Received: 18 January 2022 | Revised: 30 June 2022 | Accepted: 6 July 2022

DOI: 10.1111/ejss.13285

ORIGINAL ARTICLE

European Journal of Soil Science WILEY

Using homosoils for quantitative extrapolation of soil mapping models

Andree M. Nenkam[1] | **Alexandre M. J.-C. Wadoux**[1] | **Budiman Minasny**[1] | **Alex B. McBratney**[1] | **Pierre C. S. Traore**[2] | **Gatien N. Falconier**[3,4,5] | **Anthony M. Whitbread**[6]

[1]Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia

[2]The Resilient Farm and Food Systems (RFFS) program, International Crop Research Institute for the Semi-Arid Tropics, Dakar, Senegal

[3]UPR-Agroécologie et Intensification Durable des Cultures Annuelles (AIDA), Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, France

[4]Sustainable Agrifood Systems Program (SAS), International Maize and Wheat Improvement Centre (CIMMYT), Harare, Zimbabwe

[5]AIDA, Univ Montpellier, CIRAD, Montpellier, France

[6]Sustainable Livestock Systems Program, International Livestock Research Institute (ILRI), C/- IITA ESA Hub, Dar es Salaam, Tanzania

Correspondence
Andree M. Nenkam, Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales, Australia.
Email: andree.nenkam@sydney.edu.au

Funding information
CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), Grant/Award Number: CGIAR Trust Fund; Accelerating Impacts of

Abstract

Since the early 2000s, digital soil maps have been successfully used for various applications, including precision agriculture, environmental assessments and land use management. Globally, however, there are large disparities in the availability of soil data on which digital soil mapping (DSM) models can be fitted. Several studies attempted to transfer a DSM model fitted from an area with a well-developed soil database to map the soil in areas with low sampling density. This usually is a challenging task because two areas have hardly ever the same soil-forming factors in two different regions of the world. In this study, we aim to determine whether finding homosoils (i.e., locations sharing similar soil-forming factors) can help transferring soil information by means of a DSM model extrapolation. We hypothesize that within areas in the world considered as homosoils, one can leverage on areas with high sampling density and fit a DSM model, which can then be extrapolated geographically to an area with little or no data. We collected publicly available soil data for clay, silt, sand, organic carbon (OC), pH and total nitrogen (N) within our study area in Mali, West Africa and its homosoils. We fitted a regression tree model between the soil properties and environmental covariates of the homosoils, and applied this model to our study area in Mali. Several calibration and validation strategies were explored. We also compared our approach with existing maps made at a global and a continental scale. We concluded that geographic model extrapolation within homosoils was possible, but that model accuracy dramatically improved when local data were included in the calibration dataset. The maps produced from models fitted with data from homosoils were more accurate than existing products for this study area, for three (silt, sand, pH) out of six soil properties. This study would be relevant to areas with very little or no soil data to carry critical soils and environmental risk assessments at a regional level.

CGIAR Climate Research for Africa, Grant/Award Number: 173398; University of Sydney

Highlights

- Soil mapping models were fitted with soil data within the homosoils of Mali.
- The fitted models were applied to our study area.
- Model accuracy dramatically improved when including local data.
- Homosoil maps were more accurate for 3 out of 6 soil properties compared to global and continental maps.
- New opportunity to map the regional soil pattern of areas with limited soil data coverage.

KEYWORDS

cubist, digital soil mapping, model-based validation, soil spatial variation, soil-forming factors

1 | INTRODUCTION

Digital soil mapping (DSM) has gained importance for the last two decades (Minasny and McBratney, 2016) and soil maps have been effectively used in several applications, including precision agriculture (Shatar and McBratney, 1999), land degradation mitigation (Raina et al., 1993) and environmental and land use management (Hartemink, 2002). Digital soil maps are usually produced with statistical techniques that relate soil data collected at sites with spatially exhaustive environmental covariates known to influence soil formation. Common techniques used for DSM are geostatistics (Heuvelink and Webster, 2001) and machine learning (Wadoux et al., 2020).

Globally however, there are large disparities in the availability of soil data on which DSM models can be fitted. Soil data density varies dramatically among areas (Minasny et al., 2013). For example, Hengl et al. (2017a) had a sampling density of 1 sample per 1000 km^2 for mapping a wide range of soil properties globally, whereas Hengl et al. (2017b) had a sampling density of 3 samples per 10,000 km^2 for mapping soil minerals in the whole continent of Africa. This disparity can be attributed to the priority given to soil data collection (Minasny et al., 2013) and to the lack of funding in different countries. Several areas in the world still have a relatively low soil data coverage available publicly. This precludes the development of DSM because DSM models are data-driven and hence rely on the quantity and spatial distribution of the available soil data (Wadoux et al., 2020). One obvious solution to this problem is to collect data in the area of interest through an additional soil survey, which requires investment. Another solution that may appear cheaper in terms of new soil data collection and more readily applied is to extrapolate DSM models from one area to another. Extrapolation of DSM models relies on the assumption that the empirical relationships between the soil property and the environmental covariates are structured similarly between the two areas so that they can be transferred.

Axiomatically, two regions with similar soil-forming factors should develop similar soils. Any soil is a function of factors of soil formation (Dokuchaev, 1883). This concept was later adapted by Jenny (1941) through a state factor model of soil formation which provided a convenient theoretical basis for McBratney et al. (2003)' *scorpan* equation for DSM. Using the principle that similar soil forming factors lead to similar soils, several studies attempted to extrapolate a DSM model between two areas which were assumed to have similar soils (e.g., by Bui and Moran, 2003; Thompson et al., 2006; Lemercier et al., 2012; Cambule et al., 2013; Silva et al., 2016; Abbaszadeh Afshar et al., 2018; Du et al., 2021; Summerauer et al., 2021). Grinand et al. (2008) predicted soil classes between two adjacent areas using different environmental covariates (topography, lithology, land-cover) and concluded that upscaling the covariates to represent the regional trend of soil spatial distribution significantly improved predictions accuracy. Malone et al. (2016) predicted soil spectral indices using terrain attributes across two areas in the same region and found that the extrapolated model's accuracy was dependent on the covariate similarity between the two areas. Angelini et al. (2020) predicted quantitative soil properties across geographically remote areas using structural equation modelling that combines expert pedological knowledge and statistical correlation. They concluded that differences in the soil-covariates relationship between the two areas were the main causes for model prediction accuracy. Overall, these studies concluded that extrapolation is a challenging task and that the model's accuracy decreased when applied to the extrapolated area.

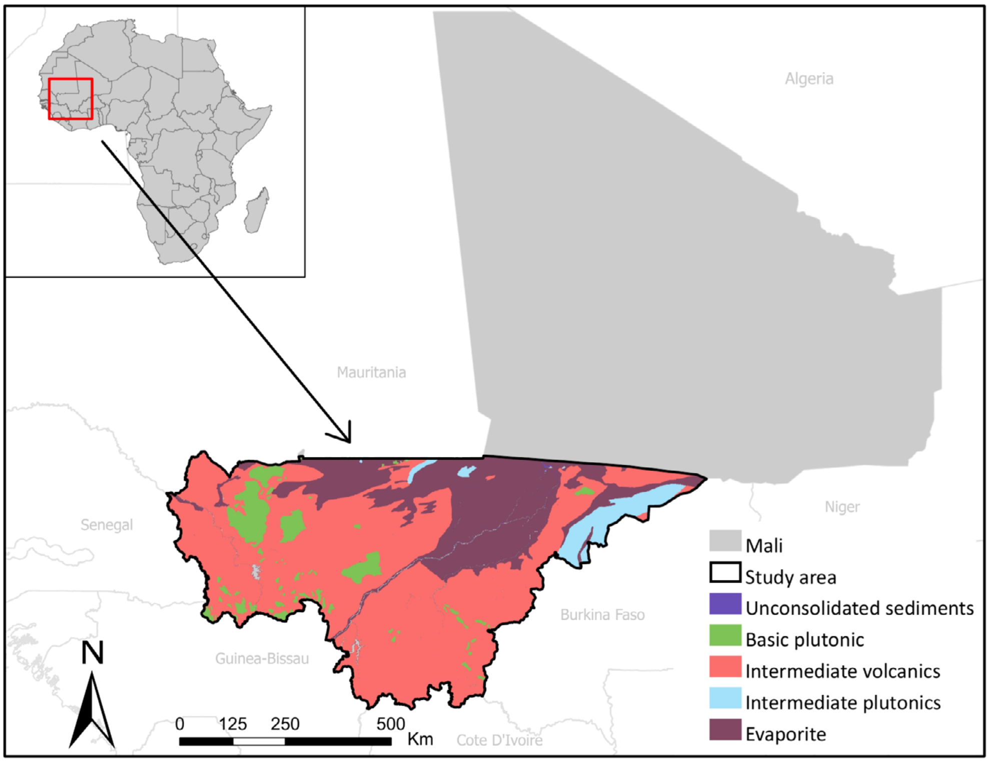

FIGURE 1 Area of interest located in the southern part of Mali. The map depicts the major parent material that spans over the study area.

In our previous study (Nenkam et al., 2022), we used the homosoil concept to find areas in the world with similar soils, with the objective of obtaining new soil data for an area of interest. Homosoils are any two soils in the world sharing similar soil-forming factors. This concept assumes that soils with similar soil-forming factors have undergone similar soil-forming processes in the past leading to similar soils today. Homosoils are relevant for DSM model extrapolation because through the mathematical calculation of covariate similarity indices, it is possible to find soils that might be similar. This may be very useful for DSM purposes, because this information can be used to delineate areas from which DSM models can be fitted or areas within which they can be transferred geographically, under the assumption that the soil-covariate relationship is similar.

In this study, we aim to determine whether finding homosoils can help the geographic extrapolation of a DSM model. We hypothesize that within areas in the world considered as homosoils, we can leverage on areas with high sampling density and fit DSM models, which can then be applied in an area with little or no data. The paper is organized as follows. First, we find homosoils of a study area and collect global soil data available within homosoils. Next, we fit a DSM model using these data, and transfer it to an area with little/no data. Finally, we validate the predicted soil maps and compare them with existing DSM products. We consider clay, organic carbon, pH, sand, silt, and total nitrogen as soil properties of interest.

2 | MATERIAL AND METHODS

2.1 | Study area

Our study area covers 440,000 km^2 in the Southern part of Mali in Western Africa (Figure 1). The area is characterized by a North–South gradient of vegetation: open shrub savanna in the north, dense shrub savanna in the center, and lightly wooded savanna and woody woodland in the south (Rian et al., 2009). The parent materials dominating the area are igneous (intermediate volcanic, basic and intermediate plutonic) and sedimentary (evaporites and unconsolidated sediments) rocks as shown in Figure 1. The semi-arid zone in the north is dominated by Aridisols and receives less than 400 mm of annual rainfall. The Sudanian zone at the center is dominated by Alfisols (annual rainfall <800 mm), while Ultisols and Vertosols dominate the Sudanian-Guinean climatic zone in the south and receive up to 1200–1400 mm of rainfall on average annually (FAO, 2000; Giannini et al., 2017). These soils are often physically and chemically degraded (erosion, low soil nutrient content, acidification, aluminium and iron toxicity) due to unsustainable management practices (Mbow et al., 2015; Lal and Stewart, 2019).

