National estimation of soil organic carbon storage potential for arable soils: A data-driven approach coupled with carbon-landscape zones

Songchao Chen, Dominique Arrouays, Denis Angers, Claire Chenu, Pierre Barré, Manuel Martin, Nicolas Saby, Christian Walter

Title: National estimation of soil organic carbon storage potential for arable soils: a data-driven approach coupled with carbon-landscape zones

Authors:

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

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

Denis A. Angers [c]. denis.angers@canada.ca

Claire Chenu [d]. claire.chenu@inra.fr

Pierre Barré, [e]. barre@geologie.ens.fr

Manuel P. Martin [a]. manuel.martin@inra.fr

Nicolas P.A. Saby [a]. nicolas.saby@inra.fr

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

Affiliations:

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

[b] UMR SAS, INRA, Agrocampus Ouest, 35042 Rennes, France

[c] Québec Research and Development Centre, Agriculture and Agri-Food Canada, Québec, GIV 2J3 Canada

[d] UMR Ecosys, INRA, AgroParisTech, Université Paris-Saclay, Campus AgroParisTech, 78850 Thiverval-Grignon, France

[e] Laboratoire de Géologie de l'ENS, PSL Research University, UMR8538 du CNRS, 75231 Paris, France

Corresponding author: Songchao Chen

E-mail: songchao.chen@inra.fr

Telephone: +33(0)602142667

Abstract

Soil organic carbon (SOC) is important for its contributions to agricultural production, food security, and ecosystem services. Increasing SOC stocks can contribute to mitigate climate change by transferring atmospheric CO_2 into long-lived soil carbon pools. The launch of the 4 per 1000 initiative has resulted in an increased interest in developing methods to quantity the additional SOC that can be stored in soil under different management options. We made a first attempt to estimate SOC storage potential of arable soils using a data-driven approach based on the French National Soil Monitoring Network. The data-driven approach was used to determine the highest reachable SOC stocks of arable soils for France. We first defined different carbon-landscape zones (CLZs) using clustering analysis. We then computed estimates of the highest possible values using percentile of 0.8, 0.85, 0.9 and 0.95 of the measured SOC stocks within these CLZs. The SOC storage potential was calculated as the difference between the highest reachable SOC stocks and current SOC stocks for topsoil and subsoil. The percentile used to determine highest possible SOC had a large influence on the estimates of French national SOC storage potential. When the percentile increased from 0.8 to 0.95, the national SOC storage potential increased by two to three-fold, from 336 to 1020 Mt for topsoil and from 165 to 433 Mt for subsoil, suggesting a high sensitivity of this approach to the selected percentile. Nevertheless, we argue that this approach can offer advantages

from an operational point of view, as it enables to set targets of SOC storage taking into account both policy makers' and farmers' considerations about their feasibility. Robustness of the estimates should be further assessed using complementary approaches such as mechanistic modelling.

Keywords: Soil organic carbon; Storage potential; Data-driven approach; Carbon-landscape zones; Gaussian mixture models; Soil management practices.

1. Introduction

Globally, the soil C pool (2500 Gt) is 3.3 times the size of the atmospheric pool (760 Gt) and 4.5 times the size of the above-ground vegetation pool (560 Gt). Therefore, soils have the potential to partly offset anthropogenic greenhouse gas emissions by sequestering SOC (Lal, 2004; Paustian et al., 2016). Moreover, increasing soil organic carbon (SOC) generally improves soil quality and functioning, and thus can potentially contribute to enhance agricultural production and food security, restore degraded land, and promote ecosystem services such as erosion mitigation, soil water provision, nutrient availability for plants, and soil biodiversity (Lal, 2004; Stockmann et al., 2013). Recognizing the importance of increasing SOC at the global scale, a voluntary action initiative “4 per 1000 carbon sequestration in soils for food security and the climate” (http://4p1000.org/) was launched at the COP21. The 4 per 1000 initiative aims at promoting land management practices (e.g., conservation agriculture, cover cropping, agroforestry) leading to an increase in SOC stocks in the 0 to 0.4m layer at an aspirational annual growth rate of 0.4% of current SOC stocks (Lal, 2016; Minasny et al., 2017).

The SOC storage potential generally refers to the maximum gain in SOC stock attainable at a given timeline by implementing changes in land use or management, and will vary under different pedoclimatic conditions (Post and Kwon, 2000; Stockmann et al., 2013; Barré et al., 2017; Chenu et al., 2018). The concept of SOC saturation has been used to estimate the maximum

amount of SOC that can be associated with the fine fraction (Hassink, 1997) and therefore considered as relatively stable. In the context of the 4 per 1000 initiative, the aspirational target of increasing SOC stocks at an annual growth rate of 0.4% relates to the total (whole-soil) SOC stocks in the 0-0.4 m layer (whole-soil, including the coarse fraction). Therefore, determining whole-soil SOC storage potential using the maximum SOC associated with the fine fraction is not appropriate because the SOC stored in the coarse fraction can represent a large percentage of the total SOC stocks. As summarized by Chen et al. (2018), under temperate climate, SOC in the coarse fraction could account, on average, for 15%, 34% and 31% of total SOC stocks under cropland, forest and grassland, respectively, in topsoil, and account for nearly 25%, 14% and 7% of SOC stocks for cropland, forest and grassland in subsoil.

For an improved quantification of SOC storage potential, Barré et al. (2017) proposed one avenue: i) First, establish the reference stocks with an estimate of the highest reachable SOC stock for a given soil; ii) second estimate possible SOC storage between the current SOC stock of a given soil and this reachable highest SOC stock under a given land-use for different land management practices. Furthermore, Barré et al. (2017) suggested that this avenue can be achieved using either a data-driven approach (empirical observation of SOC stocks and storage) or mechanistic simulation models. The data-driven approach assumes that the highest reachable SOC stocks under a specific land use/cover or land management practices for each

74 different pedoclimatic conditions could be empirically determined by the
75 highest values (e.g., by the mean of using top quantiles) among the observed
76 SOC stocks for these conditions. This hypothesis implicitly assumes that the
77 values of the top quantiles reflect the optimal management practices for SOC
78 storage and they are thus considered as 'proxies' of the maximum reachable
79 SOC stocks under these different pedoclimatic conditions.

80 Based on the detailed and extensive French Soil Monitoring Soil Network
81 data base, our objective was to test a data-driven approach for estimating
82 SOC storage of arable soils in mainland France. We developed a procedure
83 which consisted of: i) determining carbon-landscape zones by clustering the
84 data from a combination of net primary production (C input), climatic
85 decomposition index (C decomposition) and soil clay content (C protection
86 from decomposition); ii) estimating the maximum SOC stocks of arable soils
87 (topsoil and subsoil) for each carbon-landscape zone using four percentiles
88 (0.80, 0.85, 0.90 and 0.95); iii) calculating by difference with the current SOC
89 stocks, the SOC storage potential of arable topsoil and subsoil under these
90 four percentiles.

91 ## 2. Materials and methods

92 ### 2.1 Soil data

93 Covering the entire mainland France under different soil, climate, relief and
94 land cover conditions, 2092 sites from the first campaign of the French Soil

Monitoring Network (RMQS) were sampled from 2001 to 2009. The RMQS is based on a 16 km × 16 km square grid and all sites were selected at the centre of each grid cell. Topsoil (0-30 cm) and subsoil (30-50 cm) were collected using a hand auger. For each site, 25 samples were merged into a composite sample and then were air-dried (controlled at a temperature of 30 ºC and an air-moisture of 30%) and sieved to 2 mm before laboratory analysis. A soil pit was dug at 5 m from the south border of sampling sites, and the main soil characteristics were recorded and bulk density and percentage of coarse elements were measured (Martin et al., 2009). For some RMQS sites, subsoil did not exist as soils were thin at these locations. SOC was determined by dry combustion. Only these RMQS sites (n=1089) located on arable soils were used in this study (Figure 1).

The SOC stock was calculated as below:

$$SOC_{stock} = p \times SOC \times BD \times (100 - ce) \times 10^{-2} \quad (1)$$

where SOC_{stock} is the SOC stock (kg m^{-2}), p is the actual thickness (cm) of topsoil or subsoil, *SOC*, *BD* and *ce* are the content of SOC (g kg^{-1} or ‰), bulk density (kg m^{-3}), and percentage of coarse elements (%).

2.2 Net primary production, climatic data, soil clay content and SOC stocks maps

Net primary production (NPP) was extracted from the MOD17A2H version 6 Gross Primary Production product (NASA LP DAAC, 2017) from 2000 to

115 2010. It is a cumulative 8-day composite of values with 500-meter original
116 resolution. The 8-day NPP data is averaged into monthly data and resampled
117 to 1 km resolution. Cities and water-covered regions have been masked in
118 this product.

119 WorldClim Version 2 (Fick and Hijmans, 2017), which is spatially
120 interpolated using between 9000 and 60000 weather stations globally, was
121 used for climatic data: It has average monthly climate data for minimum,
122 mean, and maximum temperature and for precipitation, solar radiation, wind
123 speed and water vapour pressure for 1970-2000 at 1 km resolution.

124 Maps of soil clay content for topsoil (0-30 cm) and subsoil (30-50 cm) were
125 derived from *GlobalSoilMap* France products (Mulder et al., 2016). As these
126 were produced at six standard depth intervals (e.g., 0-5 cm, 5-15 cm, 15-30
127 cm, 30-60 cm, 60-100 cm and 100-200 cm), soil clay content maps were
128 harmonized using a depth-weighted method (Appendix Figure A1).

129 The Corine Land Cover 2006 (UE-SOeS, 2006) was used as the land
130 cover/use classification map. It has an original resolution at 100 m and was
131 resampled to 90 m in order to meet the requirement of the *GlobalSoilMap*
132 project (Sanchez et al., 2009; Arrouays et al., 2014). The Corine Land Cover
133 map was reclassified as cropland, forest, grassland and others, and only
134 cropland was presented in this study (Figure 1).

135 The current SOC stocks map for topsoil (0-30 cm) was produced using
136 RMQS dataset by a hybrid model coupling the boosted regression trees (BRT)

and robust geostatistical approaches described in Martin et al. (2014). The covariates used in modelling were explicitly documented in Chen et al. (2018). To remove the interference of the positions without SOC stocks in subsoils (where subsoil does not exist), a three-stage approach was applied for SOC stocks modelling in the subsoil (30-50 cm): 1) produce a map to identify whether subsoils exist using BRT model; 2) produce a SOC stocks map by the hybrid model, where the RMQS sites without SOC stocks are excluded; 3) merge the two maps by keeping the SOC stock values where subsoils exist and setting the locations where subsoil do not exist as NA (not available). The SOC stocks maps for topsoil and subsoil have a spatial resolution of 90 m and they can be found in the Appendix Figure A2. The national SOC stocks were 3.65 Gt and 1.04 Gt for topsoil and subsoil, respectively. Cropland contained 1.37 Gt and 0.44 Gt SOC in the topsoil and subsoil.

All the datasets were reprojected to Lambert 93, which is an official projection for mainland France.

2.3 Calculation of climatic decomposition index

Carbon decomposition generally increases with temperature and moisture, a climatic decomposition index (CDI) was used to characterise the interaction between temperature and water stress as suggested by Carol Adair et al. (2008).

