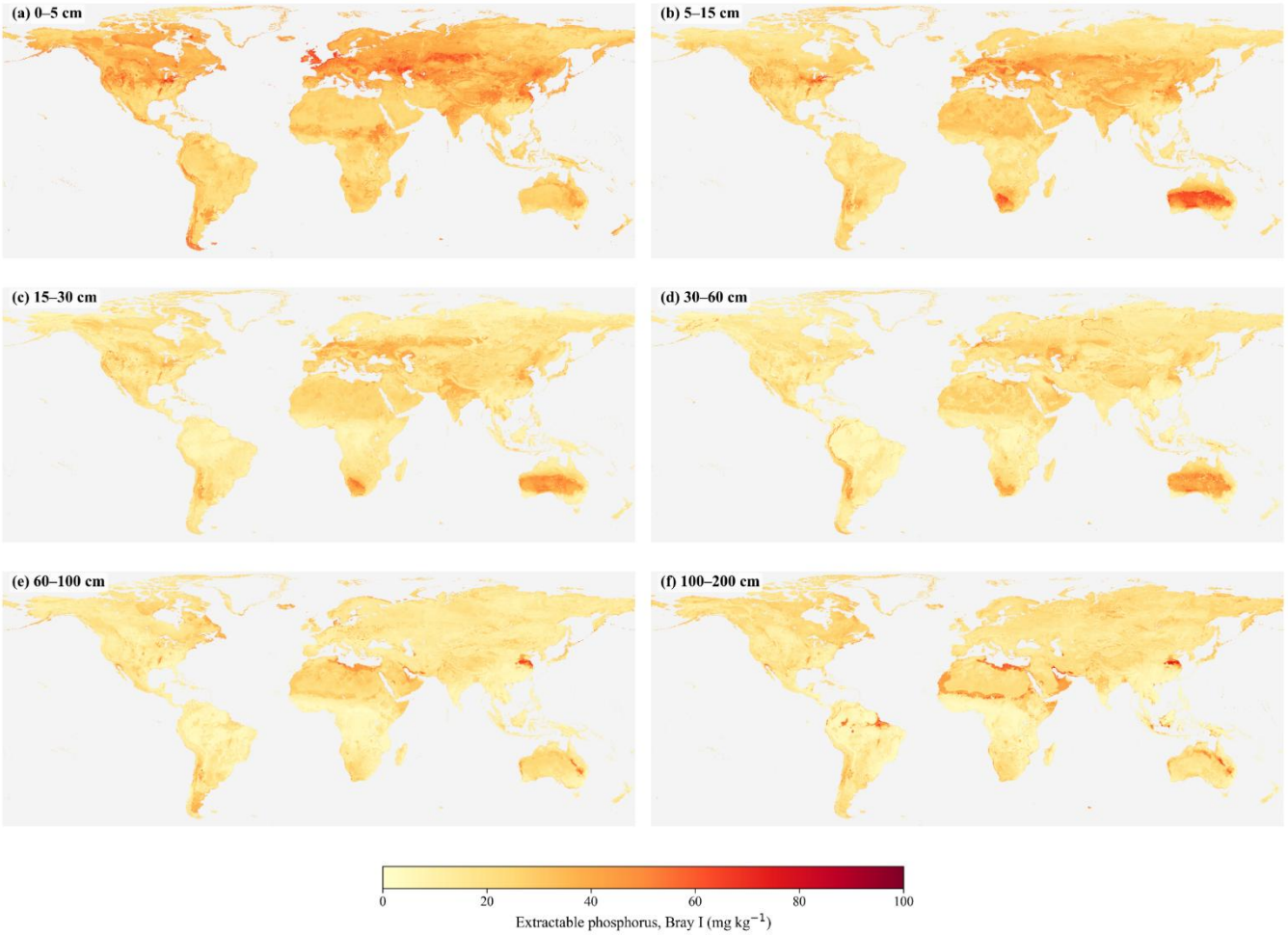


Figure S39. Final GSDEv2 mean predictions of Extractable phosphorus, Bray I (PBR) at six standard depth intervals: (a) 0–5 cm; (b) 5–15 cm; (c) 15–30 cm; (d) 30–60 cm; (e) 60–100 cm; and (f) 100–200 cm.

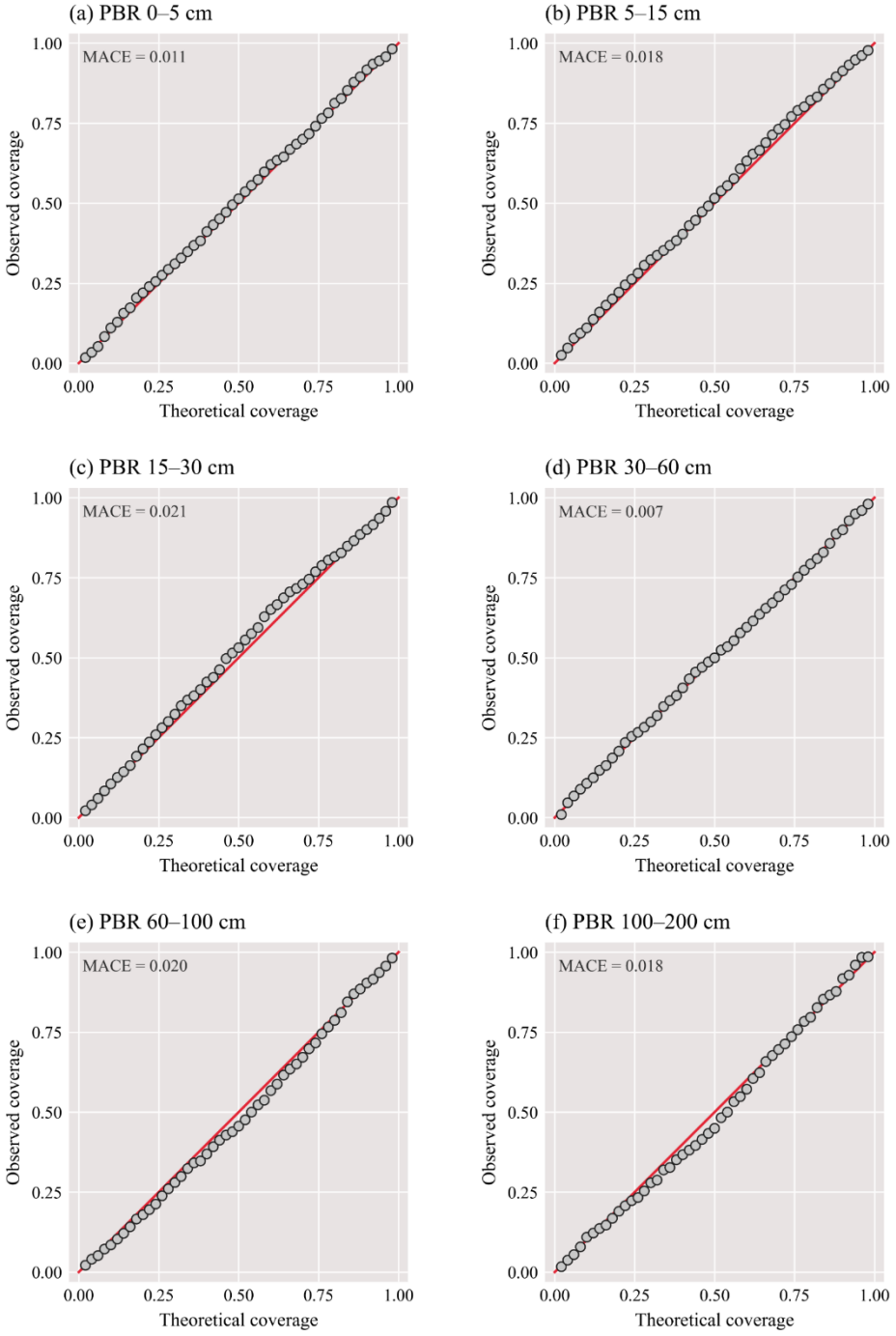


Figure S40. Calibration of validation-stage Quantile Random Forest (QRF) prediction intervals for Extractable phosphorus, Bray I (PBR) at six soil depths. Panels (a–f) correspond to 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm. The red diagonal represents perfect agreement between observed and theoretical coverage; MACE is the mean absolute calibration error.

S2. Profile harmonization and quality control

Soil observations compiled from heterogeneous global, regional, and national databases differed in variable definitions, analytical methods, units, coordinate accuracy, depth descriptions, and missing-value conventions. A hierarchical quality-control procedure was therefore applied before vertical harmonization and model development. The workflow comprised data standardization, coordinate and depth validation, physical-range screening, statistical outlier detection, pedological consistency checks, duplicate identification, and environmental-context screening. The objective was to remove records with identifiable structural, physical, or metadata errors while retaining unusual but physically plausible observations that may represent genuine global soil variability.

S2.1 Data standardization and structural checks

Variable names, analytical-method identifiers, units, profile identifiers, coordinates, and layer-depth information were converted to a common data structure. Source-specific missing-value and fill-value codes were converted to missing values. Non-numeric and infinite values, and records lacking information required for spatial modelling, were excluded.

Geographical coordinates were required to satisfy

$$-180^\circ \leq \text{longitude} \leq 180^\circ$$

and

$$-90^\circ \leq \text{latitude} \leq 90^\circ.$$

Soil horizons or layers were required to have non-negative upper depths and lower depths greater than their corresponding upper depths:

$$z_{\text{top}} \geq 0,\ z_{\text{bottom}} > z_{\text{top}}.$$

Records with missing, reversed, zero-thickness, or clearly inconsistent depth intervals were excluded or corrected only when the intended depth could be determined unambiguously from the original metadata.

Property values were converted to the common units used by GSDEv2 before subsequent screening and vertical harmonization. Where source datasets used analytically distinct definitions or measurement procedures that were not directly interchangeable, the observations were retained as separate target variables rather than merged.

S2.2 Physical-range and statistical-outlier screening

Each soil property was first screened against predefined physically plausible ranges following its definition and analytical method. For example, sand, silt, and clay fractions were required to lie between 0 and 100 %, soil pH was constrained to its valid analytical range, and bulk-density, carbon, nitrogen, phosphorus, coarse-fragment, and water-retention variables were required to satisfy their corresponding non-negative or physically meaningful limits.

After unit harmonization and physical-range screening, statistical outliers were identified separately for each soil property and target depth using the standardized score

$$z_{i,p,d} = \frac{x_{i,p,d} - \mu_{p,d}}{\sigma_{p,d}},$$

where $x_{i,p,d}$ is observation $i$for soil property p and depth interval d, and $\mu_{p,d}$and $\sigma_{p,d}$ are the corresponding mean and standard deviation. Observations satisfying

$$|z_{i,p,d}| > 3$$

were flagged for further examination.

