Model averaging for mapping topsoil organic carbon in France

Songchao Chen, Vera Leatitia Mulder, Gerard B.M. Heuvelink, Laura Poggio, Manon Caubet, Mercedes Roman Dobarco, Christian Walter, Dominique Arrouays

Title: Model averaging for mapping topsoil organic carbon in France

Authors:

Songchao Chen [a, b]. songchao.chen@inra.fr

Vera Leatitia Mulder [c]. titia.mulder@wur.nl

Gerard B.M. Heuvelink [c, d].gerard.heuvelink@wur.nl

Laura Poggio [d].laura.poggio@isric.org

Manon Caubet [a]. manon.caubet@inra.fr

Mercedes Román Dobarco [a]. mercedes.roman.dobarco@gmail.com

Christian Walter [b]. christian.walter@agrocampus-ouest.fr

Dominique Arrouays [a]. dominique.arrouays@inra.fr

Affiliations:

[a] INRAE, Unité InfoSol, 45075 Orléans, France

[b] SAS, INRAE, Agrocampus Ouest, 35042 Rennes, France

[c] Soil Geography and Landscape Group, Wageningen University, PO Box 47 6700 AA Wageningen, The Netherlands

[d] ISRIC–World Soil Information, PO Box 353 6700 AJ Wageningen, The Netherlands

Corresponding author:

Songchao Chen: songchao.chen@inra.fr

Postal address: INRAE, Unité InfoSol, 2163 Avenue de la Pomme de Pin, CS 40001 Ardon, 45075 Orléans, France

Telephone: +33(0)602142667

Abstract:

The soil organic carbon (SOC) pool is the largest terrestrial carbon (C) pool and is two to three times larger than the C stored in vegetation and the atmosphere. SOC is a crucial component within the C cycle, and an accurate baseline of SOC is required, especially for biogeochemical and earth system modelling. This baseline will allow better monitoring of SOC dynamics due to land use change and climate change. However, current estimates of SOC stock and its spatial distribution have large uncertainties. In this study, we test whether we can improve the accuracy of the three existing SOC maps of France obtained at national (IGCS), continental (LUCAS), and global (SoilGrids) scales using statistical model averaging approaches. Soil data from the French Soil Monitoring Network (RMQS) were used to calibrate and evaluate five model averaging approaches, i.e., Granger-Ramanathan, Bias-corrected Variance Weighted (BC-VW), Bayesian Modelling Averaging, Cubist and Residual-based Cubist. Cross-validation showed that with a calibration size larger than 100 observations, the five model averaging approaches performed better than individual SOC maps. The BC-VW approach performed best and is recommended for model averaging. Our results show that 200 calibration observations were an acceptable calibration strategy for model averaging in France, showing that a fairly small number of spatially stratified observations (sampling density of 1 sample per 2,500 km^2) provides sufficient calibration data. We also tested the use of model averaging in data-poor situations by reproducing national SOC maps using various sized subsets of the IGCS dataset for model calibration. The results show that model averaging always performs better than the national SOC map. However, the Modelling Efficiency dropped substantially when the national SOC map was excluded in model averaging. This indicates the necessity of including a national SOC map for model

averaging, even if produced with a small dataset (i.e. 200 samples). This study provides a reference for data-poor countries to improve national SOC maps using existing continental and global SOC maps.

Keywords: Soil organic carbon; Digital soil mapping; Bias-corrected Variance Weighted; Sample size requirement; Data-poor countries.

1. Introduction

Soils are crucial for maintaining ecosystem services such as food production, water regulation, erosion control, biodiversity, and climate regulation (Sanchez et al., 2009; Koch et al., 2013; Adhikari and Hartemink, 2016; Rumpel et al., 2018). To meet the increasing demand for up-to-date and fine-resolution soil information, Digital Soil Mapping (DSM, McBratney et al., 2003) has been widely adopted and is being rapidly developed across different spatial scales since the past decade (e.g., Grunwald et al., 2011; Poggio and Gimona. 2014; Viscarra Rossel et al., 2014; Hengl et al., 2015; Ballabio et al., 2016; Padarian et al., 2017; Sanderman et al., 2018; Chen et al., 2019;). At the global scale, different initiatives aim to deliver fine-resolution gridded soil information. The main examples are the recent Global Soil Parnership GSOC map (http://54.229.242.119/GSOCmap/), the *GlobalSoilMap* initiative (Sanchez et al., 2009; Arrouays et al., 2014a), and SoilGrids products (Hengl et al., 2017). SoilGrids adopts a “top-down” approach and produces soil property maps for the entire globe, which are freely distributed and available online (https://soilgrids.org/). *GlobalSoilMap* uses a “bottom-up approach” where each country produces soil property maps using its own national soil data and defined specifications (e.g., 3 arc second resolution, six standard depth intervals, quantified prediction uncertainty, Arrouays et al., 2014b). Then, these country-level soil maps are merged into a global map. There are also several initiatives producing soil property maps at the continental scale, such as LUCAS (Tóth et al., 2013) for Europe and AfSIS (Hengl et al., 2015) for Africa. As a result, there are often multiple maps available for a given soil property in a given area produced using various soil databases, environmental covariates, and DSM methods. Users may have multiple maps of the same property with different predictions and different map accuracy which may lead to confusion regarding which map should be

used or whether the maps could or should be combined. It is possible to select the most suitable soil property map for a specific region, when the map accuracy can be evaluated using an independent validation dataset. When deciding to combine maps, the hypothesis is that the information provided by the maps is complementary and that a more accurate map may be obtained by merging the input maps using model averaging approaches (Caubet et al., 2019). The model averaging option needs an independent validation dataset and independent calibration data to train the model averaging algorithm. Previous studies showed the potential of model averaging in improving the accuracy of soil property maps of pH, soil texture, and available water capacity (Malone et al., 2014; Padarian et al., 2014; Clifford and Guo, 2015; Román Dobarco et al., 2017; Caubet et al., 2019).

The choice between selecting a single map and combining multiple maps is not trivial, and many countries need to make this choice because of the increasing number of different prediction maps of the same soil property. It is particularly relevant to data-poor countries that may have very few or even no data to derive reliable country-based maps, and that could benefit from collecting a limited number of calibration samples to merge the national map with other existing products using model averaging.

The objectives of this study are to 1) evaluate the added value of applying model averaging in a data-rich country (e.g. France); 2) determine the most suitable model averaging approach for improving the topsoil (0-20 cm) SOC map of mainland France using three different SOC maps; 3) evaluate how well the model averaging approaches perform for different calibration sizes and optimize the calibration size required in model averaging; and 4) explore the potential of applying model averaging in data-poor situations.

2. Data

In this study, we used three SOC maps generated and harmonized from national, continental, and global DSM products and two national soil datasets in France.

2.1. French national soil organic carbon maps

Numerous maps have been generated for France following the *GlobalSoilMap* specifications. The most recent product (Mulder et al., 2016a) used all available point data for France, both from the French Soil Mapping and Inventory Program (Inventaire, Gestion et Conservation des Sols, IGCS) and an systematic grid aiming at monitoring French soil properties (RMQS). More details about these two datasets can be found in the study of Mulder et al. (2016a). For this study, we used the same *GlobalSoilMap* approach as Mulder et al. (2016a), but we set aside the RMQS grid to be used as an independent dataset for calibrating the model averaging algorithms and evaluating map accuracy (see Sections 2.3 and 2.4). A total of 30,381 soil profiles from the IGCS dataset were used to generate SOC maps at the first three *GlobalSoilMap* depth intervals (0-5, 5-15, 15-30 cm). The IGCS dataset is a compilation of soil profiles from many programs that mostly focused on agricultural soils. As a result, the soil profile density is high in some regions (Fig. 1), whereas it is low in other regions; some land uses are over- or under-represented in the calibration dataset. SOC contents at the *GlobalSoilMap* depth intervals were obtained by applying equal area quadratic splines (Bishop et al., 1999; Malone et al., 2009) to soil profile data, as outlined in Mulder et al. (2016b). Spatially exhaustive covariates, including climate zones and meteorological data, vegetation, topography, geology,

129 soils, and land management, were resampled to 90 m resolution. Details about these
130 environmental covariates are given in Mulder et al. (2016a). In this study, the national
131 SOC map (named IGCS SOC map hereafter) for the topsoil (0-20 cm) was calculated
132 from SOC maps of 0-5, 5-15, and 15-30 cm by a weighted averaging approach,
133 where the weights are proportional to the layer thickness (Fig. 2a).
134
135 2.2. Continental and global scale soil organic carbon maps
136 In addition to the aforementioned national SOC map, we also obtained SOC
137 maps for France from continental (LUCAS) and global (SoilGrids) soil map products.
138 The LUCAS SOC map (Fig. 2b) contains SOC predictions for the topsoil (0-20
139 cm) at 1 km resolution for Europe (Aksoy et al., 2016). A total of 23,835 soil samples
140 were used for model calibration. These soil samples were collected from LUCAS
141 (19,860 samples), BioSoil (3,379 plots from forest soil), and SoilTrEC (387 samples
142 from local soil data from six different critical zone observatories in Europe). From
143 these datasets, about 3,500 sites were located in France. A regression kriging model
144 was fitted to generate a SOC map using observed SOC content and 15
145 environmental covariates.
146 The SoilGrids SOC map (https://soilgrids.org, v0.5.3, Fig. 2c) was extracted
147 from the study of Hengl et al. (2017), in which SOC was mapped at seven standard
148 depths (0, 5, 15, 30, 60, 100, and 200 cm) at a resolution of 250 m for the globe.
149 These SOC maps were based on about 150,000 soil profiles along with 158 remote
150 sensing-based soil covariates. Maps were produced by fitting an ensemble prediction
151 from random forest and gradient boosting trees. From the 150,000 soil profiles,
152 nearly 3,000 were located in mainland France, mainly originating from the LUCAS
153 database. For this work, the topsoil SOC map was calculated from SoilGrids SOC

maps at 0, 5, 15, and 30 cm depth using trapezoidal numerical integration (Hengl et al., 2017).