Before determining the CDI, potential evapotranspiration (PET) was calculated using Hargreaves model (Hargreaves et al., 1985), which performs

well and requires less parameterization than the Penman-Monteith method (Hargreaves and Allen, 2003). Monthly PET (mm month^{-1}) is defined below:

$$PET = 0.0023 \times SR \times (T_{mean} + 17.8) \times \sqrt{T_{range}} \quad (2)$$

where *SR* is monthly solar radiation (mm month^{-1}, transformed from KJ m^{-2} day^{-1}), T_{mean} is monthly mean temperature (ºC) and T_{range} is the difference between the monthly maximum and minimum temperature (ºC).

The CDI is calculated as a function of the mean monthly mean temperature (*T*), monthly precipitation (*PPT*) and monthly *PET* (Carol Adair et al., 2008):

$$CDI = F_T(T) \times F_W(PPT, PET) \quad (3)$$

$$F_T(T) = 0.5766 \times e^{308.56 \times \left(\frac{1}{56.02} - \frac{1}{(273+T)-227.13}\right)} \quad (4)$$

$$F_W(PPT, PET) = \frac{1}{1 + 30 \times e^{-8.5 \times \frac{PPT}{PET}}} \quad (5)$$

where $F_T(T)$ and $F_W(PPT, PET)$ are the monthly effects of temperature and water stress on decomposition.

2.4 Delineation of carbon-landscape zones using Gaussian mixture models

Generally, SOC dynamics depend on the trade-off between the SOC input and SOC loss processes. When SOC input is greater than OC loss, the soil will accumulate C, and otherwise, soil C will decrease (Lal et al., 2015). Climatic decomposition index and NPP are here considered as proxies of C loss and input that control the SOC balance, and clay content considered as a controlling factor of SOC persistence. The underlying simplifying assumption is that decomposition mainly depends on both climate and soil characteristics.

Therefore monthly CDI and NPP, and soil clay content were used to compute the carbon-landscape zones (CLZs) using Gaussian mixture model (GMM) which is a similar approach to that used by Mulder et al. (2015). To reduce multicollinearity and computing time, principal component analysis (PCA) was performed before the clustering step on monthly CDI and NPP data separately. We retained only the first three and four principal components that explained more than 95% of the variance for CDI and NPP, respectively. Therefore, after adding soil clay content for topsoil and subsoil, a total of nine variables were used for GMM clustering. Moreover, to reduce computing complexity, we also selected 20000 pixels in France as calibration data set of the GMM clustering. The resulting clustering model was then used to predict to which CLZ each pixel of the entire territory belongs.

Gaussian mixture model was conducted to compute clusters that were considered as CLZs in this study. GMM is one of the model-based clustering techniques, which optimizes the fit between the measured data and mathematical models using a probabilistic approach. GMM is based on the assumption that the data are generated by a mixture of Gaussian distributions. Then, the parameters of GMMs are estimated by maximisation of the likelihood using the Expectation Maximization (EM) algorithm. EM algorithm starts with a random initialization and then iteratively optimizes the clustering using two steps: (i) Expectation step determines the expected probability of assignment of data to clusters using current model parameters; (ii)

Maximisation step updates the optimal model parameters of each mixture based on the new data assignment.

The number of clusters was tuned from 1 to 30 and their associated Bayesian information criterion (BIC) was calculated for the evaluation of clustering performance. The number of clusters was selected considering a trade-off between the BIC values and the available number of RMQS sites within each land use for each cluster. GMMs were performed using ClusterR package in R 3.3.2 (Mouselimis, 2016; R Core Team, 2016). The optimized CLZs map was resampled to 90 m resolution.

2.5 SOC storage potential and analysis of the sensitivity to the percentile setting

Empirical maximum SOC stock values were estimated for arable topsoil and subsoil under given CLZs using RMQS dataset (point observations). Four percentiles at 80%, 85%, 90% and 95% were tested to estimate the empirical maximum SOC stock values that could be reached under a given CLZ. A bootstrapping approach was applied to assess the uncertainty from data source both for each CLZ and tested percentiles. We repeated the bootstrapping procedure 100 times and thus obtained 100 estimates of the maximum SOC stock values for each CLZ and percentile. The mean value obtained from these one hundred estimates was used as an estimate of the maximum SOC stock value for each CLZ and percentile. We then estimated the uncertainty (90% confidence intervals, 90% CIs) of these maximum SOC

221 stock values by using the 5 and 95 percentile of the bootstrapping results.

222 The SOC storage potential was calculated as the difference between the
223 empirically-determined maximum SOC stocks and current SOC stocks (Figure
224 A2) under arable land use. Four SOC storage potential maps were produced
225 using the four tested percentiles for both topsoil and subsoil, and their
226 associated 90% CIs.

227 We evaluated the effect of percentile setting on the estimation of SOC
228 storage potential by both comparing the differences in the SOC storage
229 potential spatial distribution and national SOC storage potential estimates.

230 **3. Results**

231 3.1 Spatial distribution of CDI, NPP and their principal components

232 Figure 2 shows the spatial distribution of CDI and NPP in mainland France.
233 Globally, CDI increased gradually from January to August and then decreased
234 gradually to December. Different from CDI, NPP started to increase from
235 January and reached the peak in June, and then decreased gradually to
236 December.

237 Accounting for 98.3% and 97.0% of the total variances (95% was set as a
238 threshold), the first three and four principal components (PCs) were kept for
239 CDI and NPP, respectively. Figure 3 presents the final seven PCs used in
240 clustering. The 3 PCs of CDI showed long range spatial patterns in mainland
241 France while the spatial patterns for 4 PCs of NPP were mainly characterized

242 by median and short ranges.

243 3.2 Carbon-landscape zones

244 The BIC value decreased quickly when the number of clusters was less
245 than 10, and then it decreased slowly after 10 clusters (Figure 4). The result
246 indicated that more clusters were helpful for separating the differences within
247 clusters. However, more clusters meant less available RMQS sites falling into
248 each cluster. Figure 5 shows the number of RMQS sites located in each
249 cluster. Our aim was to avoid clusters having a number of RMQS sites less
250 than ten, which may not be enough to derive a robust estimate of the
251 quantiles. Two clusters had less than ten RMQS sites when the number of
252 clusters varied from 8 to 10. When the number of clusters increased from 11
253 to 13, three clusters were found with less than ten RMQS sites. We optimized
254 the number of clusters at ten as it appeared to be the best compromise
255 between separating the differences between clusters and keeping an
256 acceptable number of clusters having less than ten RMQS sites.

257 Figure 6 illustrates the spatial distribution of CLZs in mainland France.
258 CLZ 1 is mainly distributed in north-eastern France which is characterized by
259 a rather continental climate and relatively high soil clay contents, mostly
260 ranging from 22% to 35% in topsoil, and being even higher in subsoil
261 (Appendix Figure A3). CLZ 2 represents most of western France
262 characterized by a mild and wet oceanic climate and relatively homogeneous
263 soil clay contents (mostly ranging from 15 to 20% both in top- and subsoil,

Appendix Figure A3). CLZ 3 is located in northern France and mainly corresponds to the maximal extension of deep loess deposits. It exhibits clay contents centred around 20% for topsoil and a bit higher for subsoil, both with a rather low statistical dispersion. CLZ 4 is located in the Massif Central and the Vosges mountains, and is characterized by a rather cold climate due to elevation and rather homogeneous clay contents, mostly ranging from 15% to 20% for both layers (Appendix Figure A3). CLZ 5 is located in southern France and strictly corresponds to the area of the ‘Landes of Gascony’ which is characterized by a mild climate and nearly pure sandy aeolian deposits having clay content nearly always less than 5% (Augusto et al., 2010). CLZ 6 is located in central France and corresponds to the foothills of the Massif Central, with a lower elevation than its central part. Part of the CLZ 6 is also spread in various other locations, all of which corresponding to ancient alluvial deposits coming from these foothills. In topsoil, most clay contents range from 15% to 20% and slightly higher in subsoil. CLZ 7 is exclusively located in the highest elevations located at the top of the main mountain ranges (Pyrenees, Alps, Jura and Massif Central), with soil texture being rather clayey (around 30%). This CLZ also includes many thin soils (Lacoste et al., 2016), and thus the information on clay content of the 0.3 to 0.5 m layer is often missing. CLZ 8 occupies most of south-western France characterized by mild winters and hot summers. It is characterized by a very large range and dispersion of clay content in both layers, although a large part (interquartile range) ranges from

20% to 30%. CLZ 9 is mainly distributed in central and northern France. Its clay content in topsoil and subsoil is centred around 25% (Appendix Figure A3) and showing a small increase in subsoil. Lastly, CLZ 10 shows low NPP values in autumn, because of land use consisting mainly of vineyards and wheat crops. It is clearly located in the Mediterranean region with very hot temperatures and very low NPP in summer. The clay content is centred around 20%, with a statistical dispersion similar to the other CLZs.

Figure 7 presents the design-based estimates of SOC stocks for arable soils for the ten CLZs in topsoil and subsoil. In order to get unbiased estimates, these estimates were computed using the values obtained from the RMQS grid values within each CLZ. The median SOC stocks of topsoil ranged from 4.89 to 9.67 kg m^{-2} under the 10 CLZs. Fewer differences of SOC stocks were found in subsoil, and subsoil had much lower SOC stocks than topsoil with a range of median SOC stocks from 1.31 to 2.08 kg m^{-2}.

3.3 Empirical maximum SOC stocks under four percentile settings

As expected, there was a clear trend that the maximum SOC stocks for topsoil and subsoil increased when percentile became higher, however, the magnitude of these increases varied among different CLZs (Figure 8). In topsoil, large differences (>4 kg m^{-2}) in maximum SOC stocks between percentile of 0.95 and percentile of 0.8 were observed in CLZ 1 and CLZ 4, and the differences ranged from 0.28 to 3.65 kg m^{-2} for other CLZs. In subsoil, differences in maximum SOC stocks between percentile of 0.95 and

308 percentile of 0.8 were below 1.5 kg m^{-2} for almost all the CLZs, except for CLZ
309 4 with a value of 2.71 kg m^{-2}.

310 The 90% CIs also differed between CLZs as well as between percentiles.
311 A large percentage of high 90% CIs (upper limit minus lower limit > 10 kg m^{-2}
312 for topsoil or > 5 kg m^{-2} for subsoil) of maximum SOC stocks were found in
313 CLZ 4 and CLZ 7, which indicated large variability for these two mountainous
314 CLZs having a rather low number of sites. Besides, subsoil in arable soils had
315 lower 90% CIs than topsoil.