The 3σ criterion was used as a diagnostic flag rather than an automatic exclusion rule because global soil-property distributions can be strongly skewed and can contain valid extreme values. Flagged observations were removed only when additional evidence indicated a likely error, such as implausible units, inconsistent depth information, confirmed duplication, physically unrealistic property combinations, or contradictions with the associated metadata and environmental context. Statistically unusual observations that remained physically and pedologically plausible were retained.

S2.3 Pedological consistency checks

Related soil properties were evaluated using basic physical, compositional, and pedological consistency constraints. Individual texture fractions were required to lie between 0 and 100 %, and their sum was required to satisfy

$$90\% \leq \mathrm{SAND} + \mathrm{SILT} + \mathrm{CLAY} \leq 110\%.$$

The ±10 percentage-point tolerance accommodates rounding, analytical differences, and inconsistencies among historical data sources.

Where both soil organic carbon (SOC) and total carbon (TC) were available, the following relationship was required:

$$\mathrm{SOC} \leq \mathrm{TC}.$$

Where inorganic carbon (IC) was additionally available, substantial departures from

$$\mathrm{SOC} + \mathrm{IC} \approx \mathrm{TC}$$

were flagged as potential unit, analytical-method, or transcription errors.

Water-retention observations were checked for the physically expected decline in water content with increasing matric suction:

$$\mathrm{wv0010} \geq \mathrm{wv0033} \geq \mathrm{wv1500},$$

and

$$\mathrm{wg0010} \geq \mathrm{wg0033} \geq \mathrm{wg1500}.$$

Other pedologically expected relationships, such as lower bulk density in soils with high SOC or increasing cation-exchange capacity with clay and organic-matter content, were used as supporting diagnostic information rather than independent exclusion criteria. This prevented naturally occurring soil conditions from being rejected solely because they differed from broad global relationships.

S2.4 Duplicate and environmental-context screening

Potential duplicate observations were identified using profile identifiers, geographical coordinates, depth intervals, source information, sampling metadata, and soil-property values. Exact duplicate records were removed. When equivalent observations occurred in multiple databases, preference was given to the entry with more complete metadata, more precise coordinate and depth information, or provenance closest to the primary data source.

Co-located observations were not automatically regarded as duplicates when they represented different sampling dates, horizons, measurements, or soil-property values. This distinction was important for retaining genuinely independent observations collected at the same or nearby locations.

Profiles were additionally compared with global land, climate, terrain, vegetation, and soil-domain information. Records located in oceans or permanent water bodies, or showing strong disagreement with surrounding environmental conditions, were flagged for inspection. Environmental disagreement alone was not considered sufficient for exclusion because valid local exceptions and strong small-scale soil heterogeneity can occur. Such records were removed only when environmental inconsistency was accompanied by other evidence, such as doubtful coordinates, physically unrealistic property combinations, or inconsistent source metadata.

S2.5 Final retention

Records violating explicit geographical, depth, physical-range, or pedological constraints were excluded. Statistical outliers that remained physically plausible and were not associated with identifiable metadata or consistency problems were retained.

Quality control was applied at the individual-property level wherever possible rather than removing an entire soil profile because one variable was invalid. Consequently, observations for unaffected soil properties from the same profile could still contribute to the corresponding GSDEv2 property–depth models.

After harmonization, duplicate resolution, and quality control, 234,785 unique soil profiles or observation sites contained at least one valid GSDEv2 target-property observation. The number of available observations nevertheless varied substantially among properties and target depth intervals because of differences in source-variable availability and measured vertical support.

S2.6 Vertical harmonization of profile observations

Following quality control, equal-area quadratic smoothing splines were fitted separately for each soil property and profile (Bishop et al., 1999; Arrouays et al., 2014; Ma et al., 2019). Observations with invalid, reversed, overlapping, or zero-thickness depth intervals were removed before spline fitting.

Spline-based harmonization required at least two valid, non-overlapping measured horizons for the corresponding soil property. Profiles containing fewer than two valid horizons were not used for spline-based harmonization of that property,

although observations for other properties from the same profile could still be retained when their respective eligibility criteria were satisfied.

For a GSDEv2 target interval $[a, b]$, the corresponding interval mean was obtained by integrating the fitted depth function $f_p(z)$:

$$\bar{x}_{p,[a,b]} = \frac{1}{b-a}\int_a^b f_p\,(z)\,dz.$$

This procedure was applied to the six GSDEv2 standard depth intervals: 0–5, 5–15, 15–30, 30–60, 60–100, and 100–200 cm.

A target interval was retained only when its complete depth range was contained within the observed vertical support of the corresponding property. Extrapolation above the shallowest valid observation or below the deepest observed horizon was not permitted. Where profile-bottom or bedrock-depth information was available, target intervals extending below the reported profile or bedrock limit were also excluded.

Unsupported target intervals were assigned missing values and did not contribute response observations to the corresponding property–depth model. This strict vertical-support criterion avoids generating pseudo-observations for soil depths not constrained by measured profile information.

S3. Regional-product harmonization

Regional and national gridded soil products differed in spatial resolution, grid geometry, property definitions, measurement units, and vertical representation. Before their use as auxiliary predictors in the regional stacking framework, each product was harmonized to the corresponding GSDEv2 property definitions, units, spatial grid, and supported target-depth intervals. Regional products were used only for property–depth combinations with compatible definitions and sufficient native vertical support; otherwise, the GSDEv2 global baseline prediction was retained. Product-specific implementation details, vertical support, and provenance are summarized in Table S8.

S3.1 Spatial harmonization

All regional gridded products were transformed and aligned to the native nominal 3-arcsec GSDEv2 modelling grid before stacking. The coordinate reference system, spatial extent, grid origin, cell size, and NoData representation were matched to the corresponding GSDEv2 prediction layers.

Property values were converted to the GSDEv2 units before resampling and model construction. Cells without valid regional-product information were retained as missing and were not artificially extrapolated to areas lacking native regional support.

Spatial harmonization modifies the computational grid on which a regional product is represented but does not imply that the effective spatial resolution of the original product has been increased. The native grid spacing and original spatial support of each regional product are therefore retained as important provenance information in Tables 3 and S8.

S3.2 Vertical harmonization of layer-average products

For regional products provided as averages over depth intervals, values were harmonized to a GSDEv2 target interval $[a, b]$using overlap-thickness-weighted averaging.

For source layer i, defined over the interval $[a_i, b_i]$, the overlap thickness with the GSDEv2 target interval was calculated as

$$l_i = \max[0,\ \min(b, b_i) - \max(a, a_i)]\,.$$

The harmonized target value was then calculated as

$$z_{[a,b]} = \frac{\sum_{i=1}^{n} z_i\, l_i}{\sum_{i=1}^{n} l_i},$$

where $z_i$is the value of source layer i, and $l_i$is its overlapping thickness with the target interval.

Only source layers with $l_i > 0$contributed to the target value. A regional product was considered vertically eligible only when its native support covered the complete GSDEv2 target interval. In practice, this required

$$\sum_{i=1}^{n} l_i = b - a$$

for the contributing source layers without unsupported vertical gaps.

Partial support was not extrapolated to the remainder of a target interval. For example, a regional product extending only to 50 cm was not used to construct the 30–60 cm GSDEv2 layer because the 50–60 cm portion was unsupported. In such cases, regional fusion was not implemented for that property–depth combination and the global GSDEv2 baseline prediction was retained.

S3.3 Vertical harmonization of discrete-depth products

Some regional products provide soil-property predictions at discrete reference depths rather than as averages over predefined depth intervals. For these products, a continuous vertical function was reconstructed from the available depth-specific predictions using an equal-area spline representation consistent with the GSDEv2 vertical-harmonization framework (Bishop et al., 1999).

The harmonized value for target interval $[a, b]$was obtained by integrating the fitted regional depth function $g_p(z)$:

$$\bar{z}_{p,[a,b]} = \frac{1}{b-a}\int_a^b g_p\,(z)\,dz.$$

As for the layer-average products, integration was permitted only where the complete target interval was supported by the native vertical range of the regional product. No vertical extrapolation beyond the shallowest or deepest supported depth was allowed.

For example, SOLUS100m provides predictions at standard depths extending from 0 to 150 cm. The product could therefore support harmonization to the GSDEv2 intervals from 0–5 through 60–100 cm, whereas the 100–200 cm interval was

315 excluded because its complete vertical range was not represented by the native product. The complete product-specific vertical-harmonization rules are provided in Table S8.

S3.4 Fusion eligibility and provenance assessment

Regional fusion was implemented only when three conditions were satisfied: the regional variable was definitionally compatible with the corresponding GSDEv2 target property, its values could be converted to the required GSDEv2 units and
320 grid, and its native vertical information completely supported the target depth interval. Unsupported combinations retained the global baseline and were not filled by vertical or spatial extrapolation.

The regional products were treated as auxiliary predictors rather than as independent reference observations or direct replacements for GSDEv2. Their training-data provenance was therefore reviewed wherever documentation was available. Table S8 identifies, for each regional product, the properties included in fusion, native vertical support, harmonization method,
325 implemented and unsupported GSDEv2 intervals, source-data provenance, and the known or potential overlap between the regional-product training data and the GSDEv2 soil-profile database.

For several regional products, complete profile-level identifiers were not available, and potential overlap with soil observations included in GSDEv2 could therefore not be fully resolved. Consequently, agreement between GSDEv2 and an assimilated regional product was not treated as independent validation evidence. This source-overlap limitation was taken into
330 account when interpreting the regional spatial-block evaluation, as described in the main manuscript.

Table S1. Inventory of soil-profile and soil-observation data sources used in the development of GSDEv2.

