National Soil Resources Institute,

School of Applied Sciences,

Cranfield University,

Bedfordfordshire,

UK

11th April 2013

Dear Sir or Madam,

Thank you for your email informing us that the Open Discussion Period on our manuscript entitled “Modeling Soil Bulk Density at the Landscape Scale and its Contributions to C Stock Uncertainty" has now closed.

We have now considered the reviews and, as we have revised a number of figures and tables within the manuscript, as well as the associated text, our response is attached as a PDF supplement. We will upload a revised manuscript with changes marked, upon request.

We found the suggestions helpful and we believe that the manuscript has now been improved as a consequence.  We hope that our response and associated revisions are acceptable and we look forward to hearing from you in due course.

Yours sincerely,

Khaled Taalab

**Reviewer 1**

General comments

The methods section describes how the data is split into training and validation dataset. However in the results and discussion it is not clear to me what results pertain to the training data and which to the validation data. It even seems to me that all the presented results pertain only to the training dataset. At the current presentation it is hard to get an impression of the best fit and the danger of overfitting.

To clarify, all reported results refer to the validation dataset. Of the 342 topsoil and 339 subsoil $D_b$ samples, 103 and 101 were not used to train the models, but were used instead for independent validation for the topsoil and subsoil respectively (pg 18838, lines 22-24)

Calculating C-stocks with just one unweighted average bulk density over a broad heterogeneous area of course will lead to biased results, and I would reject such a study. One must stratify the bulk density measures or carbon stocks to areas of similar properties (e.g. soilscapes) and then aggregate the results be area weighted averages (explained e.g. with eq 1 and 2 in Wutzler et al 2006). My guess is that the carbon stocks then will be of comparable magnitude.

We agree that an area weighted average by soilscape would provide a significantly improved estimate of carbon stock. As a result of this suggestion, we have recalculated the mean bulk densities on the basis of soil great group (Avery, 1980). We have used soil groups as opposed to soilscapes as a number of soilscapes had no samples from which to calculate average $D_b$. Furthermore, an average by soil group is an established method by which to stratify soil $D_b$ (Grimm et al., 2008; Batjes, 1996).

The text in section 2.3.3 has been amended to include "As a single, unweighted mean across a heterogeneous area would lead to bias results, the mean $D_b$ was calculated for each soil great group (Avery, 1980) and weighted by area. Using a mean $D_b$ value per soil great group is a commonly used approach to representing the spatial variation of $D_b$ across the landscape (Grimm et al., 2008; Batjes, 1996)"

As a result of changes to the how the mean $D_b$ value was calculated, the results table has been amended in the manuscript. The revised table is reproduced below. Note that this replaces Table 3 in the manuscript.

Table 4: Carbon stock for the entire study area and by selected Soilscape

| Location | OC Stock (t $ha^{-1}$) estimated using great group mean $D_b$ (± 95% confidence interval) | OC Stock (t $ha^{-1}$) estimated using gridded $D_b$ (± 95% confidence interval) |
|---|---|---|
| Full study area | 86.41±15.59 | 87.01 ± 8.19 |
| Central England Plateau | 84.72 ± 15.01 | 88.25 ± 8.18 |
| Central upland spine of N England | 86.75 ± 16.98 | 71.84± 8.41 |
| Total Carbon Inventory (Tonnes) | 156834150±28295850 | 157923150 ± 14862371 |

The text relating to Table 4 in section 4.6 has also been amended to read

"To illustrate the potential improvement in OC stock estimation which could be achieved using the gridded surface of $D_b$ compared with using a stratified mean value [Mestdagh et al., 2009; Hanegraaf et al., 2009] we calculated the OC stock at each sample point using three different sets of $D_b$: the measured $D_b$, the RF gridded prediction of $D_b$ and great group mean measured value of $D_b$ calculated using all sample points in the training data. Note that results for C stock calculations using model output were produced using a calibrated RF model that used the training dataset alone, the validation data was used solely to assess model performance. The average OC stocks calculated using each $D_b$ estimate are shown in Table 2, along with the difference between the estimated and measured mean OC value, expressed as a percentage of the mean measured value. The 5$^{th}$ and 95$^{th}$ percentile errors in measured OC stocks are also shown. The gridded surface refers to a map of RF-predicted $D_b$ values (Figure 2b) produced as a raster grid with a cell size of 100 x 100 m across the entire study area. The main advantage of the gridded surface method over PTFs, which can be applied to individual points using measured soil property data for the point in question, is that the gridded method can be applied to the entire study area with the same quantifiable level of both performance and error estimation at all spatial locations. In contrast, the accuracy of predictions made using a PTF is hard to quantify beyond each sampling point.