316 3.4 Spatial distributions of SOC storage potential

317 Figure 9 and Figure 10 show the spatial distributions of SOC storage
318 potential and 90% CIs under four percentile settings for topsoil and subsoil,
319 respectively. When percentile was set at 0.8, French arable topsoil had a SOC
320 storage potential less than 2 kg m^{-2} except for a part of Brittany and south-
321 eastern France near the Mediterranean Sea. With the increasing percentile,
322 intensively cultivated plains of the central, the northern half and the
323 southwestern part of France showed a large potential to store more SOC.
324 Cropland located around mountainous regions including the Pyrenees, the
325 Alps, the Jura and the Vosges generally had a relatively low SOC storage
326 potential across all percentiles. Large differences were observed for total SOC
327 storage potential under different percentile settings (Table 2). The French
328 national SOC storage potential and 90% CIs for arable topsoil were 336
329 (203,501) Mt when percentile was 0.8. Larger increases were observed for

total SOC storage potential and 90% CIs with the increasing percentiles, which reached at 470 (308,662) Mt, 674 (434,950) Mt and 1020 (740,1283) Mt for a percentile of 0.85, 0.9 and 0.95, respectively.

The subsoil showed much lower SOC storage potential than topsoil. Most regions of mainland France had low SOC storage potential (< 1 kg m^{-2}) at percentiles of 0.8 and 0.85, and relative high SOC storage potential (1-3 kg m^{-2}) were observed in central France. Similar with topsoil, increasing percentiles resulted in higher SOC storage potential across mainland France, and fewer differences of SOC storage potential were found between cropland located around mountainous regions and other regions under four percentile settings. At percentile of 0.8, subsoil had the potential to sequester 165 Mt additional SOC with a 90% CI between 91 Mt and 250 Mt. Total SOC storage potential and their 90% CIs were 228 (150,306) Mt, 309 (226,404) Mt and 433 (331,560) Mt for percentiles of 0.85, 0.9 and 0.95, respectively.

4. Discussion

4.1 Optimizing and mapping Carbon Landscape Zones

The estimates of SOC storage potential using a data-driven approach were based on a stratification of the study area using the CLZs, therefore a procedure for optimizing the number of CLZs was necessary. We observed a negative trend between the number of clusters and BIC, which indicated that using more clusters allowed to explain more variance of our covariates.

However, as soil data was finite, creating too many clusters would have resulted in fewer soil data available for each CLZ. We assumed in this study that performing a statistical analysis with less than 10 samples was not robust; therefore we decided to optimize the number of clusters by considering a trade-off between the BIC value and the number of RMQS sites located within each cluster. Interestingly, though using a very different set of covariates and soil point data (different covariates, and a much larger number of soil point data), Mulder et al. (2015) found that the same number of clusters (10) was optimal to partition points data into soil-landscape systems relevant to SOC. Moreover, their maps showed rather similar spatial patterns (e.g. in the Mediterranean region, mountains, and western France).

4.2 National SOC storage potential

As expected, the percentile setting had a strong influence on the estimation of SOC storage potential (Table 2). If we use the national SOC storage potential at a percentile of 0.8 as a benchmark, the total SOC storage potential at percentiles of 0.85, 0.9 and 0.95 were 1.40, 2.01 and 3.04 times larger, respectively, in topsoil and were 1.38, 1.87 and 2.62 times larger, respectively, in the subsoil. Clearly, the estimates of SOC storage potential are very sensitive to the percentile chosen, especially at high values setting (e.g., 0.95).

4.3 Limitations of the data-driven approach

372 The data-driven approach has previously been implemented in a few
373 pedoclimatic regions to estimate SOC storage potential. Stolbovoy and
374 Montanarella (2008) used data from the European Soil Portal database to
375 determine the maximum observed SOC stocks for a given soil type under a
376 given climate, from which they subtracted the observed SOC stocks under
377 cultivated land. Lilly and Baggaley (2013) determined for each typological soil
378 unit the observed maximum SOC stocks, from which they subtracted the
379 observed median SOC stock under cultivated topsoils. One main difference
380 between these studies and the present one is that they did not calculate
381 percentiles but used only as reference the maximum observed values which
382 are obviously much more sensitive to the presence of very high values.
383 Another difference is that they used coarse resolution data, some of which
384 may not always be directly related to controlling factors of SOC (e.g., soil type,
385 highly aggregated data for delineating large bioclimatic regions).
386 We show here that this approach has some limitations. It is very sensitive
387 to percentile setting. This is partly attributable to the fact that the SOC
388 distributions are highly skewed with long tails at high SOC values (e.g., CLZs
389 1, 3, 4, 6, 8 and 10, see Fig.7 and Fig.8). This approach could be also
390 considered as data ‘hungry’. This sensitivity is also linked to the fact that we
391 have a rather limited number of observations for some CLZs, especially those
392 with a small crop land area (e.g. CLZs, 4 and 7, see Fig. 8), which hampers
393 the robustness of the data-driven approach. Another limitation may come from

394 the fact that some cultivated soils may have been recently converted from
395 other land uses (e.g., grassland, forest) and may not have yet reached an
396 equilibrium level, which could partly explain the long tails that we observed.
397 One alternative approach would consist in performing dedicated sampling in
398 the CLZs following a probability sampling as suggested by De Gruitjer et al.
399 (2016). In this approach, the number of sites is selected with a minimum
400 number in order to get precise estimates of the quantiles.

401 In addition, Barré et al. (2017) already mentioned two other limitations.
402 Firstly, this approach provides an estimate of soil storage potential under
403 present management practices, therefore this estimate could be largely
404 underestimated when new SOC aggrading techniques are adopted. As
405 discussed by Sparling et al. (2003), current management practices may
406 strongly affect the outcomes of a data driven approach when deriving
407 desirable soil organic carbon contents from the median of observed SOC
408 contents. Secondly, another limit of data-driven approaches would be that, for
409 most available databases, management practices are not documented, and
410 thus make it difficult to determine their influence (Barré et al., 2017). Indeed,
411 in some cases there is still a large diversity of soils within a same CLZ and
412 also very different land use histories which are not considered in this
413 approach. The influence of these two factors on the potential storage maps
414 can be easily seen for instance for western France (CLZ2, characterized by a
415 gradient linked to the date of grasslands conversion to croplands). Similarly,

the gradients observed in piedmont areas may be linked to the fact that large parts of them have been more or less recently converted from forest or grassland to cropland (e.g., Arrouays et al., 1994, 1995a, 1995b; Saby et al., 2008) and thus still have quite large SOC stocks reflecting their past land use. Finally, a CLZ may include very different agricultural production systems and in some cases reaching the storage potential would not only require to change the management practices, but the whole production system. The estimates we provide may be refined in the future taking into account the different agricultural production systems (for CLZ with enough sites).

Despite these limitations, we consider that this first national approximation of SOC storage potential is valuable in making use of a detailed and robust nation-scale database. We further point out some operational advantages of the data driven approach in section 4.6.

4.4 Complementarity with other approaches

Using a method based on the carbon saturation equation of Hassink (1997), Chen et al. (2018) estimated the SOC sequestration potential in mainland France using the same RMQS data. In their work, the concept of SOC sequestration potential referred to the additional SOC associated with soil fine fraction (< 20 µm), assumed to have pluri-decadal residence times. Their results showed that arable topsoil and subsoil could theoretically sequester 646 Mt and 752 Mt SOC, respectively. Though SOC associated with the soil fine fraction does not represent the total SOC, their estimate of

SOC sequestration potential in arable topsoil was close to the percentile of 0.9 derived SOC storage potential (674 Mt), suggesting that SOC sequestration potential can hardly be reached under current management practices. The maps of SOC sequestration potential obtained applying Hassink's equation (Chen et al., 2018) and the maps of SOC storage potential obtained through the data driven approach show rather good qualitative agreements in the western part of France. However, noticeable differences are observed in mountain areas and in the most clayey CLZs for which the data driven approach predicts a much lower additional storage potential than the theoretical SOC sequestration potential. Apart from the fact that the two maps rely on different concepts (sequestration and storage, e.g., Barré et al., 2017; Chenu et al., 2018) and different modes of calculation, this may also suggest that the pedoclimatic conditions in rather cold or clayey situations do not allow to reach the theoretical SOC sequestration potential because of insufficient plant biomass inputs. In arable subsoil, SOC storage potentials derived from a data-driven approach (under all percentiles) were much lower than C-saturation theoretical SOC sequestration potential. This may be attributed to the fact that the present data-driven estimate of SOC storage potential is based on current land management practices, while reaching the estimated SOC sequestration potential for subsoil may need more advanced land management practices with more potential to raise the SOC in both topsoil and deeper layers (Chenu et al., 2018). This may be also simply due to

460 the fact that the French pedoclimatic conditions do not allow to reach the C-
461 saturation theoretical SOC sequestration potential.

462 As suggested by Barré et al. (2017), the model-driven approach would be
463 another way of estimating SOC storage potential. In a model-driven approach,
464 process-based models are used for determining highest reachable SOC
465 stocks by simulating different management scenarios. Such an approach has
466 been applied to EU by Lugato et al. (2014). Compared to a data-driven
467 approach, this process-based model may be more suitable as it is able to
468 monitor SOC stock dynamics. However, there are also some limitations to this
469 model-driven approach: i) a lot of input data is required for modelling, for
470 instance, a CENTURY model needs site-specific precipitation, temperature,
471 soil texture, bulk density, initial SOC, land use and corresponding
472 management practice; ii) the initialization for C dynamic models is still very
473 problematic and the simulation for large dataset is time-consuming; iii) the
474 accuracy of C dynamics model prediction needs to be validated by resampled
475 soil data and (iv) the soil management options considered are limited to those
476 accounted for in current SOC dynamics models (e.g. agroforestry may not be
477 considered in most models).

478 4.5 SOC storage potential and 4 per 1000 goal

479 Based on our current SOC stock maps shown in Figure A2, the total SOC
480 stocks are estimated at 1.37 Gt and 1.81 Gt for French arable soils for the 0-
481 30 cm layer and the 0-50 cm layer, respectively. If we base these estimates

on the total area of French arable soils, reaching the 4 per 1000 aspirational target would require a storage rate of 5.48 Mt C $year^{-1}$ for 0-30 cm, or 7.24 Mt C $year^{-1}$ for 0-50 cm. According to the C storage rate for 0-30 (0-50) cm, it would take 61 (69), 85 (96), 122 (135) and 186 (200) years to reach the SOC storage potential under percentiles of 0.8, 0.85, 0.9 and 0.95, respectively. Thus our data-driven estimates of C storage potential suggest that achieving an annual rate of increase of 0.4% would have to be maintained for decades before reaching the SOC storage potential of these soils, provided that relevant management options can be implemented for such an annual SOC storage, and keeping in mind that an equilibrium level may be reached after a few decades.

4.6 The data driven approach, a potentially operational tool

We observed that that SOC storage potential is very sensitive to the percentile used in the calculation. We submit that this approach offers potential for operational purposes as it enables to set targets of SOC carbon storage for both policy makers and farmers. For instance, decision-makers may decide to implement policies aiming at reaching a minimal objective (for instance, all sites should reach the 0.6 percentile), an intermediate objective (0.8 percentile) or an ambitious objective (0.9 percentile). It could therefore be a very suitable tool to determine to which extent soils can contribute to Intended Nationally Determined Contributions (INDCs). As an additional step, more emphasis should be put both on policy and recommendations to reach

these objectives for different soils, agricultural productions systems and land use histories within each CLZ, and ultimately on developing methods to verify that the targeted objectives are reached. This approach could then be further used to improve the data-driven approach and to design future objectives. Similarly, at a local scale, farmers may compare their present SOC stocks to the theoretically reachable ones within their CLZ, and decide which goal may be reachable by implementing more or less drastic or costly changes to their management practices. They may even find out that the SOC stocks at their farm level are already close to the maximal reachable value, and thus concentrate on not losing SOC rather than on trying to increase the current stocks.