Source ID	Dataset and provider	Spatial coverage	Version / extraction date	Sampling period	Coordinate information	Unique within-source profiles/sites before screening	Unique profiles/sites retained	Properties used	Licence and reference
P01	Standardised soil profile data for the world (WoSIS snapshot – December 2023), ISRIC – World Soil Information	Global	WoSIS snapshot – December 2023; GeoPackage version (wosis_202312.gpkg); accessed 12 May 2025	1918–2022; record-specific sampling dates vary	Georeferenced point coordinates in WGS 84 (EPSG:4326); positional uncertainty is provided where available and varies among contributing sources	228,090	201,721	20 GSDEv2 properties; see Table 1 for the source–property inventory	Profile-specific CC BY licence; Calisto et al. (2023); Batjes et al. (2024)
P02	Second National Soil Survey of China soil-profile database, National Soil Survey Office of China	China	Digitized database derived from the Second National Soil Survey of China; extraction date not documented in the project archive	1979–1985	Profile locations reconstructed or georeferenced from legacy survey records; positional accuracy varies among profiles and was corrected using the procedure of Shi et al. (2024)	8,979	8,731	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC, and cfvo	Restricted access. National Soil Survey Office (1996)
P03	Regional soil-profile dataset from five study areas in China, compiled by Shangguan et al. (2012)	Five regional study areas in China	No formal version assigned. Extraction date not documented in the project archive	2008–2009	Field-recorded GPS coordinates	614	546	SAND, SILT, CLAY, SOC, and TN	Restricted access. Shangguan et al. (2012)

Source ID	Dataset and provider	Spatial coverage	Version / extraction date	Sampling period	Coordinate information	Unique within-source profiles/sites before screening	Unique profiles/sites retained	Properties used	Licence and reference
P04	A national soil organic carbon density dataset (2010–2024) in China (China Soil Organic Carbon Stock Database, Version 9); Zenodo	China	Version 9 (V9); Zenodo record 15851730; published 10 July 2025; accessed 11 December 2025	2010–2024	Georeferenced soil-profile coordinates in WGS 84	7,852	5,847	SAND, SILT, CLAY, PHH2O, BD	CC BY 4.0; Chen et al. (2025); dataset DOI: 10.5281/zenodo.15851730
P05	An Open Compendium of Soil Datasets: Soil Observations and Measurements (SoilSamples V1.1), OpenGeoHub Foundation	Global	Version 1.1; Zenodo record 19731002; published 24 April 2026; accessed 12 May 2026	Variable among contributing datasets;	Georeferenced soil-sample and soil-profile locations harmonized to WGS 84 longitude and latitude; positional precision varies among contributing datasets	216,000	17,940	SAND, SILT, CLAY, PHH2O, BD, and SOC	CC BY 4.0; individual contributing datasets may retain source-specific licence conditions; Hengl and Gupta (2025); Hengl et al. (2026)
Final compilation	Final GSDEv2 soil-profile and soil-observation compilation after cross-source duplicate screening, harmonization, and quality control	Global land area	—	Varies among contributing sources	Source-dependent; all retained coordinates standardized to WGS 84	461,535 source-level profiles/sites, including potential cross-source duplicates	**234,785**	20 GSDEv2 properties; see Table 1 for the source–property inventory	Subject to the licences and access conditions of the contributing source datasets

a) "Unique profiles/sites before GSDEv2 screening" denotes the number of unique source profiles or observation sites imported into the GSDEv2 workflow before project-specific cross-source duplicate screening, harmonization, and quality control.

b) "Unique profiles/sites retained" denotes the number of mutually exclusive profiles or sites remaining after harmonization, quality control, and cross-source duplicate resolution, with each retained profile or site contributing at least one valid observation for a GSDEv2 target property.

c) Dataset versions and extraction dates refer to the exact releases, snapshots, or database copies used in this study rather than to versions currently available from the providers.

d) Coordinate information describes the coordinate reference system and positional precision or uncertainty documented in the source metadata. "Not documented" is reported where this information was unavailable.

e) The sum of the source-level counts before screening includes potential cross-source duplicates and therefore does not represent the number of globally unique profiles or sites.

f) Source-level retained counts are mutually exclusive after cross-source duplicate resolution and therefore sum to the final total of 234,785 unique profiles or sites. Each retained record was assigned to one primary source according to the source-priority and provenance rules documented in Sect. S2.4.

g) The OpenLandMap soil-observation compendium integrates multiple underlying global and regional soil databases, including WoSIS. Records imported from this source were therefore screened against the other GSDEv2 sources before model development and validation partitioning.

Table S2. Candidate environmental covariates considered in the GSDEv2 modelling framework before property-, depth-, and soil-domain-specific predictor selection. Source products and common harmonization procedures are identified by Source ID and Processing ID and documented in Tables S3a and S3b, thereby avoiding repeated text while preserving full provenance. Native grid spacing refers to the source or derived layer rather than the final nominal 3-arcsec GSDEv2 grid.

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
GVRHWS	Fractional coverage of Vertisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GUMHWS	Fractional coverage of Umbrisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GSTHWS	Fractional coverage of Stagnosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GSNHWS	Fractional coverage of Solonetz	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GSCHWS	Fractional coverage of Solonchaks	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GRGHWS	Fractional coverage of Regosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GPZHWS	Fractional coverage of Podzols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GPTHWS	Fractional coverage of Plinthosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GPLHWS	Fractional coverage of Planosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GLPHWS	Fractional coverage of Leptosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GKSHWS	Fractional coverage of Kastanozems	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
GHSHWS	Fractional coverage of Histosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GGYHWS	Fractional coverage of Gypsisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GGLHWS	Fractional coverage of Gleysols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GFRHWS	Fractional coverage of Ferralsols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GFLHWS	Fractional coverage of Fluvisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GCRHWS	Fractional coverage of Cryosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GCMHWS	Fractional coverage of Cambisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GCLHWS	Fractional coverage of Calcisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GCHHWS	Fractional coverage of Chernozems	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GATHWS	Fractional coverage of Anthrosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GARHWS	Fractional coverage of Arenosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GANHWS	Fractional coverage of Andosols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GALHWS	Fractional coverage of Alisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GACHWS	Fractional coverage of Acrisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01
GABHWS	Fractional coverage of Albeluvisols	Soil	S01	30 arcsec (≈1 km)	WRB soil-group fractional coverage	P01

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
wc2.1_srad	Annual mean solar radiation (kJ m^{-2} day^{-1})	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual mean of 12 monthly layers	P02
wc2.1_tavg	Annual mean near-surface air temperature (°C)	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual mean of 12 monthly layers	P02
wc2.1_tmax	Annual mean of monthly maximum air temperature (°C)	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual mean of 12 monthly layers	P02
wc2.1_tmin	Annual mean of monthly minimum air temperature (°C)	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual mean of 12 monthly layers	P02
wc2.1_prec	Long-term annual precipitation (mm yr^{-1})	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual sum of 12 monthly layers	P02
wc2.1_wind	Annual mean wind speed (m s^{-1})	Climate and cryosphere	S02	30 arcsec (≈1 km)	Annual mean of 12 monthly layers	P02
TX6MOD	Mean long-term daytime land-surface temperature for October–November	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TX5MOD	Mean long-term daytime land-surface temperature for August–September	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TX4MOD	Mean long-term daytime land-surface temperature for June–July	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TX3MOD	Mean long-term daytime land-surface temperature for April–May	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TX2MOD	Mean long-term daytime land-surface temperature for February–March	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
TX1MOD	Mean long-term daytime land-surface temperature for December–January	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TNSMOD	Standard deviation of long-term nighttime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TNMMOD	Mean long-term nighttime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TNLMOD	Minimum long-term nighttime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TNHMOD	Maximum long-term nighttime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TDSMOD	Standard deviation of long-term daytime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TDMMOD	Mean long-term daytime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
TDLMOD	Minimum long-term daytime land-surface temperature	Climate and cryosphere	S03	1 km	Long-term LST statistic shown in the description	P03
PX4WCL	Long-term precipitation for August–October	Climate and cryosphere	S02	30 arcsec (≈1 km)	Three-month precipitation sum	P04
PX3WCL	Long-term precipitation for May–July	Climate and cryosphere	S02	30 arcsec (≈1 km)	Three-month precipitation sum	P04
PX2WCL	Long-term precipitation for February–April	Climate and cryosphere	S02	30 arcsec (≈1 km)	Three-month precipitation sum	P04
PX1WCL	Long-term precipitation for November–January	Climate and cryosphere	S02	30 arcsec (≈1 km)	Three-month precipitation sum	P04

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
PREGSM	Mean monthly precipitation averaged over the year	Climate and cryosphere	S04	30 arcsec (≈1 km)	Mean of 12 monthly precipitation climatologies	P04
P12DWD	Monthly precipitation for December	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 12)	P04
P11DWD	Mean precipitation of November	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 11)	P04
P10DWD	Mean precipitation of October	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 10)	P04
P09DWD	Mean precipitation of September	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 9)	P04
P08DWD	Mean precipitation of August	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 8)	P04
P07DWD	Mean precipitation of July	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 7)	P04
P06DWD	Mean precipitation of June	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 6)	P04
P05DWD	Mean precipitation of May	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 5)	P04
P04DWD	Mean precipitation of April	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 4)	P04
P03DWD	Mean precipitation for March	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 3)	P04
P02DWD	Monthly precipitation for February	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 2)	P04
P01DWD	Monthly precipitation for January	Climate and cryosphere	S04	30 arcsec (≈1 km)	Monthly precipitation climatology (month 1)	P04
MODCF	MODIS-derived cloud frequency	Climate and cryosphere	S05	30 arcsec (≈1 km)	Mean cloud-frequency climatology	P05

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
INSSRE	Standard deviation of potential incoming solar radiation	Climate and cryosphere	S06	30 arcsec (≈1 km)	Standard deviation of potential incoming solar radiation	P05
INMSRE	Mean potential incoming solar radiation	Climate and cryosphere	S06	30 arcsec (≈1 km)	Mean potential incoming solar radiation	P05
G22ESA	Permanent snow and ice	Climate and cryosphere	S08	300 m	Selected class converted to fractional coverage	P07
TDHMOD	Maximum long-term daytime land-surface temperature	Climate and cryosphere	S03	30 arcsec (≈1 km)	Maximum daytime LST over the stated temporal period	P03
B5/B7	The ratio of Band 5 (near-infrared) to Band 7 (shortwave infrared 2) surface reflectance	Organisms and land-surface conditions	S09	30 m	Computed the scaled NIR/SWIR2 reflectance ratio (SR_B5/SR_B7)	P08
NDVI	Normalized Difference Vegetation Index	Organisms and land-surface conditions	S09	30 m	Computed NDVI = (SR_B5 − SR_B4)/(SR_B5 + SR_B4)	P08
NDMI	Normalized Difference Moisture Index	Organisms and land-surface conditions	S09	30 m	Computed vegetation-water NDMI = (SR_B5 − SR_B6)/(SR_B5 + SR_B6)	P08
surR	Surface Reflectance	Organisms and land-surface conditions	S10	250 m	Long-term mean red-band surface reflectance	P05
EVI	Enhanced Vegetation Index	Organisms and land-surface conditions	S09	30 m	Computed EVI = 2.5(SR_B5 − SR_B4)/(SR_B5 + 6SR_B4 − 7.5SR_B2 + 1)	P08
SAI	Snow-cover frequency	Organisms and land-surface conditions	S09	30 m	Computed snow frequency from NDSI = (SR_B3 − SR_B6)/(SR_B3 + SR_B6), using NDSI ≥ 0.4	P17

