esa ECOSPHERE

The depth distribution of organic carbon in the soils of eastern Australia

Eleanor U. Hobley[1,†,3] and Brian Wilson[1,2]

[1]Ecosystems Management, University of New England, Armidale, New South Wales 2351, Australia
[2]New South Wales Office of Environment and Heritage, Armidale, New South Wales 2351, Australia

Citation: Hobley, E. U., and B. Wilson. 2016. The depth distribution of organic carbon in the soils of eastern Australia. *Ecosphere* 7(1):e01214. 10.1002/ecs2.1214

Abstract. Subsurface soil organic carbon (SOC) is a large but still poorly understood component of the global carbon cycle. We investigated the depth distribution of SOC in eastern Australia, testing the hypotheses that SOC content near the surface is linked with water availability, whereas the distribution of SOC with depth is linked with land use, site factors and temperature. To do this, we measured SOC concentration to 1 m at 100 sites across eastern Australia, and fitted three parameter exponential depletion models to the results. Three machine learning algorithms were used to identify predictors important to the model parameters. Multiple regression models were then created based upon the machine learning results using bootstrapped stepwise regressions and the relative importance of the selected variables was assessed using proportional marginal variance decomposition. Surface SOC concentration was influenced predominantly by climate variables, of which seasonal rainfall was by far the most important. At depth, SOC storage was most influenced by site factors (mainly bulk density and soil type), and both land use and climate contributed similar amounts to model explained variance. The depth distribution of SOC was most influenced by land use, which accounted for ~60% of model explained variance, with site and climate factors being approximately equally important. These results support our hypotheses regarding the drivers of SOC depth distribution in eastern Australia and can be used to identify regions with the potential for additional subsurface soil carbon storage.

Key words: climate; conditional inference trees; cropping; datamining; gradient boosting machines; grazing; land use; machine learning; randomForests.

Received: 7 August 2015; accepted 19 August 2015 . Corresponding Editor: D. P. C. Peters.

[3] Present address: Chair of Soil Science, Technische Universität München, Weihenstephan 85354, Germany.
† **E-mail:** nellie.hobley@wzw.tum.de

Introduction

Enhancing organic carbon storage below the soil surface ("sub-soil") is an attractive option because in most soils, organic matter content decreases with increasing depth from the surface (e.g., Gaudinski et al. 2000, Wynn et al. 2004). The C stabilization capacity of the sub-soil (Six et al. 2002) is therefore less likely to be exhausted and subsurface soils will therefore have a greater capacity for additional SOC storage. SOC age frequently increases with increasing depth and is typically more stable than SOC near the surface (e.g., Rumpel et al. 2002, Rasmussen et al. 2005, Eusterhues et al. 2007, Hobley et al. 2014), so that targeting subsurface stocks of SOC may lead to a longer retention of the additional C. Sub-SOC sequestration may therefore potentially overcome two constraints on soil carbon sequestration (Swift 2001, Powlson et al. 2011): the timeframe and magnitude of SOC storage.

Despite significant interest in subsurface SOC, our understanding of the factors influencing subsoil OC dynamics is still limited (Rumpel and Kögel-Knabner 2011), and this is particularly true in Australia. Identifying the key environmental, site and management factors driving SOC storage and depth distribution are key to the success of enhanced subsurface SOC storage, enabling us to identify sites with the greatest potential for additional subsurface SOC storage.

In addition to the potential of subsurface SOC as a sink of atmospheric carbon, it has potential as a scientific archive. Soils have a "memory" for previous environmental conditions (Janzen 2005) and it has been suggested that subsurface SOC may reflect previous management regimes and practices, rather than current ones (Wilson and Lonergan 2013). As such, a better understanding of sub-soil OC dynamics may enable us to decipher the influence of environmental and anthropogenic drivers on the global C cycle, and how this is reflected in soils and ecosystems.

Globally, subsurface SOC storage has been linked with numerous environmental and site factors. Climate is recognized as a key driver of SOC depth distribution, with mean annual precipitation (MAP) positively and mean annual temperature (MAT) negatively associated with SOC content at depth (Burke et al. 1989, Jobbagy and Jackson 2000, Wang et al. 2004). Site factors such as soil type, texture and mineralogy are important to SOC dynamics, with subsurface SOC retention positively associated with clay content (Burke et al. 1989, Jobbagy and Jackson 2000), and negatively associated with both sand (Jobbagy and Jackson 2000) and silica content (Badgery et al. 2013). Clay mineralogy also influences SOC dynamics, with the weathering of non-crystalline to crystalline clays resulting in a loss of subsurface SOC (Torn et al. 1997). However, the depth distribution of SOC also depends on other influences, such as the amount of SOC itself (Don et al. 2013) and human activities (Wiesmeier et al. 2014a, 2015).

Human influences on SOC dynamics are well documented, particularly for surface soils. In Australia, land use and land-management significantly affect SOC stocks near the surface, but effects are more difficult to detect in sub-soils (Wilson et al. 2008, 2011, Luo et al. 2010). To date no large-scale study has investigated the drivers of SOC at depth in Australia. However, in a comprehensive investigation of the drivers of SOC storage and vertical distribution in the top 30 cm of soil in eastern Australia, we found that land use was the most important indicator of the vertical distribution of SOC (Hobley et al. 2015).

In that study, we found that the drivers of SOC storage (in 0–30 cm depth) differ from those influencing its vertical distribution (Hobley et al. 2015). We developed the hypothesis that surface SOC storage in eastern Australia is controlled predominately by water availability, which limits plant growth and therefore SOC production. SOC production is greatest near the surface, so that the influence of water availability on SOC storage is most noticeable near the surface. In contrast, temperature has a greater influence on microbial activity, which determines SOC turnover throughout the soil profile, thereby controlling SOC loss (and availability for translocation) and influencing the depth distribution of SOC. Compared with native sites, anthropogenic land use limits the production of SOC (via removal of plant and animal biomass), and in the case of tillage, leads to a mixing/redistribution of SOC in the tillage layer, which results in a very strong influence of land use on SOC depth distribution. In this current study, we test these hypotheses at greater depths by investigating the depth distribution of SOC up to 1 m at 100 sites from across the State of New South Wales, Australia.

Methodology

Soils and sites

For this study, 100 sites across New South Wales (NSW, Fig. 1) were selected for analysis of profile carbon from a database of over 1500 sites across the State (Hobley et al. 2015). The data set was merged from two projects, namely the NSW Land and Soil Condition Monitoring Program (2008–2009) and the NSW component of the Soil Carbon Research Program initiated Australia-wide in 2009. An additional 22 sites were incorporated from the NSW paired site sampling for an Australian Greenhouse Office soil carbon estimation study (Murphy et al. 2003). In each of these projects, a "deep" core of up to 1 m was sampled at each site, divided into depth increments and the sub-samples air-dried prior to archiving.

We used the results from the previous data mining study to create a design matrix to select the 100 sites for deeper investigation. As climate (predominately MAP) had the greatest influence on total SOC storage in 0–30 cm depth, but land use and temperature were more important to

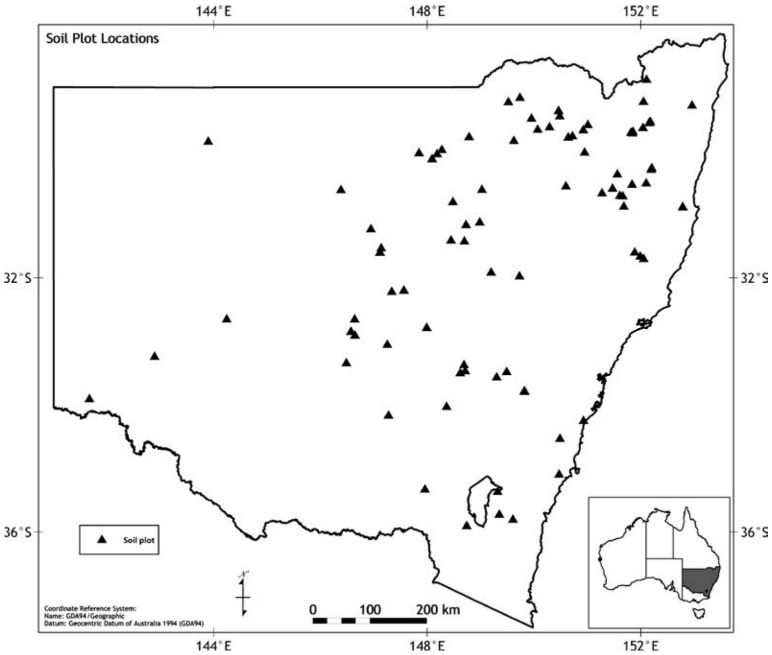

Fig. 1. Location of study sites.