5. Conclusions

We tested a data-driven approach to estimate SOC storage potential under Carbon Landscape Zones for arable soils using the French National Soil Monitoring Network. Under the trade-off between the BIC index and available data for robust statistics, the optimized number of CLZs was determined at 10, using monthly NPP, CDI, and clay content data. The national SOC storage potential varied from 336 Mt to 1020 Mt for topsoil and from 165 Mt to 433 Mt for subsoil under four percentile settings (0.8, 0.85, 0.9 and 0.95), which shows that the data-driven approach is very sensitive to the selected percentile. This sensitivity was partly attributable to a rather low number of observations in some CLZs and mainly to skewed distributions with

long tails of high SOC contents. However, we argue that this data driven approach offers meaningful advantages from an operational point of view, as it enables to adapt targets of SOC carbon storage by taking into account both policy makers' and farmers' considerations. We also argue that the data driven approach is also a convenient way to provide quantitative estimates of the SOC storage potential over large areas having widely distributed soil data. Dedicated surveys and research on management practices effects are still necessary in order to better estimate the reachable SOC stocks and the feasibility of their implementation.

Further work will focus on estimating SOC storage potential by the model-driven approach in mainland France. Producing model-driven estimates may enable to determine a more reliable percentile setting for the data-driven approach and thus provide references for the regions where exhaustive data for applying process-based models is not available.

Acknowledgement

Soil data gathering was supported by a French Scientific Group of Interest on soils: the GIS Sol, involving the French Ministry of Ecology, the French Ministry of Agriculture, the French Environment and Energy Management Agency (ADEME), the French Institute for Research and Development (IRD) and the National Institute for Agronomic Research (INRA). We thank all the people involved in sampling the sites and populating the database. Songchao Chen received the support of China Scholarship Council for three years' Ph.D.

study in INRA and Agrocampus Ouest (under grant agreement no. 201606320211). This work will also constitute a main input to the ANR (French Research National Agency) project StoreSoilC ANR-17-CE32-0005-01.

References

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Young Hong, S., … Zhang, G., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. Adv. Agrono. 125, 93-134.

Arrouays, D,. Pélissier, P., 1994. Changes in carbon storage in temperate humic loamy soils after forest clearing and continuous corn cropping in France. Plant Soil 160, 215-223.

Arrouays, D., Balesdent, J., Mariotti, A., Girardin, C., 1995a. Modelling organic carbon turnover in cleared temperate forest soils converted to maize cropping by using 13C natural abundance measurements. Plant Soil 173(2), 191-196.

Arrouays, D., Vion, I., Kicin, J.L., 1995b. Spatial analysis and modeling of topsoil carbon storage in forest humic loamy soils of France. Soil Sci. 159, 191-198.

Augusto, L., Bakker, M.R., Morel, C., Meredieu, C., Trichet, P., Badeau, V., … Ranger, J., 2010. Is grey literature a reliable source of data to characterize soils at the scale of the region? A case study in a maritime pine forest of

570 south-western France. Eur. J. Soil Sci. 61, 807-822.
571 Barré, P., Angers, D. A., Basile-Doelsch, I., Bispo, A., Cécillon, L., Chenu,
572 C., … Pellerin, S., 2017. Ideas and perspectives: Can we use the soil
573 carbon saturation deficit to quantitatively assess the soil carbon storage
574 potential, or should we explore other strategies?. Biogeosciences Discuss.
575 https://doi.org/10.5194/bg-2017-395.
576 Adair, E.C., Parton, W.J., Del Grosso, S.J., Silver, W.L., Harmon, M.E., Hall,
577 S.A., … Hart, S.C., 2008. Simple three□pool model accurately describes
578 patterns of long-term litter decomposition in diverse climates. Glob. Chang.
579 Bio. 14(11), 2636-2660.
580 Chen, S., Martin, M.P., Saby, N.P., Walter, C., Angers, D.A., Arrouays, D.,
581 2018. Fine resolution map of top-and subsoil carbon sequestration
582 potential in France. Sci. Total Environ. 630, 389-400.
583 Chenu, C., Angers, D.A., Barré, P., Derrien, D., Arrouays, D., Balesdent, J.,
584 2018. Increasing organic stocks in agricultural soils: Knowledge gaps and
585 potential innovations. Soil Till. Res.
586 https://doi.org/10.1016/j.still.2018.04.011.
587 De Gruijter, J.J., Minasny, B., McBratney, A.B., 2015. Optimizing stratification
588 and allocation for design-based estimation of spatial means using
589 predictions with error. J. Surv. Stat. Meth. 3(1), 19-42.
590 Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1 km spatial resolution
591 climate surfaces for global land areas. Int. J. Climatol. 37(12), 4302-4315.

592 Hargreaves, G.H., Allen, R.G., 2003. History and evaluation of Hargreaves
593 evapotranspiration equation. J. Irrig. Drain. E. 129, 53–63.
594 Hargreaves, G.L., Hargreaves, G.H., Riley, J.P., 1985. Irrigation water
595 requirements for Senegal River Basin. J. Irrig. Drain. E. 111, 265–275.
596 Hassink, J., 1997. The capacity of soils to preserve organic C and N by their
597 association with clay and silt particles. Plant Soil, 191(1), 77-87.
598 Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D.,
599 2016. Evaluating large-extent spatial modelling approaches: a case study
600 for soil depth for France. Geoderma Regional, 7, 137-152.
601 Lal, R., Negassa, W., Lorenz, K., 2015. Carbon sequestration in soil. Curr.
602 Opin. Env. Sust. 15, 79–86.
603 Lal, R., 2004. Soil carbon sequestration impacts on global climate change and
604 food security. Science, 304, 1623-1627.
605 Lal, R., 2016. Beyond COP 21: potential and challenges of the “4 per
606 Thousand” initiative. J. Soil Water Conserv. 71(1), 20A-25A.
607 Lilly, A., Baggaley, N.J., 2013. The potential for Scottish cultivated topsoils to
608 lose or gain soil organic carbon. Soil Use Manag. 29, 39-47.
609 Lugato, E., Bampa, F., Panagos, P., Montanarella, L., Jones, A., 2014.
610 Potential carbon sequestration of European arable soils estimated by
611 modelling a comprehensive set of management practices. Glob. Chang.
612 Biol. 20(11), 3557-3567.
613 Martin, M.P., Lo Seen, D., Boulonne, L., Jolivet, C., Nair, K.M., Bourgeon, G.,

614 Arrouays, D., 2009. Optimizing pedotransfer functions for estimating soil
615 bulk density using boosted regression trees. Soil Sci. Soc. Am.J. 73(2),
616 485-493.
617 Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P.A.,
618 Paroissien, J.B., … Arrouays, D., 2014. Evaluation of modelling
619 approaches for predicting the spatial distribution of soil organic carbon
620 stocks at the national scale. Geoderma, 223, 97–107.
621 Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D.,
622 Chambers, A., … Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma,
623 292, 59-86.
624 Mouselimis, L. 2016. Clustering using the ClusterR package.
625 http://mlampros.github.io/2016/09/12/clusterR_package/
626 Mulder, V.L., Lacoste, M., Martin, M.P., Richer□de□Forges, A. and Arrouays,
627 D., 2015. Understanding large□extent controls of soil organic carbon
628 storage in relation to soil depth and soil□landscape systems. Glob.
629 Biogeochem. Cycles 29, 1210–1229.
630 Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C. and Arrouays, D., 2016.
631 GlobalSoilMap France: High-resolution spatial modelling the soils of
632 France up to two meter depth. Sci. Total Environ. 573, 1352-1369.
633 NASA LP DAAC, 2017. MOD17A2H: MODIS/TERRA Gross Primary
634 Production. Version 6. NASA EOSDIS Land Processes DAAC, USGS
635 Earth Resources Observation and Science (EROS) Center, Sioux Falls,

636 South Dakota (https://lpdaac.usgs.gov), accessed October 30, 2017, at
637 http://dx.doi.org/ 10.5067/MODIS/MOD17A2H.006.
638 Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G.P., Smith, P.,
639 2016. Climate-smart soils. Nature, 532, 49.
640 Post, W.M., Kwon, K.C., 2000. Soil carbon sequestration and land□use
641 change: processes and potential. Glob. Chang. Biol. 6(3), 317-327.
642 R Core Team, 2016. R: A language and environment for statistical computing.
643 R Foundation for Statistical Computing, Vienna, Austria. URL
644 https://www.R-project.org/.
645 Saby, N.P.A., Arrouays, D., Antoni, V., Foucaud-lemercier, B., Follain, S.,
646 Walter, C., Schvartz, C., 2008. Changes in soil organic carbon content in
647 a French mountainous region, 1990-2004. Soil Use Manag. 24, 254-262.
648 Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising,
649 J., … Zhang, G.L., 2009. Digital soil map of the world. Science, 325, 680–
650 681.
651 Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms
652 of soil organic matter: implications for C-saturation of soils. Plant Soil,
653 241(2), 155-176.
654 Sparling, G., Parfitt, R.L., Hewitt, A.E., Schipper, L.A., 2003. Three
655 approaches to define desired soil organic matter content. J. Environ. Qual.
656 32, 760-766.
657 Stockmann, U., Adams, M., Crawford, J.W., Field, D.J., Henakaarchchi, N.,

Jenkins, M., … Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agr. Ecosyst. Environ. 164, 80-99.

Stolbovoy, V., Montanarella, L., 2008. Application of soil organic carbon status indicators for policy-decision making in the EU, In: Toth, G., Montanarella, L., Rusco, E. (Eds.), Threats to soil quality in Europe, 87–99.

UE-SOeS (2006) Corine Land Cover. Service de l'Observation et des Statistiques (SOeS) du Ministère de l’Environnement. de l’Énergie et de la Mer. Tech. rep.

Figure captions

Figure 1 RMQS sites located in arable soils.

Figure 2 Spatial distribution of monthly climatic decomposition index and net primary production. X and Y coordinates are expressed in Lambert 93 projection.

Figure 3 Spatial distribution of principal components for climatic decomposition index and net primary production.

Figure 4 Relationship between the number of clusters and BIC.

Figure 5 Number of RMQS sites located in each carbon-landscape zone.

Figure 6 Optimal 10 carbon-landscape zones in France.

Figure 7 Boxplots of SOC content in topsoil and subsoil under 10 carbon-landscape zones.

Figure 8 Empirically maximum SOC stocks in topsoil and subsoil under four percentile settings. The four colours are related to four percentiles. For each percentile, bar shows the interval between upper limit and lower limit of 90% CIs. Number of samples is shown in grey.