Using the individual measured point-based $D_b$ values gives an average OC content of 73.01 ± 0.56 t C ha$^{-1}$ compared to an average value of 71.32 ± 0.61 t C ha$^{-1}$ produced using the RF-predicted $D_b$ values and a value of 74.81 ± 0.70 t C ha$^{-1}$ generated using Great Group mean $D_b$ value. Using the OC stock calculated with measured $D_b$ as a yardstick, the gridded estimate of $D_b$ yields a marginally better C stock estimate compared with using a single (mean) $D_b$ value. In this case, the RF predictions will underestimate $D_b$ whereas using a stratified mean value will overestimate. The difference in the error associated with stock prediction using the gridded $D_b$ values compared to using the mean value of $D_b$ is particularly evident when predicting C stock levels in soils at the extremes of the expected range (i.e. the prediction errors for the 5$^{th}$ and 95$^{th}$ percentile OC stock values). The potential improvement in using the gridded estimate of $D_b$ is most evident in the 95$^{th}$ percentile, where using a stratified mean $D_b$ value will yield an error nearly two times larger.

To put the magnitude of the errors illustrated in Table 3 into context, Bellamy et al. [2005] suggest that the average annual rate of change in the OC content for UK topsoil is 0.67g kg$^{-1}$yr$^{-1}$, which equates to approximately 1.79 t C ha$^{-1}$ yr$^{-1}$. As the rate of change is comparable in magnitude to the error associated with prediction, it is clearly important to keep error to a minimum if stock changes are to be quantified accurately. The total soil OC inventory across the whole study area, calculated using both the stratified mean and gridded $D_b$ estimates, is shown in Table 4. There is a slight difference in the OC stock per unit area (0.6 t ha$^{-1}$) which equates to a over one million tonnes of carbon for this study area alone. The most notable difference between the stratified mean and gridded approaches to $D_b$ prediction is the error associated with prediction. The 95% confidence interval associated with the stratified mean model is nearly twice as large as that of the gridded model. When estimating the total C stock within the study area, this translates to a difference of over 13 million t C$^{-1}$.

To further illustrate the potential of this method, carbon stocks were calculated for the landscape as a whole and for two selected individual Soilscapes using both the stratified measured mean and gridded predictions of $D_b$. Soilscapes were selected based on the accuracy of the gridded model's $D_b$ predictions, including the Soilscapes with the most accurate and most inaccurate $D_b$ predictions. Results are shown in Table 4. The two Soilscapes; the Central Upland Spine of Northern England and the Central England Plateau show areas of relatively low and high $D_b$, respectively. These regional differences in stock calculations, particularly in the Central Upland Spine of Northern England, highlight potential errors which can be introduced to a stock calculation by using a mean $D_b$ value, depending on the scale of the study. Moreover, the gridded model has a much greater predictive accuracy, with confidence bounds nearly two times smaller compared with the stratified mean model. The mean model produced similar stock predictions for both the entire study area and the selected Soilscapes. This is a problem as, at the Soilscape scale, the stratified mean model may either under or overestimate carbon stocks. This issue does not affect the gridded model, because it is able to apply rules learned across the entire study region, to identify areas of high and low bulk density, a key advantage when working at this scale. A scale at which errors in $D_b$ estimation have shown to be highly significant to carbon stock inventory [Goidts et al., 2009]. Estimating C stocks and changes, especially at finer spatial scales requires the use of refined estimates of Db, which can be obtained using the types of landscape-scale models described in this paper. It is at these scales that many spatially distributed land-atmosphere interaction models such as JULES operate [Harrison et al., 2008]."