SOC depth distribution, we decided to base the site selection on these three variables (Hobley et al. 2015).

The original data set was divided into five MAP ranges: <500, 500–650, 650–800, 800–1100, >1100 mm/yr. Each of these MAP ranges was then subdivided into five ranges based upon (maximum) MAT, yielding 25 discrete sampling bins. Maximum MAT was used as this was found to be most important to the depth distribution of SOC in the top 30 cm. Land use was categorized into three major classes: native vegetation, grazed/cleared land and cropped/cultivated land and sites for analysis were then selected based upon land use from within each bin. In doing this, the variation in other factors important to SOC storage and depth distribution (e.g., soil type) was minimized.

Our aim was to create a balanced data set with equal representation of land use within climate ranges and thereby enable a comparison of the influence of land use and climate on the depth distribution of SOC. However, due to the uneven distribution of land use classes across the range of climates represented in the region (e.g., the lack of cropped sites in colder, high rainfall areas), as well as the availability and depth resolution of samples from within the archive, the final matrix was not perfectly balanced. Climate, land use and soil type characteristics of the 100 site data set are shown in Table 1.

Carbon analysis

Archived samples were sieved to <2 mm and ground to <100 μm prior to analysis. Carbon (C) content ($\%_{mass}$) was measured as total carbon using LECO dry combustion at the NSW Office of Environment and Heritage Yanco Soils Laboratory. C content was determined on 40 °C dry samples and adjusted for water content determined on 105 °C dry samples. The presence of carbonates in samples was tested using the HCl fizz test and samples containing carbonates were analyzed after acidification with H_2SO_3 to remove carbonates, so that all results represent the organic carbon in the soils.

Soil carbon profile modeling

As the depth increments and total sampling depth of the sites were not consistent between the different projects, negative exponential depth models (Arrouays and Pelissier 1994, Hilinski 2001, Hobley et al. 2013) were fitted to each

Table 1. Climate, land use and soil type characteristics of data set.

Variable	Statistic	Value
Mean annual precipitation (mm/yr)	Range	280–1402
	1st quartile	484
	Median	668
	Mean	695
	3rd quartile	849
Mean annual temperature (°C)		Average (minimum; maximum)
	Range	9.2–20.7 (3.6–13.7; 14.6–27.5)
	1st quartile	14.1 (3.6; 20.3)
	Median	17.7 (10.8; 24.3)
	Mean	16.5 (9.9; 23.1)
	3rd quartile	18.5 (11.7; 25.6)
Land use class	Native vegetation	35
	Grazed	34
	Cropped	31
Land-management class	Natural	36
	Native pasture	17
	Introduced pasture	16
	Crop-pasture rotation	6
	No/low-till cropping	13
	Traditional cropping	12
Soil type (Australian Soil Classification, in brackets: approximate World Reference Base; United States Department of Agriculture)	Calcarosols (Calcisols; Alfisols)	11
	Chromosols (Acrisols/Lixisols; Alfisols)	13
	Dermosols (Luvisols/Cambisols; Ultisols/Inceptisols)	4
	Ferrosols (Nitisols; Oxisols)	9
	Kandosols (Acrisols; Alfisols)	6
	Kurosols (Acrisols; Alfisols or Ultisols)	20
	Rudosols and Tenosols (Arenosols/Regosols; Entisols)	12
	Sodosols (Solonetz; Alfisols)	6
	Vertosols (Vertisols; Vertisols)	19

profile to enable comparison of the SOC profile data:

$$\mathrm{SOC}(d) = \mathrm{SOC}_{\mathrm{Inf}} + (\mathrm{SOC}_0 - \mathrm{SOC}_{\mathrm{Inf}}) \times e^{-\mathrm{d} \times k} \quad (1a)$$

$$= \mathrm{SOC}_{\mathrm{Inf}} + (\mathrm{SOC}_0 - \mathrm{SOC}_{\mathrm{Inf}}) \times e^{-\mathrm{d}/\lambda} \quad (1b)$$

Where SOC(d) is the mass concentration (%) of SOC as a function of depth below the soil surface, $\mathrm{SOC}_{\mathrm{Inf}}$ is the mass concentration (%) of residual SOC in an "infinitely" deep soil, SOC_0 is the mass concentration (%) of SOC at the soil surface, d is the depth below the soil surface (m), k is the depletion constant (m^{-1}) and λ is the length scale of depletion in SOC concentration with depth (m, $\lambda = 1/k$). Models were fitted to Eq. 1a and bootstrapped ($N = 720$) to obtain 95% confidence intervals (CI) for each parameter in the model.

Identifying variables important to SOC depth distribution

To identify variables important to SOC depth distribution, three different machine learning techniques were implemented, namely two tree ensemble methods and one gradient boosting machine. The models were used to identify the predictor variables important to each of the fit parameters (SOC_0, $\mathrm{SOC}_{\mathrm{Inf}}$, k and λ) obtained in the exponential models of the soil profiles. Modeling of both k and λ was done because, although they are directly (inversely) related, the relationship between them is non-linear. The tree-based machine learning algorithms applied here can be very good at describing non-linear relationships, but are poorer at describing linear relationships. We did not know the prior distribution of either k or λ with regards to their relationships with the potential explanatory variables, therefore, we chose to model both of the parameters and explore the capacity of the algorithms to both model them and identify their drivers.

Tree ensemble models

A preliminary investigation was undertaken to optimize tree ensemble model parameters based on model performance. Optimization was based on minimizing the mean square error prediction for out-of-bag estimates (i.e., cases not included in growing a tree). The final models were grown to 500 trees (ntree = 500). The number of splitting variables was set to half the square root of the total number of predictor variables. No restriction upon tree growth was placed (i.e., minimum number of variables in a node to split upon was set to 2 and minimum number of observations in a terminal node was set to 1). Modeling was done using the party package (Hothorn et al. 2006) in R 3.1.2 (R Core Team 2014). The possible explanatory variables used in the models were derived from numerous GIS layers relating to climate and site factors at the sites, as well as the land use categories derived from the land use recorded at sampling (Table 2).

Two different types of tree ensemble models were grown: randomForests (Breiman 2001), which recursively partition a data set into nodes of ever-increasing purity, and conditional inference tree ensembles (Hothorn et al. 2006), which differ from randomForests in that they introduce a criterion for partitioning: a split in the tree is undertaken only if there is an association between predictor and response variable, otherwise tree growth is terminated. This was set via the mincriterion parameter (set to 0.01 for randomForests and 0.95 for conditional inference tree ensembles) in the party package, where mincriterion = 1 − P-value of the association between response and predictor variables.

Tree ensemble models were grown for the mean parameter estimates as well as for 500 randomly generated parameter values from within the 95% CI obtained from bootstrapping the exponential models. In total, for each parameter 501 tree ensembles were fitted representing a total of 250 500 individual trees for each tree ensemble type. From each model, the variable importance was extracted, normalized to the total sum of variable importance in the model and weighted by dividing by the mean square error of the model. The weighted variable importance measures were then averaged across all models to identify the predictor variables important to each of the parameters obtained during exponential fitting with Eq. 1a. Variables were deemed important if their relative importance was greater than that expected from a theoretical model where all predictor variables are equally influential (Hobley et al. 2015).

Model performance was evaluated using the coefficient of determination (R^2):

$$R^2 = 1 - \frac{\text{MSE}}{\text{Variance}} \tag{2}$$

where MSE is the mean square error of the average of individual estimates in each tree and Variance the variance of the mean estimate of the response parameter.

As the models were based upon the parameters randomly generated from the 95% CI of the exponential fit models, the MSE was assessed using both the square distance to the mean parameter estimate as well as the square distance to the 95% CI bounds. The MSE was calculated for both fitted and predicted values (calculated using the out-of-bag estimates in each tree, i.e., the cases not used to fit the models).