TABLE 1 Soil forming factors and their corresponding environmental covariates and sources.

State factor	Environmental covariates	Source
Climate	A set of 10 first principal components obtained from 24 continuous covariates	Fick and Hijmans (2017)
Lithology	Lithological classes - 1st level elevation	Hartmann and Moosdorf (2012)
Topography	Slope	Rabus et al. (2003)
	Multiscale position topographic index	Theobald et al. (2015)
Bio-vegetal	First three principal components of 12 continuous bio-vegetal covariates	Justice et al. (2002)

Note: The set of covariates used for finding homosoils are obtained after dimension reduction. Categorical covariates were transformed to quantitative covariates prior to this step. We refer to the supplementary material in Nenkam et al. (2022) for the complete set of covariates before dimension reduction.

The main crop types in the study area are rice, cotton, maize, peanut, sorghum and pearl millet.

2.2 | Environmental covariates

We used an initial set of 40 environmental covariates as proxies for the state factors of soil formation. Note that, state factors refer to soil-forming factors where each factor comprises one or more environmental covariates, as shown in Table 1. The covariates consisted of 1 parent material covariate (16 classes), 3 topographic covariates, 12 bio-vegetal covariates and 24 climate covariates. The original spatial resolution of these covariates spans from 30 m × 30 m to 1 km × 1 km. All covariates were resampled to the same spatial resolution with grid cells of 1 km × 1 km resolution using the pyramiding policy toolbox of Google Earth Engine (Gorelick et al., 2017), which computes the mean or the mode (category that appears most frequently within the 1 km × 1 km cell) of lower-level pixels for quantitative or categorical covariates, respectively. Next, we applied principal components analysis to reduce the set of climate and bio-vegetal covariates and selected a number of components that explained at least 97% of the original dataset variance. The parent material covariate (lithology) was converted from categorical to continuous using non-metric multidimensional-scaling (nMDS). More information on nMDS and how it was used can be found in the Supplementary Material. Table 1 shows the final list of 18 covariates after dimension reduction.

2.3 | Homosoils

Homosoils were found by the method described in Nenkam et al. (2022), which we summarize here. The study area was grouped based on the environmental covariates described in Table 1 using the *k*-means clustering method. This was done for computational efficiency, because a homosoil is found for a spatial location and there are hundreds of thousands spatial locations in the area. Therefore, to limit the number of spatial locations for which we find homosoils, we classified our study area into five homogeneous clusters whose pattern was inline with the major agro-climatic regions (Andrieu et al., 2017). We proceeded by finding the homosoil of each cluster centroid, leading to five homosoils corresponding to five spatial locations. Finding the homosoil for a spatial location is done in three steps:

2.3.1 | First, find the homoclime

We sought the homoclime (any two locations in the world with similar climatic conditions) to the spatial locations by computing the similarity indices between its climate and that of each node of a fine grid (1 km × 1 km) of climate covariates covering the world. We consider as homoclime the set of nodes whose similarity indices were smaller than a threshold value. A threshold equivalent to the 10th percentile of the similarity indices was used. This helped remove areas with different climatic conditions from being considered as homoclime.

2.3.2 | Second, build a covariate database

We built a numerical database of environmental covariates within the homoclime's spatial extent. Covariates corresponding to each state factor were obtained based on their global availability (Table 1) and expert pedological knowledge on their impact on soil-forming processes.

2.3.3 | Third, identify homosoils

We identified homosoils by (i) computing the similarity indices between each state factor of the spatial location and that of each node of the fine grid of covariates within the spatial extent of the homoclime, and (ii) average these similarity indices with each state factor being equally weighted. We consider as homosoil the set of nodes whose similarity indices are smaller than a threshold value. The threshold equivalent to the 20th percentile of the similarity indices was selected as a cut-off value based on trial and error. As discussed in Nenkam et al. (2022), the cut-off value mainly controls the spatial extent of the resulting homoclime and homosoils. The Mahalanobis distance (Webster, 1977; De Maesschalck et al., 2000) was used as a similarity index to find both the homoclime and the homosoils.

2.4 | Collecting soil data

Within homosoils, we collected soil data from the World Soil Information Services (WoSIS, Batjes et al., 2020). The soil depth intervals were harmonized using a mass preserving soil-depth function spline (Bishop et al., 1999), to the standard GlobalSoilMap specification depth intervals 0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm. The soil properties collected are clay, silt and sand (%), organic carbon (OC, g/kg), pH (H_20) and total nitrogen (Total N, g/kg). The clay, silt and sand values were harmonized so that their sum equals 100%. These soil properties were selected primarily for their importance for crop growth and availability in the WoSIS dataset.

2.5 | Extrapolation

2.5.1 | Model calibration and extrapolation

We used a model-tree, cubist, based on the M5 algorithm of Quinlan (1992) to define the relationship between the soil data and the environmental covariates. The tree is built by partitioning the covariates into different rules called nodes. Each node consists of a covariate-based condition and an ordinary least square regression model used to make predictions. The condition at a node can be nested with the condition of another node (child node) up to a terminal node referred to as a leaf. Predictions are made at each node, and smoothed using the predictions of the parent node. Predictions made by the model at the terminal node are the final predictions. A cubist model has two primary parameters: committees and neighbour. Committees is a boosting-like parameter used to adjust predictions by creating an iteration of rule-based models so that predictions from one model is adjusted by the predictions from the previous model. The estimates from these individual models are then averaged to generate the final prediction. On the other hand, neighbour defines the number of neighbouring points from the training set and is used to adjust predictions when predicting new samples. Model trees like cubist are advantageous in that the logical construct of the model rules require little data pre-processing and can handle non-linear relationships between the explanatory and the response variables (Kuhn and Johnson, 2013).

Cubist models were calibrated for each soil property and horizon depth interval using two calibration strategies. In the first calibration strategy, models were calibrated using WoSIS soil data within homosoils excluding those of the study area. Models were then geographically extrapolated to predict the soil property within the study area. In the second calibration strategy, models were calibrated using the WoSIS soil data collected for both within homosoils and within the study area, and then used to predict the soil property in the study area. The objective of the first strategy is to evaluate the prediction accuracy when the model (built within homosoils only) is extrapolated geographically, while the second strategy evaluates the prediction accuracy of the model when the training dataset include both the data collected within homosoils and within the study area. Maps of soil properties at 1 km × 1 km resolution were generated for each depth interval. The map predictions of clay, silt and sand at any prediction location were normalized to satisfy the condition that their sum should be equal to 100%.

The cubist models were calibrated with 10 committees for the two calibration strategies, using 0 and 9 neighbours for the prediction of the first and second calibration strategies, respectively. The implementation was provided by the Cubist package (Kuhn and Quinlan, 2020) in the R programming language.

2.5.2 | Map validation

The maps generated from the two models were validated using the following three approaches: the first approach was used to validate the maps from the first calibration strategy. This approach is the most common method for map validation. The second and third approaches were used to validate maps from the second calibration strategy. When we used the first calibration strategy, we validated the predictions using the WoSIS data for the study area (within Mali). Recall that in this calibration strategy, the WoSIS data within the study area were not used for model calibration. Predictions made at validation locations were compared to the measured values with statistical indices.

The second approach for validation made use of models calibrated with the second calibration strategy.

Predictions were obtained through a 10-fold cross-validation of the WoSIS data within the study area, and compared to the measured values of the soil property using statistical indices.

Finally, maps, from models calibrated with the second strategy, are validated using an independent dataset obtained from multiple sources (i.e., Doumbia et al., 2009; Benjaminsen et al., 2010; Verbree et al., 2015; Degerickx et al., 2016; Falconnier et al., 2016; Bayala et al., 2020; Birhanu et al., 2020; Huet et al., 2020). The units of the soil properties in this dataset were harmonized to that of the WoSIS dataset. Their depth intervals were also standardized. However, because most of the sources measured soil properties at 0–15 cm depth interval, the values of the 0–5 and 5–15 cm samples were combined to 0–15 cm where appropriate, and only this dataset was used for validation. Since this independent data exhibit spatial clustering, we use a model-based validation approach to obtain the validation statistics (Brus et al., 2011; de Bruin et al., 2022). This is done by computing the residuals at validation location, and using them to estimate a variogram of the residuals. The sample variogram is fitted by the Methods-of-Moment with a spherical correlation function. We use the fitted variogram to generate 500 simulations of the residuals over a fine grid (1 km × 1 km) covering the study area using sequential Gaussian simulation (Webster and Oliver, 2007). From the 500 residuals fields, we compute the expected values of the validation statistics and their 0.05 and 0.95 quantiles.

The validation statistics used to evaluate and compare the maps are the mean error (ME):

$$\mathrm{ME}=\frac{1}{n}\sum_{i=1}^{n} z(\mathbf{s}_i)-\hat{z}(\mathbf{s}_i), \tag{1}$$

where $z(\mathbf{s}_i)$ and $\hat{z}(\mathbf{s}_i)$ are observed and predicted soil property at location $\mathbf{s}_i(i=1,\ldots,n)$, respectively, and n is the number of validation locations. The ME has an optimal value of 0. Positive and negative ME values indicate under-prediction and over-prediction, respectively.

The root mean squared error (RMSE):

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z(\mathbf{s}_i)-\hat{z}(\mathbf{s}_i)\right)^2}. \tag{2}$$

The RMSE is a nonnegative statistic with no upper bound and optimal value of 0. It indicates the magnitude of the error in the soil property's unit.

The Pearson's r correlation coefficient:

$$r=\frac{\sum_{i=1}^{N}\left(z(\mathbf{s}_i)-\bar{z}\right)\left(\hat{z}(\mathbf{s}_i)-\bar{\hat{z}}\right)}{\sqrt{\sum_{i=1}^{N}\left(z(\mathbf{s}_i)-\bar{z}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\hat{z}(\mathbf{s}_i)-\bar{\hat{z}}\right)}}. \tag{3}$$

where $\bar{z}$ and $\bar{\hat{z}}$ are the mean of the measured and predicted values, respectively.

The modelling efficiency coefficient (MEC, Janssen and Heuberger, 1995) which quantifies the improvement made by the model over using the mean of the validation data as prediction,

$$\mathrm{MEC}=1-\frac{\sum_{i=1}^{n}\left(z(\mathbf{s}_i)-\hat{z}(\mathbf{s}_i)\right)^2}{\sum_{i=1}^{n}\left(z(\mathbf{s}_i)-\bar{z}\right)^2}. \tag{4}$$

A value of 1 indicates a perfect prediction, while a value of 0 indicates that the model prediction is as accurate as using the mean of the validation data as prediction. Note that the MEC can be negative if the residual variance is larger than the variance of the validation data.