The LUCAS and SoilGrids SOC maps were resampled to 90 m using bilinear interpolation and reprojected to the Lambert 93 coordinate system to match these with the national SOC map.

2.3. Independent soil data for model averaging calibration and SOC map validation

To evaluate the accuracy of the input and merged maps, an independent validation dataset and an independent dataset for calibration of the model averaging algorithm were needed. These datasets were derived from the RMQS French systematic grid, which covers different soil, climate, relief, and land cover conditions (Fig. 1). The RMQS dataset is a 16 km × 16 km square grid where sampling sites are at the centre of each grid cell, covering mainland France (Jolivet et al., 2006). For each site, 25 individual core samples were collected by a hand auger and mixed into a composite sample, both for 0–30 cm and 30–50 cm depth intervals. For more detailed information about the soil sampling design and laboratory analyses, refer to Martin et al. (2009). Because there were no SOC measurements for a depth of 0-20 cm for the RMQS sites, we calculated these values depending on land use: 1) for most agricultural soils, SOC concentration decreases at a small rate with depth in the topsoil because of ploughing; thus, SOC content at 0-20 cm is close to that of 0-30 cm (Arrouays et al., 2001). We therefore used SOC at 0-30 cm to represent the SOC at 0-20 cm for RMQS sites under agricultural soils; 2) for natural soils (grassland and forest), SOC usually decreases with depth in the topsoil. Therefore, we first calculated SOC at 0-20 cm and at 0-30 cm by equal area quadratic splines using 5785 grassland and forest soil profiles from the IGCS dataset. We then fitted a linear

model between SOC at 0-20 cm and SOC at 0-30 cm ($SOC_{0\text{-}20\ cm}$ =1.04×$SOC_{0\text{-}30\ cm}$ +0.26, R^2=0.986). We used this model to derive SOC at 0-20 cm from SOC at 0-30 cm for all RMQS sites under natural soils.

3. Methods

3.1. Generic framework for model averaging

Fig. 3 shows the generic framework for model averaging, which includes four steps. We first explain the procedure used for selecting the calibration and validation subsets from the RMQS dataset. To obtain spatially representative calibration and validation datasets, equal-size clustering (iterative nearest neighbour approach, Monlong, 2018) was applied to the RMQS sites (Step 1), which resulted in spatially compact clusters. This was done for five cluster sample sizes (4, 10, 20, 50, and 100). Note that the cluster sample size is only approximately the same for all clusters because the total number of observations (i.e., 1996) is not always a multiple of the cluster sample size. Fig. 4 shows the spatial distribution of the clusters. In Step 2, a *k*-fold cross-validation framework (*k* = 4, 10, 20, 50, 100) was used to separate a calibration set by randomly allocating one observation per cluster to each fold. Thus, the sample size of each fold was approximately 500, 200, 100, 40, and 20, for *k*=4, 10, 20, 50 and 100, respectively. In each of the *k* times, one of the folds was used to calibrate the model averaging approaches (Step 3), whereas the remaining *k-1* folds were used for model validation (Step 4, as explained in Section 3.2). By performing this analysis for different values of *k*, we could also evaluate the performance of the model averaging approaches for different calibration sizes (i.e. 500, 200, 100, 40, and 20). Note that the cross-validation procedure used here has some similarities with spatial cross-validation (Roberts et al., 2017).

204

205 3.2. Model averaging approaches

206 Five model averaging approaches were compared in this study. They are
207 Granger-Ramanathan (Granger and Ramanathan, 1984), Variance Weighted (Bates
208 and Granger, 1969; Heuvelink and Bierkens, 1992), Bayesian model averaging
209 (Hoeting et al., 1999), Piecewise linear decision tree (Quinlan, 1992), and Residual-
210 based piecewise linear decision tree.

211

212 *3.2.1. Granger-Ramanathan*

213 The Granger-Ramanathan (GR) approach was proposed by Granger and
214 Ramanathan (1984). It assumes that a combination of different model predictions can
215 be approached using a traditional Ordinary Least Square (OLS) method. In our case,
216 a linear regression model was fitted between the measured SOC contents of the
217 calibration set and the SOC predictions of the three SOC maps. The outcome SOC_{GR}
218 from the GR approach can be calculated as

219 $$SOC_{GR} = \sum_{i=1}^{p}(\alpha_i \cdot SOC_i) + \beta \quad (1)$$

220 where α_i and SOC_i are the regression coefficient and SOC prediction of the *i*-th SOC
221 map (p=3 in this study), and β is the intercept. The α and β coefficients are solved by
222 the OLS method, and the sum of the α_i is not necessarily equal to 1.

223

224 *3.2.2. Variance Weighted*

225 We used the Bias-corrected Variance Weighted (BC-VW) approach from Ge et
226 al. (2014), which is based on the error variance-covariance matrix that is estimated
227 by comparing model predictions with observations. Thus, the outcome $SOC_{BC\text{-}VW}$ is
228 calculated as

$$SOC_{BC\text{-}VW} = \sum_{i=1}^{p} \alpha_i \cdot (SOC_i - \beta_i) \quad (2)$$

where α_i and SOC_i are the weight and SOC prediction of SOC map *i*, respectively, and β_i is the bias correction coefficient for SOC map *i*. The latter is calculated as

$$\beta_i = \frac{1}{m}\sum_{k=1}^{m}(SOC_{i,k} - SOC_{obs,k}) \quad (3)$$

where *m* is the number of calibration observations, and $SOC_{i,k}$ and $SOC_{obs,k}$ are the SOC prediction of SOC map *i* and the SOC observation at the *k*-th calibration site, respectively.

As described in Ge et al. (2014), the vector $\alpha = [\alpha_1 \cdots \alpha_p]^T$ is calculated by minimizing the error variance of the model predictions:

$$\alpha^T = (\mathbf{1}^T \boldsymbol{V}^{-\mathbf{1}} \mathbf{1})^{-1} \mathbf{1}^T \boldsymbol{V}^{-\mathbf{1}} \quad (4)$$

where **1** is the *p*-dimensional identity matrix (recall that *p*=3 in this study), and $\boldsymbol{V}$ is the *p*-dimensional variance-covariance matrix of the prediction error. The elements of $\boldsymbol{V}$ are determined as

$$\widehat{\boldsymbol{V}}_{\boldsymbol{ij}} = \frac{1}{m}\sum_{k=1}^{m}(SOC_{i,k} - SOC_{obs,k})(SOC_{j,k} - SOC_{obs,k}) \quad (5)$$

where *i,j* = 1,…,*n* represent SOC maps, and *m* is the number of calibration observations. Note that the correlations between SOC map errors are considered in the BC-VW approach.

3.2.3. Bayesian Model Averaging

The Bayesian Model Averaging (BMA) approach assigns a conditional probability density function (PDF) to each model prediction (Hoeting et al., 1999). The BMA posterior distribution of the final output (SOC_{BMA}) can be expressed as (Raftery et al., 2005):

$$p(SOC_{BMA}|SOC_{obs}) = \sum_{i=1}^{p} p(SOC_{BMA}|SOC_{obs}, SOC_i)p(SOC_i|SOC_{obs}) \quad (6)$$

where SOC_{obs} are the SOC observations, p is the number of SOC maps (in this study p=3), and SOC_i denote the values of SOC extracted from the SOC map i at the locations of observations. Therefore, the BMA posterior distribution of SOC_{BMA} is a weighted average of the posterior distributions of SOC_{BMA} under each of the SOC maps, weighted by their posterior model probabilities.

The posterior model probability of SOC_i is expressed as (Raftery et al., 2005)

$$p(SOC_i|SOC_{obs}) = \frac{p(SOC_{obs}|SOC_i)p(SOC_i)}{\sum_{l=1}^{p} p(SOC_{obs}|SOC_l)p(SOC_l)} \quad (7)$$

where $p(SOC_{obs}|SOC_i)$ is the integrated likelihood of SOC_i, and it can be calculated by BIC approximation (more details can be found in Raffery et al., 2005).

We used the R package “BMA” (Raftery et al., 2005) to apply BMA in our case study.

3.2.4. Piecewise linear decision tree

The Piecewise linear decision tree approach (Cubist) is based on the M5 algorithm (Quinlan, 1992). It partitions the dataset into several subsets within which inputs (independent variables) are similar. In a given subset, the standard deviation of the target values is treated as a measure of error and is used as a node splitting criterion. Every potential split is evaluated by the reduction in standard deviation. After evaluating all possible splits, Cubist chooses the one split that maximizes the reduction in error. Then, pruning and smoothing processes are performed to get the final model. More details are given in Quinlan (1992).