Eq. 5 describes error propagation for a single measurement point or a single pixel. How were the errors propagated to the estimates of soilscapes or total area? One cannot assume independence of the single pixels as they are predictions of the same model with uncertain parameters. One approach how to deal with this with an ordinary mixed nonlinear regression model in a different conext see e.g. appendix A2 of Wutzler et al. 2008). How was covariance between OC and D in eq. 5 derived?

In accordance with this suggestion, the covariance of the mean great group $D_b$ value and OC was determined using a mixed-effects model. In the gridded model, covariance was determined using the predicted $D_b$ values and the measured OC values. The text in section

2.3.3 of the manuscript had been amended to include “In the gridded model, covariance was determined using the predicted $D_b$ values and the measured OC values. In the stratified model, the covariance between the mean great group $D_b$ and OC was determined using a mixed-effects model (Wutzler et al., 2008)”.

**Specific comments**

P18836 L12: Why two numbers of landscape subdivisions? Which one was used both? Table 1: What are Greatgroup, AT 0 annual , F CD med ? A table with a short description of all predictors would be helpful (maybe include two columns for their ranking in the RF-A and RF-S models).

We tested two landscape classification algorithms because they base their classification on different attributes. No model included both subdivisions. They were tested separately and the best performing classification was included within the predictive model. Great group refers to the soil group, AT0 Annual is the accumulated temperature above 0 °C and FCD med is the median number of field capacity days. We have produced a table detailing the predictors (below).

Table 1: Predictor variables used in the ANN and RF model. The variables are lised in order of importance for the RF model predicting A horizon $D_b$.

| Name | Description | Number of classes/ Range |
|---|---|---|
| Land Use | Land use derived from the 1 km x 1 km Land Cover Map 2000 produced by the Centre for Ecology and Hydrology (CEH) (Fuller et al., 2002) | 8 |
| Soil Association | Soils grouped to the association level (Avery, 1973) derived from a 1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996). | 24 |
| Elevation | Elevation above sea-level derived from a 10m DEM (Childs, 2004) | -2 - 558.9 m |
| Great group | 1:250,000 scale National Soil map of England and Wales (NATMAP; Hallett et al., 1996) classified into soil Great Groups (Avery, 1980) | 5 |
| AT0_Annual | Average accumulated temperature above 0°C derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005) | 2564 - 3871 °C |
| Parent Material | Soil parent material derived from a 1:250,000 scale Soil map of England and Wales (NATMAP; Hallett et al., 1996) | 18 |
| PSMD | Potential soil moisture deficit related to the balance between rainfall and potential evapotranspiration (Jones and | 50 - 261 mm |

| | | |
|---|---|---|
| | Thomasson, 1985) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005) | |
| PT | Potential evapotranspiration is the amount of evaporation which would occur if water was not limited (Hess, 2000) derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005) | 480 – 708 mm $y^{-1}$ |
| AAR | Average annual rainfall derived from average monthly reports from the UK Meteorological Office on a 5km x 5km grid (Perry & Hollis, 2005) | 548 – 1347 mm $y^{-1}$ |
| RCS | Bedrock geology derived from 1:50,000 scale British Geological Survey rock classification scheme map, detailing bedrock lithology | 27 |
| FCD_MED | Median number field capacity days derived from average monthly reports from the UK Meteorological Office on a 5 km x 5 km grid (Perry & Hollis, 2005) | 107-290 days |
| Curvature | Surface curvature derived from a 10m DEM (Childs, 2004) | -74.8 – 66.4 |
| Iwahashi | Iwahashi landform classification uses a terrain classification algorithm based on slope, surface texture and local convexity (Iwahashi & Pike, 2007) derived from a 10m DEM | 8 |
| Pennock | Pennock landform classification uses a terrain classification algorithm based on slope, curvature and catchment size (Pennock et al., 1987) derived from a 10m DEM | 7 |
| STI | Sediment transport index derived from a 10m DEM | -67.4 - 0 |
| Slope | Slope derived from a 10m DEM (Childs, 2004) | 0 – 74.9 |
| SWI | Saga Wetness Index, a terrain-derived index of soil moisture derived from a 10m DEM (Böhner et al., 2001) | 9.8 – 19.7 |
| Aspect | Aspect derived from a 10m DEM (Childs, 2004) | -1 - 360 |

P 18838 L4ff. It did not become clear who derived the soilscapes I miss discussion of covariance between predictions. Especially for the soil classifications of high predictor importance, please discuss their derivation. It is only based on texture or are there topographic arguments, vegetation etc, which factors do they covary with?