2.5.3 | Map comparison with existing products

The maps generated using the second calibration strategy were further compared against existing digital soil maps produced at continental (i.e., iSDAsoil, Hengl et al., 2021) and at global (i.e., SoilGrids, Hengl et al., 2017a) extents. These maps were also generated using the WoSIS database. We compared them using two approaches: first, we carried a visual comparison to evaluate the predicted spatial pattern of the soil property at 0–15 cm soil depth interval. Second, we validated the maps using the third validation approach on an independent dataset and compared the resulting validation statistics. Because of the computational demand to carry the third validation strategy, the global and continental maps were brought to a common resolution of 1 km × 1 km.

3 | RESULTS

3.1 | Homosoils

The homosoil areas for the five cluster centroids were merged and the final map is shown in Figure 2. The colored area represents the spatial extent of the homosoils and shows that many areas in the world have similar soils as the study area in Mali. These areas include Mexico, Eastern Brazil and Northern Argentina in America, the Sahelian band in Africa, Southern Africa, Yemen, Pakistan, India, Myanmar, Thailand and northern

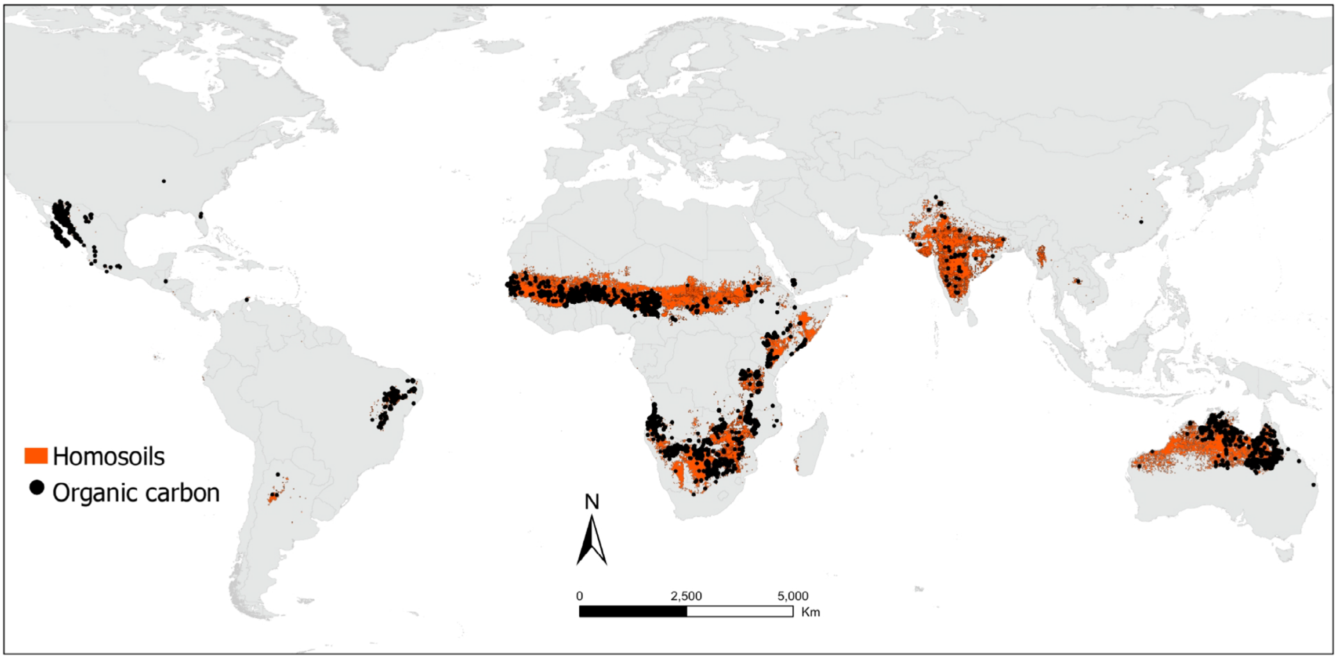

FIGURE 2 Homosoils of the study area. These homosoils extend from the far west in Western Mexico, through the Sahelian band and southern Africa to India and northern Australia. The black dots are locations of soil OC samples collected from WoSIS for the 5–15 cm depth interval which fall within the homosoil areas

Australia. The Homosoils of our study area, therefore, fall within semi-arid and tropical regions of the world. Figure 2 also shows negligible homosoil areas in Eastern American, Guatemala, Venezuela and East China.

3.2 | Data collected

The black dots in Figure 2 are the soil OC samples collected from WoSIS for the soil depth interval 5–15 cm. Many areas within homosoils have high sampling density, such as in inter-alia Western Mexico (365 samples), Eastern Brazil (102 samples), the center of the Sahelian band (Burkina Faso: 648, Niger: 451, and Nigeria: 408). India and the eastern part of the Sahelian band (Tchad, Central Republic, Sudan, Ethiopia and Kenya), conversely, have few samples sparsely distributed. Despite the large areas covered by the latter, they amount to 197 samples.

The total number of soil samples collected for both the study area and the homosoils for all soil properties and depth intervals is presented in Figure 3. Clay had the largest number of soil samples in both areas (565 and 6123 sample for the 5–15 cm intervals within the study area and homosoils, respectively), while OC had the lowest number of samples (450 samples at 5–15 cm) within the study area and total N the lowest number in the homosoils (3748 samples at 5–15 cm). Total N was, overall, the property with the lowest number of samples. For all properties, the number of available samples decreased with depth. The deepest depth interval (100–200 cm) always had the smallest number of samples.

Figure 4 shows the boxplots of the soil properties within the study area and within homosoils. The spread of the soil property values within homosoils was nearly always larger than the spread of the soil property within the study area. Figure 4 also indicates that the average value of soil properties within homosoils and the study area were different. The pH, sand and total N content were larger for homosoils, whereas soils in the study area had a relatively high and constant silt content across all depth intervals. For example, 75% of the measured values of silt were larger than 19%. The average value of silt content in Mali was also 60% higher than that within homosoils. Clay content increased with depth, while soil OC and total N content decreased with depth. Overall, the soil properties in the study area in Mali and in its homosoils showed disparities in the range of values but similar trend across depth intervals.

3.3 | Model calibration and validation

Figure 5 shows the validation statistics (i.e., ME, RMSE, *r* and MEC) for the first and second validation approaches. Validation approaches 1 and 2 refer to validation of models calibrated using the homosoils only,

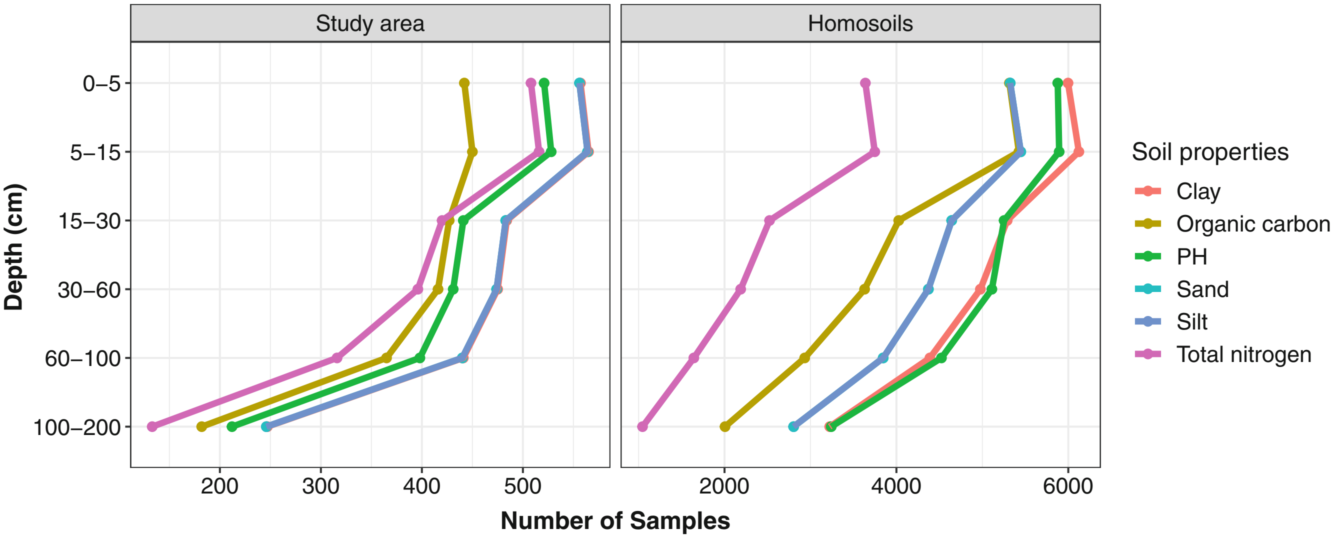

FIGURE 3 Number of soil samples collected from WoSIS within the study area and the homosoils for each soil properties and for six depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, 100–200 cm). Lines are added for visualisation purposes