In the final Cubist model, partitions are defined by a list of rules, which are arranged in a hierarchy. Each rule has the following form:

if [condition] **then** [linear regression model]

else [apply next rule].

A rule indicates that whenever a case satisfies the condition of one rule, the corresponding linear regression model is used to predict the output. In this study, we used the R package "Cubist" (Kuhn et al., 2012).

3.2.5. Residual-based piecewise linear decision tree

The framework of Residual-based piecewise linear decision tree (Residual-based Cubist, revised from Tao et al., 2018) is as follows: 1) calculate the arithmetic mean SOC value (SOC_{mean}) extracted from IGCS (SOC_{IGCS}), LUCAS (SOC_{LUCAS}), and SoilGrids ($SOC_{SoilGrids}$) SOC maps at locations of soil observations; 2) calculate the residuals (RES_{IGCS}, RES_{LUCAS}, and $RES_{SoilGrids}$) between SOC_{mean} and SOC_{IGCS}/SOC_{LUCAS}/$SOC_{SoilGrids}$, which are used as predictors in the Cubist model; 3) calculate the residuals (RES_{obs}) between SOC_{mean} and SOC observations (SOC_{obs}), which are used as the target variable in the Cubist model ; and 4) once the Cubist model is fitted, calculate the final SOC predictions of the Residual-based Cubist by summing up the RES_{obs} (derived from Cubist) and SOC_{mean}.

3.3. Evaluation of three SOC maps and five model averaging approaches using different calibration sizes

The performance of three individual soil SOC maps was assessed using all RMQS data. Based on a *k*-fold cross-validation framework explained in Section 3.1, we evaluated the five model averaging approaches using different calibration sample sizes (from 20 to 500). Three indicators, the Modelling Efficiency (ME), the Root Mean Square Error (RMSE), and Bias, were used to evaluate prediction accuracy.

$$\mathrm{ME} = 1 - \frac{\sum_{i=1}^{n}(\hat{z}_i - z_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \qquad (8)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(\hat{z}_i - z_i)^2} \quad (9)$$

$$\mathrm{Bias} = \frac{1}{n}\sum_{i=i}^{n}(\hat{z}_i - z_i) \quad (10)$$

where *n* is the size of the cross-validation dataset, z_i and $\hat{z}_i$ are measured and predicted values for the *i*-th observation in the cross-validation dataset, respectively, and $\bar{z}$ is the mean of the observations in the cross-validation datset. A negative ME means that the model performs worse than using the average of the observations as a prediction.

3.4. The effect of national SOC maps on model averaging

The IGCS map was generated using the entire IGCS dataset (about 30,000 soil profiles), which is very large and hence is an example of a case study in a data-rich country (1 profile per 18 km^2). To assess the usefulness of model averaging in data-poor situations, we applied model averaging to a case in which the national SOC map (IGCS) was generated from a much smaller number of soil profiles. To do so, we generated IGCS SOC maps by randomly selecting 10,000, 5,000, 1,000, 800, 600, 400, and 200 soil profiles from the whole IGCS dataset. To filter out random sampling effects, we repeated this procedure 100 times for each sample size and reported the average results. These IGCS SOC maps with LUCAS and SoilGrids were finally merged only with the best model averaging approach and using the minimum necessary number of calibration sites as previously estimated. Using the same minimum necessary number of calibration sites, we tested the assumption of SoilGrids and LUCAS providing additional information that is not captured in IGCS SOC map by removing these two SOC maps in model averaging and only using GR

approach to calibrate the generated IGCS SOC maps (using 200 to 10,000 soil profiles). We also tested model averaging using only SoilGrids and LUCAS to test the assumption that no national SOC map was available.

4. Results

4.1. Summary of IGCS, RMQS, and LUCAS datasets

Table 1 summarises SOC statistics of the IGCS, RMQS, and LUCAS (located in France) datasets. About 80% (24,596) of IGCS soil profiles were located in arable soils, and 20% (5,785) were located in forest and permanent grassland soils. In the IGCS soil database, grassland and forest soils (mean SOC of 24.88 g kg^{-1}) had higher SOC values than arable soils (mean SOC of 16.66 g kg^{-1}). Nearly half (985) of the RMQS sampling sites were located in permanent grasslands or forest soils, and the remaining half (1011) were under arable soils. In the RMQS dataset, the mean SOC was 18.19 g kg^{-1} for arable soils and 35.51 g kg^{-1} for permanent grassland and forest soils. LUCAS observations had a mean SOC of 26.20 g kg^{-1} for permanent grassland and arable soils.

4.2. Evaluation of SOC maps from IGCS, LUCAS, and SoilGrids datasets

The IGCS SOC map has the lowest RMSE (18.86 g kg^{-1}) and highest ME (0.25) among the three SOC maps (Fig. 5). The negative Bias (-6.17 g kg^{-1}) indicates that SOC is underestimated in the IGCS SOC map. When the performance of the IGCS SOC map for arable and forest/grassland soils was separately evaluated, arable soils (ME of 0.19 and RMSE of 10.02 g kg^{-1}) were found to have higher accuracy than forest/grassland soils (ME of 0.09 and RMSE of 24.85 g kg^{-1}). SOC maps of LUCAS and SoilGrids have a much higher RMSE of 30.62 and 32.75 g kg^{-1},

and a negative ME of -1.18 and -1.27, respectively. Positive Bias of LUCAS (6.73 g kg^{-1}) and SoilGrids (21.81 g kg^{-1}) showed that these two maps overestimated SOC. The overestimation was larger in SoilGrids than in the LUCAS SOC map.

4.3. Comparison of five model averaging approaches using different calibration sizes

The BC-VW approach performed best among the five model averaging approaches across different calibration sizes, with the lowest RMSE (16.77-18.71 g kg^{-1}) and highest ME (0.23-0.38) (Fig. 6). The GR and BMA ranked second and third when the calibration size was large (100, 200 or 500), with an ME between 0.33 and 0.38. The performance of GR substantially decreased when using a calibration sample size of 40 and 20, whereas BMA was more stable (and ranked third) when using a small calibration sample size. Cubist performed worst in the case of a large calibration sample size (100, 200, or 500) but ranked second when the calibration sample size was small (20 or 40). Residual-based Cubist did not perform well across the different calibration sample sizes. It should be noted that BC-VW, GR, and BMA had a Bias close to 0 under different calibration sample sizes, while Cubist and Residual-based Cubist had a large negative ME.

All model averaging approaches showed better performance metrics than using the individual LUCAS and SoilGrids SOC maps for all calibration sample sizes. Improvement on the IGCS SOC map only occurred when the calibration sample size was large (100, 200, or 500), while the model averaging approaches performed worse than the IGCS SOC map when the calibration sample size was 20 or 40.

In general, the model performance of the five model averaging approaches declined when the calibration size decreased (Fig. 6). Being the best model averaging approach, BC-VW had better performance than the IGCS SOC map when

374 calibration samples were 500, 200, and 100, and it was still slightly better when only
375 40 calibration samples were used. However, 20 calibration samples were not
376 sufficient to improve SOC maps using any of the five model averaging approaches.
377 GR and BMA could improve SOC predictions when calibration sample sizes were
378 500, 200, and 100. However, Cubist and Residual-based Cubist only performed
379 better than the IGCS SOC map when using a calibration sample size of 200 or more.
380 As shown in Fig. 6, only slight differences (ME of 0.37-0.38, and RMSE of
381 16.77-16.90 g kg^{-1}) were observed between 500 and 200 calibration sample sizes
382 when using BC-VW, which was the best model averaging approach. Nevertheless,
383 the model performance of BC-VW showed a steady decline when the calibration
384 sample size decreased from 200 to 20.
385
386 4.4. SOC maps using five model averaging approaches
387 Fig. 7 shows SOC maps obtained from the five model averaging approaches
388 using all RMQS data for calibration. The general spatial patterns of these five SOC
389 maps were quite close, which is consistent with their similar model performance (in
390 the case of a 500 calibration sample size) in Fig. 6. In comparison with the IGCS
391 SOC map (Fig. 2a), these five SOC maps have higher SOC in mountainous regions
392 (e.g., the Alps, the Central Massif, the Pyrenees), forests, and grasslands (e.g., the
393 Landes of Gascony, western Brittany). As shown in Fig. 7f to Fig. 7o, SOC maps
394 derived from GR, BC-VW, and BMA had slightly higher SOC contents than Cubist
395 and Residual-based Cubist. This is particularly visible in Fig. 7k to Fig. 7o, which
396 zooms in on a square area in the Landes of Gascony forest.
397

4.5. Influence of national SOC maps on model averaging performance

The performance (ME and RMSE) of the IGCS SOC maps derived from different sample sizes showed a slight decline when the number of soil profiles used decreased from 10,000 to 800 (Fig. 8). A stronger decline in performance was observed when the number of soil profiles decreased further from 800 to 200, with ME values dropping from 0.23 to 0.16 and RMSE increasing from 19.11g kg^{-1} to 19.89 g kg^{-1}. The performance of the BC-VW approach on the three SOC maps and the GR approach only on IGCS SOC map showed similar declining trends as the IGCS SOC maps. However, the BC-VW maps always performed better than the IGCS maps ($\Delta ME > 0.1$ and $\Delta RMSE < -2$ g kg^{-1}) and GR maps ($\Delta ME > 0.04$ and $\Delta RMSE < -1$ g kg^{-1}). When using only LUCAS and SoilGrids for model averaging, BC-VW performed much worse than all other SOC maps produced using IGCS, LUCAS, and SoilGrids in model averaging, with a ME of -0.24 and a large RMSE of 23.65 g kg^{-1}.