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
LAI	Leaf Area Index	Organisms and land-surface conditions	S09	30 m	Landsat-derived vegetation-index proxy	P08
NPP	Net Primary Productivity	Organisms and land-surface conditions	S11	500 m	Mean annual net primary productivity	P05
Landuse	Global land-cover class	Organisms and land-surface conditions	S12	1 km	Categorical class	P06
CanopyHeight	Canopy height	Organisms and land-surface conditions	S13	10 m	Area-weighted mean canopy height	P10
Sentinel2B2	Sentinel-2 blue-band reflectance	Organisms and land-surface conditions	S14	10 m	Mean of quality-screened Level-2A B2 observations	P08
Sentinel2B3	Sentinel-2 green-band reflectance	Organisms and land-surface conditions	S14	10 m	Mean of quality-screened Level-2A B3 observations	P08
Sentinel2B4	Sentinel-2 red-band reflectance	Organisms and land-surface conditions	S14	10 m	Mean of quality-screened Level-2A B4 observations	P08
Sentinel2B8	Sentinel-2 near-infrared reflectance	Organisms and land-surface conditions	S14	10 m	Mean of quality-screened Level-2A B8 observations	P08
Sentinel2B9	Sentinel-2 water-vapour-band reflectance	Organisms and land-surface conditions	S14	60 m	Mean of quality-screened Level-2A B9 observations	P08
SR_B4	Landsat red-band surface reflectance	Organisms and land-surface conditions	S09	30 m	Median of quality-screened red-band surface reflectance observations	P08

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
SR_B5	Landsat near-infrared surface reflectance	Organisms and land-surface conditions	S09	30 m	Median of quality-screened near-infrared surface reflectance observations	P08
SR_B6	Landsat SWIR1 surface reflectance	Organisms and land-surface conditions	S09	30 m	Median of quality-screened SWIR1 surface reflectance observations	P08
SR_B7	Landsat SWIR2 surface reflectance	Organisms and land-surface conditions	S09	30 m	Median of quality-screened SWIR2 surface reflectance observations	P08
ST_ATRAN	Atmospheric Transmittance	Organisms and land-surface conditions	S09	30 m distributed grid; thermal support coarser	Median of quality-screened atmospheric-transmittance observations	P09
ST_B10	Band 10 Surface Temperature	Organisms and land-surface conditions	S09	30 m distributed grid; thermal support coarser	Median of quality-screened Landsat surface-temperature observations	P09
ST_EMSD	Emissivity standard deviation	Organisms and land-surface conditions	S09	30 m distributed grid; thermal support coarser	Median of quality-screened emissivity-standard-deviation observations	P09
ST_TRAD	Thermal Radiance	Organisms and land-surface conditions	S09	30 m distributed grid; thermal support coarser	Median of quality-screened thermal-radiance observations	P09
ST_URAD	Upwelled radiance	Organisms and land-surface conditions	S09	30 m distributed grid; thermal support coarser	Median of quality-screened upwelled-radiance observations	P09
RANGE0105	Range of Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
PDMGPW	Average population density	Organisms and land-surface conditions	S16	2.5 arcmin (≈5 km)	Population density	P12

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
MAX0105	Maximum of Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
LCEE10	ESA land cover	Organisms and land-surface conditions	S17	300 m	Categorical class	P06
L16IGB	Barren or sparsely vegetated	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L15IGB	Snow and ice index	Climate and cryosphere	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L14IGB	Cropland/natural vegetation mosaic	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L13IGB	Urban and built-up	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L12IGB	Croplands index	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L11IGB	Permanent Wetlands	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L10IGB	Grasslands	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L09IGB	Savannas	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
L06IGB	Closed shrublands	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L05IGB	Mixed forests index	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L04IGB	Deciduous broadleaf forest	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L03IGB	Deciduous needleleaf forest	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L02IGB	Evergreen broadleaf forest	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
L01IGB	Evergreen needleleaf forest	Organisms and land-surface conditions	S07	500 m	Fractional coverage of IGBP class 1 in 2010	P07
GLCJRC	Global Land Cover map for the year 2000	Organisms and land-surface conditions	S18	1 km	Categorical class	P06
GLCESA	Land Cover classes	Organisms and land-surface conditions	S08	300 m	Categorical class	P06
GACGEM	Global accessibility map	Organisms and land-surface conditions	S19	30 arcsec (≈1 km)	Travel time to the nearest major city	P12
G21ESA	Water bodies	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
G20ESA	Bare areas	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G19ESA	Artificial surfaces and associated areas	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G18ESA	Closed to open grassland or woody vegetation on regularly flooded or waterlogged soil - Fresh, brackish or saline water	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G17ESA	Closed broadleaved forest or shrubland permanently flooded - Saline or brackish water	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G16ESA	Closed to open broadleaved forest regularly flooded - Fresh or brackish water	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G15ESA	Sparse (<15%) vegetation	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G14ESA	Closed to open shrubland	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G12IGB	Land cover types for 2012	Organisms and land-surface conditions	S07	500 m	Categorical class	P06
G11IGB	Land cover types for 2011	Organisms and land-surface conditions	S07	500 m	Categorical class	P06

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
G11ESA	Mosaic forest or shrubland / grassland	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G10IGB	Land cover types for 2010	Organisms and land-surface conditions	S07	500 m	Categorical class	P06
G10ESA	Closed to open mixed broadleaved and needleleaved forest	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G09ESA	Open needleleaved deciduous or evergreen forest	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G08ESA	Closed needleleaved evergreen forest	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G07ESA	Open broadleaved deciduous forest/woodland	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G05ESA	Closed to open broadleaved evergreen or semi-deciduous forest	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G04IGB	Land cover types for 2004	Organisms and land-surface conditions	S07	500 m	Categorical class	P06
G04ESA	Mosaic vegetation / cropland	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G03ESA	Mosaic cropland / vegetation	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
G02IGB	Land cover types for 2002	Organisms and land-surface conditions	S07	500 m	Categorical class	P06
G02ESA	Rainfed croplands	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
G01IGB	Land cover types for 2001	Organisms and land-surface conditions	S07	500 m	Categorical class	P06
G01ESA	Post-flooding or irrigated croplands (or aquatic)	Organisms and land-surface conditions	S08	300 m	Selected class converted to fractional coverage	P07
EVSMOD	Standard deviation MODIS Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
EVMMOD	Mean MODIS Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
EVENNESS	Evenness of Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
ENTROPY	Disorderliness of Enhanced Vegetation Index	Organisms and land-surface conditions	S15	30 arcsec (≈1 km)	Published EVI heterogeneity metric	P11
COVER	Tree-canopy cover in 2000	Organisms and land-surface conditions	S20	30 m	Percentage tree-canopy cover	P10
TERECO	Terrestrial ecosystem class	Organisms and land-surface conditions	S21	250 m	Categorical class	P06

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
L3POBI	Physiographic landform units (SCALA project)	Relief and terrain	S22	30 arcsec (≈1 km)	Categorical physiographic landform unit	P06
DEM	Land-surface elevation	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	Land-surface elevation	P13
Slope	Terrain slope	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	Horn 3×3 slope algorithm	P14
DEM_crd	Local downslope curvature	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	local downslope curvature from a 3×3 neighbourhood	P14
DEM_cru	Local upslope curvature	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	local upslope curvature from a 3×3 neighbourhood	P14
DEM_crv	Downslope Curvature	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	downslope/vertical curvature from a 3×3 neighbourhood	P14
DEM_dvm	Deviation from Mean Value (surface roughness)	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	deviation from mean elevation within a 13-cell-radius window	P14
DEM_mrn	Melton Ruggedness Number	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	Melton ruggedness number within a 13-cell-radius window	P14
DEM_nopn	Negative Topographic Openness	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	negative topographic openness using a 10-cell search radius	P14
DEM_popn	Positive Topographic Openness	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	positive topographic openness using a 10-cell search radius	P14
DEM_prof	Profile curvature	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	profile curvature from a 3×3 neighbourhood	P14
DEM_tpi	Topographic Position Index	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	topographic position index within a 13-cell-radius window	P14
DEM_twi	Topographic Wetness Index	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	topographic wetness index using D8 contributing area and local slope	P14

Covariate code	Description	Covariate group	Source ID	Native grid spacing	Variable-specific derivation or statistic	Processing ID
DEM_vbf	Multiresolution Index of Valley Bottom Flatness	Relief and terrain	S23	3 arcsec (≈90 m at the equator)	multiresolution index of valley-bottom flatness	P14
RTMUSG15	Lithological class	Parent material	S24	250 m derived raster	First-level lithological class	P15
SEDDEP	Mean thickness of soil, regolith and sedimentary deposits	Parent material	S25	30 arcsec (≈1 km)	Total unconsolidated-material thickness	P16

Table S3. Source datasets, versions, temporal support, references, and common covariate-harmonization protocols.

Table S3a. Source datasets and provenance.

Source ID	Source product and version	Temporal coverage	Covariate role	Reference / permanent source
S01	Harmonized World Soil Database v1.2 (HWSD); WRB soil-group component fractions	Static legacy soil database	Soil-group fractional-cover predictors	FAO/IIASA/ISRIC/ISSCAS/JRC, HWSD v1.2; https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/
S02	WorldClim v2.1 climate normals	1970–2000	Temperature, precipitation, radiation and wind climatologies	Fick and Hijmans (2017); https://www.worldclim.org/data/worldclim21.html
S03	MODIS/Terra MOD11A2 Collection 6.1, 8-day land-surface temperature	2001–2020; variable-specific long-term mean, minimum, maximum, and standard-deviation statistics	Daytime and nighttime LST statistics	NASA MOD11A2 v6.1, DOI:10.5067/MODIS/MOD11A2.061; https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD11A2
S04	CHELSA climatologies v2.1	1981–2010	Monthly precipitation climatologies	Karger et al. (2021)
S05	EarthEnv Global 1-km Cloud Frequency v1.0	2000–2014	Cloud-frequency climatology	Wilson and Jetz (2016); https://www.earthenv.org/cloud.html
S06	WorldGrids potential incoming solar-radiation layers v1.0	Long-term terrain-adjusted climatology	Potential incoming solar-radiation statistics	WorldGrids environmental layers; https://www.worldgrids.org/
S07	MODIS MCD12Q1 Collection 6.1, IGBP land-cover classification	Annual; years specified by the individual covariate	Categorical land cover and class-fraction predictors	NASA MCD12Q1 v6.1, DOI:10.5067/MODIS/MCD12Q1.061; https://lpdaac.usgs.gov/products/mcd12q1v061/
S08	ESA GlobCover 2009 v2.3	2009	Categorical land cover and class-fraction predictors	ESA GlobCover 2009 v2.3; https://due.esrin.esa.int/page_globcover.php
S09	USGS Landsat 8 Collection 2 Level-2	2017–2020; median of valid observations	Surface reflectance, spectral indices and surface-temperature-related bands	USGS Landsat Collection 2 Level-2, DOI:10.5066/P9OGBGM6; https://www.usgs.gov/landsat-missions/landsat-collection-2
S10	MODIS/Terra MOD09Q1 Collection 6.1	2001–2020 multi-year mean	Surface-reflectance climatology	NASA MOD09Q1 v6.1, DOI:10.5067/MODIS/MOD09Q1.061; https://www.earthdata.nasa.gov/data/catalog/lpcloud-mod09q1-061
S11	MODIS/Terra MOD17A3HGF Collection 6.1	2001–2020 multi-year mean	Net primary productivity	NASA MOD17A3HGF v6.1, DOI:10.5067/MODIS/MOD17A3HGF.061; https://lpdaac.usgs.gov/products/mod17a3hgfv061/