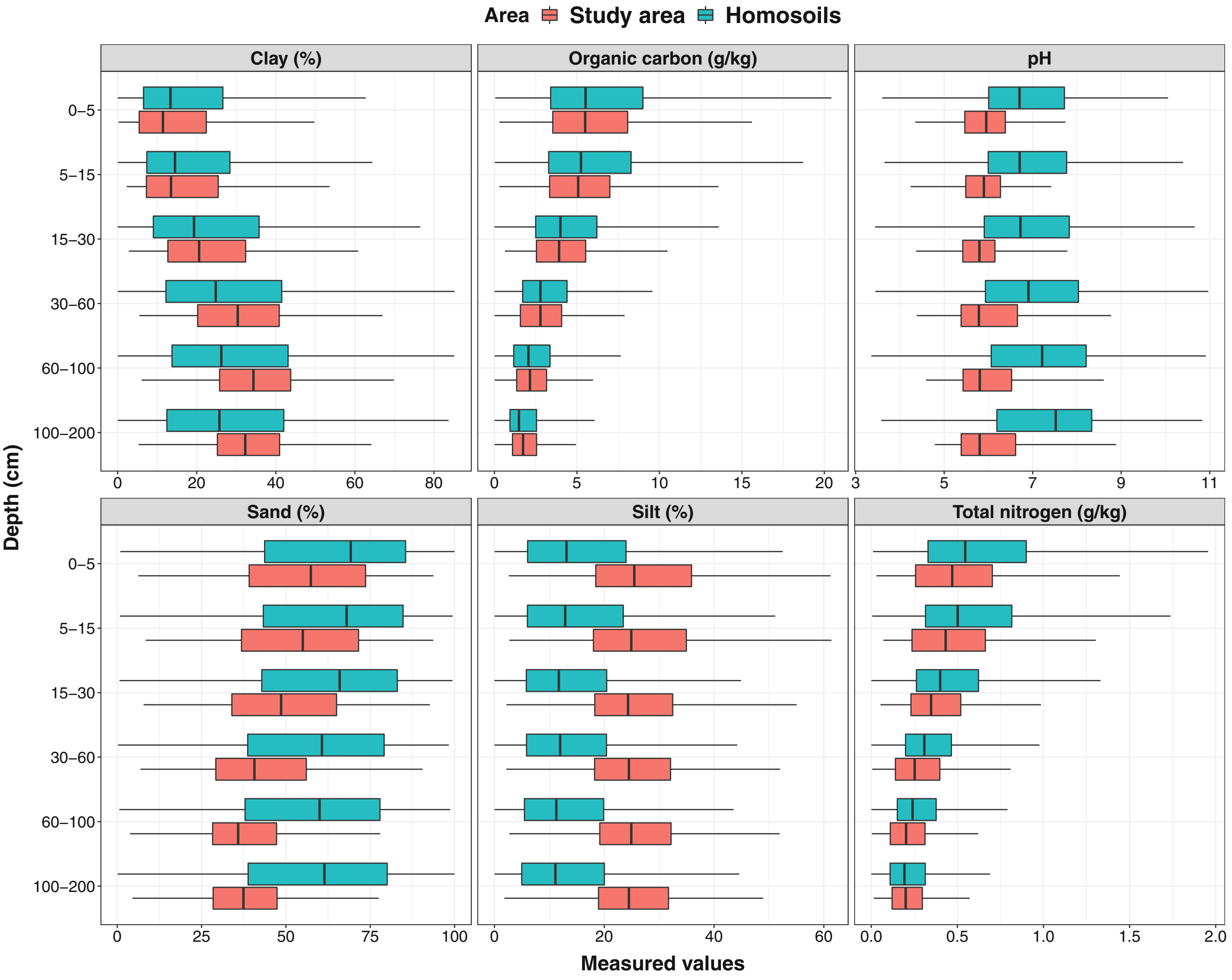

FIGURE 4 Boxplot of clay (%), OC (g/kg), pH, sand (%), silt (%) and total N (g/kg) collected within the study area and its homosoils. The summary statistics are available in the supplementary material

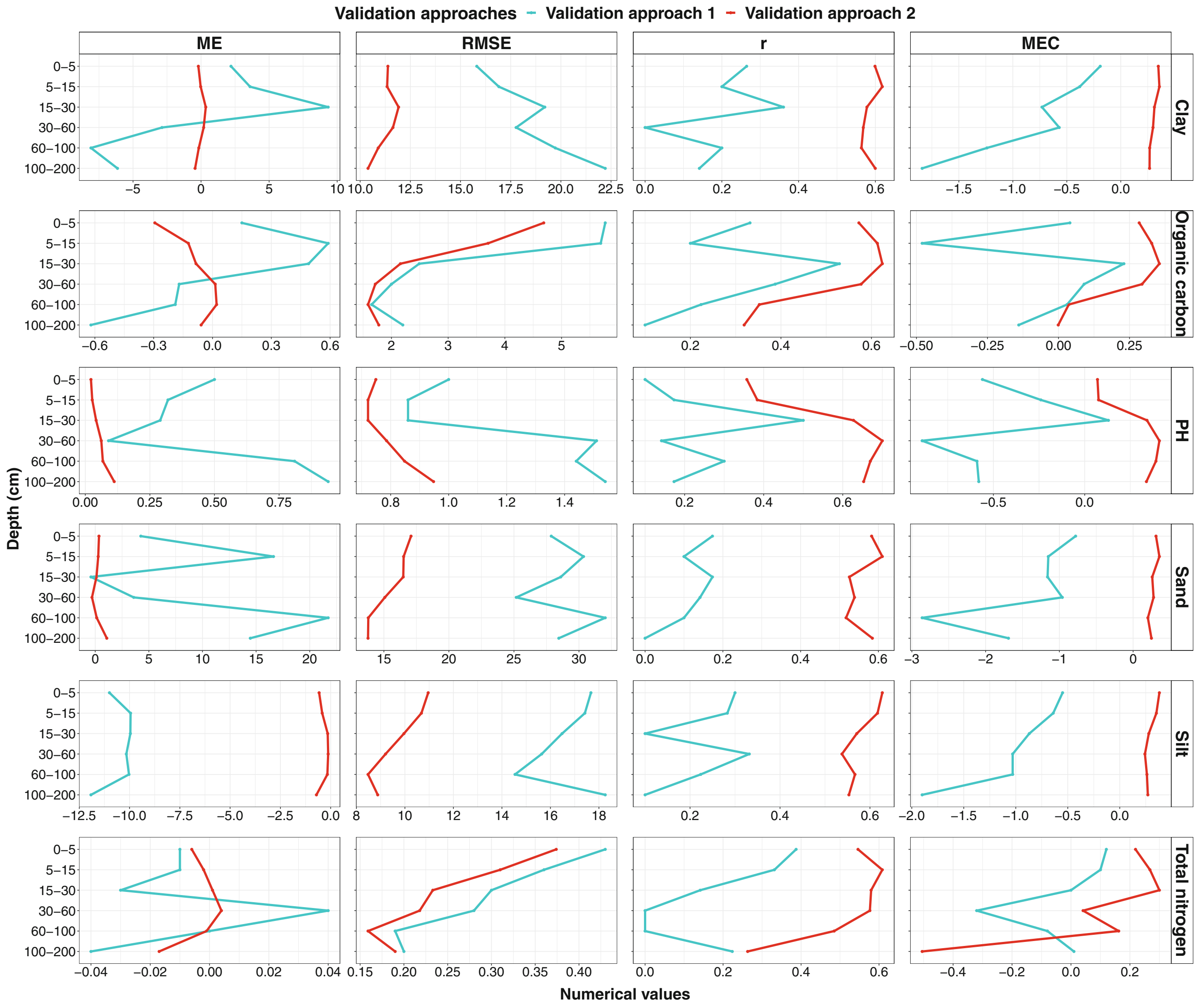

FIGURE 5 Validation statistics of the two validation approaches. The first and second validation approaches refer to the validation of the first and second calibration strategies, respectively. Note that lines are shown to ease visualisation. The red and blue lines represent the validation statistics for the models from calibration strategy 1 and 2, respectively

and homosoils plus data in Mali, respectively, for each soil property and soil depth interval. Figure 5 shows a significant difference between the two model calibration strategies. The models validated using the second validation approach overall performed better than those validated using the first validation approach, as indicated by the ME and RMSE which were always closer to zero than those of the first calibration strategy. The ME and RMSE from validation approach 2 were on average 100% and 32% smaller than those from validation approach 1. The Pearson's *r* correlation values of the second validation approach were on average 190% higher than those of the first validation approach, indicating a better linear relationship between the measured and predicted values by the model. Finally, the MEC of the first validation strategy was nearly always negative for all soil properties and all horizon depth intervals. This suggests that the models from the first calibration strategy performed poorly, and that using the mean of the measured values as an estimate would be a better predictor than the predictions from the models from the first calibration strategy. Models from the second calibration strategy, conversely, had positive MEC values (on average 160% larger than that from the first calibration strategy). Thus, Figure 5 shows that the performance of an extrapolated model calibrated on homosoils values of soil properties dramatically changed whether or not it included training dataset from the area of study within Mali.

While maps generated by the second calibration strategy were on average more accurate, Figure 5 further

TABLE 2 Validation statistics of the maps using the third validation approach

	ME	$q_{0.05}$	$q_{0.95}$	RMSE	$q_{0.05}$	$q_{0.95}$	MEC	$q_{0.05}$	$q_{0.95}$
Clay	−0.16	−4.74	4.12	11.9	10.25	14.06	0.17	−0.23	0.47
Silt	−2.27	−8.39	3.24	13.89	12.32	16.24	0.26	−0.09	0.52
Sand	2.32	−4.72	10.05	19.28	17.25	22.08	0.3	0.03	0.5
OC	−1.4	−3.36	0.47	3.5	2.75	4.98	0.24	−0.56	0.64
pH	−0.14	−0.35	0.07	0.69	0.64	0.77	0.15	−0.1	0.34
Total N	−0.13	−0.24	−0.01	0.27	0.2	0.37	0.53	0.12	0.76

Note: $q_{0.05}$ and $q_{0.95}$ are the 0.05 and 0.95 quantiles of the validation statistics from 500 realisations of the validation residuals.

shows that the accuracy of the model predictions validated with this second approach varied greatly between soil properties and depth intervals. The accuracy of models for clay, silt and sand was rather constant across soil depth intervals while it was not the case for that of pH, OC and total N. The ME values of the second validation approach show that predictions were unbiased. However, the model predictions for silt and pH always overestimated and underestimated the measured values, respectively. The model accuracy of pH increased with depth, while that of OC and total N showed an opposite trend, as indicated in Figure 5 by the increase of the average error indices (ME and RMSE) and the decrease in *r* and MEC values for pH, and by the validation statistics of OC and total N showing an opposite trend with depth. For instance, for OC, the ME and RMSE values at soil depth 0–5 cm were 0.47 and −0.3 g/kg respectively, while they were −0.06 and 1.8 g/kg at soil depth 100–200 cm. This trend can be attributed to the variance of the measured values which was increasing with depth for pH, whereas decreasing with depth for OC and total N (Figure 4). Overall, for the second validation strategy, clay had the best prediction accuracy as indicated by the relatively high and stable values across depth intervals of the MEC (the average value is 0.31 across depth intervals) and *r* (the average value is 0.59 across depth intervals). Conversely, total N had the lowest prediction accuracy with an average MEC and r values of 0.08 and 0.51, respectively.

Table 2 shows the validation statistics obtained by the third validation approach for the 0–15 cm soil depth interval along with their 90% interval. The lower and upper limits of the 90% intervals were represented by the 0.05 and 0.95 quantiles obtained from 500 realizations of the residuals and denoted $q_{0.05}$ and $q_{0.95}$. Recall that this validation strategy refers to the validation of maps obtained by the second calibration strategy. The negative ME values show that the maps are overestimating the soil properties, with the exception of sand, whose ME value was positive. The maps explained at least 15% of the variation of the measured values, for all soil properties. Positive values of the MEC are large for total N (MEC is 0.53) and relatively low for pH (MEC is 0.15). The 90% interval of the validation statistics, however, showed large variation in the range of values. The magnitude of the MEC variation, for example, was largest for OC (1.40) and lowest for pH (0.44).

3.4 | Maps and comparison with existing products

Figure 6 shows the maps of clay content predicted with models from the second calibration strategy for all soil depths. The maps show considerable spatial variation in clay content over the area. Large clay content (i.e., clay >40%) was consistently found in the East and West of the center of the study area across all soil depth intervals, while the northern part of the area consistently had the lowest clay content (clay ¡ 10%). Moreover, clay content in the study area in Mali increased with depth; the 0–5 cm soil depth interval contains 18% clay on average, while the 60–100 and 100–200 cm depth intervals contain 31% and 28%, respectively.