5. Discussion

5.1. Performance evaluation of SOC maps from IGCS, LUCAS, and SoilGrids

The IGCS SOC map had the best performance indicators among the three source SOC products. However, it showed a slight overall underestimation and a clear tendency to underestimate large SOC values. This may be because the calibration data for generating the IGCS SOC map are dominated by cultivated soils (80% of IGCS dataset), which typically have low SOC values because of management practices (Table 1). As natural soils occupy 45% of the total area of mainland France (Chen et al., 2018), high SOC values are under-represented in the dataset for producing the IGCS SOC map. It consequently resulted in

423 underestimating the effect of some controlling factors driving high SOC values (e.g.,
424 forest or grassland land uses, high elevations). Although the effects of land use and
425 elevation are still clearly visible (Fig. 2a), the spatial patterns of the resulting map are
426 too smooth, as was already described by Mulder et al. (2016a; 2016b). In the French
427 *GlobalSoilMap* product, Mulder et al. (2016a) produced national SOC maps at the
428 first three depth intervals (0-5, 5-15, and 15-30 cm) using both IGCS and RMQS
429 data. The ME evaluated using 10-fold cross-validation ranged from 0.26 to 0.36 for
430 the first three depth intervals. This shows that including RMQS data into national
431 SOC modelling improves model performance. Nevertheless, SOC was still slightly
432 underestimated because the IGCS dataset is almost 15 times larger than the RMQS
433 dataset and IGCS data generally have low SOC content (Table 1).
434 The predictive performance of the LUCAS map and SoilGrids map was much
435 worse than that of the IGCS map, as illustrated in Fig. 2. They both have a tendency
436 to overestimate SOC, either slightly (LUCAS) or largely (SoilGrids). The LUCAS map
437 also exhibited more contrasted and irregular patterns than the IGCS map. Moreover,
438 the LUCAS map showed some areas with artificially rounded boundaries (mainly in
439 southwest France), suggesting a bias linked to the environmental covariates,
440 predictive model, and/or interpolation method used. The SoilGrids map clearly
441 overestimated SOC for the large majority of situations (Fig. 5). It also clearly missed
442 the effect of some land use types on decreasing SOC (e.g., intensively cultivated
443 plains in northern and southwestern parts of France, vineyards in southern France).
444 This suggests that the covariates used for global modelling could not capture these
445 effects; e.g., land use/land cover classes used as covariates for SoilGrids were
446 limited to cultivated land, forests, grasslands, shrublands, wetlands, tundra, artificial
447 surfaces, and bare land cover.

Homogenising data to a common depth of 0-20 cm may have induced some additional uncertainty (Laborczi et al., 2018). We also acknowledge that resampling SoilGrids and LUCAS to 90 m resolution may have added a source of discretionality and potential uncertainty.

5.2. Potential and limitations of model averaging approaches

Our results demonstrate the ability of model averaging approaches to improve national SOC maps (Fig. 5 and Fig. 6). The improvement strongly depends on the calibration sample size used for model averaging. It is encouraging that 200 spatially stratified samples (1 sample per 2,500 km^2) were enough for producing a sufficiently accurate national SOC map (ME of 0.37 for BC-VW approach) when applying model averaging in France. Note also that the performance of this SOC map is comparable to that of the *GlobalSoilMap* SOC map using IGCS and RMQS datasets (Mulder et al., 2016a).

We note that we did not map the uncertainty of SOC predictions when applying model averaging. Prediction uncertainty should be considered in future studies because it is crucial for assessing model quality and robustness. It is also a strongly recommended product outcome, as indicated in the *GlobalSoilMap* specifications (e.g., Arrouays et al., 2014a; Heuvelink, 2014). We could use the method proposed by Ge et al. (2014) to estimate uncertainty when using BC-VW for merging multiple SOC maps.

In addition to deriving SOC predictions using model averaging, it would be beneficial to also explicitly quantify the uncertainties associated with these predictions. This can be done using uncertainty propagation techniques such as the Taylor series method and Monte Carlo simulation (Heuvelink, 2018; Román Dobarco

et al., 2019) provided that the uncertainties of the input maps and their correlations are quantified. This may be a useful extension of the work presented here. If it is done, it would be useful to also evaluate the validity of the uncertainty maps by computing statistics of the standardised squared prediction error (Lark, 2000) and accuracy plots (Goovaerts, 2001; Wadoux et al., 2018).

5.3. Comparison with previous model averaging studies

Our results suggest that map performance improves when using model averaging approaches and that the BC-VW method is the best approach for SOC mapping in mainland France. Previous studies also showed that model averaging improves map predictions, but different approaches tend to have similar performance (e.g., Malone et al., 2014; Román Dobarco et al., 2017; Caubet et al., 2019). Caubet et al. (2019) applied two model averaging approaches (GR and BC-VW) to improve soil texture maps (clay and sand) and showed that both model averaging approaches improved the accuracy and that GR outperformed BC-VW. Similar results were found by Román Dobarco et al. (2017) for mapping soil texture, and Malone et al. (2014) on pH mapping. Indeed, the best-performing algorithm for model averaging may vary between study areas and for different soil properties, and thus optimization of model averaging methods is case-specific.

Caubet et al. (2019) also mentioned the potential use of non-linear models for improving model averaging. However, in our study, non-linear models like Cubist and Residual-based Cubist did not perform better than a linear model like GR. Perhaps this is because three SOC products are not sufficient for calibrating a regression tree or machine learning approach, and that other additional covariates (e.g., elevation, land use, and climatic variables) may be helpful to improve model performance.

Especially, the example of the Landes of Gascony (see Fig. 7k to Fig. 7o) shows that the model does not capture the effect of forest land use well in many areas when using a rule-based model such as Cubist.

Caubet et al. (2019) found that around 200 to 300 calibration samples were sufficient for model averaging of soil texture over mainland France. This result is consistent with our finding that 200 calibration samples (1 sample per 2,500 km^2 for a total area of 550,000 km^2 and a country having a high pedodiversity (Minasny et al., 2010)) selected from equal-size clustering are enough to improve existing SOC maps using model averaging. In our case, it is promising that adding rather few samples improves the SOC maps considerably. This suggests that adding some soil observations uniformly spread over the geographic space helps to correct the bias of the original maps.

5.4. Contribution of model averaging approaches to data-poor countries

We tested model averaging on a situation that may be considered “rich” concerning the amount of available data (Arrouays et al., 2017). In this study, we used 30,000 samples for national SOC mapping, which is 1 sample per 18 km^2. Although France has numerous point soil data, these data are rather clustered and irregularly cover the territory. They also over-represent some agro-pedoclimatic conditions (e.g., low elevations and intensively cultivated areas). These conditions (irregularity and non-representativeness of samples) are likely to be similar in most data-rich countries that use legacy data for DSM.

The fact that the number of samples needed to calibrate the averaging model is rather low is encouraging, i.e. 200 samples for mainland France. This is cost-

522 effective given the limited effort required to gather a fairly small number of soil
523 samples to improve national soil maps.
524 The results shown in Fig. 8 indicate that model averaging always has a
525 substantial added value in terms of model performance compared to using the IGCS
526 SOC map alone. Surprisingly, the added value of model averaging does not depend
527 on the sample size (200 to 10,000 samples) used for producing the national map.
528 This might be due to the fact that our calibration sample size for model averaging (i.e.
529 200 spatially stratified observations) is large enough to capture the main variations of
530 SOC in mainland France. The results shown in Figure 8 also show that removing the
531 LUCAS and SoilGrids SOC maps (GR) decreases the map accuracy in model
532 averaging (BC-VW) which implies that these two SOC maps are complementary to
533 the IGCS SOC map for model averaging. Moreover, the added value of model
534 averaging is larger than that of only increasing the number of profiles used for
535 producing the IGCS SOC map. For example, using 200 samples for model averaging
536 calibration results in an ME increase of 0.12, whereas the ME only increases by 0.07
537 when the number of profiles used for producing the IGCS SOC map increases from
538 200 to 10,000. This indicates that adding a relatively small regular grid of soil
539 samples to merge several maps might be more efficient than expanding the database
540 with a large number of soil samples for which the sample locations are not controlled.
541 In many countries, soil mapping activities are frequently guided by local needs and
542 interests. This explains why national soil datasets are often clustered and why adding
543 more legacy data may sometimes lead to increasing sources of bias (e.g., Poggio et
544 al., 2019). Overall, our study advocates merging predictions in both data-rich and
545 data-poor situations and demonstrates that the added value of merging is relatively
546 higher in data-poor situations. However, notably, the performance of BC-VW drops