Boosted regression tree models

The gradient boosting machines (Friedman 2001) were optimized for number of trees, learning rate, tree depth (i.e., number of splits in each step of the tree) and minimum number of observations in each node using five-fold cross-validation based on fitting models to the mean parameter estimates of the exponential models. Using these optimized model parameters (Table A1), gradient boosted models were then grown for the estimated SOC depth distribution parameters obtained with the exponential fit models, as well as for 250 000 randomly generated values from within the 95% CI of the exponential fit models. For each individual model, the data set was randomly subdivided into 75 training data points and 16 test data points. The former were used to grow the models and calculate the MSE of the fit, the latter used for cross-validation and to

Table 2. Predictor variables used in machine learning algorithms.

Variable class/name	Type† (Levels)	Explanation	Source and further information
Climate			
ClmZone	*F* (4)	Climate zone defined according to temperature and humidity	a: Australian Bureau of Meteorology (BOM) 2006 Published scale: ~2 × ~2 km (0.025°) Based on a standard 30-yr climatology (1961–1990)
Evaporation	*C*	Evaporation (mm/yr)	b: Australian Department of Environment Published scale: ~1 × ~1 km 1970–2012 monthly averages
Koppen code	*F* (8)	Detailed classification of Köppen climate zone based on a modified Köppen classification system defined for Australia	a
Koppen group	*F* (4)	Major Köppen climate classification	a
MAP	*C*	Mean annual precipitation (mm/yr)	b
MARH	*C*	Mean annual relative humidity (%, 9 a.m. data: $\mathrm{MARH}_{9\ \mathrm{a.m.}}$, 3 p.m. data: $\mathrm{MARH}_{3\ \mathrm{p.m.}}$)	BOM 2008 Published scale: ~10 × ~10 km (0.1°) Based on a 30-yr climatology (1976–2005)
MAT	*C*	Mean annual temperature (°C, average: $\mathrm{MAT}_{\mathrm{ave}}$, minimum: $\mathrm{MAT}_{\mathrm{min}}$ and maximum: $\mathrm{MAT}_{\mathrm{max}}$)	b
T_{diff}	*C*	$\mathrm{MAT}_{\mathrm{max}}$–$\mathrm{MAT}_{\mathrm{min}}$	Authors
T_{diffnorm}	*C*	T_{diff} divided by $\mathrm{MAT}_{\mathrm{ave}}$	Authors
T_{ratio}	*C*	Ratio of $\mathrm{MAT}_{\mathrm{max}}$ to $\mathrm{MAT}_{\mathrm{min}}$	Authors
VPD	*C*	Vapor pressure deficit (kPa)	b
Rain season	*F* (9)	Seasonal rainfall zones based on median annual rainfall (November to April and May to October) and seasonal incidence of rainfall	BOM 2006 Published scale: ~25 × ~25 km (0.25°) Based on a 100-yr period (1900–1999)
Land use			
LM codes	*F* (6)	Land-management categories based upon original land use recorded during sampling.	Authors
LU class	*F* (3)	Land-use classes based upon land use recorded during sampling and via visual cross-checking using satellite imagery from Bing and Google maps.	Authors
Site and soil			
ASC	*F* (9)	Australian soil class from map of soil types across NSW using the Australian Soils Classification at Order level	NSW Office of Environment and Heritage 2012 Published scale: 1:250 000
AWC	*C*	Available water capacity (%) in a given soil depth	c: Soil and landscape grid of Australia CSIRO 2014 Published scale 3 arc s (approx. 90 m)
$\mathrm{AWC}_{\mathrm{ratio}}$	*C*	Ratio of available water capacity (%) in 0–5 to 100–200 cm depth	Authors
Clay	*C*	Clay mass fraction (%) in a given soil depth	c
$\mathrm{Clay}_{\mathrm{ratio}}$	*C*	Ratio of clay content in 0–5 to 100–200 cm depth	Authors
DepthReg	*C*	Depth of regolith (m)	c
DepthSoil	*C*	Depth of soil A and B horizons (m)	c
ECE	*C*	Cations extracted using BaCl_2 plus exchangeable H + Al (m_{eq}/100 g)	c
$\mathrm{ECE}_{\mathrm{ratio}}$	*C*	Ratio of cation exchange capacity in 0–5 to 100–200 cm depth	Authors

Table 2. Continued.

Variable class/name	Type† (Levels)	Explanation	Source and further information
Elevation	*C*	Elevation (m above sea level)	Surface Geology of Australia Geoscience Australia 2012 Published scale: 1:1 000 000
N	*C*	Total N concentration ($\%_{mass}$) in a given soil depth	c
N_{ratio}	*C*	Ratio of N concentration in 0–5 to 100–200 cm depth	Authors
P	*C*	Total P concentration ($\%_{mass}$) in a given soil depth	c
P_{ratio}	*C*	Ratio of P concentration in 0–5 to 100–200 cm depth	Authors
pH	*C*	pH in $CaCl_2$	c
pH_{ratio}	*C*	Ratio of pH in 0–5 to 100–200 cm depth	Authors
ρ	*C*	Bulk density of soil (g/cm^3) in a given depth	c
ρ_{ratio}	*C*	Ratio of bulk density in 0-5 to 100-200 cm depth	Authors
Sand	*C*	Sand mass fraction (%) in a given soil depth	c
$Sand_{ratio}$	*C*	Ratio of sand content 0–5 to 100–200 cm depth	Authors
Silt	*C*	Silt mass fraction (%) in a given soil depth	c
$Silt_{ratio}$	*C*	Ratio of silt content in 0–5 cm to 100–200 cm depth	Authors
TWI	*C*	Topographic wetness index (dimensionless units)	CSIRO 2012 Published scale: 1-s of arc (approx. 30 m)

Note: †*F*: Factor variable (number of categories), *C*: continuous variable.

calculate the MSE of prediction. Modeling was done using the gbm package in R (Ridgeway 2013).

Corresponding to the methodology applied to the tree ensemble model results, variable importance was extracted from each boosted model, weighted, normalized and averaged over all models to obtain the final variable importance. Variables with a relative importance greater than expected in a random model with all variables equally influential were deemed to be important.

Assessing influence of variables important to SOC depth distribution

To assess the influence of the important variables identified using the machine learning techniques, multiple regression models were created for each parameter of the exponential fits. Our aim was not to create predictive models but to assess the relative contributions and influence of individual variables identified within the tree models. The predictor variables in the regression models were selected from the important variables identified with the three data mining methods using bidirectional stepwise regression, based upon log-likelihood (Akaike Information Criterion) with the initial model specified as the full model containing all variables. Stepwise regressions were built using a bootstrap method (N = 1000) to select predictors for the final models based upon frequency of inclusion (>50%) and stability of coefficients (>75%) of each predictor in the individual bootstrapped models. Investigations of the best correlations between continuous, positive, natural and log-transformed response and predictor variables, as well as the final model goodness-of-fit were used to decide whether to log-transform variables for the models.

The variance attributable to each predictor variable in the final models was then determined using the proportional marginal variance decomposition method of Feldman (2005, as cited in Groemping 2006), in which the relative contribution to variance of a given predictor is calculated based upon its likelihood of appearing in a given position in the model. Variance decomposition was done using the relaimpo package in R (Groemping 2006).

From these models, the influence of the continuous variables was assessed via the model coefficients. To assess the influence of categorical variables, a combination of ANOVA, partial regression and post hoc analysis (using the Games-Howell post hoc test) was applied (Hobley et al. 2015). For the most important (categorical) variables in each regression, a simple ANOVA combined with Games-Howell post hoc analysis was undertaken. For less important variables, the model was de-trended of more influential variables using partial regression and Games-Howell post hoc applied to an ANOVA of the residuals of the partial regression and the modeled variable. Results with probability of error (*P*-value) <0.1 are reported.

Results

Exponential fits

Exponential fitting of SOC to soil depth was found to have adequate results (based upon R^2 of fit and visual assessment of fitted function vs. measured values) for 91 of the 100 profiles (for examples see Fig. 2). Of these 91 profiles, 77 had an $R^2 > 0.9$ and 66 had an $R^2 > 0.95$, indicating a good to excellent model fit. For the remaining profiles, although only having a poor to moderate fit ($R^2 = 0.80 \pm 0.12$), the model fit was visually deemed adequate and so they were included in subsequent analyses. Nine profiles were inadequately described using an exponential model and were excluded from the analyses. Of these nine profiles, some failed to converge during fitting, whereas others had a very poor fit ($R^2 \sim 0$).