Figure 7 shows maps of all soil properties for the topsoil (0–5 cm). The maps show substantial magnitude in spatial variation but a similar pattern for all soil properties. There is a south–north decreasing gradient except sand and pH which increase towards the North. The southern part of the study area is relatively more acidic (pH <5.5). On average, the topsoil contained 57% of sand, 25% of silt, a pH of 6.2, an OC content of 7.6 g/kg and a total N of 0.64 g/kg. We refer to the Supplementary material for the maps of all soil properties at all depth intervals.

The maps from three different products: homosoils maps made by the second calibration strategy, and two existing products: SoilGrids and iSDAsoil, are shown in Figure 8 for the depth interval 0–15 cm. The maps had differences in spatial pattern, both in terms of magnitude and spatial variation. Spatial pattern of clay maps varied greatly between products, but was similar in magnitude:

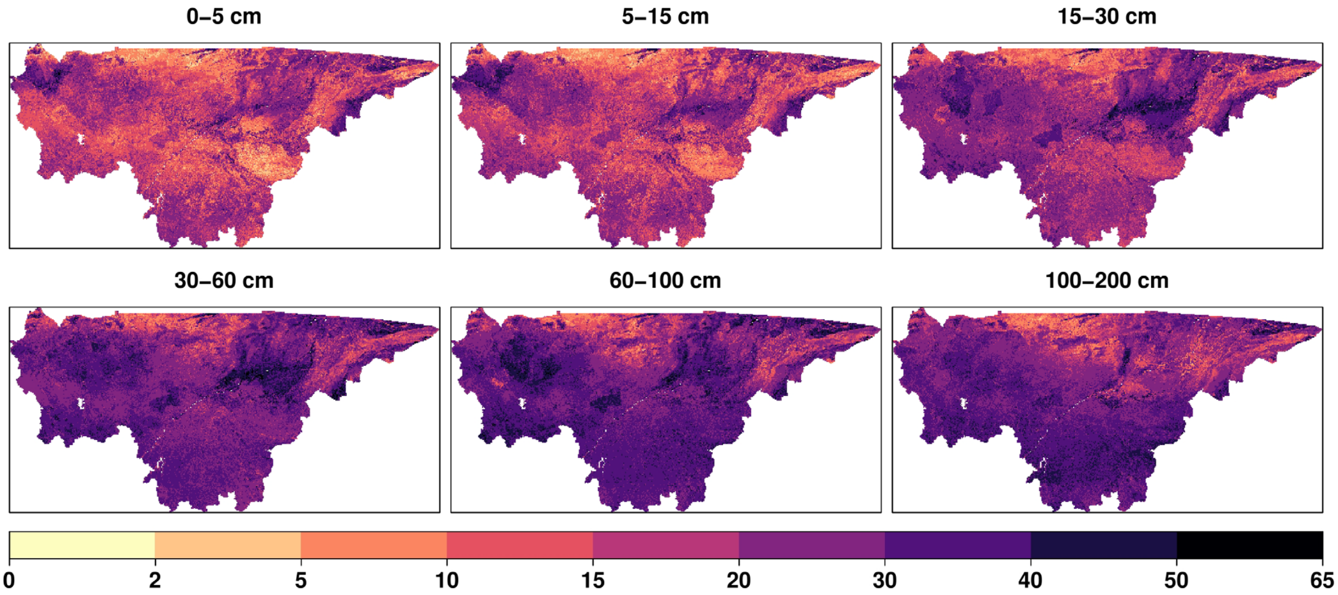

FIGURE 6 Maps of clay (%) for the six soil depth intervals (0–5, 5–15, 15–30, 30–60, 60–100, 100–200 cm)

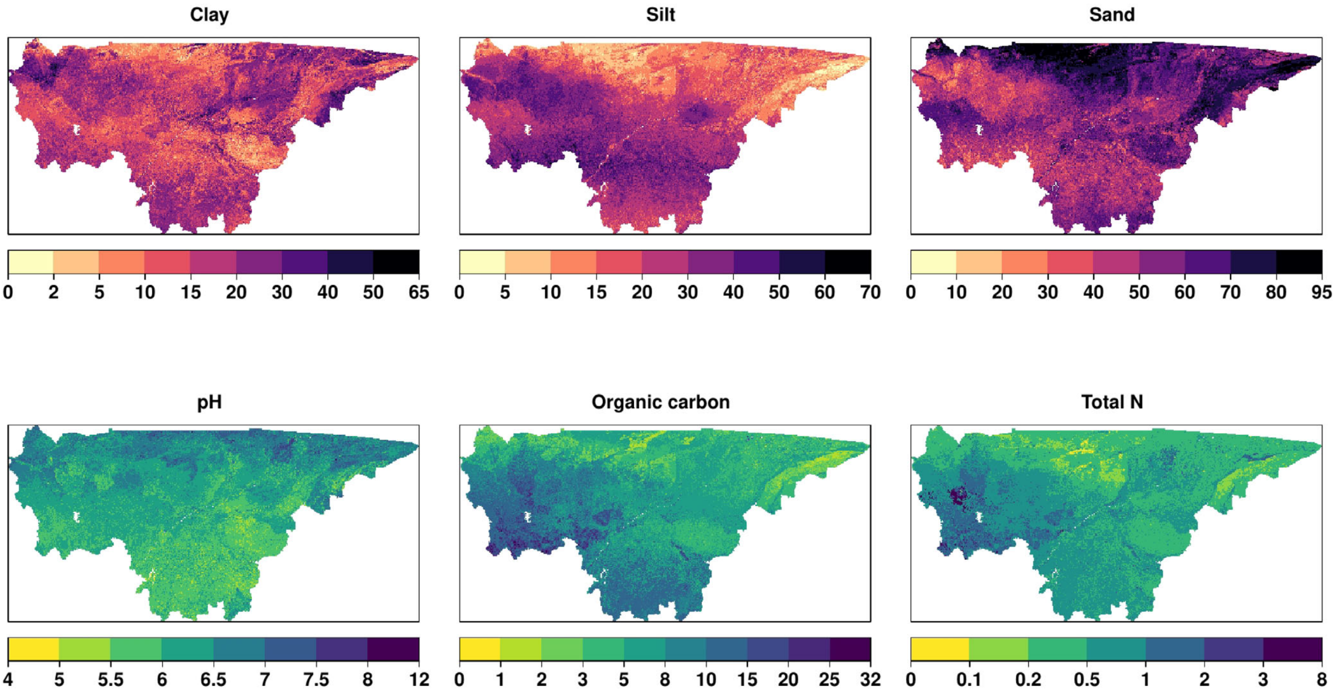

FIGURE 7 Maps of clay, silt and sand (%), OC and total N in g/kg for the topsoil (0–5 cm)

minimum and maximum values of clay content are similar. Maps of silt, sand, OC, pH and Total N have both similar spatial pattern and magnitude of values, but maps from iSDAsoil were consistently smoother than those from homosoils and SoilGrids.

The validation statistics of the homosoils, SoilGrids and iSDAsoils maps using the third validation approach are shown in Figure 9 for all soil properties at a soil depth interval of 0–15 cm, along with their 90% interval. Figure 9 showed significant differences between the accuracy of the different maps. Maps from homosoils had a higher MEC for silt, sand, pH and total N. The MEC of sand, for example, is 0.3 for homosoils, 0.22 for SoilGrids, while it is 0.05 for iSDAsoil. The SoilGrids map explained the largest amount of variation for clay (MEC is 0.22), while the largest amount of variation explained for OC is by the iSDAsoil map (MEC is 0.33). These MEC values were supported by

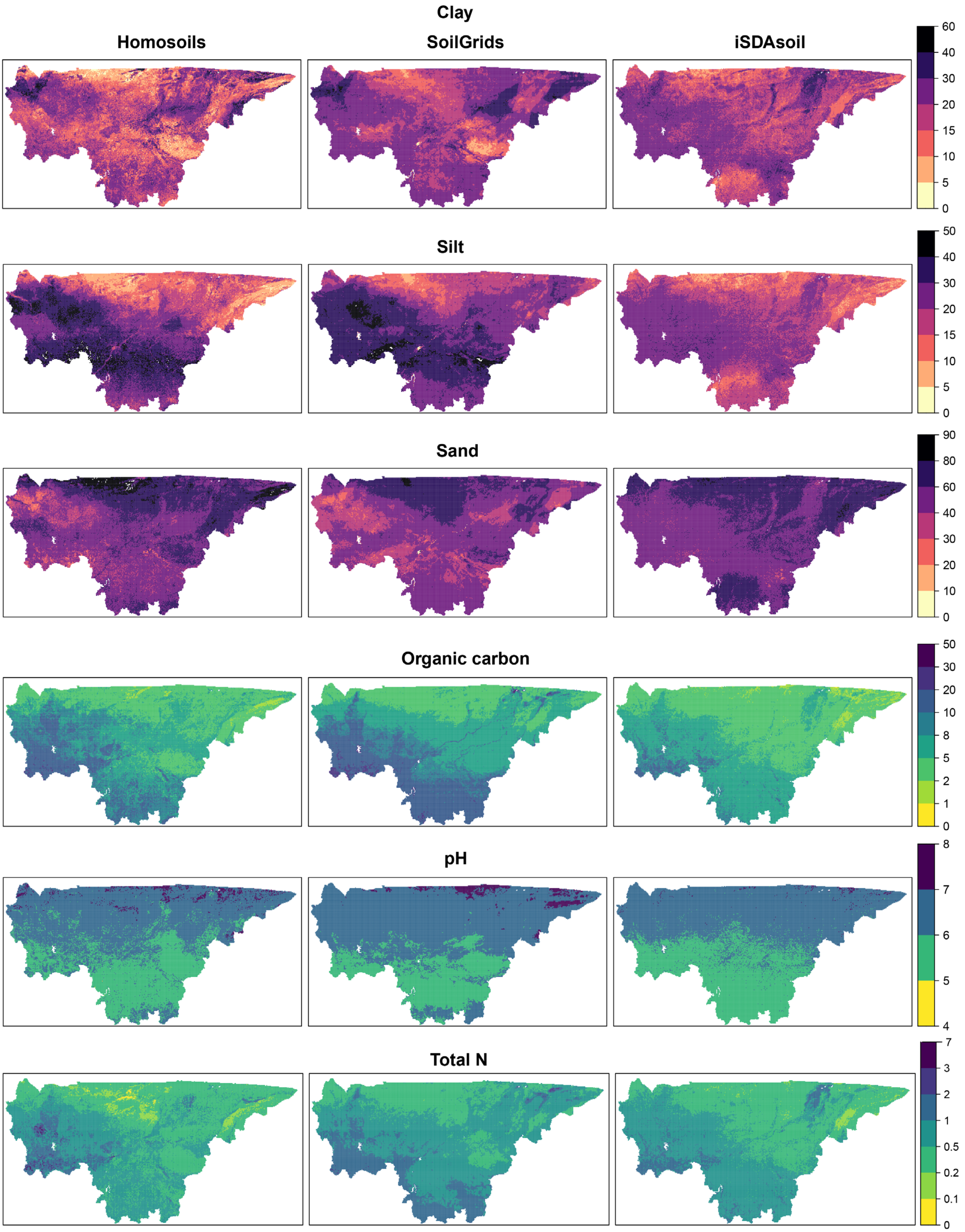

FIGURE 8 Maps of from homosoils, SoilGrids and ISDAsoils (1 km × 1 km) for the soil depth of 0–15 cm. Clay, silt and sand are in percent, OC and Total N are in g/kg