547 substantially when excluding the IGCS SOC map and when it only uses LUCAS and
548 SoilGrids for model averaging. This indicates the importance of a national SOC map
549 in model averaging, even if this SOC map is produced with a small dataset (i.e. 200
550 samples).
551
552 **6. Conclusion**
553 We tested the ability of five model averaging approaches for improving
554 existing SOC maps by merging national, continental, and global SOC products. All
555 five model averaging approaches could improve the national SOC map when more
556 than 100 soil samples were used for calibration of the model averaging approaches.
557 The BC-VW approach performed better than the other four approaches. Model
558 averaging approaches using a rather small calibration dataset (i.e. 200 observations
559 uniformly spread over mainland France) for calibration proved to be efficient. The
560 national SOC map was very important and drove performance when merging all SOC
561 maps, however SoilGrids and LUCAS SOC maps had added value by capturing
562 relevant patterns additional to the national SOC map. By reducing the number of
563 national soil samples in France for producing the national SOC map, we found that
564 merging maps using model averaging is also applicable to data-poor situations and
565 might thus be attractive to data-poor countries, provided sufficient soil data are
566 available for calibration of the model averaging approach.
567
568 **Acknowledgements**
569 Soil data collection was supported by the French Scientific Group of Interest
570 on soils: the GIS Sol, involving the French Ministry of Ecology, the French Ministry of
571 Agriculture, the French Environment and Energy Management Agency (ADEME), the

French Institute for Research and Development (IRD), the French National Geographic and Forest Inventory Institute (IGN) and the French National Institute for Agronomic Research (INRA). This work was partly funded by the Environment & Agronomy Department of INRA, in the framework of its calls for innovative research (grant no. 6282), and parly funded by the project Coordination of International Research Cooperation on soil CArbon Sequestration in Agriculture (CIRCASA) (grant no. 774378) under H2020-EU.3.2.1.1. Vera Laetitia Mulder, Laura Poggio and Dominique Arrouays are members of a Research Consortium supported by LE STUDIUM Loire Valley Institute for Advanced studies. We also thank all colleagues involved in soil sampling and populating the soil database. Songchao Chen received support from the China Scholarship Council (grant no. 201606320211).

584 **References**

585 Adhikari, K., Hartemink, A.E., 2016. Linking soils to ecosystem services—A global
586 review. Geoderma 262, 101–111.
587 Aksoy, E., Yigini, Y., Montanarella, L., 2016. Combining soil databases for topsoil
588 organic carbon mapping in Europe. PloS One 11, e0152098.
589 Arrouays, D., Deslais, W., Badeau, V., 2001. The carbon content of topsoil and its
590 geographical distribution in France. Soil Use Manage. 17, 7–11.
591 Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M.,
592 Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J.,
593 Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez,
594 P.A., Thompson, J.A., Zhang, G.-L., 2014a. Chapter Three — GlobalSoilMap:
595 Toward a Fine-Resolution Global Grid of Soil Properties. Adv. Agron. 125, 93–
596 134.
597 Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C.,
598 Greve, M., Grundy, M., Guerrero, E., Hempel, J.W., Hengl, T., Heuvelink,
599 G.B.M., et al., 2017. Soil legacy data rescue via GlobalSoilMap and other
600 international and national initiatives. GeoResJ 14, 1–19.
601 Arrouays, D., McKenzie, N.J., Hempel, J., Richer de Forges, A.C., McBratney, A.B.,
602 2014b. GlobalSoilMap: Basis of the Global Spatial Soil Information System. 1st
603 ed. CRC Press Taylor & Francis Group, pp. 9–12.
604 Ballabio, C., Panagos, P., Monatanarella, L., 2016. Mapping topsoil physical
605 properties at European scale using the LUCAS database. Geoderma 261, 110–
606 123.
607 Bates, J.M., Granger, C.W., 1969. The combination of forecasts. Oper. Res. Soc. 20,
608 451–468.

609 Bishop, T.F.A., McBratney, A.B., Laslett, G.M., 1999. Modelling soil attribute depth
610 functions with equal-area quadratic smoothing splines. Geoderma 91, 27–45.
611 Caubet, M., Román Dobarco, M., Arrouays, D., Minasny, B., Saby, N.P.A., 2019.
612 Merging country, continental and global predictions of soil texture: Lessons from
613 ensemble modelling in France. Geoderma 337, 99–110.
614 Chen, S., Arrouays, D., Angers, D.A., Chenu, C., Barré, P., Martin, M.P., Saby,
615 N.P.A., Walter, C., 2019. National estimation of soil organic carbon storage
616 potential for arable soils: A data-driven approach coupled with carbon-landscape
617 zones. Sci. Total Environ. 666, 355–367.
618 Chen, S., Martin, M.P., Saby, N.P.A., Walter, C., Angers, D.A., Arrouays, D., 2018.
619 Fine resolution map of top-and subsoil carbon sequestration potential in France.
620 Sci. Total Environ. 630, 389–400.
621 Clifford, D., Guo, Y., 2015. Combining two soil property rasters using an adaptive
622 gating approach. Soil Res. 53, 907–912.
623 Ge, Y., Avitabile, V., Heuvelink, G.B.M., Wang, J., Herold, M., 2014. Fusion of pan-
624 tropical biomass maps using weighted averaging and regional calibration data.
625 Int. J. Appl. Earth Obs. 31, 13–24.
626 Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science.
627 Geoderma 103, 3–26.
628 Granger, C.W., Ramanathan, R., 1984. Improved methods of combining forecasts. J.
629 Forecasting 3, 197–204.
630 Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital soil mapping and
631 modeling at continental scales: Finding solutions for global issues. Soil Sci. Soc.
632 Am. J. 75, 1201–1213.

633 Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić,
634 A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara,
635 M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E.,
636 Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil
637 information based on machine learning. PLoS One 122, e0169748.
638 Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd,
639 K.D., Sila, A., MacMillan, R.A., de Jesus, J.M., Tamene, L., Tondoh, J.E., 2015.
640 Mapping soil properties of Africa at 250 m resolution: Random forests
641 significantly improve current predictions. PloS One 10, e0125814.
642 Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. In
643 GlobalSoilMap: basis of the global spatial soil information system. 1st ed. CRC
644 Press Taylor & Francis Group, pp. 335–340.
645 Heuvelink, G.B.M., 2018. Uncertainty and uncertainty propagation in soil mapping
646 and modelling. Pedometrics, pp.439–461.
647 Heuvelink, G.B.M., Bierkens, M.F.P.,1992. Combining soil maps with interpolations
648 from point observations to predict quantitative soil properties. Geoderma 55, 1–
649 15.
650 Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model
651 averaging: a tutorial. Stat. Sci. 14, 382–401.
652 Jolivet, C., Arrouays, D., Boulonne, L., Ratié, C., Saby, N.P.A., 2006. Le réseau de
653 mesures de la qualité des sols de France (RMQS). Etat d'avancement et
654 premiers résultats. Etude et Gestion des Sols 13, 149–164.
655 Koch, A., McBratney, A.B., Adams, M., Field, D., Hill, R., Crawford, J., Minasny, B.,
656 Lal, R., Abbott, L., O'Donnell, A., Angers, D., 2013. Soil security: solving the
657 global soil crisis. Glob. Policy 4, 434–441.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., 2012. Cubist models for regression. R package Vignette R package version 0.0, 18.

Laborczi, A., Szatmári, G., Kaposi, A.D., Pásztor, L., 2018. Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma https://doi.org/10.1016/j.geoderma.2018.01.020

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51, 137–157.

Loiseau, T., Chen, S., Mulder, V.L., Román Dobarco, M., Richer-de-Forges, A.C., Lehmann, S., Bourennane, H., Saby, N.P.A., Martin, M.P., Vaudour, E., Gomez, C., Lagacherie, P., Arrouays, D., 2019. Satellite data integration for soil clay content modelling at a national scale. Int. J. Appl. Earth Obs. 82, 101905.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152.

Malone, B.P., Minasny, B., Odgers, N.P., McBratney, A.B., 2014. Using model averaging to combine soil property rasters from legacy soil maps and from point data. Geoderma 232, 34–44.

Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G., Arrouays, D., 2009. Optimizing pedotransfer functions for estimating soil bulk density using boosted regression trees. Soil Sci. Soc. Am. J. 73, 485–493.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.

Minasny, B., McBratney, A.B., Hartemink, A.E., 2010. Global pedodiversity, taxonomic distance and the World Reference Base. Geoderma 155, 132–139.

682 Monlong, J., 2018. Hippocamplus, Github repository,
683 https://github.com/jmonlong/Hippocamplus/blob/master/content/post/2018-06-09-
684 ClusterEqualSize.Rmd
685 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Arrouays, D., 2016a.
686 GlobalSoilMap France: High-resolution spatial modelling the soils of France up
687 to two meter depth. Sci. Total Environ. 573, 1352–1369.
688 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016b.
689 National versus global modelling the 3D distribution of soil organic carbon in
690 mainland France. Geoderma 263, 16–34.
691 Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: a
692 contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28.
693 Padarian, J., Minasny, B., McBratney, A.B., Dalgliesh, N., 2014. Predicting and
694 mapping the soil available water capacity of Australian wheatbelt. Geoderma
695 Reg. 2, 110–118.
696 Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon
697 stocks with uncertainty propagation—an example from Scotland. Geoderma 232,
698 284–299.
699 Poggio, L., Lassauce, A., Gimona, A., 2019. Modelling the extent of northern peat
700 soil and its uncertainty with Sentinel: Scotland as example of highly cloudy
701 region. Geoderma 346, 63–74.
702 Quinlan, J.R., 1992. Learning with continuous classes. In 5th Australian Joint
703 Conference on Artificial Intelligence, 92, 343–348.
704 Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian
705 model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133,
706 1155–1174.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40 (8), 913–929.