Tree models

The fit and prediction performance of the ensemble models (randomForests and conditional inference tree ensembles) was generally similar (Table 3). On average, randomForests performed slightly better at fitting, but not necessarily at predicting the exponential fit parameters than the conditional inference tree ensembles. The gradient boosting machines exhibited a very wide spread of values, from extremely negative R^2 (i.e., MSE ≫ Variance) to near perfect predictive performance (Table 3). Although the poorest gradient boosted models were significantly worse than the poorest performing ensemble models, the best predictive

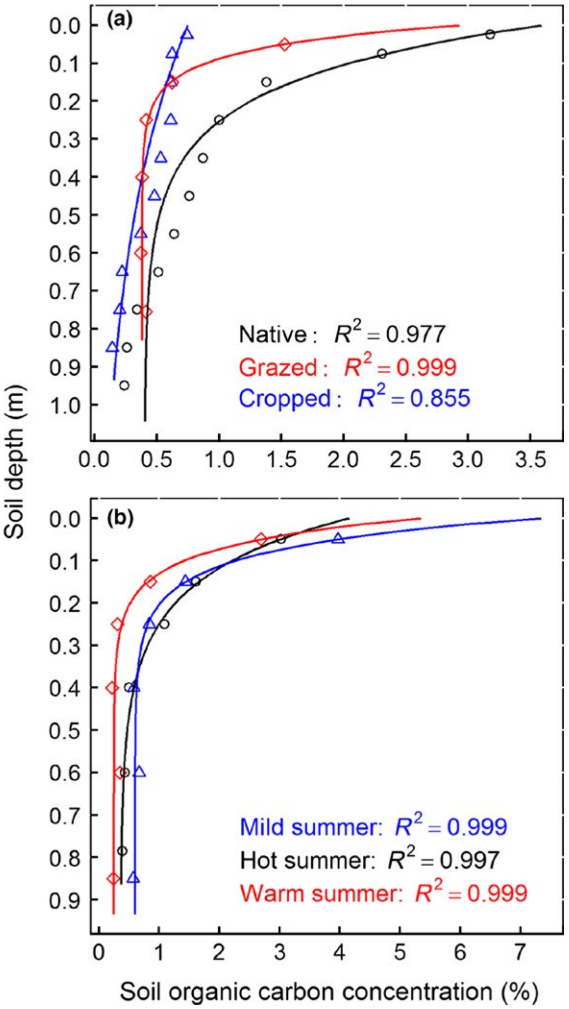

Fig. 2. Exponential fits of soil organic carbon as a function of depth using Eq. 1a. (a) Land-use comparisons within a given climate (Temperate zone, no dry season, hot summer, MAP 610 ± 40 mm/yr, MAT 17.9 ± 0.3 °C) and soil type (Australian Soil Classification: Vertosol, WRB and USDA: Vertisol). (b) Sites with different climate (MAT and MAP): hot summer (MAT 15.2 °C, MAP 717 mm/yr), warm summer (MAT 13.7 °C, MAP 785 mm/yr) and mild summer (MAT 12.6 °C, MAP 996 mm/yr) under a single land use (grazed) and soil type (Australian Soil Classification: Chromosols, WRB: Acrisols, USDA: Alfisols).

Table 3. Performance of the tree-based models.

Model and variable	R^2 (mean ± standard deviation, minimum and maximum in brackets)			
	Expected value of exponential model		95% CI of exponential model	
	Fit	Prediction	Fit	Prediction
randomForest				
SOC_0	0.86 ± 0.04 (0.71;0.92)	0.51 ± 0.07 (0.36;0.69)	0.98 ± 0.02 (0.92;0.99)	0.82 ± 0.05 (0.72;0.92)
k	0.74 ± 0.05 (0.56;0.84)	0.17 ± 0.05 (−0.05;0.28)	0.98 ± 0.0.1 (0.93;0.99)	0.72 ± 0.04 (0.51;0.81)
λ	0.65 ± 0.11 (0.33;0.83)	0.07 ± 0.11 (−0.30;0.26)	0.95 ± 0.02 (0.89;0.99)	0.75 ± 0.10 (0.45;0.94)
SOC_{lim}	0.50 ± 0.15 (0.05;0.86)	0.29 ± 0.06 (−0.03;0.41)	0.97 ± 0.01 (0.96;0.98)	0.80 ± 0.02 (0.70;0.86)
Conditional inference trees				
SOC_0	0.65 ± 0.07 (0.49;0.80)	0.35 ± 0.05 (0.18;0.41)	0.89 ± 0.03 (0.81;0.96)	0.74 ± 0.03 (0.66;0.83)
k	0.32 ± 0.04 (0.19;0.43)	0.13 ± 0.03 (0.05;0.19)	0.82 ± 0.03 (0.74;0.88)	0.74 ± 0.02 (0.68;0.78)
λ	0.38 ± 0.09 (0.18;0.64)	0.13 ± 0.03 (0.01;0.23)	0.97 ± 0.01 (0.93;0.99)	0.92 ± 0.03 (0.82;0.97)
SOC_{lim}	0.42 ± 0.05 (0.27;0.56)	0.25 ± 0.03 (0.16;0.34)	0.85 ± 0.02 (0.79;0.90)	0.76 ± 0.01 (0.71;0.80)
Gradient boosting machines				
SOC_0	0.68 ± 0.18 (−1.95;0.95)	−1.10 ± 3.27 (−160;0.99)	0.84 ± 0.12 (−1.55;0.99)	−0.43 ± 2.81 (−153;0.99)
k	0.36 ± 0.21 (−2.07;0.83)	−0.28 ± 1.33 (−42.78;0.97)	0.73 ± 0.14 (−1.39;0.96)	0.33 ± 0.99 (−32.84;0.99)
λ	0.19 ± 0.51 (−5.54;0.86)	−1.71 ± 4.22 (−201.18;0.97)	0.86 ± 0.06 (0.31;0.99)	−0.55 ± 3.35 (−187.19;0.99)
SOC_{lim}	0.55 ± 0.12 (−0.09;0.89)	−0.48 ± 1.46 (−35.89;0.97)	0.96 ± 0.02 (0.70;0.99)	0.34 ± 0.78 (−20.78;0.99)

performance obtained with individual models were gradient boosting models.

The variables indicated as important to SOC depth distribution were nearly identical in rank (though not relative values) for both tree ensemble model methods. In contrast, the important variables identified using the gradient boosting machines varied substantially in rank and value from those identified using the tree ensemble methods. The variables identified as important were generally the same using the two different weighting methods (fit and prediction) although some alterations in the rankings occurred, although not for the top three variables in the models. Figs. 3–6 show the variable importance averaged over the fitted and predicted weightings.

Variables important to surface SOC: SOC_0

The variables identified as important to SOC_0 using the tree ensemble models were almost exclusively related to climate, predominately seasonal rainfall. The notable exception to the climate signal on SOC_0 was bulk density at 0–5 cm, which was the most important variable for modeling SOC_0, accounting for around 25% of the variable importance (Fig. 3).

The gradient boosting machines also predominately identified climate variables as important to SOC_0, although the results differed substantially compared with the tree ensembles, with temperature more important than rainfall. In addition to climate, elevation, bulk density and N content at 0–5 cm were identified as the most important variables to SOC_0 (Fig. 3). Elevation was highly collinear with temperature ($R = -0.93$) and can be viewed as a proxy for temperature within the data set.

Variables important to depth depletion of SOC: k and λ

Land use, soil depth and land-management were the variables most important to the SOC depletion constant (k) within the tree ensemble models. The normalized MAT difference was identified as being of minor importance to k (Fig. 4) with the tree ensembles, whereas the gradient boosting machines identified $T_{\text{diff, norm}}$ as the most important variable to k. Compared with the tree ensembles, the gradient boosting machines identified several additional variables important to k, including pH_{ratio}, ECE_{ratio}, soil type and $silt_{ratio}$, MAT_{ave} and elevation (Fig. 4).