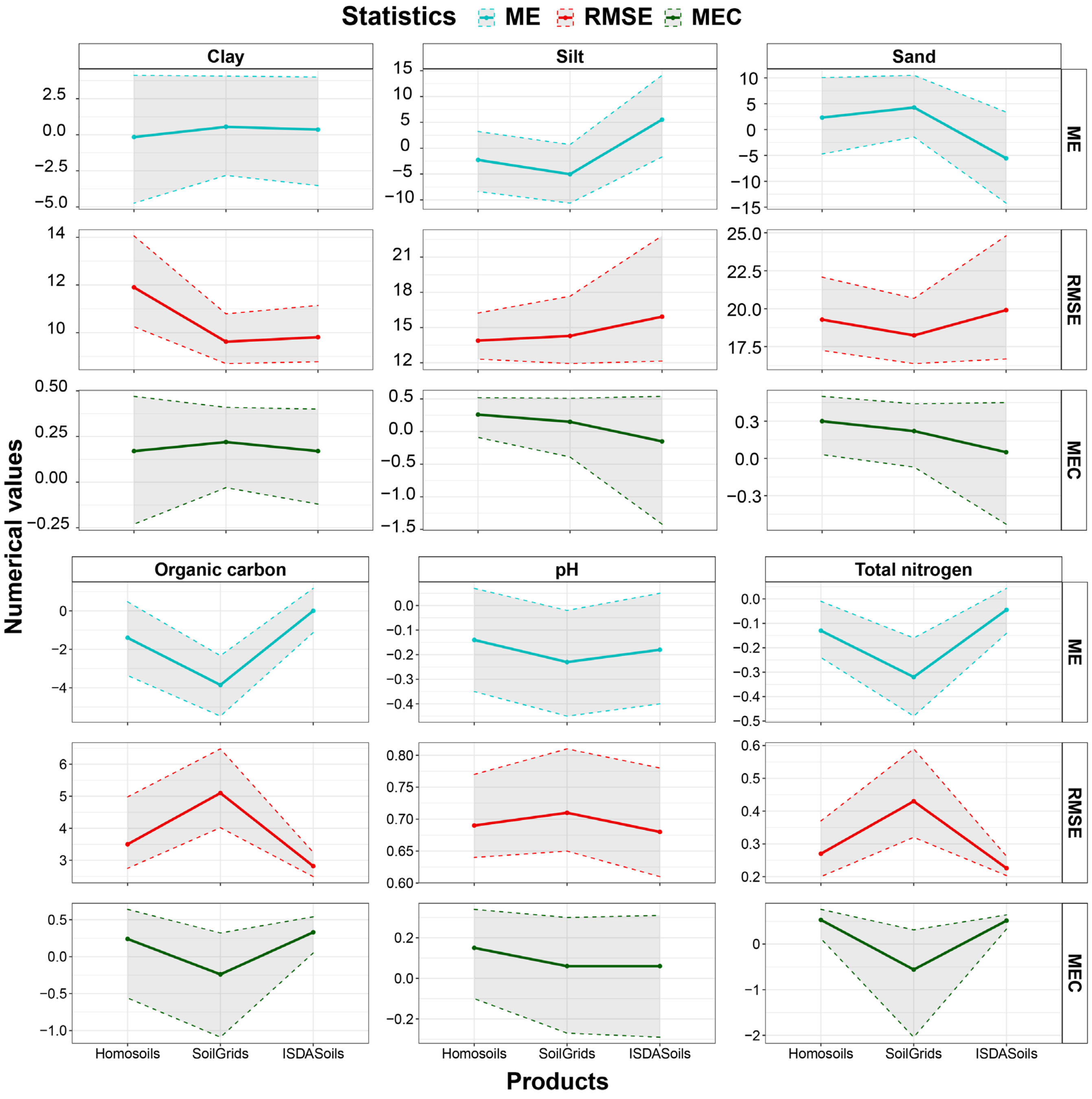

FIGURE 9 Validation statistics of the homosoils, ISDAsoils and SoilGrids maps using the third validation approach and for the depth interval 0–15 cm. The shaded area represents the 90% interval

low ME and RMSE values which indicated relatively small bias and error between the predictions and the observed values except for total N. Despite the map from homosoils having a slightly higher MEC value than that of iSDAsoil for total N, predictions from iSDAsoil have the smallest bias. The 90% interval of the 500 realizations of the validation statistics showed variable range across the maps and soil properties. The magnitude of the range varied greatly between the maps for MEC and RMSE, while it was relatively constant for ME as shown by the greyed area of MEC, RMSE and ME in Figure 9.

4 | DISCUSSION

4.1 | Homosoils and soil data

We found that many areas in the world have soil forming factors similar to those from our study area in Mali, which suggest that there could also be similar soils. Within these areas, large regions in India and the Eastern part of the Sahelian band had relatively low sampling density, whereas areas within Australia, Brazil, Burkina Faso, Bostwana, Mexico, Niger and Nigeria had higher sampling density. Mexico, for example, has 0.64% of its area covered by homosoils (i.e., 125,213 km^2), from which we sourced 365 OC samples from WoSIS for the 5–15 cm depth interval. These samples have a relatively good spatial coverage of the area and represented a density of 3 samples per 1000 km^2. This sampling density is higher than usual density of many large-scale digital soil mapping studies (see also the studies reported in Wadoux et al., 2020). A similar density was found in Burkina Faso, where homosoils cover 78% of the country (i.e., 213,823 km^2), from which 648 samples were collected with a sampling density equivalent to 2 samples per 1000 km^2. In India, conversely, 48% of the country was covered by homosoils (i.e., 1,504,226 km^2), but with only 0.3 samples per 1000 km^2. One reason for the dense soil sampling density in Mexico and Burkina Faso, among others, was the presence of well-established soil survey systems through which soil data are consistently and regularly collected (Van Wesemael et al., 2011). Another reason may be the development of specific soil-related projects, in the 1960s, which allowed soil data collection for many decades (Van Ranst et al., 2010). For instance, the WoSIS soil data gathered for Burkina Faso were collected between 1966 and 2000. The low sampling density, as found in India, reflects the difficulty in accessing country data (Reddy et al., 2021). Overall, it was possible to find some areas in the world that share similar soil-forming factors as our area of study, which suggest they might also have similar soils, and these areas also have a relatively high sampling density.

The soil data collected within the homosoils and the study area showed a similar vertical pattern with different variability (Figure 4) for most soil properties. The average pH content, for example, was generally constant across soil depth intervals, but higher in homosoils compared to our study area. This is expected because of the dominating climate regime (semi-arid to arid) covering the homosoils. In semi-arid to arid environments, the presence of evaporites or carbonate rocks, or the accumulation of salt due to the evaporative behaviour of the soil may be the main source of high pH in the soil (Weil and Brady, 2018; Lal and Stewart, 2019). Silt content was also constant across depth intervals, but systematically higher in our study area than in the homosoils. This is because Mali is exposed to local environmental factors such as airborne dust storm which originates from the Sahara-desert from which aeolian sediments (particles of 10–50 μm) are transported to Mali, thus contributing to the accumulation of silt (Nickling and Gillies, 1993; Schütz, 1980). On the other hand, for both the homosoils and the study area, clay content increased with depth as a result of vertical clay movement (eluviation/illuviation), whereas sand content decreased with depth. This opposite trend suggested the presence of contrast soil texture within the areas. Moreover, homosoils had larger sand content compared to our study area, because they span over a wide range of semi-arid to arid environments whose soils experience large sand particles accumulation (Department, 2014). The homosoils defined in our study did not take into account aeolian processes. For other properties such as OC and total N content, there was more variability in homosoils than in our study area, but the trend with depth was similar (i.e., total N and OC content decreased with depth). The large variability in homosoils can be ascribed to differences in land management practices which greatly influenced the dynamics of both OC and total N in the soils.

4.2 | Model extrapolation

Our study showed that extrapolating soil mapping models between homosoil areas was a challenging task. We tested two extrapolation strategies and found that nearly all the models built using only homosoils data had quasi-null accuracy when extrapolated to the study area (except for 19% of these models which included one to three soil depth intervals of the dynamic properties – OC, total N and pH). These results confirm the results reported in Angelini et al. (2020), in which it was found that the predictive performance of the models for mapping OC, cation exchange capacity (CEC), and clay was quasi-null when extrapolated between two geographically remote areas considered as homosoils. On the other hand, a more recent study found contradictory results. Du et al. (2021) extrapolated topographical random forest models from one area to a geographically close area (15 km away), and showed that the models could explain most variance (i.e., 73%) of the measured soil OC without including local data (i.e the data within the extrapolated area). Our results and the studies of Angelini et al. (2020) and Du et al. (2021) suggest that geographical proximity plays an important role when transferring soil mapping models. This might be a reason for the low predictive power of the models. This is reflected in Figure 4 by the high variation in the soil properties between the study area and the homosoils which certainly led to differences in the soil-covariate relationship between the two regions, and which then affected the predictions when the models were extrapolated. Another possible reason for the low predictive power is probably the influence of other soil forming factors which we disregarded in this study. Our results suggest that the global soil-forming factors defined by the homosoils may miss important soil processes that affect local soil variation. Such processes include anthropogenic activities and/or site-specific environmental factors (e.g., aeolian sediments deposits mentioned earlier), among others, which greatly vary per region and can influence the soil dynamics differently. A third reason is that homosoils are defined for a single spatial location which highly depends on the conditions of surrounding spatial locations following the *soilscape* concept presented in Lagacherie et al. (2001). Further work could integrate this concept at a local scale by including the surrounding conditions of the spatial locations when defining homosoils, to *homosoilscape*.

We found that including soil data from within the study area increased prediction accuracy. In particular, adding local soil samples dramatically improved the MEC and RMSE by 160% and 32%, respectively. These results corroborate previous studies (e.g., from Lemercier et al., 2012; Du et al., 2021). Lemercier et al. (2012) extrapolated models to a region that included the area within which the models were built, and found that the models could explain up to 49% of the variation of parent material, and 52% of the variation of soil drainage. Similarly, Du et al. (2021) increased the amount of variance explained for OC by 8% and decreased the prediction error by 25% through the addition of local soil sample information. Our study suggests, similarly, that extrapolation between homosoil areas is possible but that local samples from within the extrapolated area are necessary.