Source ID	Source product and version	Temporal coverage	Covariate role	Reference / permanent source
S12	USGS Global Land Cover Characterization v2.0	1992–1993	Global categorical land-cover class	USGS GLCC v2.0, DOI:10.5066/F7GB230D; https://www.usgs.gov/centers/eros/science/usgs-eros-archive-land-cover-products-global-land-cover-characterization-glcc
S13	ETH Global Canopy Height 10 m 2020, version 1	2020	Canopy-height predictor	Lang et al. (2023), DOI:10.3929/ethz-b-000609802
S14	Copernicus Sentinel-2 MSI Level-2A	2018–2022; multi-year median composite of quality-screened valid observations	Multispectral surface-reflectance bands	Sentinel-2 User Handbook, DOI:10.5270/S2_-znk9xsj; https://sentinels.copernicus.eu/documents/247904/685211/Sentinel-2_User_Handbook
S15	EarthEnv Global Habitat Heterogeneity metrics derived from MODIS EVI	2001–2005	Vegetation heterogeneity and texture statistics	Tuanmu and Jetz (2015); https://www.earthenv.org/texture
S16	Gridded Population of the World v4.11	2015	Population-density predictor	CIESIN GPW v4.11; https://www.earthdata.nasa.gov/data/projects/gpw
S17	ESA Climate Change Initiative Land Cover annual map v2.0.7	2010	Categorical land-cover predictor	ESA CCI Land Cover; https://climate.esa.int/en/projects/land-cover/
S18	Joint Research Centre Global Land Cover 2000 v1.1	2000	Categorical land-cover predictor	Bartholomé and Belward (2005); https://forobs.jrc.ec.europa.eu/glc2000
S19	Global accessibility to cities	2015	Travel-time / accessibility predictor	Weiss et al. (2018), DOI:10.1038/nature25181
S20	Global Forest Change tree-canopy cover product	Tree-canopy cover for 2000	Percentage tree-canopy cover	Hansen et al. (2013), DOI:10.1126/science.1244693
S21	World Terrestrial ecosystem class, version 1.0	Static categorical ecosystem map	Categorical terrestrial-ecosystem class	World Terrestrial ecosystem class; https://catalog.data.gov/dataset/world-terrestrial-ecosystems-wte-2020
S22	WorldGrids / SCALA physiographic landform units v1.0	Static	Categorical physiographic landform class	WorldGrids environmental layers; https://www.worldgrids.org/
S23	MERIT DEM v1.0.3	Static	Elevation and terrain derivatives	Yamazaki et al. (2017), DOI:10.1002/2017GL072874; https://global-hydrodynamics.github.io/res_topography/
S24	Global Lithological Map v1.1 (GLiM), rasterized lithological classes	Static	Parent-material / lithological class	Hartmann and Moosdorf (2012), DOI:10.1594/PANGAEA.788537
S25	Global soil, regolith and sedimentary-deposit thickness dataset	Static	Unconsolidated-material thickness	Pelletier et al. (2016), DOI:10.1002/2015MS000526; https://daac.ornl.gov/SOILS/guides/Global_Soil_Regolith_Sediment.html

Table S3b. Common preprocessing and resampling protocols.

Processing ID	Applicable data type	Common preprocessing and resampling protocol	Scientific rationale
P01	Fractional soil-class coverage	Derived class fractions from source map-unit components and harmonized them using area-weighted resampling.	Area-weighted resampling preserves the fractional-cover interpretation of soil-group predictors while avoiding the creation of artificial categorical values.
P02	Continuous climate normals	Calculated the stated annual statistic from monthly climatologies and aligned the resulting continuous layer using bilinear interpolation.	Resampling aligns the grid but does not increase the source product's effective spatial resolution.
P03	MODIS land-surface temperature	Applied scale factors and quality screening, calculated the stated long-term statistic from valid 8-day observations, and aligned the continuous layer using bilinear interpolation.	Only pixels with mandatory QA flags indicating good-quality LST retrievals were retained.
P04	Monthly or seasonal precipitation	Selected or aggregated the relevant monthly climatologies to the statistic shown in Table S1 and aligned the continuous layer using bilinear interpolation.	Monthly and seasonal precipitation aggregates characterize complementary components of the long-term water-supply regime.
P05	Coarse continuous environmental layer	Applied source-specific scale and validity rules where required and aligned the continuous layer using bilinear interpolation.	Used for cloud, solar-radiation, reflectance and similar continuous coarse-grid predictors.
P06	Categorical map	Reprojected and aligned the categorical class layer using nearest-neighbour resampling.	Nearest-neighbour resampling prevents creation of artificial class codes.
P07	Binary class indicator / fractional cover	Converted the selected class to a binary indicator and aggregated it as fractional coverage using area-weighted resampling.	The resulting layer represents sub-grid class coverage rather than a categorical code.
P08	Fine-resolution optical reflectance or spectral index	Official scale factors were applied, and clouds, cloud shadows, snow, saturated pixels, and other invalid observations were removed using the relevant quality-assurance layers. Spectral indices were calculated for individual valid observations, after which a multi-year temporal median was derived. The resulting fine-resolution composite was then aggregated to the nominal 3-arcsec grid using an area-weighted mean.	Quality screening reduces contamination by clouds, shadows, snow, and invalid observations. Temporal median compositing limits the influence of residual outliers, whereas area-weighted spatial aggregation preserves the mean fine-scale spectral information represented within each target grid cell.
P09	Landsat surface-temperature product band	Applied official scale/offset and quality masks, calculated a multi-year median, and aggregated the valid distributed-grid values using an area-weighted mean.	The distributed 30 m grid should not be interpreted as the effective thermal sensor resolution.
P10	Fine-resolution static continuous layer	Removed invalid or NoData values and aggregated valid source pixels to the target grid using an area-weighted mean.	Used for canopy height and comparable fine-grid continuous products.

Processing ID	Applicable data type	Common preprocessing and resampling protocol	Scientific rationale
P11	Published vegetation-heterogeneity metric	Used the published continuous metric and aligned it to the target grid using bilinear interpolation.	The precise metric definition and temporal window remain encoded by the individual covariate description.
P12	Continuous human-pressure layer	Applied valid-value and unit checks and aligned the continuous surface using bilinear interpolation.	Population density and travel time should retain their original units and reference year.
P13	Native-grid elevation	Used the native MERIT DEM grid and aligned tile origins and extents directly to the common nominal 3-arcsec grid.	Three arcsec is a nominal angular grid spacing; ground dimensions vary with latitude.
P14	Terrain derivative	Derived the stated terrain metric from MERIT DEM on the native grid and retained it as a continuous layer on the common 3-arcsec tile grid.	Deriving terrain attributes directly from the native DEM avoids additional interpolation before geomorphometric calculation and provides complementary representations of local terrain form, position, roughness, and hydrological convergence.
P15	Categorical lithological class	Rasterized or used the categorical lithological class and aligned it using nearest-neighbour resampling.	Nearest-neighbour resampling preserves discrete lithological class labels without generating intermediate class values. The first-level GLiM classification was used.
P16	Continuous subsurface-thickness layer	Applied source units and validity rules and aligned the continuous thickness layer using bilinear interpolation.	The continuous thickness layer characterizes broad spatial variation in unconsolidated material depth; bilinear alignment preserves its continuous representation without implying increased effective resolution.
P17	Multi-temporal snow-cover indicator	Applied reflectance scaling and masked clouds, cloud shadows, saturated and invalid observations while retaining snow pixels. NDSI was calculated from green and SWIR1 reflectance, and snow-cover frequency was calculated as the proportion of valid observations with NDSI ≥ 0.4.	Retaining snow observations while excluding invalid observations allows the covariate to represent the temporal frequency of snow cover rather than cloud contamination.

Table S4. HWSD soil-type codes used to define the GSDEv2 soil-forming domains.

Modelling domain	HWSD pixel values	Corresponding HWSD classes or map units	Assignment priority
Histosol domain	86–91	Histosols, including Folic, Terric, Fibric, Thionic, and Gelic Histosols	1
Fluvisol–Arenosol–Gypsisol–dune domain	1–8; 38–45; 81–85; 194	Fluvisols, Arenosols, Gypsisols, and dunes and shifting sands	2
Other mineral-soil domain	All remaining valid soil-type codes	Remaining valid soil classes not assigned to the two domains above	3

Note: Domain assignment was hierarchical. Histosol cells were assigned first, followed by cells belonging to the selected Fluvisol–Arenosol–Gypsisol–dune classes; all remaining valid soil cells were assigned to the other mineral-soil domain. Pixel value 0 and non-soil or invalid map units were excluded using the valid soil and land masks.

Table S5. Hyperparameters used for the considered soil properties at the 0–5 cm depth interval. See Table 1 for abbreviations and units of the soil properties considered.

Property	Number of covariates	Number of trees	mtry
SAND	35	350	13
SILT	28	350	12
CLAY	30	350	12
PHH2O	29	350	12
BD	30	350	12
SOC	28	350	12
TC	25	350	13
TN	31	350	13
CEC	28	300	13
CECPH7	27	300	9
cfvo	26	300	10
cfgr	27	300	12
bdfi33	27	300	12
wv0033	25	350	12
wv1500	25	350	13
wv0010	25	350	13
wg0033	26	350	13
wg1500	27	350	13
wg0010	27	350	13
PBR	28	350	12

Table S6. Descriptive statistics of the soil-profile observations for the 20 GSDEv2 properties at six standard depth intervals. Refer to Table 1 for the abbreviations and units of the soil properties considered.