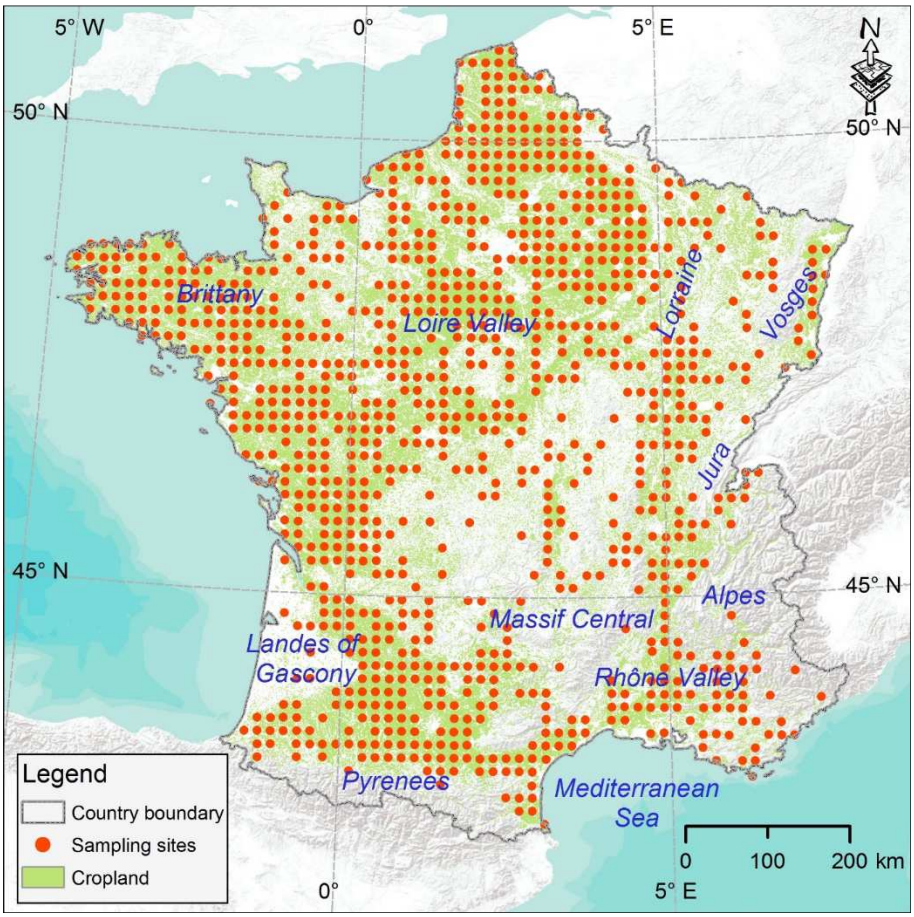
Figure 9 SOC storage potential for arable topsoil under four percentile settings.

Figure 10 SOC storage potential for arable subsoil under four percentile settings.

Table captions

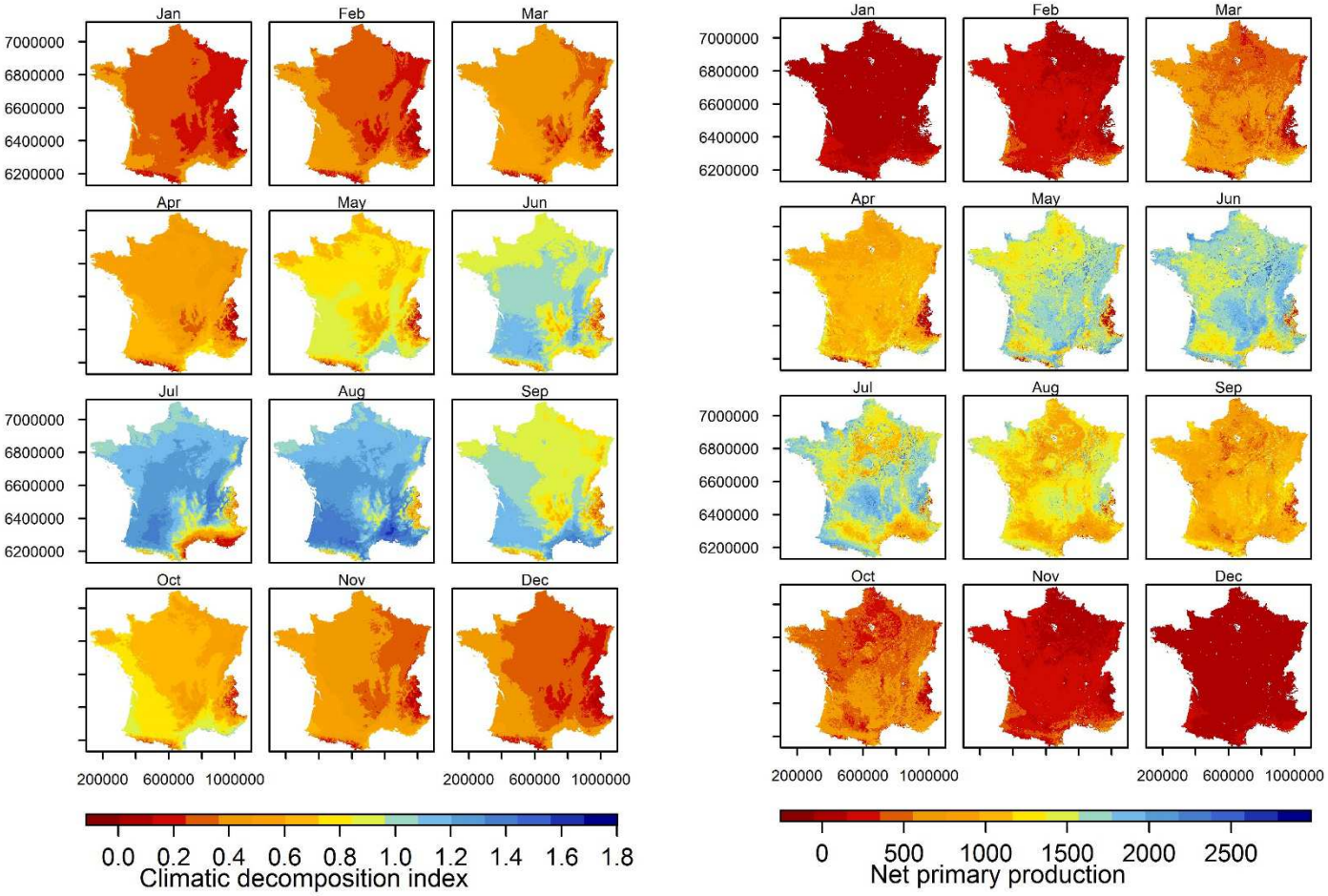
Table 1 National SOC storage potential stocks of French arable soils under different percentile settings. Lower limit and upper limit of 90% CIs are also provided.