The soilscapes were derived by the National Soil Resources Institute of Cranfield University. Soil scientists Dick Thompson and John Hollis were two of the leading figures in the creation of the Soilscapes dataset. The Soilscapes themselves were based on soil drainage characteristics, land use and geology, hence will co-vary with the soil textural properties, land use, bedrock geology and parent material data layers. (Farewell et al., 2011)

I also miss a discussion about good models by chance. How many predictors/models did you compare? If you compare enough models, several will be good despite they are actually not related to the observed patterns.

To limit the potential of good models by chance, the models have been validated using independent data meaning that it is extremely unlikely that models reported in this study produce good results due to the fact that they happen to have fit the data particularly well. There is always a possibility that explanatory variables will be included even when they have no causal relationship with the predicted variable, however, this is why we have reported the most influential variable for each model.

Fig 3b and 3c: In order to appreciate the information contained in the mean residual, can you plot a distribution of residuals?

We have created a plot for the residuals of figure 3b and 3c (see Figure below). However, we do not feel it that this adds much value to the paper. If the editor thinks that they should be included we are happy to redraft the paper accordingly.

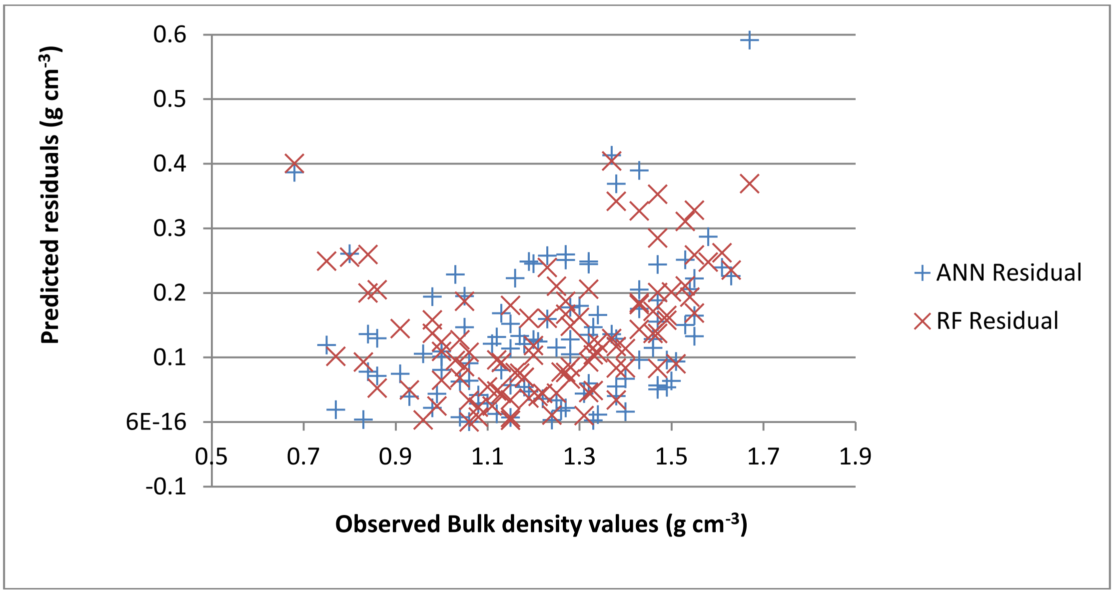



Table 2: I would prefer quantiles in absolute values instead of percentages.

Percentiles have now been included in the revised ms (See Table 3 below). Note that in the manuscript this replaces Table 2.