Depth interval (cm)	Number	Max	Min	Mean	SD	CV	Skewness	Kurtosis
SAND								
0-5	96404	99.76	0.25	43.7	27.19	0.62	0.19	-1.11
5-15	107530	99.47	0.23	43.71	26.59	0.61	0.18	-1.06
15-30	105233	99.54	0.24	41.82	26.61	0.64	0.27	-1.01
30-60	101302	98.91	0.24	40.7	26.89	0.66	0.32	-0.97
60-100	90617	99.72	0.25	41.02	27.4	0.67	0.33	-0.97
100-200	69835	99.74	0.23	41.98	28.43	0.68	0.33	-1.01
SILT								
0-5	112077	96.11	0.00	32.29	21.71	0.67	0.55	-0.69
5-15	128194	96.44	0.00	32.02	21.09	0.66	0.56	-0.60
15-30	125794	95.00	0.00	31.11	20.67	0.66	0.56	-0.62
30-60	117968	95.00	0.00	29.13	19.69	0.68	0.57	-0.60
60-100	104942	94.26	0.01	28.14	19.49	0.69	0.62	-0.51
100-200	78672	94.18	0.01	28.51	20.15	0.71	0.62	-0.46
CLAY								
0-5	116921	96.69	0.21	20.86	15.63	0.75	1.20	1.22
5-15	135170	95.00	0.18	21.23	15.65	0.74	1.14	1.08
15-30	130689	95.00	0.15	24.31	16.84	0.69	0.90	0.40
30-60	122905	94.79	0.16	27.86	18.39	0.66	0.65	-0.12
60-100	110166	93.36	0.16	28.80	18.98	0.66	0.60	-0.19
100-200	83544	93.22	0.15	27.57	19.27	0.70	0.73	-0.01
PHH2O								
0-5	120957	10.82	1.09	6.18	1.21	0.20	0.14	-0.47
5-15	130326	11.20	1.30	6.20	1.22	0.20	0.18	-0.52
15-30	124695	10.99	1.33	6.25	1.25	0.20	0.29	-0.59
30-60	122713	11.24	1.77	6.35	1.33	0.21	0.31	-0.77
60-100	110378	11.76	1.71	6.46	1.41	0.22	0.25	-0.95

Depth interval (cm)	Number	Max	Min	Mean	SD	CV	Skewness	Kurtosis
100-200	85199	11.11	1.88	6.59	1.44	0.22	0.11	-1.06
BD								
0-5	19659	2.43	0.02	1.23	0.40	0.33	-1.10	1.00
5-15	22371	2.28	0.04	1.28	0.37	0.29	-1.17	1.38
15-30	23054	2.27	0.03	1.35	0.35	0.26	-1.29	1.99
30-60	22119	2.52	0.04	1.42	0.35	0.25	-1.43	2.60
60-100	19257	2.62	0.04	1.49	0.34	0.23	-1.63	3.93
100-200	15248	2.43	0.03	1.54	0.33	0.22	-1.76	4.91
SOC								
0-5	98940	845.31	0.07	49.6	97.16	1.96	3.51	12.66
5-15	109537	805.53	0.04	32.09	62.87	1.96	5.36	33.70
15-30	115917	789.67	0.05	24.39	50.18	2.06	6.36	52.90
30-60	94716	795.60	0.02	13.23	43.70	3.30	9.42	102.02
60-100	79577	793.19	0.02	10.23	42.38	4.14	10.09	115.09
100-200	57024	724.25	0.03	8.50	40.60	4.78	10.74	130.30
TC								
0-5	19326	867.39	0.27	108.08	153.19	1.42	1.62	1.20
5-15	29995	841.98	0.12	49.95	94.35	1.89	3.63	13.21
15-30	17876	730.22	0.15	42.29	95.21	2.25	3.93	15.49
30-60	16899	635.38	0.23	31.01	86.01	2.77	4.77	23.34
60-100	13830	700.00	0.24	24.73	80.21	3.24	5.47	30.79
100-200	10206	878.00	0.27	15.43	56.02	3.63	8.18	75.87
TN								
0-5	55340	98.8	0.08	3.24	4.92	1.52	4.00	28.12
5-15	64402	87.31	0.05	2.33	3.87	1.66	5.68	52.19
15-30	55506	79.96	0.03	1.73	3.47	2.01	6.47	56.90
30-60	49382	85.64	0.04	1.29	3.24	2.50	6.88	59.69
60-100	36529	83.45	0.06	1.07	3.22	3.02	7.63	77.00
100-200	22134	75.29	0.05	0.88	2.95	3.37	8.33	91.63
CEC								
0-5	24030	108.71	0.01	11.79	13.12	1.11	2.63	8.79

Depth interval (cm)	Number	Max	Min	Mean	SD	CV	Skewness	Kurtosis
5-15	27759	99.07	0.02	11.02	12.77	1.16	2.85	10.32
15-30	27114	102.05	0.01	10.07	12.23	1.21	2.88	10.72
30-60	27644	106.66	0.03	11.10	12.82	1.16	2.53	8.40
60-100	25361	99.79	0.02	11.75	13.14	1.12	2.37	7.48
100-200	19588	110.41	0.01	12.03	13.54	1.13	2.41	7.64
CECPH7								
0-5	41918	179.88	0.04	20.85	19.14	0.92	2.52	8.60
5-15	44034	155.34	0.03	18.93	16.42	0.87	2.68	11.27
15-30	43913	144.46	0.03	17.30	15.17	0.88	2.69	12.48
30-60	43248	167.71	0.02	17.00	14.63	0.86	2.62	12.69
60-100	38160	194.42	0.01	16.61	14.20	0.86	2.62	13.30
100-200	31553	146.64	0.01	15.79	13.46	0.85	2.57	12.93
cfvo								
0-5	25563	98.57	0.23	7.42	12.07	1.63	3.06	11.44
5-15	30224	94.77	0.24	7.61	12.48	1.64	3.00	10.88
15-30	29892	93.17	0.26	9.48	14.72	1.55	2.62	7.81
30-60	30703	93.91	0.24	11.90	17.42	1.46	2.17	4.71
60-100	28330	92.41	0.23	14.42	19.56	1.36	1.87	3.16
100-200	22391	84.84	0.21	15.68	20.64	1.32	1.79	2.82
cfgr								
0-5	18121	98.52	0.51	13.29	17.67	1.33	1.81	3.03
5-15	21517	98.00	0.23	13.26	17.44	1.32	1.74	2.68
15-30	21506	97.98	0.24	15.19	19.48	1.28	1.61	2.05
30-60	21690	97.52	0.24	16.87	21.53	1.28	1.51	1.48
60-100	19088	96.51	0.21	18.51	23.06	1.25	1.42	1.10
100-200	14856	97.07	0.23	19.38	23.77	1.23	1.39	0.94
bdfi33								
0-5	9782	2.40	0.05	1.37	0.29	0.21	-1.16	2.19
5-15	11802	2.23	0.06	1.39	0.27	0.19	-1.15	2.18
15-30	12402	2.29	0.04	1.45	0.25	0.17	-1.38	3.64
30-60	12439	2.25	0.06	1.50	0.23	0.15	-1.47	4.93

Depth interval (cm)	Number	Max	Min	Mean	SD	CV	Skewness	Kurtosis
60-100	11466	2.43	0.04	1.56	0.23	0.14	-1.34	5.52
100-200	9721	2.49	0.04	1.60	0.22	0.14	-0.98	4.27
wv0033								
0-5	2639	84.21	0.10	23.34	15.09	0.65	0.68	-0.08
5-15	3672	81.29	0.77	22.91	13.85	0.60	0.67	0.09
15-30	2758	79.00	0.76	23.46	13.92	0.59	0.62	0.10
30-60	2675	78.83	0.52	24.40	13.46	0.55	0.60	0.35
60-100	2368	78.39	0.82	25.21	13.32	0.53	0.58	0.44
100-200	1594	78.70	0.48	24.70	13.51	0.55	0.67	0.71
wv1500								
0-5	2602	70.12	0.17	13.35	10.38	0.78	1.33	2.19
5-15	3747	69.18	0.09	13.93	10.38	0.75	1.20	1.57
15-30	2888	68.70	0.11	14.77	10.55	0.71	1.18	1.60
30-60	2791	69.87	0.08	16.23	10.74	0.66	1.09	1.53
60-100	2422	71.63	0.84	17.01	10.72	0.63	1.06	1.42
100-200	1462	64.64	0.27	16.8	10.73	0.64	1.07	1.34
wv0010								
0-5	2100	82.96	1.00	38.30	16.12	0.42	0.22	-0.44
5-15	2122	80.87	1.00	36.44	14.01	0.38	0.09	-0.27
15-30	2095	75.57	1.00	35.88	12.68	0.35	-0.01	-0.05
30-60	2055	74.89	3.96	36.03	11.98	0.33	0.02	0.26
60-100	2023	75.32	3.78	36.01	11.94	0.33	0.27	0.71
100-200	1084	76.39	3.66	35.53	11.69	0.33	0.30	1.19
wg0033								
0-5	14138	98.66	0.44	28.00	13.34	0.48	1.00	1.54
5-15	16908	99.5	0.08	27.28	12.74	0.47	1.02	1.73
15-30	17550	91.51	0.08	26.17	12.06	0.46	1.06	2.01
30-60	17503	82.41	0.01	25.95	11.98	0.46	1.01	1.96
60-100	15357	86.46	0.02	25.34	12.15	0.48	0.98	1.74
100-200	11689	85.96	0.00	23.88	11.75	0.49	1.02	2.00
wg1500								

Depth interval (cm)	Number	Max	Min	Mean	SD	CV	Skewness	Kurtosis
0-5	26068	95.71	0.01	13.63	11.51	0.84	2.61	8.85
5-15	28263	97.74	0.04	12.63	9.36	0.74	2.51	10.28
15-30	28419	92.93	0.01	12.21	8.50	0.70	2.31	10.27
30-60	28115	86.98	0.02	12.73	8.32	0.65	1.85	7.50
60-100	25225	83.74	0.00	12.54	8.22	0.66	1.56	5.44
100-200	19903	80.74	0.02	11.85	8.02	0.68	1.53	5.10
wg0010								
0-5	1650	88.99	2.12	27.46	13.59	0.49	0.88	1.36
5-15	2123	83.36	1.24	26.99	13.32	0.49	0.86	1.32
15-30	2278	82.93	1.55	26.06	12.57	0.48	0.72	1.10
30-60	2333	83.28	0.30	24.84	12.56	0.51	0.77	1.25
60-100	1973	82.62	0.18	23.01	12.94	0.56	0.92	1.50
100-200	1603	79.54	0.19	22.31	12.92	0.58	0.80	0.88
PBR								
0-5	7609	99.38	0.21	24.26	30.41	1.25	2.00	4.02
5-15	8511	98.43	0.23	20.62	27.21	1.32	2.16	4.84
15-30	8010	98.58	0.51	14.24	21.25	1.49	2.70	8.60
30-60	7574	96.61	0.45	9.95	17.48	1.76	3.46	14.71
60-100	6074	97.82	0.43	10.46	18.78	1.80	3.44	14.32
100-200	4124	97.83	0.42	12.52	20.87	1.67	3.19	12.56

CV: coefficient of variation; SD: standard deviation;

Table S7. Accuracy evaluation of the soil properties prediction with GSDEv2 global baseline predictions before regional fusion, GSDEv1, SoilGrids 2.0, and HWSD 2.0, based on the randomly held-out soil profiles. The "N" column indicates the number of samples used during testing. See Table 1 for the abbreviations and units of the soil properties considered.