Román Dobarco, M., Arrouays, D., Lagacherie, P., Ciampalini, R., Saby, N.P.A, 2017. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. Geoderma 298, 67–77.

Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N.P.A., Cousin, I., Martin, M.P., 2019. Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma 344, 14–30.

Rumpel, C., Amiraslani, F., Koutika, L., Smith, P., Whitehead, D., Wollenberg, D., 2018. Put more carbon in soils to meet Paris climate pledges. Nature 564, 32–34.

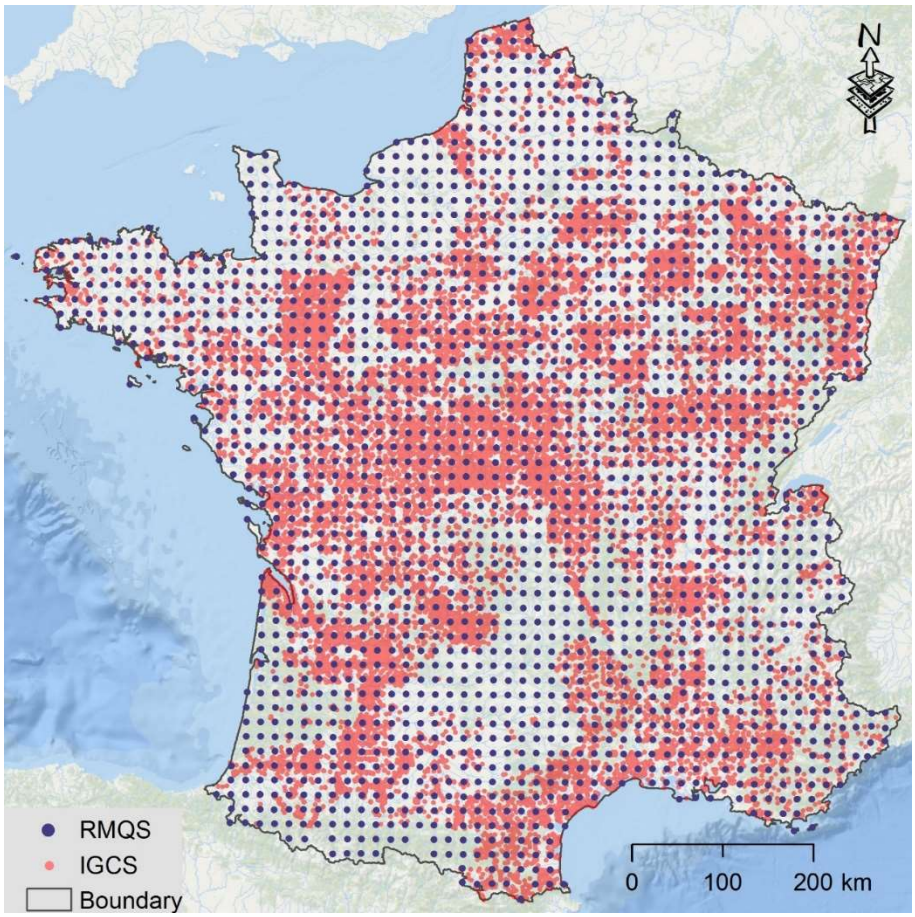
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., de Lourdes Mendonça-Santos, M., Minasny, B., 2009. Digital soil map of the world. Science 325, 680–681.

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M. F., Benson, L., Bukoski, J.J., Carnell, P., Cifuentes-Jara, M., Donato, D., Duncan, C., Eid, E.M., zu Ermgassen, P., Lewis, C.J.E., Macreadie, P.I., Glass, L., Gress, S., Jardine, S.L., Jones, T.G., Nsombo, E.N., Rahman, M.M., Sanders, C.J., Spalding, M., Landis, E., 2018. A global map of mangrove forest soil carbon at 30 m spatial resolution. Environ. Res. Lett. 13, 055002.

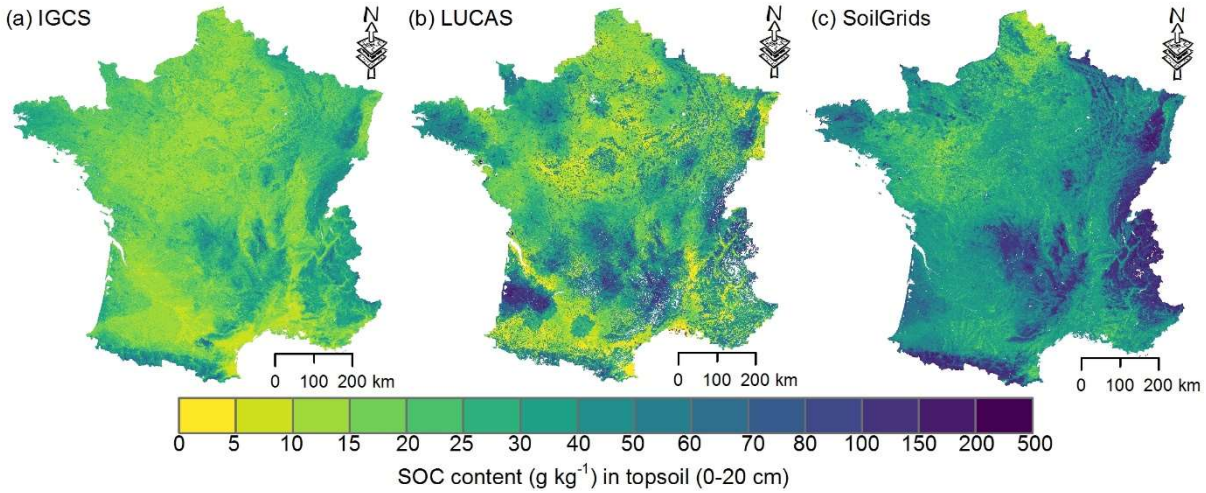
730 Tao, Y., Yang, T., Faridzad, M., Jiang, L., He, X., Zhang, X., 2018. Non-stationary
731 bias correction of monthly CMIP5 temperature projections over China using a
732 residual‐based bagging tree model. Int. J. Climatol. 38, 467–482.
733 Tóth, G., Jones, A. Montanarella, L., 2013. The LUCAS topsoil database and derived
734 information on the regional variability of cropland topsoil properties in the
735 European Union. Environ. Monit. Assess. 185, 7409–7425.
736 Viscarra Rossel, R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of
737 organic carbon in Australian soil to support national carbon accounting and
738 monitoring under climate change. Glo. Change Biol. 20, 2953–2970.
739 Wadoux, A.M.C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary
740 variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

Figures

Fig. 1 Study area (Mainland France) and soil sampling sites from IGCS and RMQS datasets.



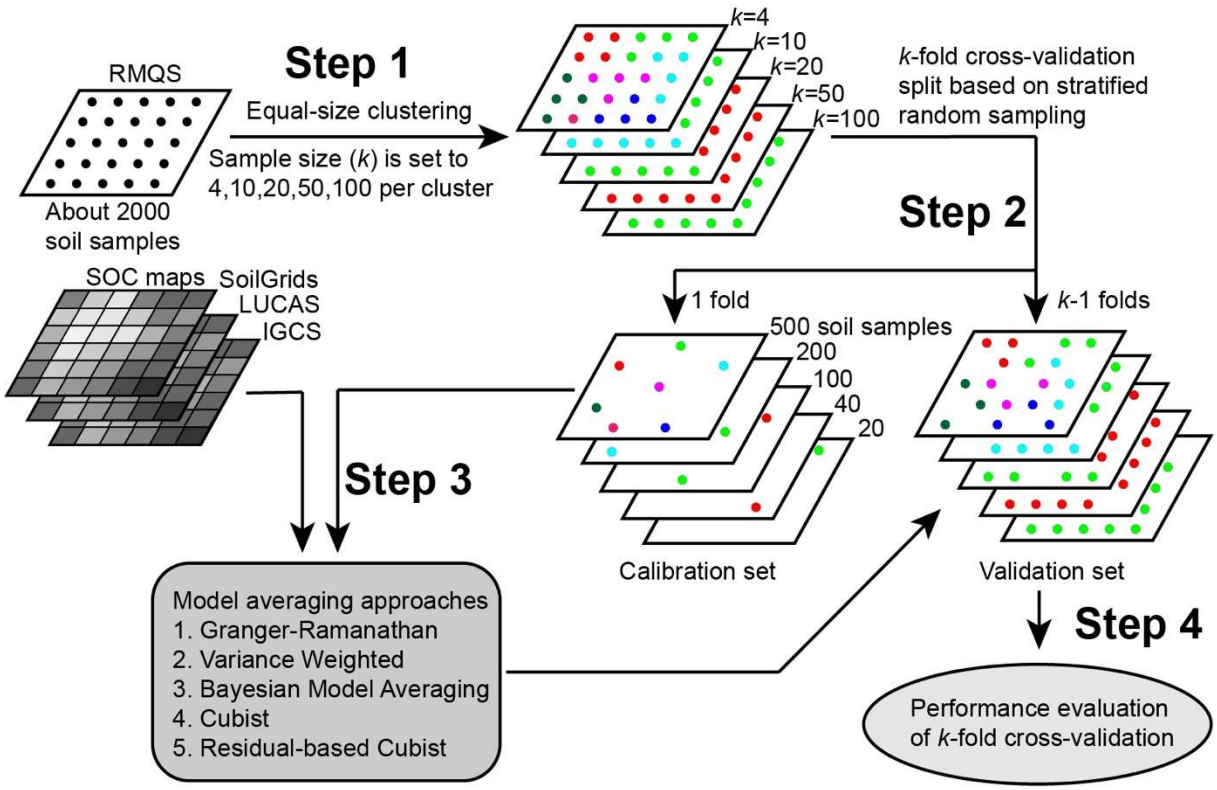
746 Fig. 2 SOC maps of mainland France from IGCS (a), LUCAS (b) and SoilGrids (c).