We recognise the similarity between our approach for local mapping using homosoils and approaches based on machine learning models calibrated with global data. It is likely that machine learning models such as used in SoilGrids and iSDAsoil assign higher weights when predicting to observations that come from similar environments (i.e., observations that are close to each other in the covariate space), so effectively applying the homosoils concept developed in this study for DSM. Comparing the two approaches could be the purpose of further research. In our study, when comparing the maps made with models fitted on data within homosoil areas and SoilGrids and iSDAsoil maps, we found that, generally, our approach performed slightly better. This is, however, made at the expense of adding an extra step before model calibration to find the homosoils. Moreover, finding homosoils leads to more soil science discovery (such as knowing where similar soils might be and what are their properties) rather than the global DSM approach.

4.3 | Interpretation of the maps

Here, we summarize the most striking map features in our study area. The pattern of soil texture is linked to the dominating soil types. Clay accumulated in deeper soil horizons due to the dominating soils types alfisols and ultisols (USAID, 1983). Moreover, clay content was consistently higher (>35%) in the east of the center of the study area. This region corresponds to the inland delta of the Niger river in Mali (Thom and Wells, 1987) and is dominated by hydromorphic soils which are generally silty with clayey alluvial deposits (Diarra et al., 2004; Ajayi et al., 2012). A large portion of the study area has high (i.e., >25%) silt content which mainly originated from aeolian sediments deposits. The pattern of sand content in the north east of the study area was characterized by the presence of aridisols (Nettleton and Peterson, 1983). Climate regime controls the spatial variation of pH with Alkaline soils (pH >7) being common in arid regions due to the presence of carbonate rocks or the accumulation of soluble salt (Weil and Brady, 2018). OC and total N followed the distribution of climate and landcover, respectively. High and low values of the dynamic properties (OC and Total N) were found in humid (in the south) and arid (in the north) regions of the study area, respectively, which reflect higher and lower net-primary production. The distribution of OC in our study corroborates findings from Akpa et al. (2016) in Nigeria. The soil maps presented in our study area share similar spatial pattern to the coarse-resolution global maps from SoilGrids (Poggio et al., 2021) and the African maps from iSDAsoil (Hengl et al., 2021) products for Mali (shown in Figure 8). These soil maps were calibrated mainly using the WoSIS dataset as used in this study, however because they were generated with limited soil data coverage, they may only be useful at a regional scale, inhibiting their application at a local scale where availability of soil information is more critical. In such circumstances, only new soil data collection might reverse this issue.

4.4 | Validation with clustered data

We acknowledge that the comparison of the maps (homosoils, SoilGrids, iSDAsoils) was made using an independent and spatially clustered sample, which may result in biased estimates of the map accuracy. Clustered data in the geographic space often lead to a clustering in the covariate space (Elliott and Valliant, 2017) and failure in assessing map accuracy in areas with zero sampling density, thus leading to over-optimistic map accuracy estimates. However, unbiased estimates of digital soil maps accuracy can only be obtained through the collection of a post-mapping independent soil dataset with probability sampling and design-based statistical inference methods (Brus et al., 2011). Due to the pandemic, we could not collect an additional probability sampling for map validation and therefore used a model-based validation approach using a geostatistical model, where the map residuals were kriged with ordinary kriging and the sampling distribution of the map accuracy indices was computed using 500 simulations of the residuals. We considered this approach suitable to deal with the clustered data. Moreover, in case of clustered data, model-based validation approaches based on weighted cross-validation had smaller bias than conventional cross-validation (de Bruin et al., 2022). In our case, the model-based approach showed a high uncertainty of the map accuracy indices as shown by the 90% interval. Validating digital soil maps using clustered data is not straightforward and needs further research.

4.5 | Limitations

One major limitation of this study is the inaccuracy of the covariate that describes the state factor lithology. The global lithology dataset (Hartmann and Moosdorf, 2012) is described with multiple orders: first, second and third lithology orders. The highest order, wherever present, provides further granular differentiation within a given first-order lithological unit. This surely has had an effect on the soil classes and properties due to the different mineral content and morphology (particularly texture) that it represents. In our study, we only used lithological information of the first order because higher orders were not publicly available. The effect of lithology is reflected in the statistical indices (RMSE, *r* and MEC) presented in Figure 5, where those of the texture maps (i.e., stable properties clay, sand and silt) significantly improve for validation approach 2, whereas that of the dynamic soil properties (OC, pH and total N) shows no significant difference. This may reflect differences in lithology between our study area and its homosoils, and part of these differences could be ascribed not only to the unavailability of the highest order lithological units but also, probably, to the spatial resolution at which the homosoils were found. Recall that the lithology variable was upscaled from 250 m to 1 km which may have decreased the granularity of the information and thus contributed to these differences. Therefore, accessing lithological units at a finer resolution would require finding homosoils at a much higher resolution (e.g., 250 m and below), besides accessing higher order lithological units below 100 m may be practically challenging when working at a global scale.

Another limitation of this study was the omission of the anthropogenic soil forming factor in both identifying the homosoils and in generating the maps, because these were not available at global scale. Several studies have stressed that human activities greatly influence soil dynamics (Amundson and Jenny, 1991; Hooke, 2000; Wilkinson, 2005; Richter Jr et al., 2007) and may be the main soil forming factor (Kuzyakov and Zamanian, 2019) because of their critical influence on soil-forming processes. This implies that data on anthropogenic activities are critical for digital soil mapping exercise. However, both actual and historical management practices are needed, which is practically impossible to obtain. Besides, data on actual anthropogenic activities are barely available at local scale, and much less at the continental and global scale. Generating such spatially exhaustive information at large scale and on a time-scale would certainly make a valuable extension to future soil mapping studies.

5 | CONCLUSION

We tested the geographic extrapolation of a model to map soil properties. The model was applied to our area of interest in Mali after being calibrated with data from its homosoil area. We tested different calibration and validation strategies, including or not local data for calibration. From the results and discussion, we draw the following conclusions:

- Within areas considered as homosoils, we can leverage on areas with relatively high sampling density and build a soil mapping model which can be applied on areas with limited soil data.
- The soil data collected within the homosoils showed a similar vertical pattern but large variability compared to our study area.
- Homosoils help transfer soil information from one area to another by means of DSM model extrapolation methods. The model built within the homosoils performed poorly when extrapolated to our study area, however this accuracy increased dramatically when local soil samples were also used to calibrate the model.
- The maps generated from homosoils were more accurate than those generated at the continental and global scale for our study area in Mali, for three (silt, sand, and pH) out of six soil properties. However, the spatial pattern was similar.
- Homosoils represent an opportunity to generate digital soil maps for areas that are scarce in soil data, because it is cheap and usually fast to implement compared to new soil surveys.
- Validating the soil maps in this study was challenging due to the lack of reliable soil data and poor spatial

coverage of the existing observations. The collection of new soil data (through the establishment of sustainable soil survey systems) remains necessary and is required to tackle the issue with soil data gap. That is the one and only way accurate soil maps can be generated and confidently used to mitigate soil-environmental issues.

AUTHOR CONTRIBUTIONS

Andree M. Nenkam: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); project administration (lead); validation (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Alexandre M. J.-C. Wadoux:** Conceptualization (equal); methodology (supporting); project administration (supporting); supervision (lead); validation (supporting); visualization (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Budiman Minasny:** Conceptualization (equal); investigation (supporting); methodology (supporting); project administration (supporting); supervision (lead); writing – original draft (supporting); writing – review and editing (supporting). **Alex B. McBratney:** Conceptualization (equal); funding acquisition (lead); investigation (supporting); methodology (supporting); project administration (supporting); supervision (lead); writing – review and editing (supporting). **Pierre C. S. Traore:** Funding acquisition (supporting); supervision (supporting); writing – review and editing (supporting). **Gatien N. Falconier:** Resources (supporting); writing – review and editing (supporting). **Anthony M. Whitbread:** Funding acquisition (supporting); resources (supporting); supervision (supporting); writing – review and editing (supporting).

ACKNOWLEDGEMENTS

This research and salary support of the ICRISAT authors were in part funded by the CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS) carried out with support from the CGIAR Trust Fund and through bilateral funding agreements. For details, visit https://ccafs.cgiar.org/donors. The World Bank-funded AICCRA project (Accelerating Impacts of CGIAR Climate Research for Africa) Project ID 173398 is acknowledged for funding the involvement of Anthony Whitbread in this study. Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

Public data are available on request from the authors.

ORCID

Andree M. Nenkam https://orcid.org/0000-0001-8921-4969
Alexandre M. J.-C. Wadoux https://orcid.org/0000-0001-7325-9716
Budiman Minasny https://orcid.org/0000-0002-1182-2371

REFERENCES

A. Department. (2014). Keys to soil taxonomy, Government Printing Office.

Abbaszadeh Afshar, F., Ayoubi, S., & Jafari, A. (2018). The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma*, *315*, 36–48.

Ajayi, O. C., Diakit'e, N., Konate, A. B., & Catacutan, D. (2012). Rappid assessment of the inner Niger Delta of Mali, in: ICRAF Working Paper No. 144, World Agroforestry Centre Nairobi.

Akpa, S. I. C., Odeh, I. O. A., Bishop, T. F. A., Hartemink, A. E., & Amapu, I. Y. (2016). Total soil organic carbon and carbon sequestration potential in Nigeria. *Geoderma*, *271*, 202–215.

Amundson, R., & Jenny, H. (1991). The place of humans in the state factor theory of ecosystems and their soils. *Soil Science*, *151*, 99–109.

Andrieu, N., Sogoba, B., Zougmore, R., Howland, F., Samake, O., Bonilla-Findji, O., Lizarazo, M., Nowak, A., Dembele, C., & Corner-Dolloff, C. (2017). Prioritizing investments for climate-smart agriculture: Lessons learned from Mali. *Agricultural Systems*, *154*, 13–24.

Angelini, M. E., Kempen, B., Heuvelink, G. B. M., Temme, A. J. A. M., & Ransom, M. D. (2020). Extrapolation of a structural equation model for digital soil mapping. *Geoderma*, *367*, 114226.

Batjes, N. H., Ribeiro, E., & van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*, 299–320.

Bayala, J., Sanou, J., Bazié, H. R., Coe, R., Kalinganire, A., & Sinclair, F. L. (2020). Regenerated trees in farmers' fields increase soil carbon across the Sahel. *Agroforestry Systems*, *94*, 401–415.

Benjaminsen, T. A., Aune, J. B., & Sidibé, D. (2010). A critical political ecology of cotton and soil fertility in Mali. *Geoforum*, *41*, 647–656.

Birhanu, B. Z., Traoré, K., Sanogo, K., Tabo, R., Fischer, G., & Whitbread, A. M. (2020). Contour bunding technology-evidence and experience in the semiarid region of southern Mali. *Renewable Agriculture and Food Systems*, *35*, 1–9.

Bishop, T. F. A., McBratney, A. B., & Laslett, G. M. (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, *91*, 27–45.

Brus, D. J., Kempen, B., & Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science*, *62*, 394–407.

Bui, E. N., & Moran, C. J. (2003). A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray–Darling basin of Australia. *Geoderma*, *111*, 21–44.