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

Table S7. Continued.

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
CEC	0-5	2592	0.51	8.86	-0.23	0.32	10.44	-0.45	0.45	9.39	-0.32	0.28	10.74	-0.38
	5-15	2920	0.52	8.07	-0.37	0.31	9.68	-0.41	0.42	8.87	-0.34	0.28	9.88	0.28
	15-30	2837	0.48	8.94	-0.01	0.31	10.30	-0.12	0.44	9.28	-0.24	0.25	10.74	0.12
	30-60	2822	0.48	8.37	0.10	0.28	9.85	-0.23	0.40	8.99	-0.25	0.24	10.12	0.15
	60-100	2502	0.51	8.05	0.02	0.25	9.96	-0.57	0.38	9.06	-0.27	0.18	10.41	0.37
	100-200	1727	0.44	8.93	0.22	0.25	10.33	-0.32	0.37	9.47	-0.18	0.19	10.74	0.21
cfvo	0-5	2449	0.37	12.20	-0.27	0.16	14.09	-0.57	0.30	12.86	-0.62	0.13	14.34	-0.32
	5-15	2860	0.36	10.76	-0.31	0.12	12.62	-0.51	0.28	11.41	-0.63	0.13	12.55	-0.37
	15-30	2825	0.33	10.34	-0.25	0.12	11.85	-0.48	0.25	10.94	-0.67	0.12	11.85	-0.51
	30-60	2880	0.31	10.68	-0.47	0.11	12.13	-0.41	0.25	11.13	-0.52	0.11	12.13	-0.51
	60-100	2620	0.30	14.55	-0.88	0.10	16.50	-0.52	0.24	15.16	-0.51	0.10	16.50	-0.15
	100-200	1995	0.29	15.60	-0.67	0.10	17.56	-0.68	0.22	16.35	-0.67	0.11	17.47	-0.46
TN	0-5	5279	0.54	4.49	-1.26	0.21	5.88	-2.21	0.41	5.09	-2.15	0.19	5.96	-3.21
	5-15	6066	0.53	5.44	0.05	0.20	7.10	-2.45	0.47	5.78	-2.11	0.17	7.23	0.95
	15-30	5376	0.47	6.61	1.01	0.20	8.12	-2.12	0.45	6.73	-2.18	0.17	8.27	-0.12
	30-60	4807	0.48	5.73	-0.17	0.18	7.20	-1.48	0.43	6.00	-1.92	0.15	7.33	-0.23
	60-100	3595	0.45	2.45	-0.09	0.15	3.03	1.12	0.44	2.46	-1.94	0.14	3.05	-0.24
	100-200	2012	0.43	1.97	0.14	0.12	2.45	-0.92	0.41	2.00	-2.21	0.09	2.49	-0.59
SILT	0-5	10120	0.71	11.68	-0.11	0.39	16.94	0.23	0.65	12.83	0.56	0.31	18.08	0.37
	5-15	11383	0.68	11.85	-0.04	0.37	16.63	0.21	0.63	12.74	0.48	0.30	17.53	0.28
	15-30	11376	0.68	11.63	-0.20	0.35	16.58	0.10	0.58	13.32	-0.31	0.28	17.45	0.21
	30-60	10600	0.63	11.96	0.05	0.33	16.09	0.19	0.58	12.74	-0.23	0.27	16.80	0.84
	60-100	9353	0.57	12.88	-0.04	0.29	16.55	0.25	0.55	13.18	-0.24	0.19	17.68	0.34
	100-200	6766	0.54	13.97	0.02	0.17	18.77	0.23	0.53	14.13	0.31	0.12	19.32	-0.51
TC	0-5	1388	0.55	87.09	-1.10	0.23	113.92	-5.21	-	-	-	-	-	-
	5-15	2349	0.53	53.55	-3.19	0.22	68.92	-3.18	-	-	-	-	-	-
	15-30	1392	0.43	58.61	-2.03	0.23	68.12	-4.12	-	-	-	-	-	-
	30-60	1342	0.44	36.04	-2.64	0.20	43.08	-3.18	-	-	-	-	-	-
	60-100	1145	0.47	59.90	-0.91	0.19	74.05	-2.22	-	-	-	-	-	-
	100-200	870	0.33	43.63	0.04	0.17	48.56	-3.14	-	-	-	-	-	-

Table S7. Continued.

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
CECPH7	0-5	4192	0.52	13.16	-0.39	-	-	-	-	-	-	-	-	-
	5-15	4403	0.44	11.04	-0.24	-	-	-	-	-	-	-	-	-
	15-30	4391	0.41	12.68	-0.35	-	-	-	-	-	-	-	-	-
	30-60	4325	0.39	10.79	-0.46	-	-	-	-	-	-	-	-	-
	60-100	3816	0.39	11.90	-0.52	-	-	-	-	-	-	-	-	-
	100-200	3155	0.41	10.28	-0.08	-	-	-	-	-	-	-	-	-
bdfi33	0-5	977	0.46	0.21	0.01	-	-	-	-	-	-	-	-	-
	5-15	1180	0.42	0.19	0.02	-	-	-	-	-	-	-	-	-
	15-30	1240	0.43	0.18	0.00	-	-	-	-	-	-	-	-	-
	30-60	1244	0.37	0.18	0.00	-	-	-	-	-	-	-	-	-
	60-100	1146	0.32	0.19	0.00	-	-	-	-	-	-	-	-	-
	100-200	972	0.36	0.19	0.00	-	-	-	-	-	-	-	-	-
PBR	0-5	980	0.34	14.43	-0.96	-	-	-	-	-	-	-	-	-
	5-15	972	0.34	18.84	-0.53	-	-	-	-	-	-	-	-	-
	15-30	1012	0.32	22.32	1.05	-	-	-	-	-	-	-	-	-
	30-60	988	0.31	18.38	-0.78	-	-	-	-	-	-	-	-	-
	60-100	982	0.30	17.45	-0.62	-	-	-	-	-	-	-	-	-
	100-200	965	0.32	17.68	-1.35	-	-	-	-	-	-	-	-	-
cfgr	0-5	1874	0.38	16.86	-0.24	-	-	-	-	-	-	-	-	-
	5-15	1848	0.37	13.22	-1.02	-	-	-	-	-	-	-	-	-
	15-30	1754	0.34	16.21	-0.31	-	-	-	-	-	-	-	-	-
	30-60	1648	0.33	14.14	0.97	-	-	-	-	-	-	-	-	-
	60-100	1641	0.31	18.14	-1.24	-	-	-	-	-	-	-	-	-
	100-200	1554	0.32	18.54	-1.19	-	-	-	-	-	-	-	-	-
wv0010	0-5	210	0.62	9.66	0.60	0.18	14.19	1.32	-	-	-	-	-	-
	5-15	212	0.67	8.19	-0.31	0.17	12.99	-0.58	-	-	-	-	-	-
	15-30	210	0.51	7.75	0.02	0.13	10.33	0.76	-	-	-	-	-	-
	30-60	206	0.49	7.56	0.30	0.14	9.82	-0.35	-	-	-	-	-	-
	60-100	202	0.45	7.95	-0.52	0.11	10.11	0.21	-	-	-	-	-	-
	100-200	108	0.43	9.20	0.71	0.08	11.69	0.34	-	-	-	-	-	-

Table S7. Continued.

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
wv0033	0-5	264	0.67	8.47	0.37	0.26	12.68	1.47	-	-	-	-	-	-
	5-15	367	0.62	8.99	0.11	0.23	12.80	1.78	-	-	-	-	-	-
	15-30	276	0.61	8.35	-1.13	0.22	11.81	0.95	-	-	-	-	-	-
	30-60	268	0.54	9.38	-0.25	0.22	12.21	0.72	-	-	-	-	-	-
	60-100	237	0.50	8.47	-1.07	0.20	10.71	0.63	-	-	-	-	-	-
	100-200	159	0.50	8.86	-0.46	0.20	11.21	1.15	-	-	-	-	-	-
wv1500	0-5	260	0.62	5.99	-0.98	0.28	8.25	1.11	-	-	-	-	-	-
	5-15	375	0.58	6.50	-0.47	0.25	8.69	1.45	-	-	-	-	-	-
	15-30	289	0.52	7.24	-0.77	0.27	8.93	1.12	-	-	-	-	-	-
	30-60	278	0.50	7.71	-0.31	0.22	9.63	1.65	-	-	-	-	-	-
	60-100	242	0.55	7.23	0.88	0.21	9.58	1.22	-	-	-	-	-	-
	100-200	146	0.51	7.54	-0.15	0.21	9.57	0.83	-	-	-	-	-	-
wg0010	0-5	384	0.45	9.33	0.58	-	-	-	-	-	-	-	-	-
	5-15	374	0.40	10.19	-0.85	-	-	-	-	-	-	-	-	-
	15-30	376	0.39	9.76	-0.27	-	-	-	-	-	-	-	-	-
	30-60	365	0.42	9.12	-0.33	-	-	-	-	-	-	-	-	-
	60-100	346	0.41	8.68	-1.61	-	-	-	-	-	-	-	-	-
	100-200	289	0.41	10.16	-1.93	-	-	-	-	-	-	-	-	-
wg0033	0-5	354	0.41	10.17	-0.54	-	-	-	-	-	-	-	-	-
	5-15	345	0.41	9.40	-0.48	-	-	-	-	-	-	-	-	-
	15-30	341	0.42	9.29	-0.01	-	-	-	-	-	-	-	-	-
	30-60	332	0.38	9.16	-0.16	-	-	-	-	-	-	-	-	-
	60-100	351	0.40	8.91	-0.32	-	-	-	-	-	-	-	-	-
	100-200	295	0.38	8.76	-0.15	-	-	-	-	-	-	-	-	-
wg1500	0-5	312	0.41	9.14	-0.32	-	-	-	-	-	-	-	-	-
	5-15	289	0.42	8.91	0.83	-	-	-	-	-	-	-	-	-
	15-30	271	0.38	6.88	0.08	-	-	-	-	-	-	-	-	-
	30-60	276	0.38	6.88	0.08	-	-	-	-	-	-	-	-	-
	60-100	277	0.41	5.94	-0.39	-	-	-	-	-	-	-	-	-
	100-200	291	0.39	6.15	-0.36	-	-	-	-	-	-	-	-	-

"-" indicates that the soil property was not mapped or not included in the dataset.