In contrast, the most important variable to the length scale of depletion (λ) identified with the tree ensembles was clay content at 0–5 cm, followed by sand content (0–5 cm) and the N_{ratio} from topsoil to depth. Land-management, land use and surface pH were also important

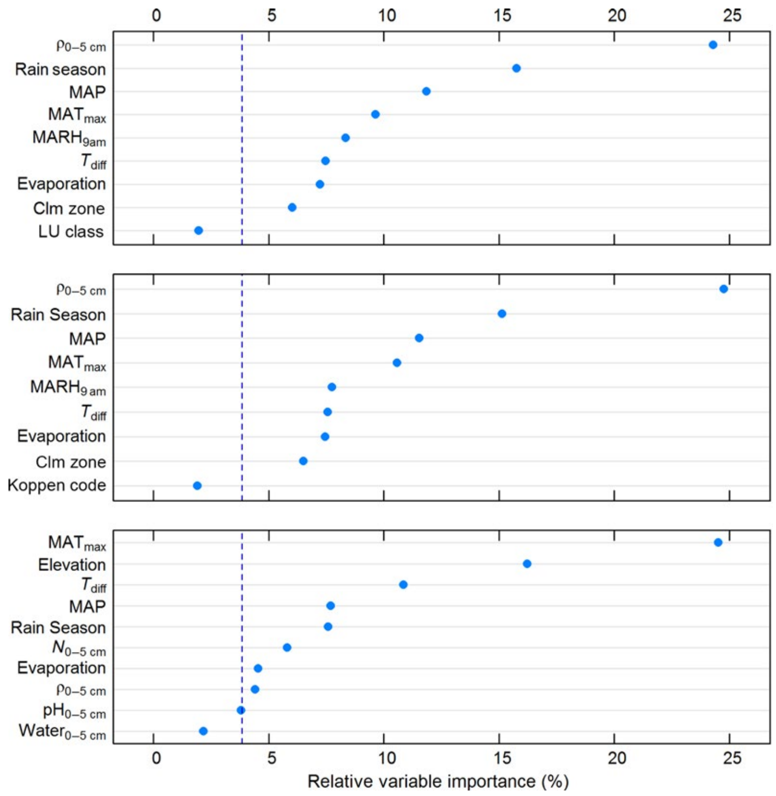

Fig. 3. Variables identified as important to surface soil organic carbon (SOC) storage (SOC_0) using three machine learning algorithms: (a) randomForests, (b) conditional inference tree ensembles, (c) gradient boosting machines. Variable importance was weighted using the goodness-of-fit obtained during fitting the models and model validation using predictive performance and averaged over the two results. The horizontal blue line indicates the expected variable importance in a random model, where all predictors are equally influential. See Table 2 for an explanation of variables.

to λ. The gradient boosting machines identified different variables important to λ, above all land-management, $\mathrm{ECE}_{\mathrm{ratio}}$ and soil type (Fig. 5).

Variables important to residual SOC at depth: SOC_{Inf}

The tree ensembles identified site variables ($\mathrm{pH}_{\mathrm{ratio}}$ and $\mathrm{ECE}_{\mathrm{ratio}}$, $\varrho_{100\text{–}200\ \mathrm{cm}}$, $\mathrm{Sand}_{100\text{–}200\ \mathrm{cm}}$) and land use as most important to the content of SOC retained at depth in the soils. Climate variables were also important to $\mathrm{SOC}_{\mathrm{inf}}$, although not as influential as land use and soil depth (Fig. 6). The gradient boosting machines identified soil type as the most important variable to SOC content at depth, followed by land-management and other site and climate variables (Fig. 6).

Multiple regression models

The bootstrapping of stepwise regressions resulted in final models which explained a similar amount of variance in the expected model parameters as the tree ensemble models. In total the multiple regression models accounted for 50–65% of observed variance (Table 4).

The final model explaining SOC_0 was dominated by climate variables, which accounted for 73% of model explained variance (Table 4). The remaining 27% was attributable to surface bulk density. Seasonal rainfall was the most influential variable, accounting for over 50% of the total explained variance (Fig. 7).

Bulk density and evaporation were negatively associated with surface SOC, whereas MAP and T_{diff} were positively associated with SOC_0.

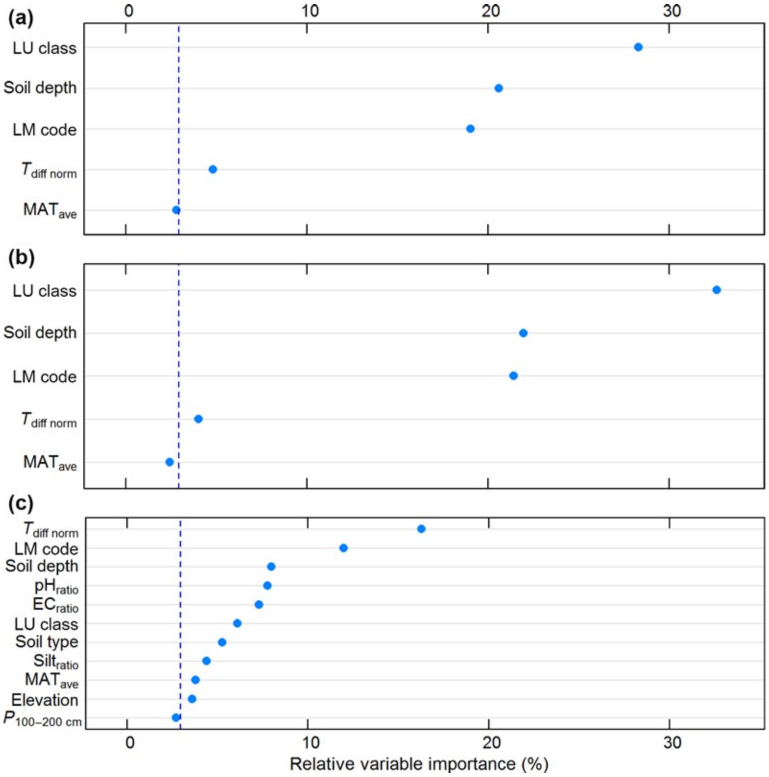

Fig. 4. Variables identified as important to SOC depletion constant (*k*) using three machine learning algorithms: (a) randomForests, (b) conditional inference tree ensembles, (c) gradient boosting machines. Variable importance was weighted using the goodness-of-fit obtained during fitting the models and model validation using predictive performance and averaged over the two results. The horizontal blue line indicates the expected variable importance in a random model where all predictors are equally influential. See Table 2 for an explanation of variables.

The Games-Howell post hoc analysis identified several significant differences in SOC_0 between seasonal rainfall zones. Generally, SOC at the surface was higher in higher rainfall areas for a given seasonality. In areas of similar annual rainfall, uniform annual rainfall distribution generally resulted in higher surface SOC than summer dominant rainfall (Table 5). Within temperature and humidity zones, no significant differences were apparent using the Games-Howell test after de-trending for other influential variables via partial correlation.

In contrast to the predictors influencing the content of SOC at the surface, the most influential variables to the depth depletion of SOC (both *k* and *λ*) were related to land use and management, which accounted for 52–67% of the explained variance of the model (Table 4). Site variables contributed similar amounts to the model explained variance (~30%), whereas climate variables were less influential (4–16% explained variance, Fig. 7).

The SOC depletion constant, *k*, was negatively associated with temperature and the $\text{ECE}_{\text{ratio}}$. The length scale of depletion, *λ*, was positively associated with surface clay content and the ratios of silt content, available water capacity and exchangable cations from soil surface to depth, but negatively associated with the ratios of N and clay content.