Cambule, A. H., Rossiter, D. G., & Stoorvogel, J. J. (2013). A methodology for digital soil mapping in poorly-accessible areas. *Geoderma*, *192*, 341–353.

de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst, T., & Wadoux, A. M. J.-C. (2022). Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics*, *69*, 101665.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18.

Degerickx, J., Almeida, J., Moonen, P. C. J., Vervoort, L., Muys, B., & Achten, W. M. J. (2016). Impact of land-use change to Jatropha bioenergy plantations on biomass and soil carbon stocks: A field study in Mali. *GCB Bioenergy*, *8*, 443–455.

Diarra, S., Kuper, M., & Mahé, G. (2004). Mali: Flood management-Niger River inland delta, Integrated Flood Management, Case Study. WMO/GWP Associated Programme on Flood Management, Edited by Technical Support Unit.

Dokuchaev, V. V. (1883). The russian chernozem report to the free economic society, Imperial University of St. Petersburg: St. Petersburg [in Russian].

Doumbia, M., Jarju, A., Sène, M., Traoré, K., Yost, R., Kablan, R., Brannan, K., Berthe, A., Yamoah, C., Querido, A., Traoré, P. C. S., & Ballo, A. (2009). Sequestration of organic carbon in west African soils by aménagement en courbes de niveau. *Agronomy for Sustainable Development*, *29*, 267–275.

Du, L., McCarty, G. W., Li, X., Rabenhorst, M. C., Wang, Q., Lee, S., Hinson, A. L., & Zou, Z. (2021). Spatial extrapolation of topographic models for mapping soil organic carbon using local samples. *Geoderma*, *404*, 115290.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*, 249–264.

Falconnier, G. N., Descheemaeker, K., Van Mourik, T. A., & Giller, K. E. (2016). Unravelling the causes of variability in crop yields and treatment responses for better tailoring of options for sustainable intensification in southern Mali. *Field Crops Research*, *187*, 113–126.

FAO. (2000). Quatorzième Réunion du Sous-Comité Ouest et Centre Africain de Corrélation des Sols pour la Mise en Valeur des Terres, Technical Report, Food and Agriculture Organization (FAO), Roma, Italy.

Fick, S. E., & Hijmans, R. J. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302–4315.

Giannini, A., Krishnamurthy, P. K., Cousin, R., Labidi, N., & Choularton, R. J. (2017). Climate risk and food security in Mali: A historical perspective on adaptation. *Earth's Future*, *5*, 144–157.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, *202*, 18–27.

Grinand, C., Arrouays, D., Laroche, B., & Martin, M. P. (2008). Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma*, *143*, 180–190.

Hartemink, A. E. (2002). Soil science in tropical and temperate regions—Some differences and similarities. *Advances in Agronomy*, *77*, 269–292.

Hartmann, J., & Moosdorf, N. (2012). The new global lithological map database (GLiM): A representation of rock properties at the earth surface. *Geochemistry, Geophysics, Geosystems*, *13*, Q12004.

Hengl, T., Leenaars, J. G., Shepherd, K. D., Walsh, M. G., Heuvelink, G., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I., & Kwabena, N. A. (2017b). Soil nutrient maps of sub-saharan africa: Assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, *109*, 77–102.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017a). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*, e0169748.

Hengl, T., Miller, M. A., 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., … Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, *11*, 1–18.

Heuvelink, G. B. M., & Webster, R. (2001). Modelling soil variation: Past, present, and future. *Geoderma*, *100*, 269–301.

Hooke, R. L. B. (2000). On the history of humans as geomorphic agents. *Geology*, *28*, 843–846.

Huet, E. K., Adam, M., Dembele, A. O., & Descheemaeker, K. (2020). Analysis of soil samples from the Koutiala region 2017-2019. project: Pathways to agroecological intensification in crop-livestock farming systems in southern Mali. CCRP McKnight foundation, CRP GLDC, AfricaRISING; unpublished.

Janssen, P. H. M., & Heuberger, P. S. C. (1995). Calibration of process-oriented models. *Ecological Modelling*, *83*, 55–66.

Jenny, H. (1941). *Factors of soil formation: A system of quantitative Pedology*. McGrawHill.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., Roy, D. P., & Morisette, J. T. (2002). An overview of MODIS land data processing and product status. *Remote Sensing of Environment*, *83*, 3–15.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling*. Springer.

Kuhn, M., & Quinlan, R. (2020). Cubist: Rule- and instance-based regression modeling. R package version 0.2.3. Retrieved from 10.09.2021 https://CRAN.R-project.org/package=Cubist

Kuzyakov, Y., & Zamanian, K. (2019). Reviews and syntheses: Agropedogenesis–humankind as the sixth soil-forming factor and attractors of agricultural soil degradation. *Biogeosciences*, *16*, 4783–4803.

Lagacherie, P., Robbez-Masson, J.-M., Nguyen-The, N., & Barthès, J. (2001). Mapping of reference area representativity using a mathematical soilscape distance. *Geoderma*, *101*, 105–118.

Lal, R., & Stewart, B. A. (2019). *Soil degradation and restoration in Africa*. CRC Press.

Lemercier, B., Lacoste, M., Loum, M., & Walter, C. (2012). Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. *Geoderma*, *171-172*, 75–84.

Malone, B. P., Jha, S. K., Minasny, B., & McBratney, A. B. (2016). Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma*, *262*, 243–253.

Mbow, C., Brandt, M., Ouedraogo, I., De Leeuw, J., & Marshall, M. (2015). What four decades of earth observation tell us about land degradation in the Sahel? *Remote Sensing*, *7*, 4048–4067.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52.

Minasny, B., & McBratney, A. B. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma*, *264*, 301–311.

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, *118*, 1–47.

Nenkam, M. A., Wadoux, A. M. J. C., Minasny, B., McBratney, A. B., Traore, P. C. S., & Whitbread, A. M. (2022). Using homosoils to enrich sparse soil data infrastructure, Under Review.

Nettleton, W. D., & Peterson, F. F. (1983). Aridisols. In L. P. Wilding, N. E. Smeck, & G. F. Hall (Eds.), *Developments in soil science* (Vol. 11, pp. 165–215). Elsevier.

Nickling, W. G., & Gillies, J. A. (1993). Dust emission and transport in Mali, West Africa. *Sedimentology*, *40*, 859–868.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil*, *7*, 217–240.

Quinlan, J. R. (1992). Learning with continuous classes. 5th Australian joint conference on artificial intelligence, *92*, pp. 343–348.

Rabus, B., Eineder, M., Roth, A., & Bamler, R. (2003). The shuttle radar topography mission—A new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, *57*, 241–262.

Raina, P., Joshi, D., & Kolarkar, A. (1993). Mapping of soil degradation by using remote sensing on alluvial plain, Rajasthan, India. *Arid Land Research and Management*, *7*, 145–161.

Reddy, N. N., Chakraborty, P., Roy, S., Singh, K., Minasny, B., McBratney, A. B., Biswas, A., & Das, B. S. (2021). Legacy databased national-scale digital mapping of key soil properties in India. *Geoderma*, *381*, 114684.

Rian, S., Xue, Y., MacDonald, G. M., Touré, M. B., Yu, Y., De Sales, F., Levine, P. A., Doumbia, S., & Taylor, C. E. (2009). Analysis of climate and vegetation characteristics along the savanna-desert ecotone in Mali using Modis data. *GIScience & Remote Sensing*, *46*, 424–450.

Richter, D. D., Jr. (2007). Humanity's transformation of Earth's soil: Pedology's new frontier. *Soil Science*, *172*, 957–967.

Schütz, L. (1980). Long range transport of desert dust with special emphasis on the Sahara. *Annals of the New York Academy of Sciences*, *338*, 515–532.

Shatar, T. M., & McBratney, A. B. (1999). Empirical modeling of relationships between sorghum yield and soil properties. *Precision Agriculture*, *1*, 249–276.

Silva, S. H. G., de Menezes, M. D., Owens, P. R., & Curi, N. (2016). Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in southeastern Brazil. *Geoderma*, *267*, 65–77.

Summerauer, L., Baumann, P., Ramirez-Lopez, L., Barthel, M., Bauters, M., Bukombe, B., Reichenbach, M., Boeckx, P., Kearsley, E., Van Oost, K., Vanlauwe, B., Chiragaga, D., Heri-Kazi, A. B., Moonen, P., Sila, A., Shepherd, K., Mujinya, B. B., Ranst, E. V., Baert, G., … Six, J. (2021). The central African soil spectral library: A new soil infrared repository and a geographical prediction analysis. *The Soil*, *7*, 693–715.

Theobald, D. M., Harrison-Atlas, D., Monahan, W. B., & Albano, C. M. (2015). Ecologically-relevant maps of landforms and physiographic diversity for climate adaptation planning. *PLoS One*, *10*, e0143619.

Thom, D. J., & Wells, J. C. (1987). Farming systems in The Niger inland delta, Mali. *Geographical Review*, *77*, 328–342.

Thompson, J. A., Pena-Yewtukhiw, E. M., & Grove, J. H. (2006). Soil landscape modeling across a physiographic region: Topographic patterns and model transportability. *Geoderma*, *133*, 57–70.

USAID. (1983). Projet Inventaire des Ressources Terrestres (PIRT): Rapport Technique, Technical Report, United States Agency for International Development (USAID), Washington, D.C., United States.

Van Ranst, E., Verdoodt, A., & Baert, G. (2010). Soil mapping in Africa at the crossroads: Work to make up for lost ground. *Bulletin des Séances d'Académie Royale des Sciences d'Outre-Mer*, *56*, 147–163.

Van Wesemael, B., Paustian, K., Andrén, O., Cerri, C. E. P., Dodd, M., Etchevers, J., Goidts, E., Grace, P., Kätterer, T., McConkey, B. G., Ogle, S., Pan, G., & Siebner, C. (2011). How can soil monitoring networks be used to improve predictions of organic carbon pool dynamics and CO2 fluxes in agricultural soils? *Plant and Soil*, *338*, 247–259.

Verbree, C. L., Aitkenhead-Peterson, J. A., Loeppert, R. H., Awika, J. M., & Payne, W. A. (2015). Shea (*Vitellaria paradoxa*) tree and soil parent material effects on soil properties and intercropped sorghum grain-Zn in southern Mali, West Africa. *Plant and Soil*, *386*, 21–33.

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359.

Webster, R. (1977). *Quantitative and numerical methods in soil classification and survey*. Oxford University Press.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

Weil, R. R., & Brady, N. C. (2018). *Elements of the nature and properties of soils*. Pearson Education.

Wilkinson, B. H. (2005). Humans as geologic agents: A deep-time perspective. *Geology*, *33*, 161–164.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nenkam, A. M., Wadoux, A.-C., Minasny, B., McBratney, A. B., Traore, P. C. S., Falconier, G. N., & Whitbread, A. M. (2022). Using homosoils for quantitative extrapolation of soil mapping models. *European Journal of Soil Science*, *73*(5), e13285. https://doi.org/10.1111/ejss.13285