Table 3: Point estimates of OC stock. Average stock was calculated using Equation 6. Of the prediction methods, ‘Measured’ uses measured $D_b$ values, ‘Gridded’ uses the gridded predicted $D_b$ values and ‘Mean’ uses the measured mean $D_b$ per soil great group.

| Prediction method | Average OC stock (tC $ha^{-1}$) (± 95% confidence interval) | Error from average OC stock (tC $ha^{-1}$) (% in brackets) | $5^{th}$ percentile error (tC $ha^{-1}$) (% in brackets) | $95^{th}$ percentile error (tC $ha^{-1}$) (% in brackets) |
|---|---|---|---|---|
| Measured | 73.01±0.56 | NA | NA | NA |
| Gridded | 71.32±0.61 | 1.69 (-2.31%) | 5.71 (-15.43%) | 10.79 (8.37%) |
| Great Group mean | 74.81±0.70 | 1.80 (2.47%) | 6.34 (-17.14%) | 19.31 (14.99%) |

Is it possible to compare results to studies that make use of neighbourhood of sampling points, e.g. Kriging? Can this somehow be combined with the presented machine learning approaches?

We believe that this would be a very interesting approach to creating a spatial prediction of soil bulk density and we would definitely consider using this technique in future work. However, we did not pursue this approach as we believe that across this particular study area, the assumption of stationarity would not hold true.

**Technical comments**

18849 L11ff: The locations of table 3 have not been introduced yet and I got confused. Maybe move this section further down.

Fig 4 Panels b and c are redundant to Table 3, Shapes of Fig 4a might be included in Fig 2. Hence, Fig 4 is not necessary.

We have amended Figure 2 as directed (see below) and removed Figure 4.