693 Figure 1



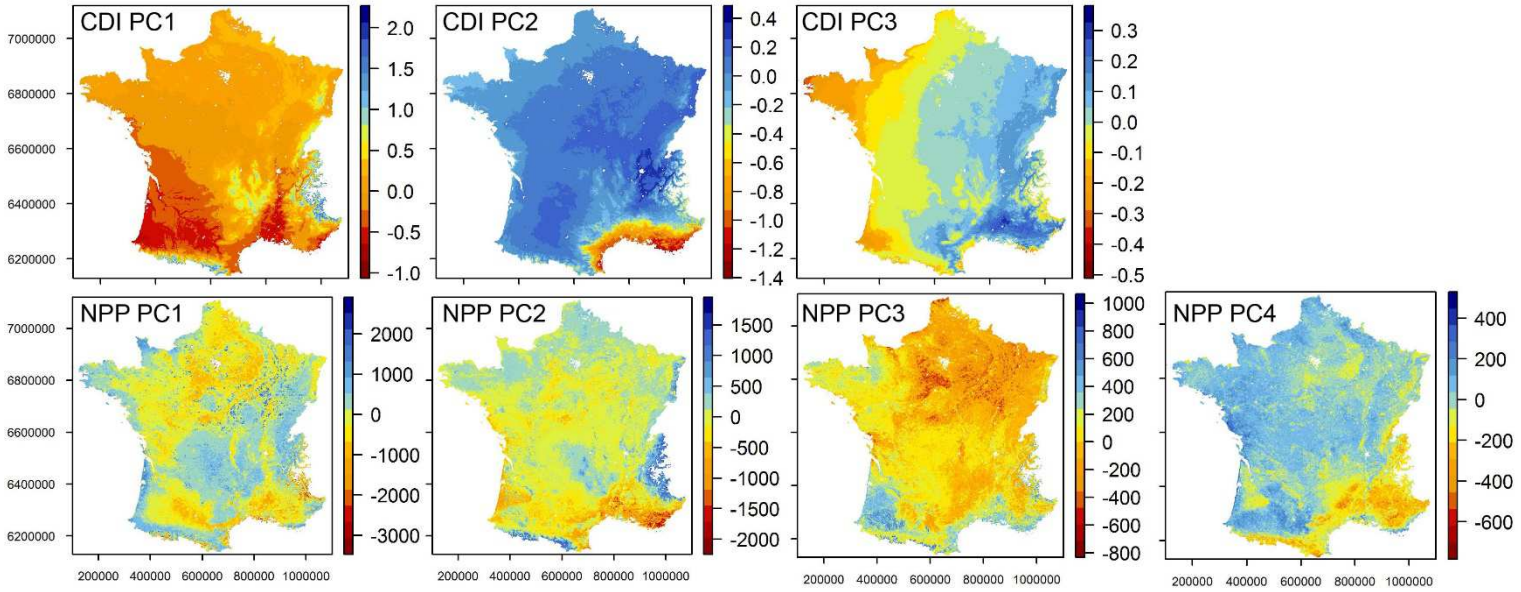
694

695 Figure 2



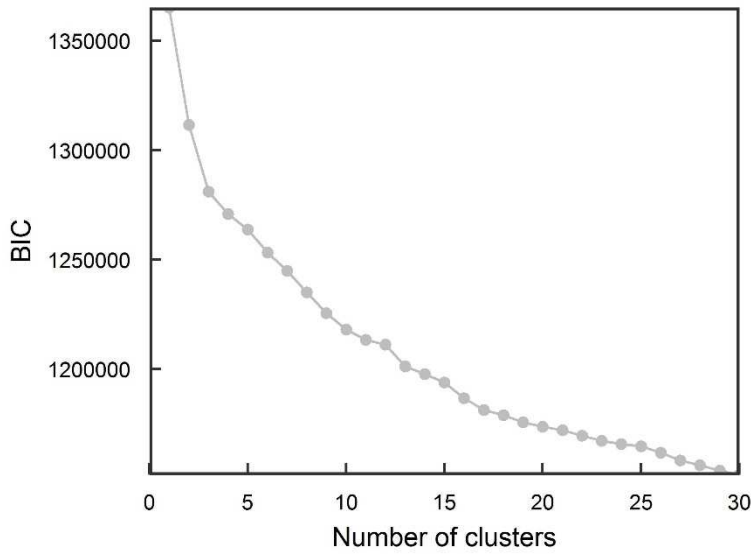
696

697 Figure 3



698

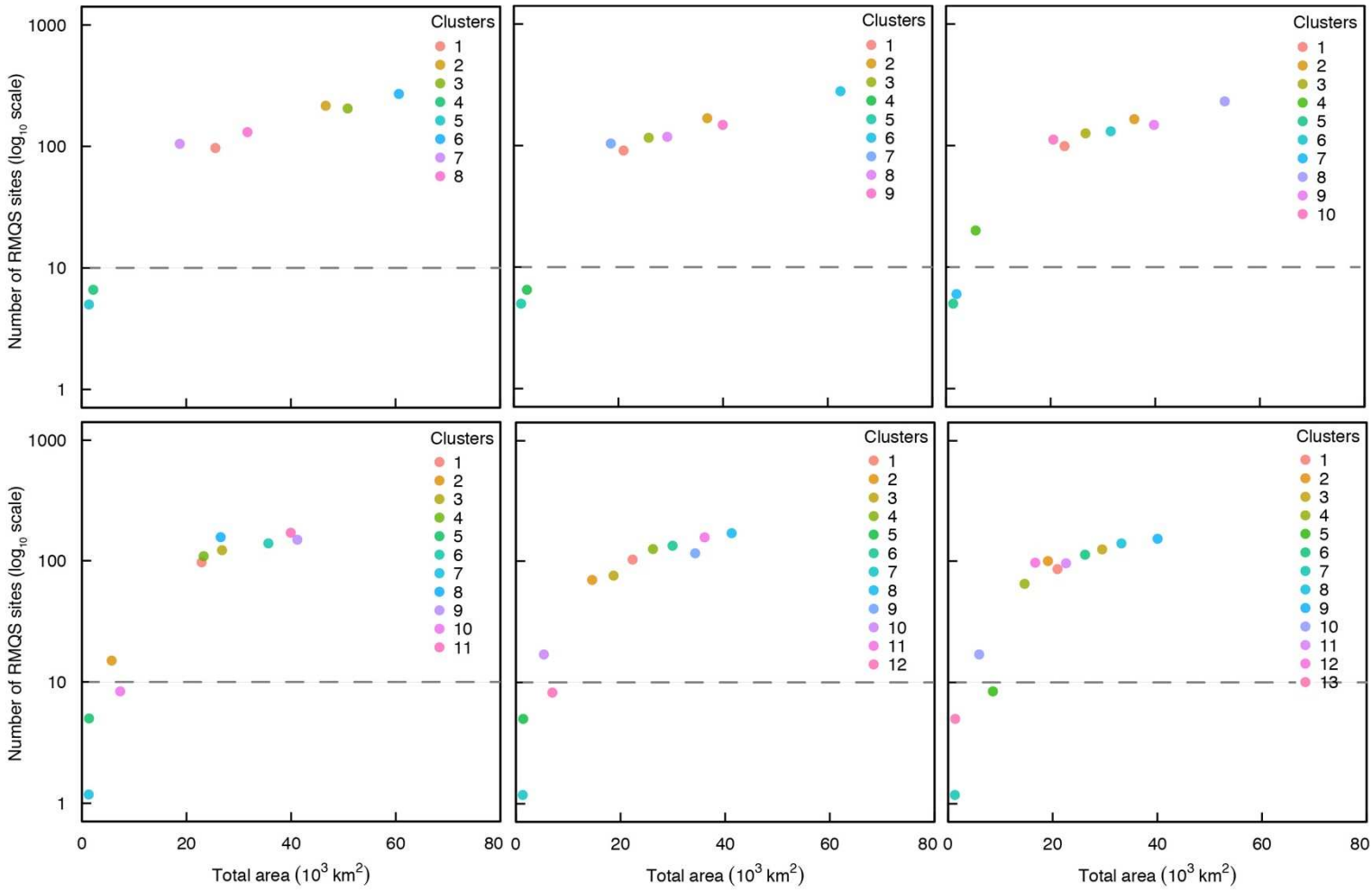
699 Figure 4



700

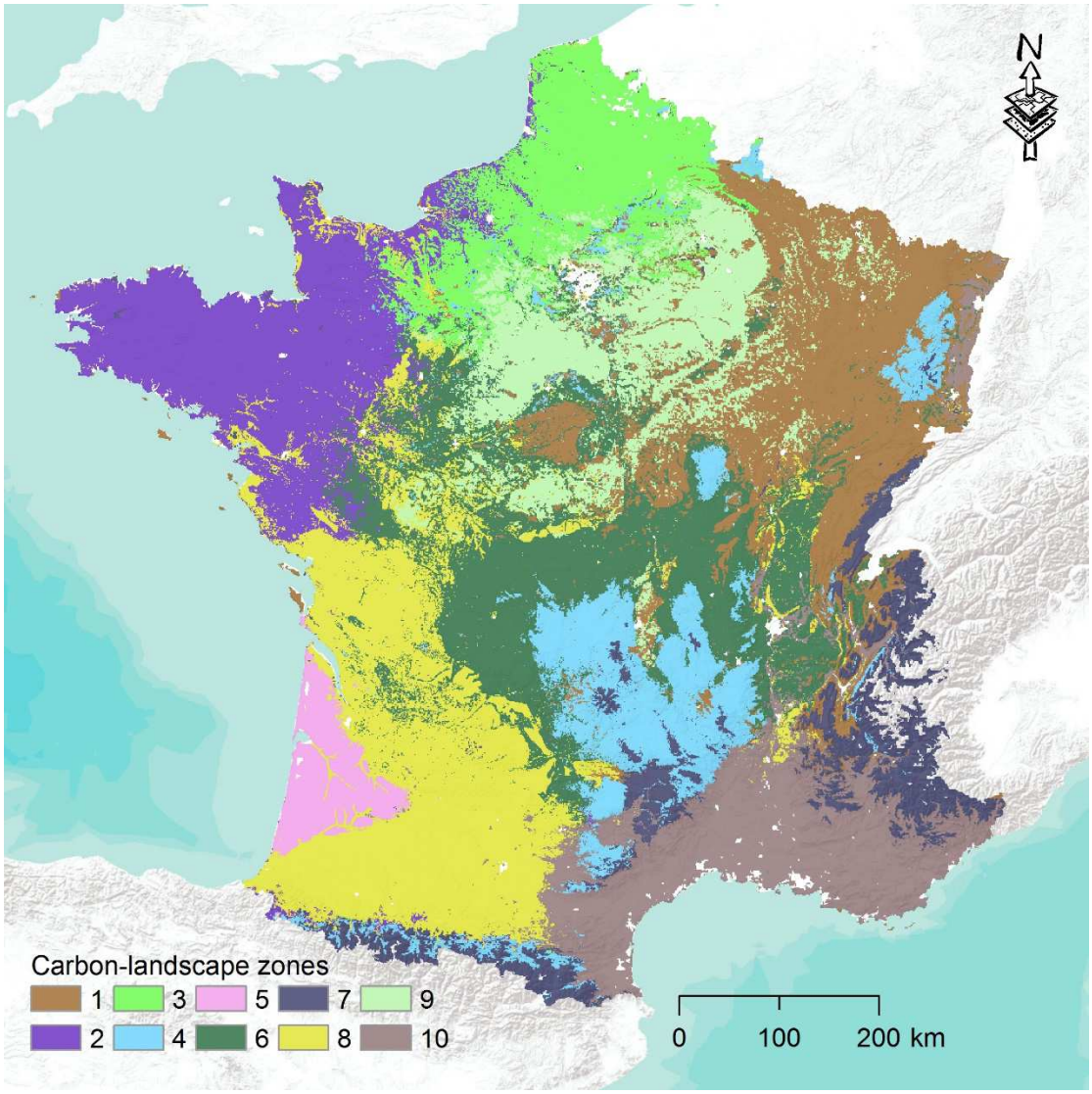
701

Figure 5


Number of RMQS sites (log10 scale)
1000
100
10
1
Clusters
1
2
3
4
5
6
7
8
9
10
11
12
13
0
20
40
60
80
Total area (10^3 km^2)