747

748

749 Fig. 3 Model averaging workflow.



750
751

Fig. 4 Spatial cluster distribution of RMQS sites, using equal-size clustering. The cluster sample sizes are 4 (a), 10 (b), 20 (c), 50 (d) and 100 (e).

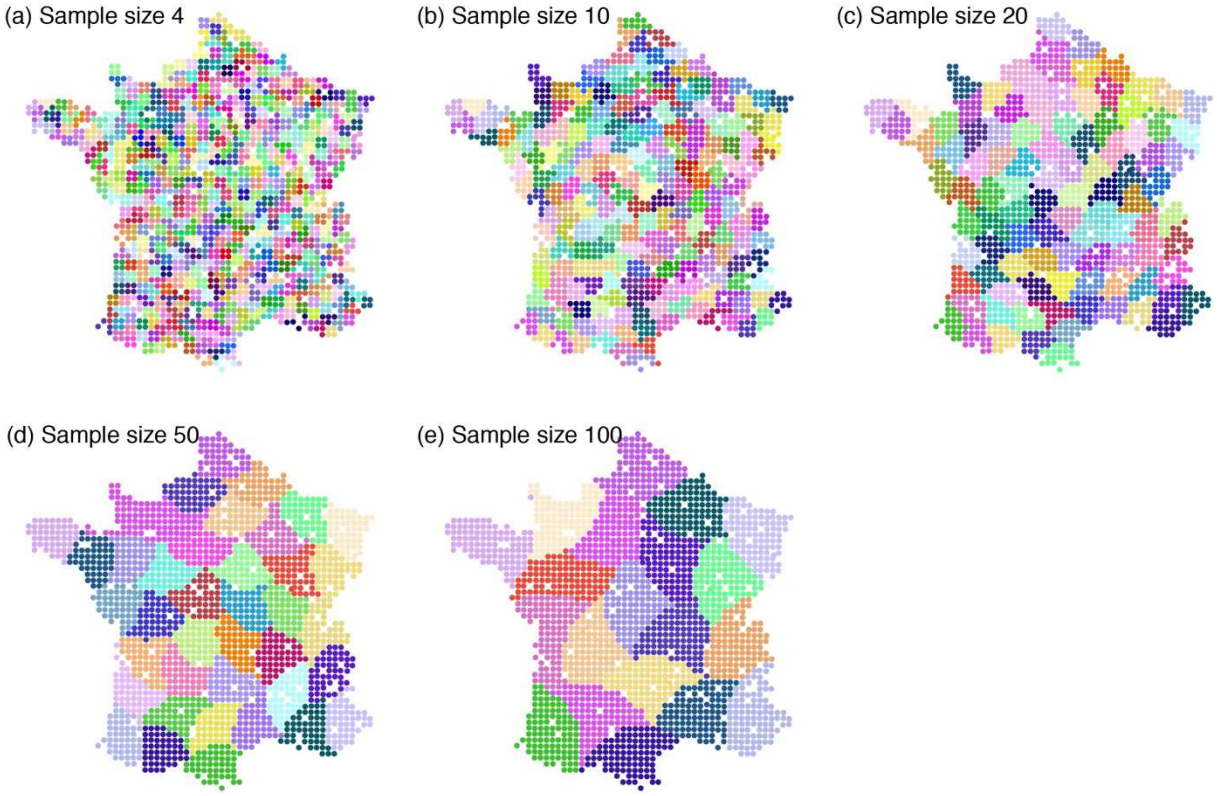
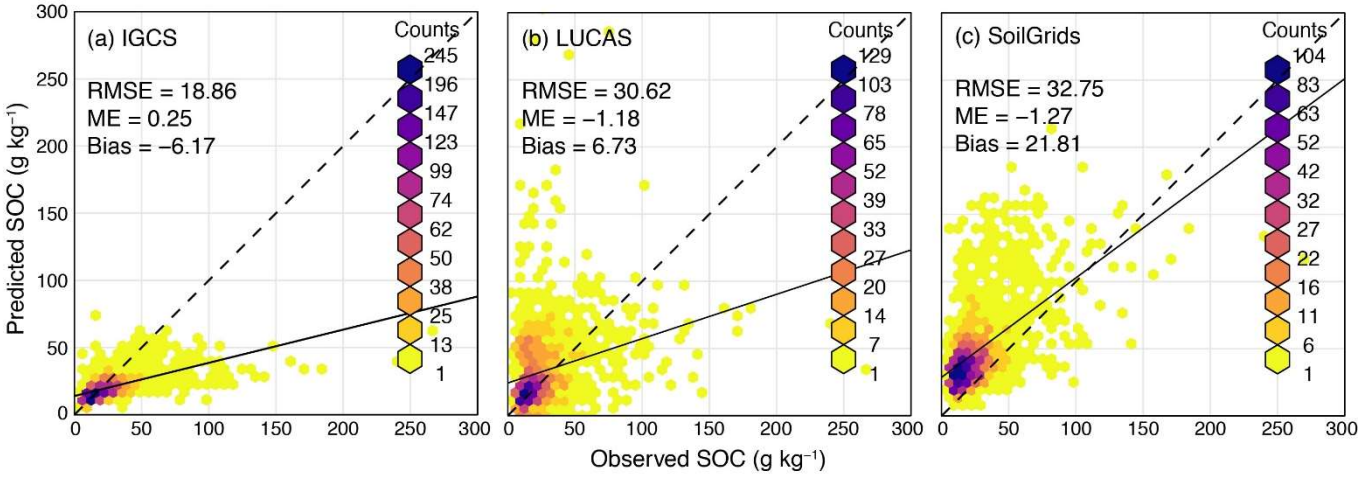
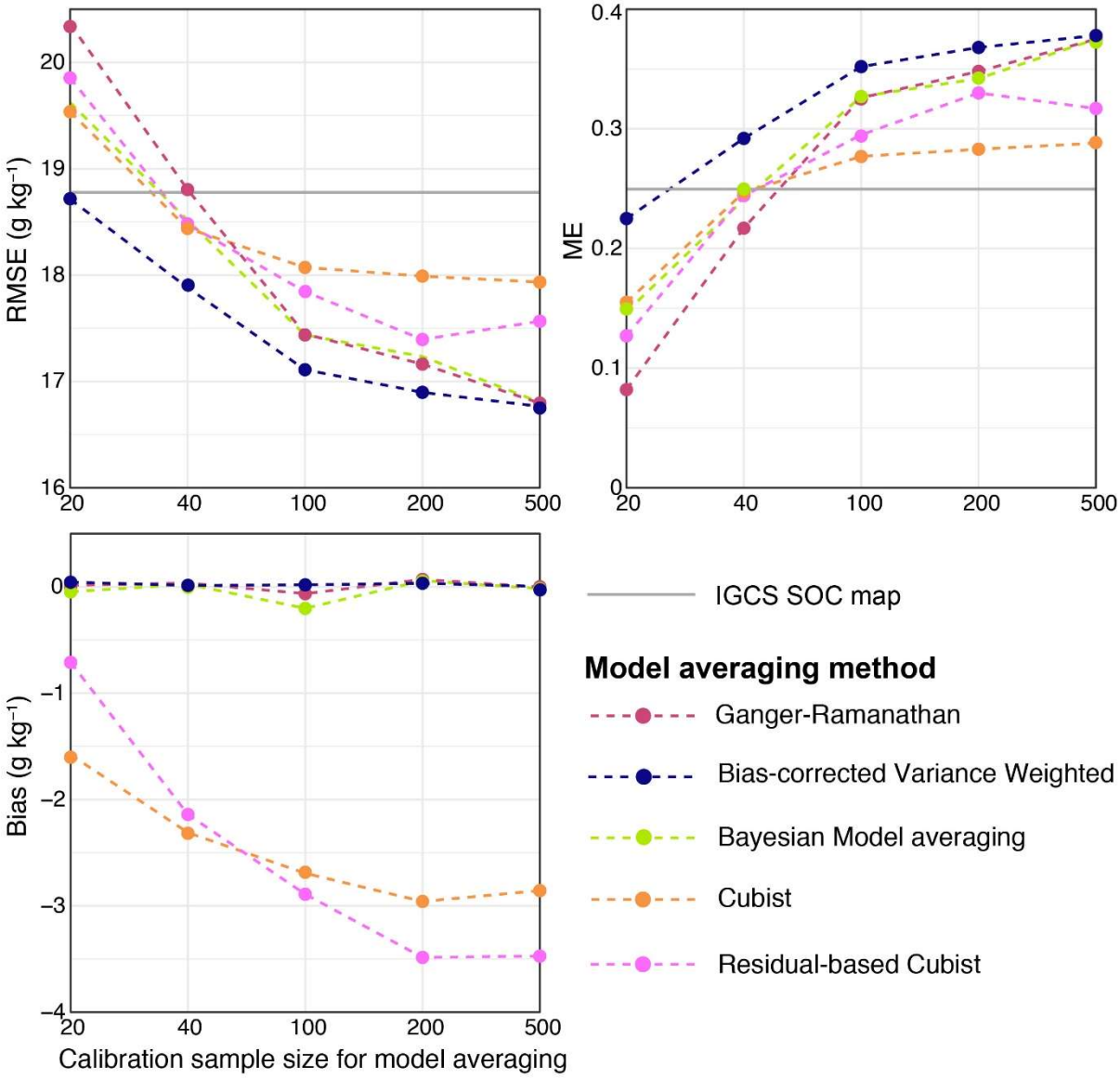


Fig. 5 Performance of IGCS (a), LUCAS (b) and SoilGrids (c) SOC maps.