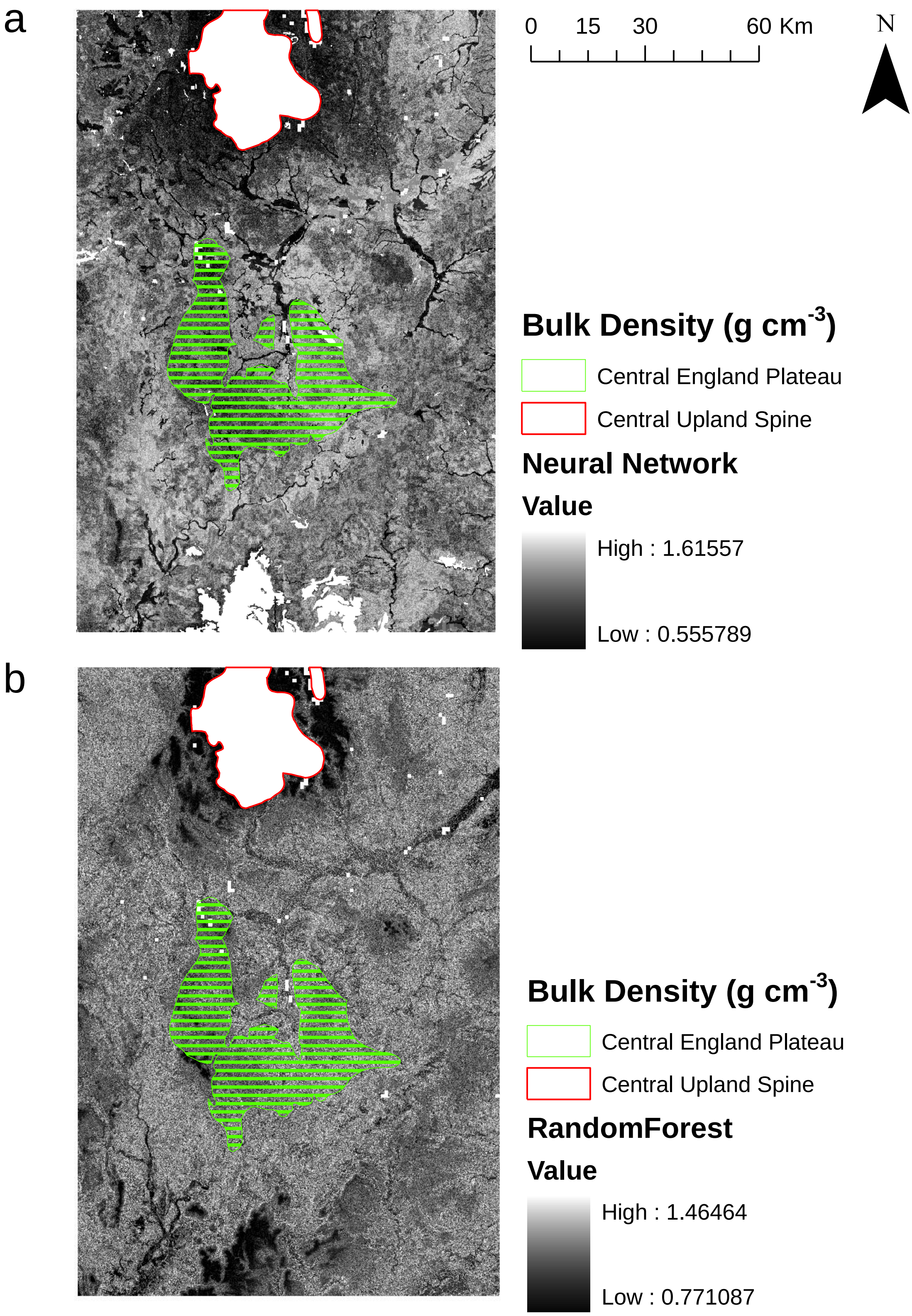
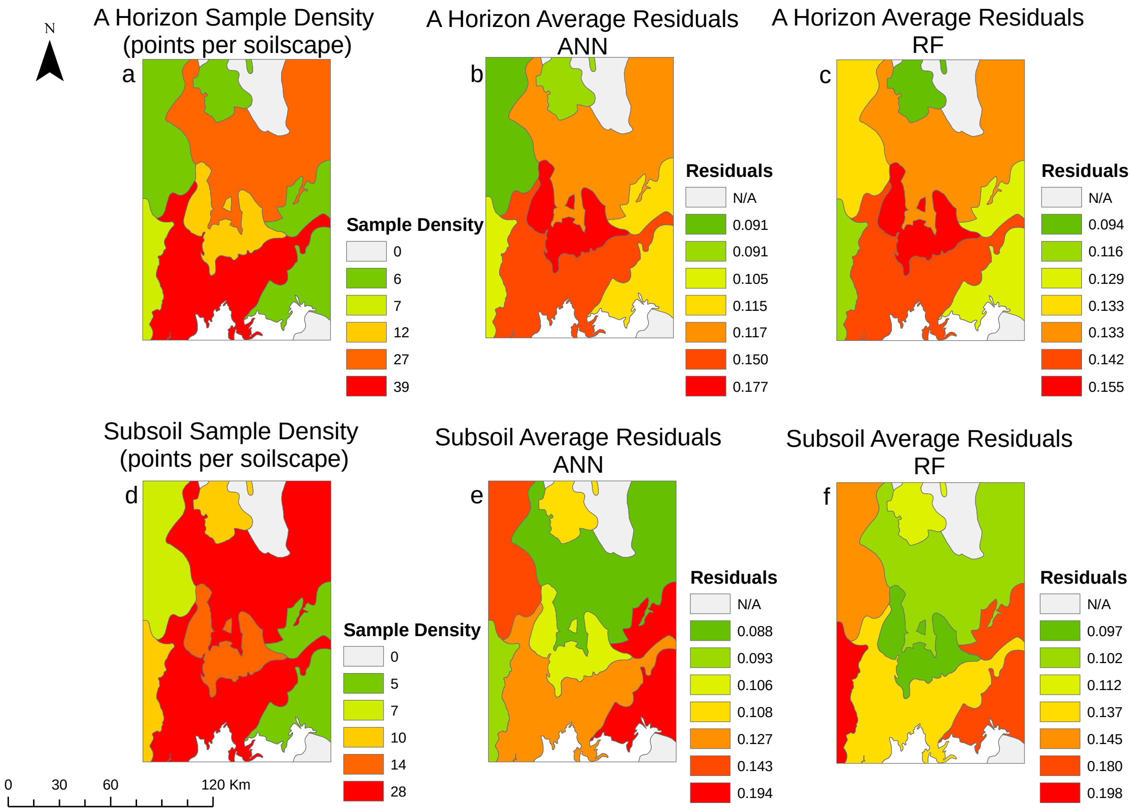