703

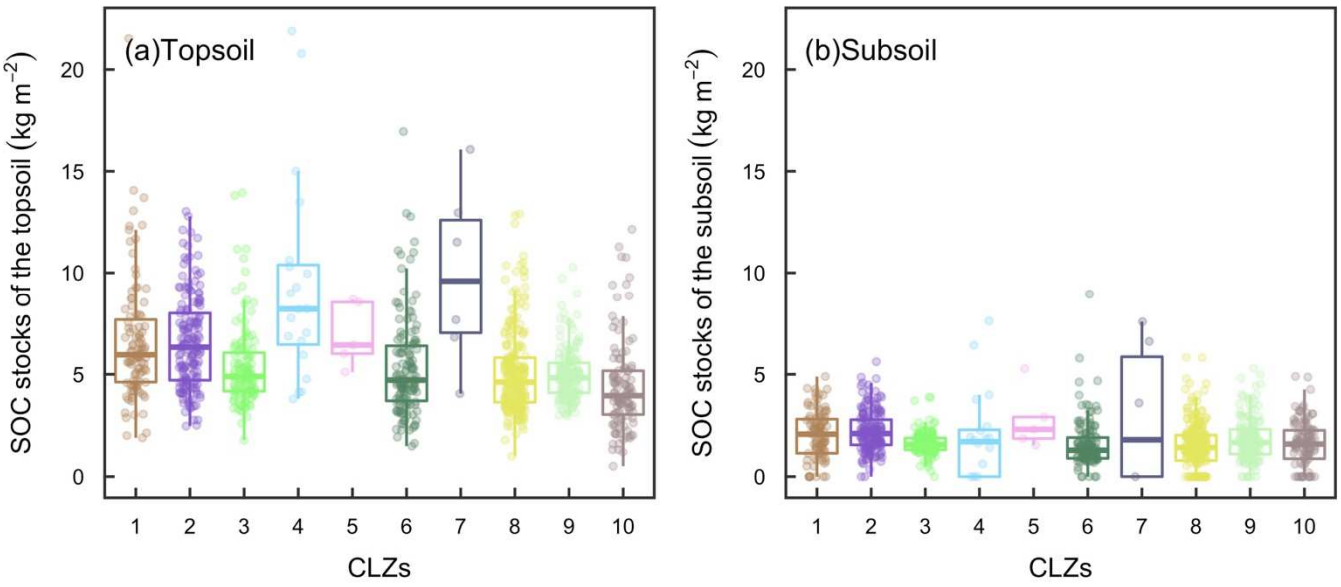
704 Figure 6



705

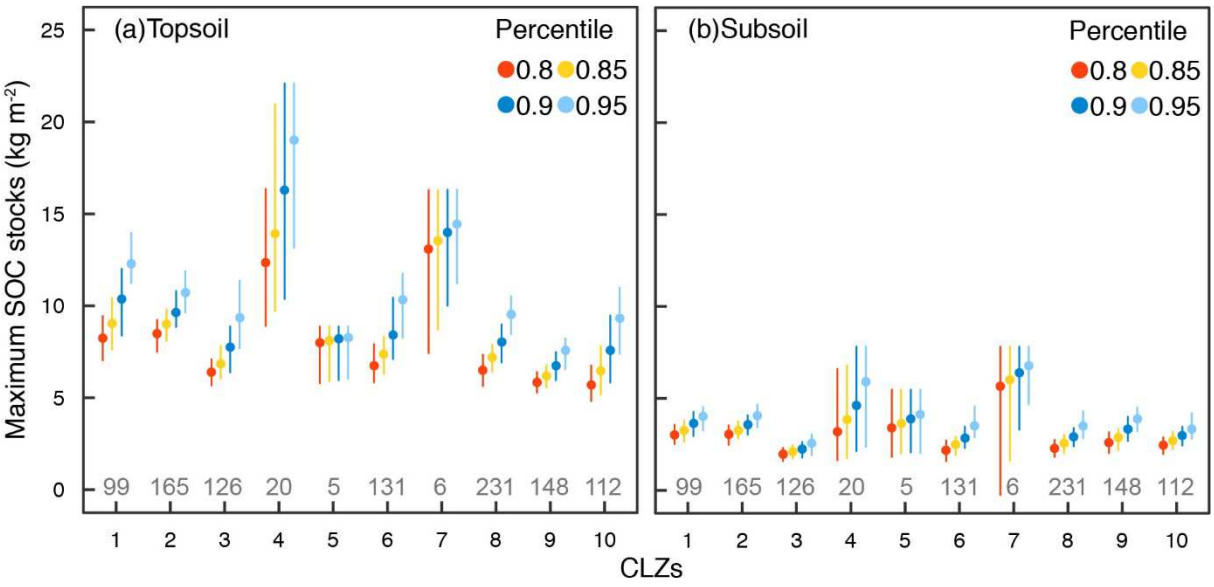
706

707 Figure 7



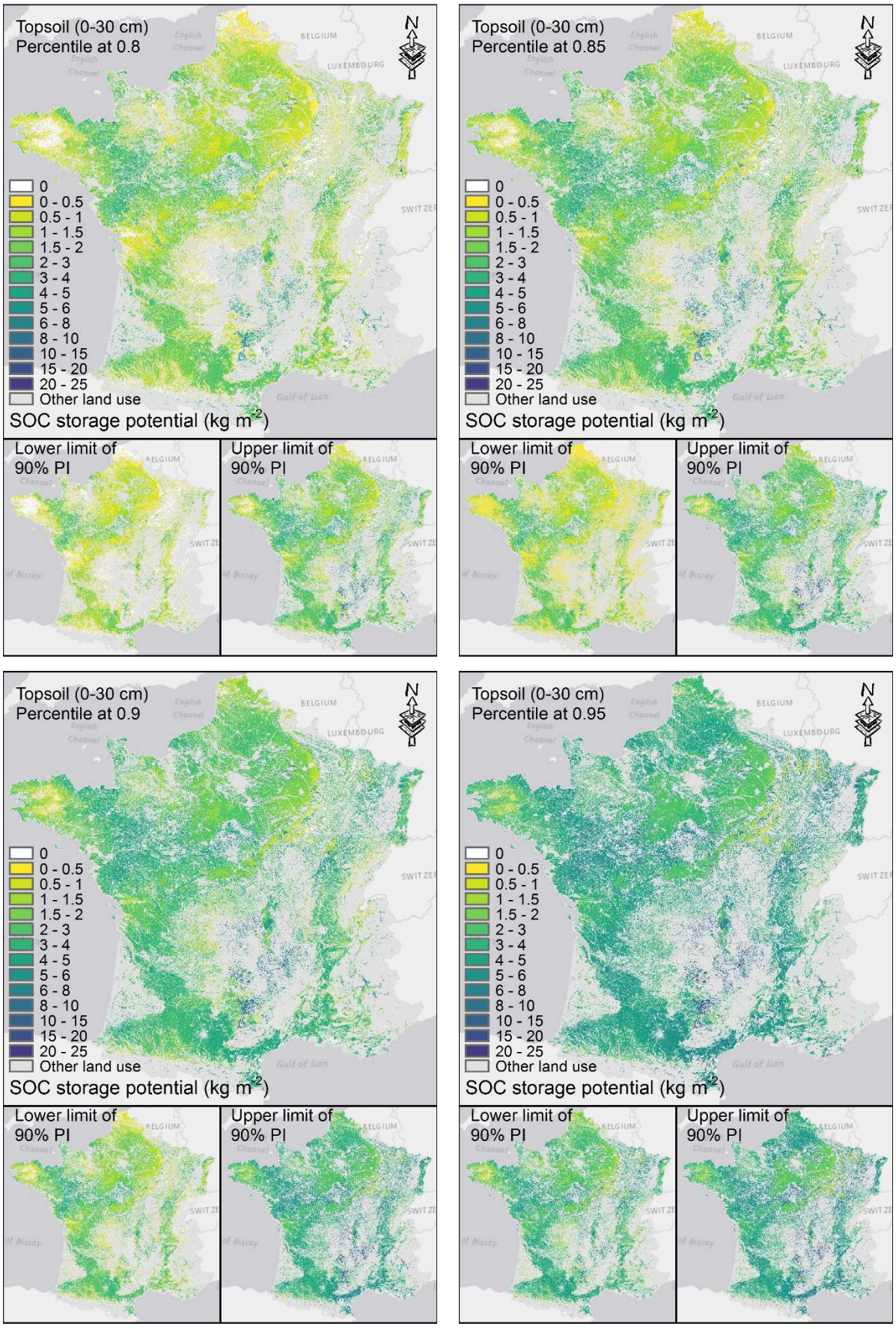
708

709 Figure 8

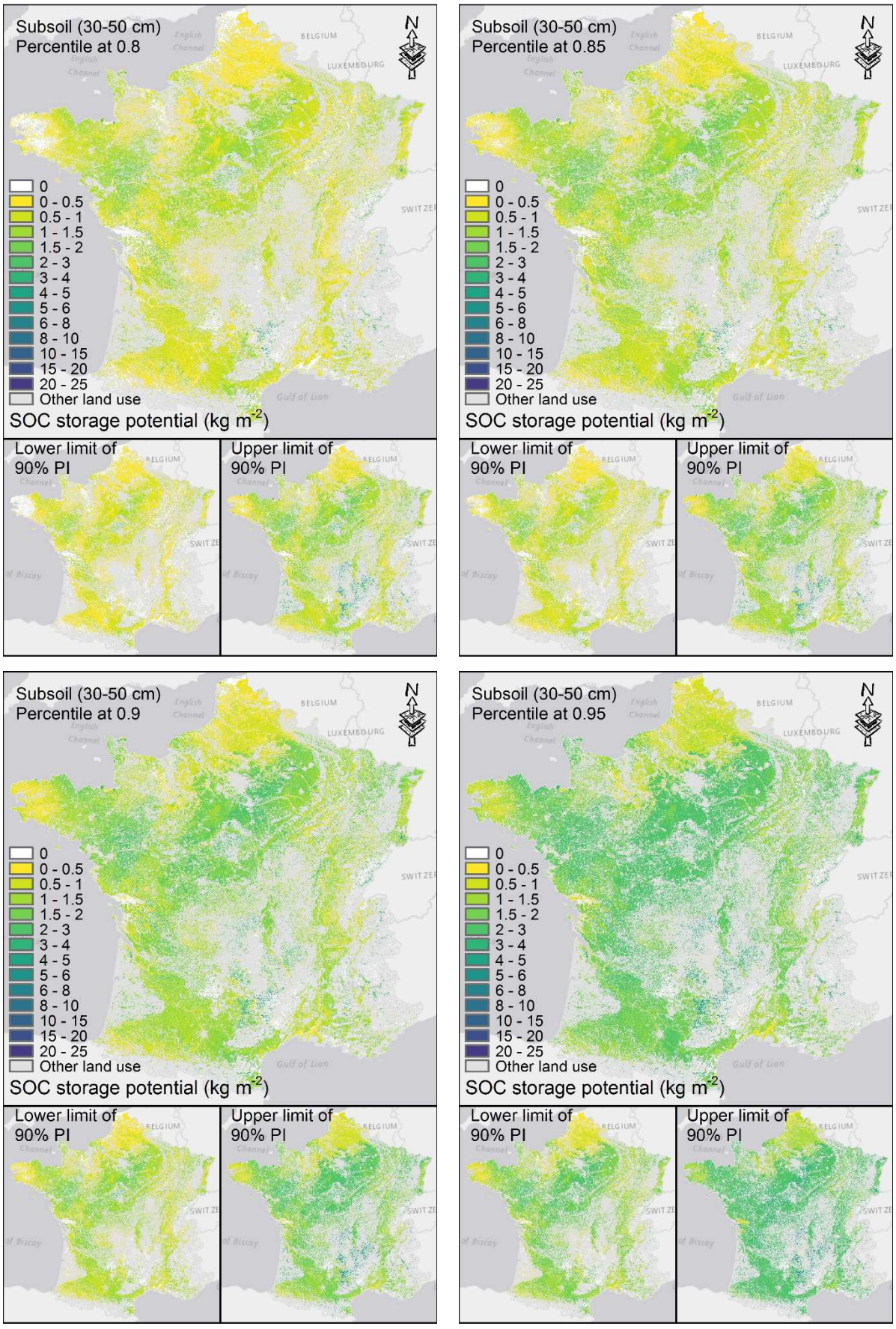


710

711 Figure 9


Topsoil (0-30 cm)
Percentile at 0.8
0
0 - 0.5
0.5 - 1
1 - 1.5
1.5 - 2
2 - 3
3 - 4
4 - 5
5 - 6
6 - 8
8 - 10
10 - 15
15 - 20
20 - 25
Other land use
SOC storage potential (kg m-2)
Lower limit of 90% PI
Upper limit of 90% PI
Topsoil (0-30 cm)
Percentile at 0.85
SOC storage potential (kg m-2)
Lower limit of 90% PI
Upper limit of 90% PI
Topsoil (0-30 cm)
Percentile at 0.9
SOC storage potential (kg m-2)
Lower limit of 90% PI
Upper limit of 90% PI
Topsoil (0-30 cm)
Percentile at 0.95
SOC storage potential (kg m-2)
Lower limit of 90% PI
Upper limit of 90% PI

712 Figure 10



713 Table 1

Soil horizon	Area (km^2)	Total SOC storage potential under four percentile settings (Mt)			
		0.8	0.85	0.9	0.95
Topsoil	239395	336(203,501)	470(308,662)	674(434,950)	1020(740,1283)
Subsoil	228467	165(91,250)	228(150,306)	309(226,404)	433(331,560)