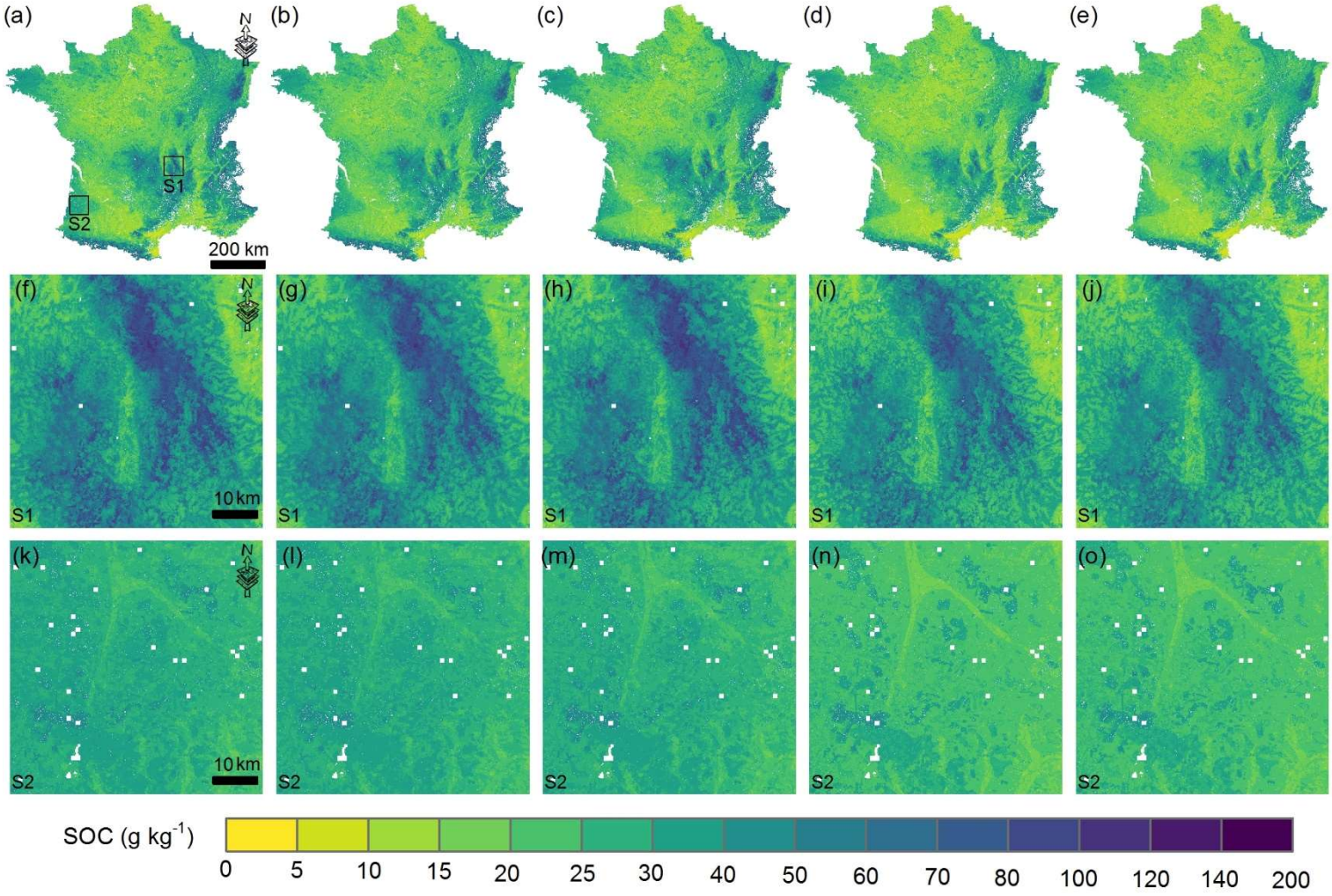


758 Fig. 6 Model performance of the five model averaging approaches using different
759 calibration sample sizes.



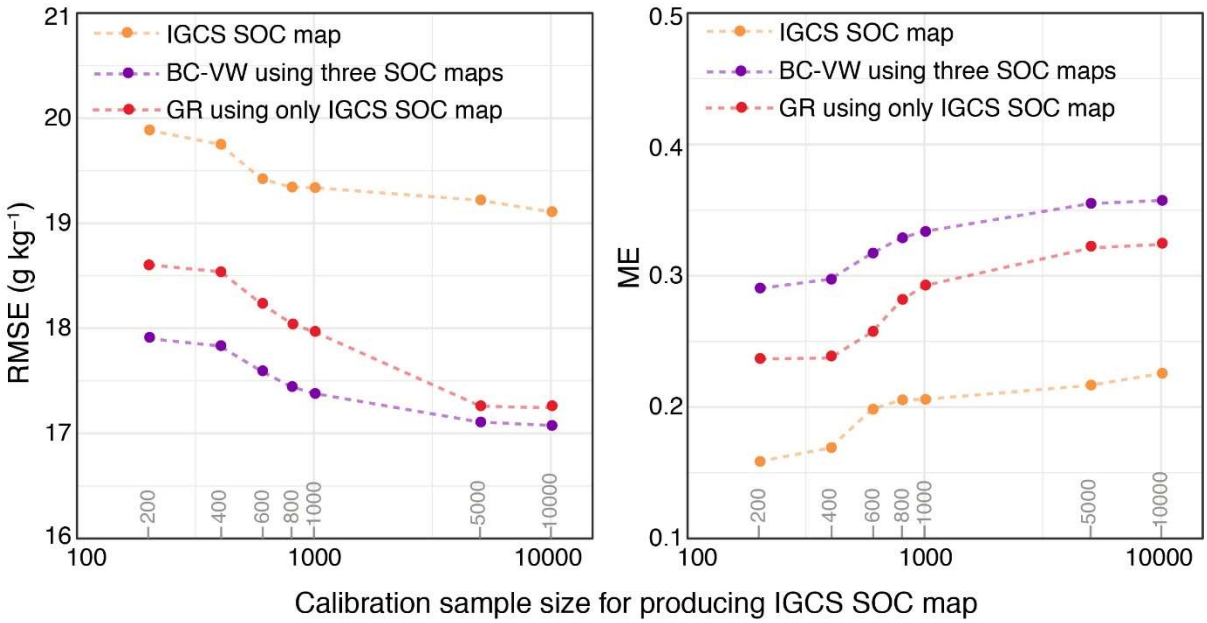
760
761

762 Fig. 7 SOC maps obtained from the Granger-Ramanathan (a), Bias-corrected
763 Variance Weighted (b), Bayesian Model averaging (c), Cubist (d) and Residual-
764 based Cubist (e) model averaging approaches, using all RMQS data for calibration.
765 Local comparisons in areas S1 (f, g, h, I and j) and S2 (k, l, m , n and o) are also
766 shown for all five model averaging approaches.



767

Fig. 8 Model performance of the Bias-corrected Variance Weighted (BC-VW) model averaging (using 200 calibration samples for three SOC maps) and Granger-Ramanathan (GR) model (using the same 200 calibration samples for only calibrating IGCS SOC map) when using different calibration sample sizes (200 to 10,000) for generating IGCS SOC map. Using only the LUCAS and SoilGrids SOC maps for BC-VW model averaging leads to an RMSE of 23.65 g kg^{-1} and ME of -0.24 (points not shown). The x-axis is on log10 scale.



Tables

Table 1 Summary statistics of SOC content (g kg^{-1}) in topsoil (0-20 cm) for IGCS, RMQS and LUCAS datasets.

Dataset	Land use*	N	Min.	Q1	Median	Mean	Q3	Max.	Sk.	SD
IGCS	F & G	5,785	0.39	12.75	19.86	24.88	30.83	373.00	3.42	20.97
	A	24,596	0.09	9.70	13.68	16.66	19.75	354.05	4.92	12.88
RMQS	F & G	985	3.78	18.86	28.37	35.51	44.00	266.60	2.81	26.01
	A	1,011	2.58	11.10	15.40	18.19	22.30	133.00	3.01	11.16
LUCAS	A & G	2,950	1.00	13.20	19.99	26.20	31.30	472.10	6.11	23.93

N, dataset size ; Min., minimum; Q1, first quantile; Q3, third quantile; Max., maximum; Sk., skewness; SD, standard deviation. * F, forest; G, permanent grasslands; A, arable.