Fig 3: Adding labels directly to the fig instead of the caption would be helpful (A horizon, subsoil, RF model, ANN model)

We have amended Fig. 3 as suggested by the referee (see below)




**Reviewer 2**

**Specific comments**

P18836. 2.1 Data may be renamed to Data and predictor variable

We have amended the title of 2.1 to ‘Data and Predictor Variables’

P18839. 2.3 Statistical methods should be renamed to Modeling methods in compatible with your title; In addition, it is necessary to indicate this empirical modeling approach relies on two different statistical methods.

As suggested, the title of 2.3 has been modified to ‘Modeling Methods’.

The abstract has been amended to read ‘The models were constructed using two distinct statistical methods: random forests and artificial neural networks’.

P18842. 2.3.3 Line 9. Add d

This has now been amended to read ‘d is depth of topsoil’

Line 19 delete from OC*Db

$$Variance(OC\ Stock) = (OC\ Stock)^2 \cdot \left(\frac{(\sigma OC)^2}{(OC)^2} + \frac{(\sigma D_b)^2}{(D_b)^2} + 2\frac{covOC - D_b}{OC \cdot D_b}\right)$$

The equation and the text have now been amended to read $covOC - D_b$

P18845 Line 5. Recorded to recoded

To clarify, the land use was recorded at the time of sampling and recoded at a later date. The text has been modified accordingly.

Line 6 delete use of Table 1. Some of these variables are not clearly indicated in either method section or the Table. In addition, how do you rank these variables? What is your criterion for determining their importance? If this was derived from a scoring system please give these quantities explicitly so that one may compare the importance of certain variable as derived in different methods, such as land use.

The text in Section 2.1.1 has been amended to include 'The mapped soil data were taken from the 1:250000 scale national soil map (NATMAP) of England and Wales (Soil Survey Staff, 1983; Hallet et al., 1996), where soils are displayed in a hierarchal classification scheme (Avery, 1980)'.

The text in Section 2.1.4 has been amended to read 'The parent material was derived from the 1:250000 scale National Soil Map (NATMAP) of England and Wales (Soil Survey Staff, 1983)'.

Discussion about how each model ranks variables is included on on p. 18840, lines 10-14 for Random Forest and p18841, lines 18-22 for ANNs. Each model ranks variables internally, using different metrics and the scoring systems are on different scales. We feel that including these numbers explicitly would confuse the results as the ranking uses the validation data and hence does not necessarily reflect the predictive capabilities of the models, which are demonstrated using independent data.

Table 2 and table 3 should be combined together and the error type should be explicitly indicated (standard error or standard deviation). What is Eq. (6)?

We feel that keeping Tables 2 and 3 separate maintains a clear distinction between results relating to the sample point data alone and results regarding the whole study area.

The error type is the 95% confidence interval. The tables have been amended to show this more clearly.

Figure 3. Why do the panels b and e look identical to each other?

This is an error, the amended figure is shown above.

## Additional References

Avery, B. W.: Soil Classification for England and Wales (Higher Categories), Soil Survey Technical Monograph, 14, Harpenden, pp. 67, 1980.

Childs, C.: Interpolating surfaces in ArcGIS spatial analyst, ArcUser, July-September, 32-35, 2004.

Hallett, S. H., Jones, R. J. A., & Keay, C. A.: Environmental information systems developments for planning sustainable land use. International Journal of Geographical Information Science, 10, 47-64, 1996.

Batjes, N. H.: Total carbon and nitrogen in the soils of the world, Eur. J. Soil Sci., 47, 151–163, 1996.

Farewell, T.S., Truckell, I.G., Keay, C.A. Hallett, S.H.: Use and applications of the Soilscapes datasets, Cranfield University, 2011.

Grimm, R., Behrens, T., Maerker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island – Digital soil mapping using Random Forests analysis, Geoderma, 146, 102–113, 2008.

Soil Survey Staff.: The National Soil Map of England and Wales, I :250000 Scale (In Six Sheets) (Southampton: Ordnance Survey), 1983.