In general, under different land-management, *k* decreased in order natural ≥ grazed > crop-pasture rotations > no/low-till ≥ conventional cropping systems (and correspondingly *λ* increased in this order), although not all differences were significant (Table 6). There were no significant

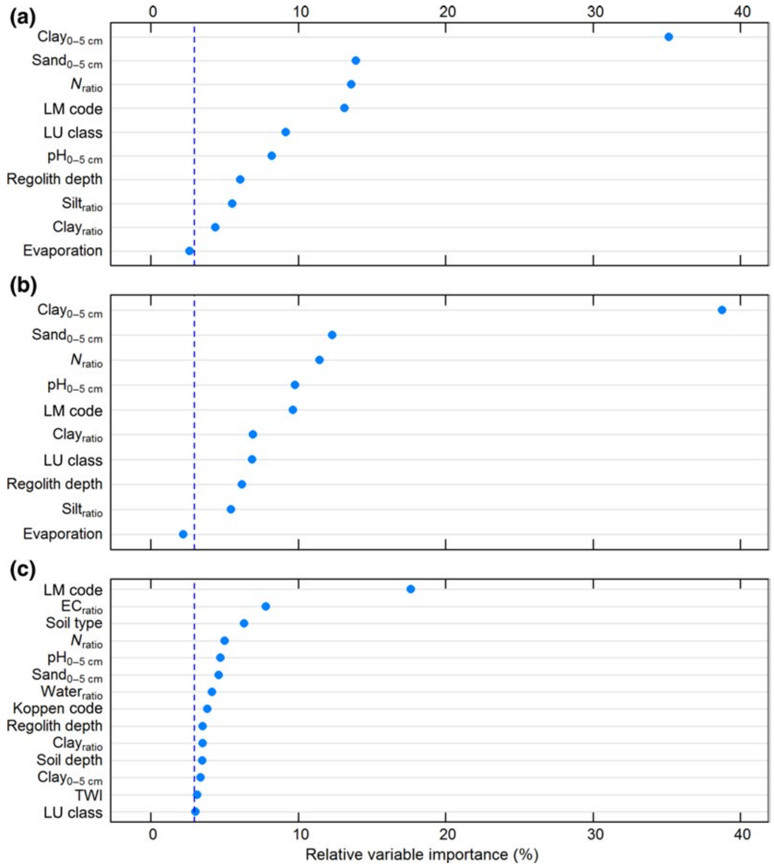

Fig. 5. Variables identified as important to length scale of SOC depletion (λ) using three machine learning algorithms: (a) randomForests, (b) conditional inference tree ensembles, (c) gradient boosting machines. Variable importance was weighted using the goodness-of-fit obtained during fitting the models and model validation using predictive performance and averaged over the two results. The horizontal blue line indicates the expected variable importance in a random model where all predictors are equally influential. See Table 2 for an explanation of variables.

differences in depletion constants between natural and grazed sites (Table 6). After de-trending for other important signals, neither of the depth depletion parameters differed significantly between soil types nor between climate categories.

Residual SOC at depth was most influenced by site factors, which accounted for ~50% of explained variance, followed by land use and climate (Table 4). Land-management and bulk density at 100–200 cm were equally important to the model, each accounting for over 25% of model explained variance (Fig. 7). The influential climate variable (Köppen classification) was related to seasonality as opposed to absolute indices (e.g., MAP or MAT).

The SOC_{Inf} was negatively associated with bulk density, sand content and pH at 100–200 cm, and positively associated with the ratios of P and bulk density from soil surface to depth. The highest residual SOC at depth was found under subtropical climates without a dry season (Table 5). The relative amount of SOC_{Inf} under different land-management classes mimicked closely the order of the depth depletion constant, k (Table 6).

Discussion

Modeling the depth distribution of SOC

The generally good to excellent results obtained while fitting Eq. 1a indicate that a simple, three

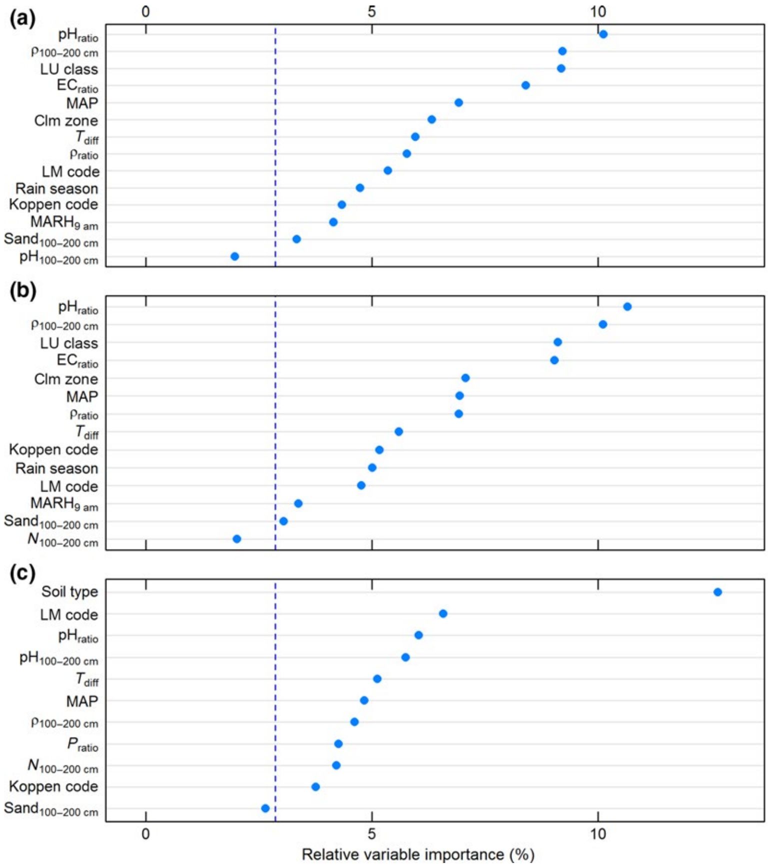

Fig. 6. Variables identified as important to residual SOC content at depth (SOC_{Inf}) using three machine learning algorithms: (a) randomForests, (b) conditional inference tree ensembles, (c) gradient boosting machines. Variable importance was weighted using the goodness-of-fit obtained during fitting the models and model validation using predictive performance and averaged over the two results. The horizontal blue line indicates the expected variable importance in a random model where all predictors are equally influential. See Table 2 for an explanation of variables.

Table 4. Multiple regression models.

Final model	R^2 (adj. R^2)	Proportion of variance		
		Climate	Site	Land use
SOC_0 ~ Rain season + $\varrho_{0–5\ cm}$ + ClmZone + T_{diff} + Evaporation + MAP	0.65 (0.60)	0.73	0.27	0
log(k) ~ LM codes + Soil type + LU class + log(ECE_{ratio}) + MAT_{ave}	0.50 (0.39)	0.04	0.29	0.67
log(λ) ~ LM codes + KöppenCode + Soil type + LU class + log($Water_{ratio}$) + log(ECE_{ratio}) + log(N_{ratio})+ $Clay_{0–5\ cm}$ + log($Silt_{ratio}$) + $Clay_{ratio}$	0.64 (0.49)	0.16	0.32	0.52
SOC_{Inf} ~ LM codes + $\varrho_{100–200\ cm}$ + KöppenCode + $Sand_{100–200\ cm}$ + log($pH_{100–200\ cm}$) + P_{ratio} + ϱ_{ratio}	0.63 (0.55)	0.20	0.52	0.28

parameter exponential model (Hilinski 2001) can be applied to adequately describe the depth distribution of SOC in a large number of samples from across the region. The main advantage of this model compared with other models (e.g., power functions) is the interpretability of the model parameters to ecologically relevant processes (e.g., SOC production or translocation). However, as 9% of the sites analyzed here were poorly described by the models, prudence is

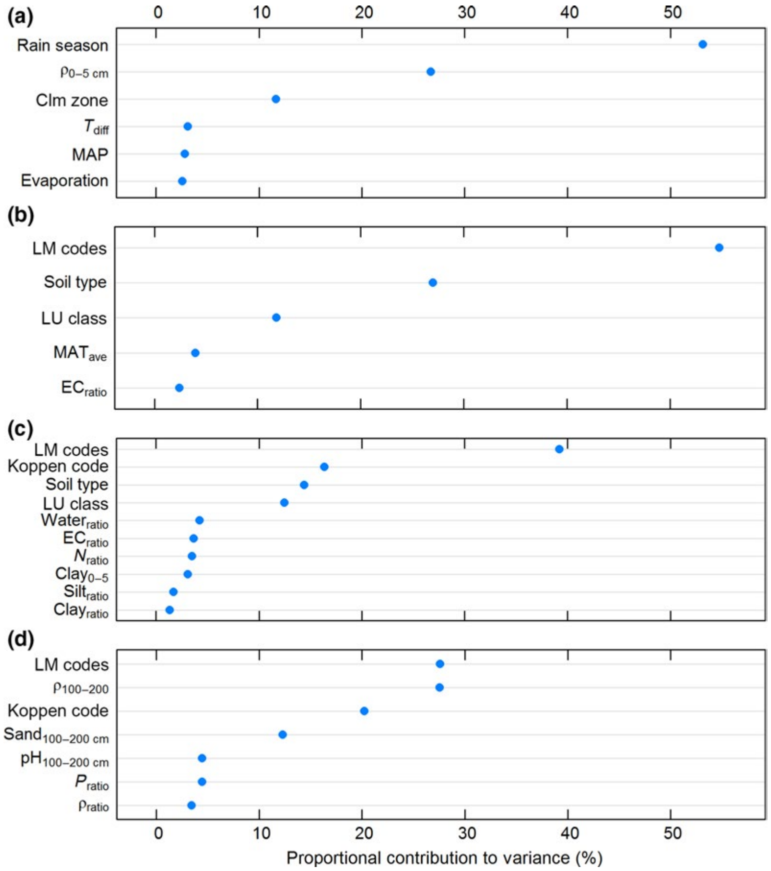

Fig. 7. Relative importance of variables influential to: (a) surface SOC (SOC$_0$); (b) SOC depletion constant (k); (c) length scale of SOC depletion (λ); and (d) residual SOC at depth (SOC$_{\mathrm{Inf}}$) calculated with the proportional marginal variable decomposition analysis on multiple linear regression created from the combined machine learning results. See Table 2 for an explanation of variables.