714

715 **Appendix**

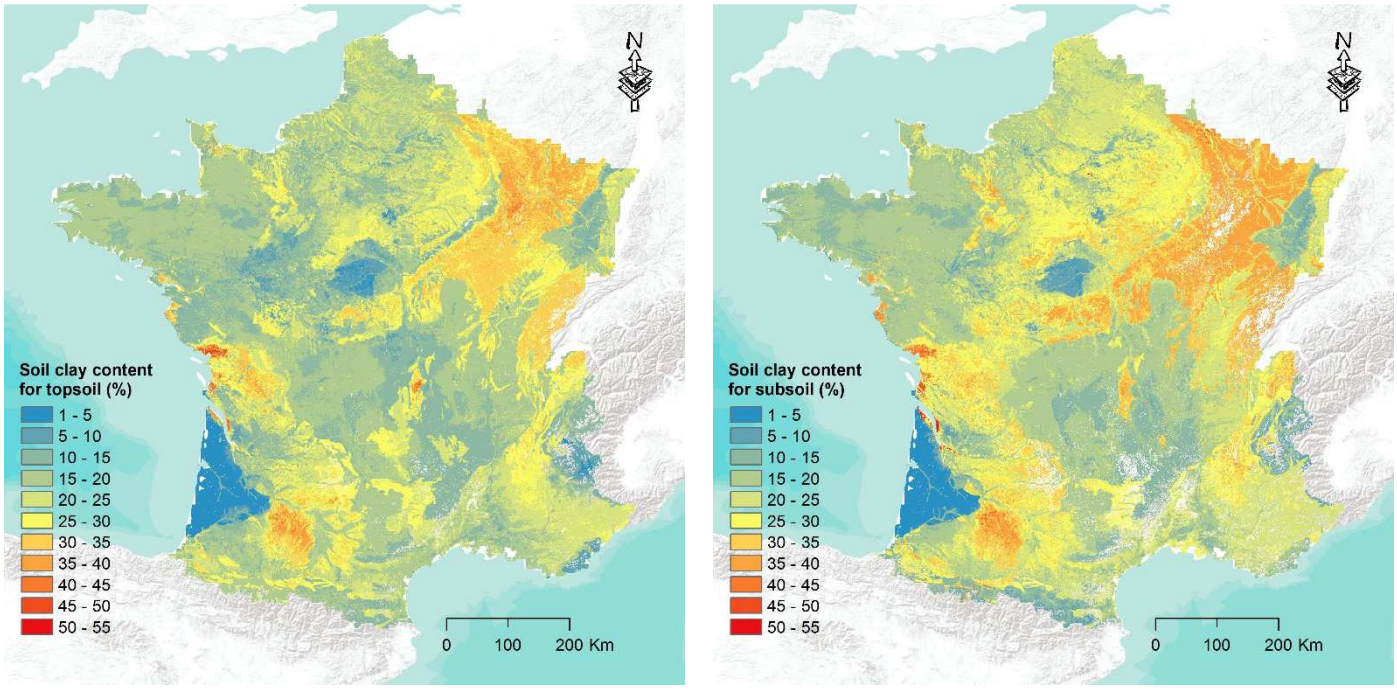


Figure A1 Soil clay content for topsoil and subsoil

716

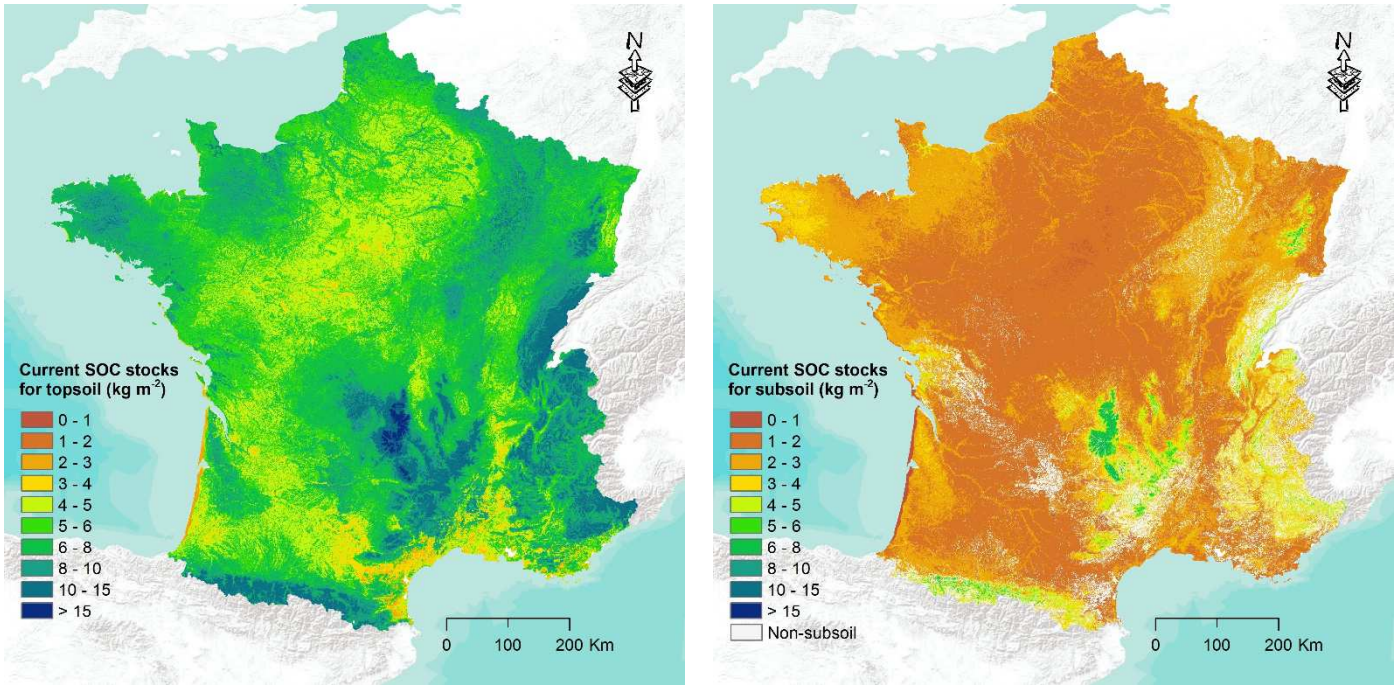


Figure A2 Current SOC stocks for topsoil and subsoil

717

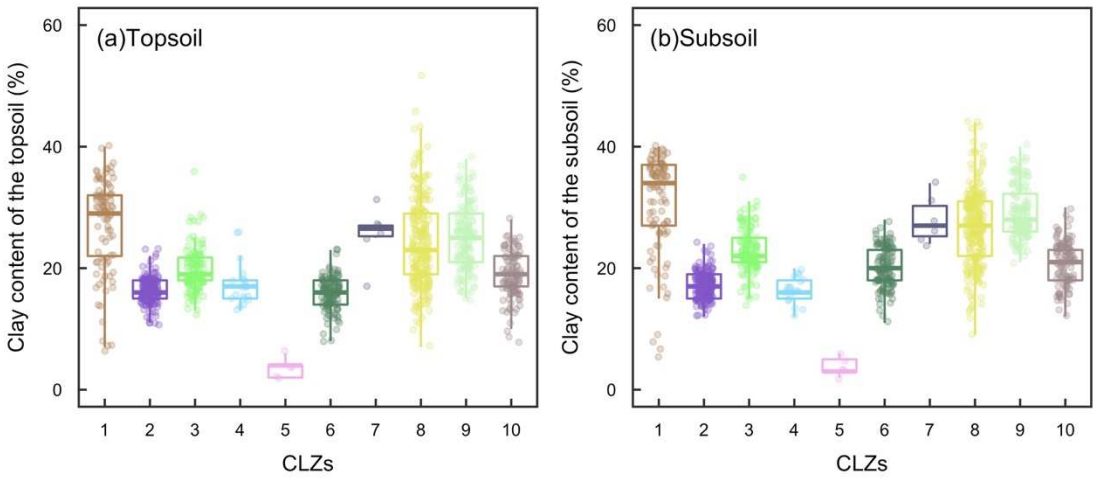
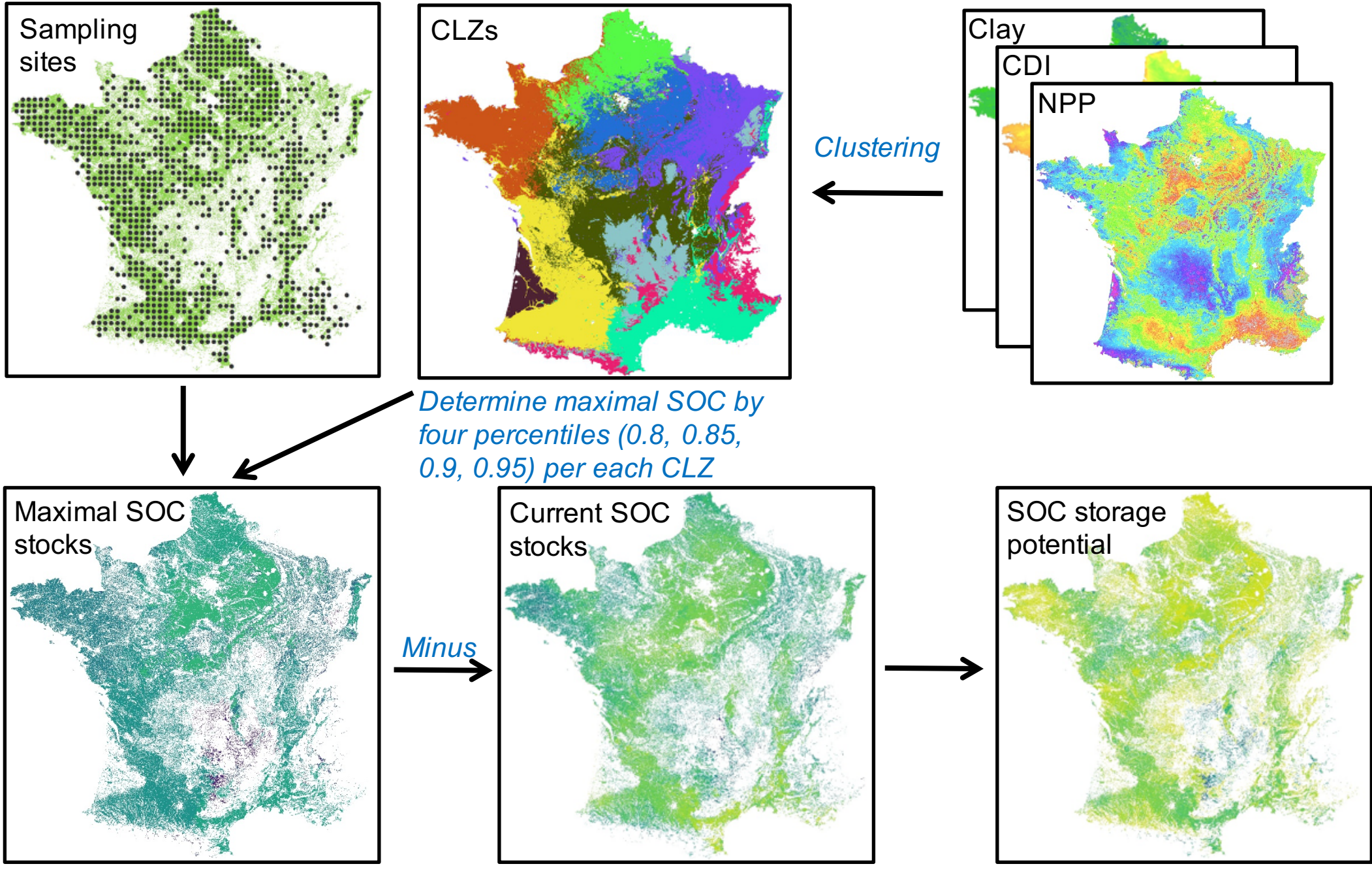


Figure A3 Boxplots of clay content in topsoil and subsoil under the 10 carbon-landscape zones.

718



CLZs, carbon-landscape zones; CDI, climatic decomposition index; NPP, net primary production.