Table S8. Regional-fusion implementation, vertical harmonisation, and source-overlap assessment for the regional and national soil products used in GSDEv2.

Fusion region	Regional product	Properties used	Native depths	Harmonisation	Implemented intervals	Unsupported intervals	Training-data provenance	Potential overlap	Interpretation
China	CSDLv2, Shi et al. (2025)	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC, cfvo	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Direct interval matching; no vertical extrapolation	All six GSDEv2 intervals for the listed properties	None for the listed properties.	SNSSC, WoSIS, regional Chinese profiles, and FNSSC	Confirmed overlap with P01, P02, and P03	Source overlap with P01, P02, and P03 was confirmed and could not be fully removed from the regional product. The assessment was therefore spatially blocked but not fully source-independent.
China	National Soil Information Grids of China, Liu et al. (2020, 2022)	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC, cfvo	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Direct interval matching; no vertical extrapolation	All six GSDEv2 intervals for the listed properties	None for the listed properties.	4,844 representative profiles from the National Soil Series Survey and Compilation of Soil Series of China conducted during 2009–2019	Unresolved; no profile-level identifiers were available to determine possible overlap with P04	Possible profile overlap could not be excluded. This product and the China soil-texture product were developed from the same recent national soil-survey framework and were therefore treated as related rather than independent regional predictors.
United States	SOLUS100m, Nauman et al. (2024)	SAND, SILT, CLAY, PHH2O, BD, SOC, CEC, CECPH7, cfvo	Predictions at standard depth points of 0, 5, 15, 30, 60, 100, and 150 cm	A mass-preserving equal-area spline was fitted to the discrete depth estimates and integrated over each supported GSDEv2 target interval; no extrapolation beyond the native depth range	0–5, 5–15, 15–30, 30–60, and 60–100 cm	100–200 cm was rejected because native vertical support ended at 150 cm and did not cover the complete target interval	USDA–NRCS Soil Characterization Database laboratory observations, NASIS pedons linked to soil-survey map-unit components, and sampled gNATSGO map-unit weighted-average estimates	Possible overlap with US profiles included in P01 and/or P05	Complete training-profile identifiers were unavailable for cross-product matching. Potential overlap could not be excluded, and the assessment was interpreted as spatially blocked but not fully source-independent.

Fusion region	Regional product	Properties used	Native depths	Harmonisation	Implemented intervals	Unsupported intervals	Training-data provenance	Potential overlap	Interpretation
Africa	iSDAsoil, Hengl et al. (2021)	SAND, SILT, CLAY, PHH2O, BD, SOC, TC, TN, CEC, cfgr	Regional layers representing 0–20 and 20–50 cm soil-depth support	Overlap-thickness-weighted averaging of the native layers; no extrapolation beyond 50 cm	0–5, 5–15, and 15–30 cm	30–60 cm was rejected because only the 30 – 50 cm portion was natively supported; 60 – 100 and 100 – 200 cm were rejected because native information ended at 50 cm	Approximately 150,000 soil samples compiled from AfSIS, ISRIC African soil-profile collections, LandPKS, IFDC, AfricaRice/TAMASA, and additional national, governmental, NGO, and project datasets	Possible overlap with African observations included in P01 and P05	Complete profile-level source matching was unavailable. Potential overlap could not be excluded, and the regional assessment was treated as a spatially blocked comparative evaluation.
Canada	Canadian soil grids, Geng et al. (2025)	SAND, SILT, CLAY, BD, SOC, CEC	0-5, 5-15, 15-30, 30-60, and 60-100 cm	Direct interval matching; no vertical extrapolation	0–5, 5–15, 15–30, 30–60, and 60–100 cm	100–200 cm was rejected because the regional product did not extend below 100 cm	Soil-point data from the Canadian Soil Information Service, including the National Pedon Database, and the Canadian Forest Service upland forest soil-profile database	Possible overlap with Canadian profiles included in P01 and/or P05	Potential profile overlap could not be excluded from the available provenance information. The assessment was spatially blocked but was not regarded as fully source-independent.
Australia	Soil and Landscape Grid of Australia, Version 2 (SLGA), Malone et al. (2025)	SAND, SILT, CLAY, PHH2O, BD, SOC, TN, CEC	0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm	Direct interval matching; no vertical extrapolation	All six GSDEv2 intervals for the listed properties	None for the listed properties.	Australian National Soil Site Data Collation and SoilDataFederator records, state and territory soil-information systems, CSIRO soil databases, and property-specific laboratory, field-observation, and spectral datasets; Malone et al. (2025)	Possible overlap with Australian profiles included in P01 and/or P05	Complete product-specific training records were unavailable for cross-product matching. Potential overlap could not be excluded, and the assessment was interpreted as spatially blocked but not fully source-independent.

a) Regional products were included as auxiliary predictors in the stacking models rather than being directly substituted for the GSDEv2 global baseline predictions.

b) Fusion was implemented only when the complete GSDEv2 target interval was supported by the native vertical range of the regional product. Partial vertical support and extrapolation beyond the shallowest or deepest native support were not permitted.

c) "Confirmed overlap" indicates that the regional-product publication explicitly reported one or more profile sources also included in the GSDEv2 compilation. "Possible overlap" indicates that a regional product used national, continental, or federated soil-profile collections that may contain profiles also distributed through WoSIS or the OpenGeoHub soil-observation compendium, but profile-level identity could not be verified..

d) The nested spatial-block design prevented leakage from the newly fitted GSDEv2 baseline predictions into the stacking meta-model but could not remove information already embedded in regional products trained using overlapping observations. Regional performance was therefore interpreted as a spatially blocked comparative assessment rather than fully source-independent validation.

e) The National Soil Information Grids of China and the high-resolution three-dimensional soil-texture maps of China were both developed using profiles from the National Soil Series Survey and Compilation of Soil Series of China conducted during 2009–2019. They were treated as related regional products, and their agreement was not interpreted as independent evidence.

f) P01–P05 refer to the GSDEv2 soil-profile and soil-observation sources defined in Table S1.

g) Region-specific boundary-transition widths were 10 km for China, 20 km for the United States, 50 km for Africa, 20 km for Canada, and 50 km for Australia. These widths were predefined operational settings applied only to supported property–depth combinations.

Table S9. Configuration of the CoLM2024 soil-input sensitivity experiments driven by GSDEv2, GSDEv1, and SoilGrids 2.0.

Configuration	CoLM-GSDEv2	CoLM-GSDEv1	CoLM-SoilGrids 2.0
CoLM version	CoLM2024, https://github.com/CoLM-SYSU/CoLM	same	same
Domain	global	global	global
Horizontal grid	0.25°	0.25°	0.25°
Simulation period	2013–2018	same	same
Spin-up endpoint	31-Dec-17	same	same
Repeated spin-up cycles	0	0	0
Time step	1800 s	same	same
Atmospheric forcing	ERA5 configuration	same	same
LAI	monthly, no yearly change	same	same
History output	daily	daily	daily
Changed input	GSDEv2 raw soil data	GSDEv1 raw soil data	SoilGrids raw soil data
Evaluation year	2018	2018	2018
Evaluated variable	first-layer f_h2osoi	same	same

Note: All model, atmospheric-forcing, initialisation, temporal, and output settings were identical among the three experiments; only the soil-input dataset differed.

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, Adv. Agron., 125, 93–134, https://doi.org/10.1016/B978-0-12-800137-0.00003-0 , 2014.

Bartholomé, E. and Belward, A. S.: GLC2000: a new approach to global land cover mapping from Earth observation data, Int. J. Remote Sens., 26, 1959–1977, https://doi.org/10.1080/01431160412331291297, 2005.

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.

Calisto, L., de Sousa, L. M., and Batjes, N. H.: Standardised soil profile data for the world (WoSIS snapshot – December 2023), ISRIC – World Soil Information [data set], https://doi.org/10.17027/isric-wdcsoils-20231130, 2023.

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.

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

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.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G.: High-resolution global maps of 21st-century forest cover change, Science, 342, 850–853, https://doi.org/10.1126/science.1244693 , 2013.

Hartmann, J. and Moosdorf, N.: Global Lithological Map Database v1.0 (gridded to 0.5° spatial resolution), PANGAEA [data set], https://doi.org/10.1594/PANGAEA.788537, 2012.

Hengl, T. and Gupta, S.: An Open Compendium of Soil Datasets: Soil Observations and Measurements, OpenGeoHub Foundation [data set], version 0.2, https://doi.org/10.5281/zenodo.4748499, 2025.

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 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.

Karger, D. N., Böhner, J., Conrad, O., Linder, H. P., Kessler, M., Soria-Auza, R. W., Kawohl, T., Kreft, H., and Zimmermann, N. E.: CHELSA V2.1, EnviDat [data set], https://doi.org/10.16904/envidat.228.v2.1 , 2021.

Lang, N., Jetz, W., Schindler, K., and Wegner, J. D.: A high-resolution canopy height model of the Earth, Nat. Ecol. Evol., 7, 1778–1789, https://doi.org/10.1038/s41559-023-02206-6 , 2023.

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, Sci. Bull., 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), Eur. J. Soil Sci., 70, 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.

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

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 Sci. Soc. Am. J., 88, 2046–2065, https://doi.org/10.1002/saj2.20769 , 2024.

Pelletier, J. D., Broxton, P. D., Hazenberg, P., Zeng, X., Troch, P. A., Niu, G.-Y., Williams, Z., Brunke, M. A., and Gochis, D.: A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling, J. Adv. Model. Earth Syst., 8, 41–65, https://doi.org/10.1002/2015MS000526, 2016.

Shangguan, W., Dai, Y., Liu, B., Ye, A., and Yuan, H.: A soil particle-size distribution dataset for regional land and climate modelling in China, Geoderma, 171–172, 85–91, https://doi.org/10.1016/j.geoderma.2011.01.013 , 2012.

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.

Tuanmu, M.-N. and Jetz, W.: A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling, Global Ecol. Biogeogr., 24, 1329–1339, https://doi.org/10.1111/geb.12365, 2015.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S., Kang, S. Y., Cameron, E., Bisanzio, D., Battle, K. E., Bhatt, S., and Gething, P. W.: A global map of travel time to cities to assess inequalities in accessibility in 2015, Nature, 553, 333–336, https://doi.org/10.1038/nature25181 , 2018.

Wilson, A. M. and Jetz, W.: Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions, PLoS Biol., 14, e1002415, https://doi.org/10.1371/journal.pbio.1002415 , 2016.

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., Sampson, C. C., Kanae, S., and Bates, P. D.: A high-accuracy map of global terrain elevations, Geophys. Res. Lett., 44, 5844–5853, https://doi.org/10.1002/2017GL072874 , 2017.