Table 5. Games-Howell post hoc analyses of climate categorical factors.

Response variable (predictor)	Comparison	Mean difference (%$_C$)	Significance level
SOC$_0$ (Seasonal rainfall zones)	Summer (650–1200 mm/yr) > Summer (350–650 mm/yr)	2.8	<0.001
	Summer (650–1200 mm/yr) > Uniform (500–800 mm/yr)	1.6	<0.08
	Summer (650–1200 mm/yr) > Uniform (250–500 mm/yr)	2.5	<0.001
	Uniform (250–500 mm/yr) > Summer (350–650 mm/yr)	0.3	<0.01*
	Uniform (500–800 mm/yr) > Uniform (250–500 mm/yr)	0.9	<0.08*
SOC$_{\mathrm{Inf}}$ (Köppen climate zone)	Subtropical, no dry season > Grassland, warm, persistently dry	0.6	<0.001*
	Subtropical, no dry season > Grassland, hot, persistently dry	0.4	<0.02*
	Subtropical, no dry season > Subtropical, moderately dry winter	0.4	<0.04*
	Subtropical, no dry season > Temperate, no dry season, warm summer	0.4	<0.001*
	Subtropical, no dry season > Temperate, no dry season, hot summer	0.4	<0.001*

Note: *Significant difference after partial correlation to de-trend for more important influences.

Table 6. Games-Howell post hoc analyses of land-management and land use categories.

Variable	Comparison	Mean difference†	Significance level
k (λ in brackets)	Natural > Crop-pasture rotation (no significant difference in λ)‡	2.8	<0.05*
	Natural > no/low-till cropping (Natural < no/low-till cropping)‡	9.2 (0.16)	<0.01
	Natural > conventional cropping (Natural< conventional cropping)‡	9.9 (0.43)	<0.02
	Native pasture > no/low-till cropping (Native pasture < no/low-till cropping)‡	5.0 (0.14)	<0.04
	Native pasture > conventional cropping (Native pasture< conventional cropping)‡	5.7 (0.42)	<0.07
	Crop-pasture rotation > introduced pasture (no significant difference in λ)‡	1.8	<0.02*
	Crop-pasture rotation > no/low-till cropping (Crop-pasture rotation < no/low- till cropping)‡	6.4 (0.16)	<0.07
	Crop-pasture rotation > conventional cropping (Crop-pasture rotation < conventional cropping)‡	7.1 (0.42)	<0.05
	Cleared/grazed > cultivated/cropped (Cleared/grazed < cultivated/cropped)§	8.4 (0.21)	<0.01
	Native > cultivated/cropped (Native < cultivated/cropped)§	4.3 (0.24)	<0.01
$\mathrm{SOC}_{\mathrm{Inf}}$	Natural > Crop-pasture rotation‡	0.2	<0.001*
	Natural > no/low-till cropping‡	0.3	<0.001
	Natural > conventional cropping‡	0.3	<0.001
	Native pasture > no/low-till cropping‡	0.3	<0.04
	Native pasture > conventional cropping‡	0.3	<0.04
	Crop-pasture rotation > no/low-till cropping‡	0.1	<0.2*
	Crop-pasture rotation > conventional cropping‡	0.1	<0.04*

Notes:†Difference in means of k in m^{-1} (λ in m) and $\mathrm{SOC}_{\mathrm{Inf}}$ in $\%_{\mathrm{C}}$.
‡Land-management as comparative factor.
§Land use as comparative factor.
*Significant difference after partial correlation to de-trend for more important influences.

advisable when generalizing about the depth distribution of SOC, especially when the aim is to predict the amount of SOC stored at depth from surface samples. Future research should focus on identifying the factors that determine whether or not these simple exponential models are applicable, which would then enable predictions of SOC at depth from surface samples.

When viewed in relationship with their distance from the 95% CI, the tree ensemble models generally performed well at both fitting and predicting all parameters of the exponential models (cf. Wiesmeier et al. 2014b, Hobley et al. 2015). The randomForests usually explained a greater proportion of variance than the conditional inference ensembles while fitting, although not necessarily for predicting, indicating that they may have a propensity to overfit (Strobl et al. 2007). Despite this, the two tree ensemble methods resulted in nearly identical important variables, so that for identification purposes they are equal.

In contrast, the gradient boosting machines gave a wide array of results, from very poor ($R^2 < 0$) to near perfect ($R^2 = 0.99$). This variation in performance is possibly the result of tuning the models using the mean parameter estimates and then using these tuning parameters to model the random estimates from within the 95% CI. Similar to the randomForests, there was a wide discrepancy between mean fit and prediction performance, indicating a tendency of the models to overfit. When using gradient boosting machines for predictive purposes, cross-validation of the models is essential and close attention must be paid to the effect of tuning parameters on model performance (Elith et al. 2008).

The performance of the multiple regressions was worse than the randomForests for fitting the surface SOC content, but similar for describing the depth depletion and residual SOC at depth. However, as indicated above, randomForests have the propensity to overfit and are therefore not necessarily a good benchmark for model comparison. In contrast, the linear models explained similar (or even greater) amounts of variance to both the conditional inference trees and gradient boosting machines, suggesting that the multiple regression models are adequate and can be used to assess and compare variable influence.

The variables selected as important using the gradient boosting machines were frequently different from those selected using the tree ensemble methods, which is potentially attributable to the differing model algorithms. The individual trees in the ensemble models repetitively partition a random subset of the data set into ever purer nodes (based upon the best random subset of predictors) and the results are then amalgamated into the ensemble. The boosting machines create an initial (usually quite small) tree, shrink it, and then repeatedly partition the *residuals* of the previous tree, in essence, similar to incorporating partial regression into a decision tree. The optimal tuning parameters of the gradient boosting machines indicated both a small tree depth (2–3) and small number of trees (1–15), so that the final models were considerably smaller than the tree ensembles (with unlimited splits and 500 trees per model). Our approach to separately model individual, random parameters and then weight and combine the variable importance is an attempt to overcome these differences in variable identification purposes.

The multiple regressions tended to identify and rank variables more similar to the tree ensembles, whereas the most important variables in the gradient boosting machines were not always included in the multiple regressions. This suggests that, if assessment of variable importance using variance decomposition in linear models is an adequate methodology, tree ensemble methods are preferable to gradient boosting machines. Despite this, variables from both methods were selected using the stepwise regressions to create the multiple regressions, indicating that both methods are informative.

It is notable that the variables important to SOC depth depletion parameters k and λ differed substantially, although they indicated an overall similar contribution to the three variable classes (climate, site, land use) in the multiple linear regressions. This discrepancy between results is probably a result of the non-linear relationship between these variables. We recommend testing numerous modeling approaches to maximize the information content of the models.

Drivers of the depth distribution of SOC

The main driver of SOC concentration at the surface was the seasonal rainfall, accounting for over 50% of model explained variance, whereas temperature was less important to surface SOC. This is consistent with our understanding of the positive relationship between precipitation and SOC production (Post et al. 1982) and confirms our hypothesis that SOC storage at the surface is driven primarily by water availability (Hobley et al. 2015). Within a seasonal rainfall zone, higher precipitation results in greater surface SOC, and where there is a large difference in annual rainfall, the influence of seasonality is not pronounced (Table 5). However, in contrast to the importance of *total* precipitation in our previous study both the *amount* and *seasonality* of rainfall were important in this smaller data set. This difference is potentially attributable to differences in the distribution of precipitation in the different data sets (i.e., the original, larger data set exhibited a strong positive skew, whereas precipitation is more normally distributed within this smaller data set).

Nevertheless, it is interesting that both seasonality and amount of precipitation affect surface SOC in the study region. It appears that temperature is not very important to SOC in the region, despite global relationships between temperature and net primary production (Michaletz et al. 2014). This is probably because winters are rarely long or severe in the region—mean winter temperature is ~10 °C (4.3–16.2 °C minimum–maximum) (Australian Bureau of Meteorology)—and so temperature will not decisively affect plant growth in many localities. Therefore, although the growing season may be affected by daylight hours (mainly in the south), in areas of uniformly distributed rainfall, plant growth (and SOC production) can occur for most of the year. In contrast, in locations where rainfall is seasonal, plant growth will be limited by water availability in drier periods of the year.

This implies that C dynamics are non-linearly related with water availability (Sierra et al. 2015), i.e., the relative rates of C production and turnover in wetter months are not equal to those in the drier months. Specifically, the lower SOC in regions with seasonal rainfall compared with uniformly distributed rainfall suggests that (1) C turnover is less affected by water availability than C production is and/or (2) that the difference between C production rates between rainfall seasons is greater than the difference between C turnover rates between rainfall seasons, resulting in a net loss of C due to seasonality of rainfall. Future investigation should focus on seasonality and SOC dynamics to help elucidate these relationships.

This influence of seasonality is also evident in the residual amount of SOC at depth, where subtropical climates without a dry season contain significantly more SOC than regions with seasonality as well as in warmer and drier climates. Targeting sites of low SOC in regions of higher, non-seasonal rainfall may therefore result in the greatest likelihood of success with regards to SOC sequestration throughout the soil profile.

Consistent with our previous findings (Hobley et al. 2015), temperature does not appear to be a major driver of SOC content, but was identified as important to the SOC depth depletion constant. This supports our hypothesis that temperature influences the depth distribution of SOC, possibly due to its effects on decomposition (Lützow and Kögel-Knabner 2009) and the dissolution of organic matter (Toosi et al. 2014). However, temperature has a much smaller impact on SOC depth distribution than human influence and site characteristics. Nevertheless, the negative association of temperature with the depth depletion constants of SOC indicates that proportionally more subsurface SOC is retained in hotter than in cooler climates. Although this is potentially due to a lower surface SOC in warmer regions compared with cooler regions, subsurface SOC sequestration may have the best potential in degraded sites in warmer regions. Further investigations are needed into the relationships between temperature, microbial activity and SOC turnover and depth distribution, especially in light of changing climate regimes.

In contrast to surface SOC content, SOC depth depletion appeared to be driven predominately by land-management and, to a lesser extent, site factors, with climate only playing a minor role. This supports our hypotheses that land use is the main driver of SOC depth distribution. The lowest depletion constants of SOC were found under traditional cropping, implying a proportionally smaller SOC reduction with depth than under grazing or natural systems. However, both the surface and residual SOC content at depth were significantly lower under cropped systems than under other land uses, so that conventional cropping can be seen as being the most detrimental for SOC concentration. Not only are our results consistent with the current understanding into the effects of land-management on SOC but also they imply that the application of novel statistical methods may help to overcome the traditional difficulties encountered in identifying effects of land-management and land use on SOC at depth (Guo and Gifford 2002, Wilson et al. 2008, 2011, Luo et al. 2010).

The negative effects of tillage on surface SOC content are well documented (Lal 1997) but effects in the sub-soil are often assumed to be negligible (see above) or even positive (Blanco-Canqui and Lal 2008). However, our results indicate that reducing soil disturbance by introducing minimum or no till cropping or rotational cropping and grazing management could potentially be used to enhance sub-soil OC stocks. This will only have a net positive effect on the total soil C budget if changes in land-management at one site are not offset by negative changes at another site (e.g., conversion of cropped sites to rotational cropping and grazing is not matched by conversion of grazed sites to rotational cropping and grazing). Other studies suggest afforestation and fallow periods (Don et al. 2011) may be used for subsurface SOC sequestration. Future investigations into the physico-chemical characteristics of SOC under different land-management schemes are needed to elucidate which C pools at depth are affected by land-management changes, thereby enabling insights into the stability of this dynamic subsurface C. Investigating the age distribution of subsurface C will help to decipher the timeframes of C dynamics and the stability of the C lost or gained after land-management change.

Of the site factors relevant to the depth distribution of SOC, soil physical properties—above all bulk density, but also available water capacity and texture—were more important than soil chemical properties. The negative association of bulk density with SOC—both at the surface and at depth—is consistent with our previous findings (Hobley et al. 2015) and implies that soil physical condition is an important driver and indicator of SOC dynamics. If enhancing SOC concentration is a goal, is advisable to limit soil compaction.

The inverse relationship between bulk density and SOC has long been recognized (Russell 1960), but it is important to note that an increase in SOC concentration will only lead to an increase in SOC stocks if it is not compensated for by the concurrent decrease in bulk density. This will be the case if the relationship between these two variables is non-linear and the relative change in SOC concentration is greater than the relative change in bulk density, which appears to be the case (Ruehlmann and Körschens 2009). Regardless of whether or not this relationship is valid, decreasing bulk density and increasing SOC concentration in compacted soils will also benefit available water capacity and air capacity (Archer and Smith 1972, Hudson 1994) and plant root growth (Passioura 1991), resulting in an overall improvement in soil health. Our results imply that this is true both at the surface and at depth, so that reducing soil compaction may be a means of subsurface SOC sequestration. Cover crops have been shown to have potential for reducing soil compaction (Chen and Weil 2010) and have been suggested as a measure to increase soil carbon stocks (Lal 2004) and we recommend investigations into the effects of cover crops as a potential means to facilitate a reduction in sub-soil bulk density and concurrent increase in sub-soil SOC storage.

Australian soils are generally quite old and highly weathered, predominately due to the lack of rejuvenating factors (such as uplift, glaciation, or volcanism) within recent geological times over most of the continent (Charman and Murphy 2007). The Australian soil classification scheme (Isbell 2002) is predominately based upon morphological and chemical soil characteristics. As such, soil type is largely a proxy for factors such as texture, mineralogy or pH, rather than pedogenesis. That soil type is important to the depth distribution of OC (but significant differences were not identified using the Games-Howell post hoc tests of the partial regressions) therefore likely reflects the influence of texture and mineralogy on the retention and storage of SOC (e.g., Krull et al. 2003, Hobley et al. 2013, 2014), rather than the processes of soil genesis.

Lastly, the variables identified as important to SOC depth distribution with the data mining methods and the variance decomposition results using multiple regressions were very similar to the trends identified in our previous analysis (Hobley et al. 2015), although the studies used different predictor variables, had different sample sizes and investigated different depths. This implies that, in the study region and regions with similar soils, we can identify the drivers of subsurface SOC dynamics from investigations in the top 30 cm of soil. Future research into the potential of predicting subsurface SOC from surface SOC content, environmental, management, and site characteristics is required to develop and inform global carbon models.

Outlook

The data mining algorithms used in this study to identify the drivers of the depth distribution of SOC in eastern Australia yielded at times substantially different results, but all were informative. Our results support the hypotheses developed from our previous investigation of the top 30 cm soil, indicating that in these soils, the drivers of SOC content and depth distribution can largely be identified by investigating the near surface depths. Surface SOC is primarily driven by climate, above all rainfall—both amount and seasonality—which we attribute to the influence of water availability on plant growth. In contrast, subsurface SOC depletion and residual SOC at depth are influenced largely by land use and site factors (mainly soil physical characteristics such as bulk density, sand content and soil type), and to a minor extent temperature and climate seasonality. We recommend targeting areas of higher, non-seasonal rainfall for enhancing subsurface SOC. Reducing tillage intensity and soil compaction will most likely have the greatest chances of success. Our results help to elucidate the effects of climate, land-management and site on subsurface SOC content and, importantly, our methods successfully identified significant differences in SOC concentrations well below the soil surface, which have been largely unsuccessful using traditional statistical analyses. Future research should focus on which pools of deep SOC are affected by land use and environmental factors, as well as the timeframes of changes in these pools.

Acknowledgments

Thank you to Brian Murphy for providing samples, to Arjan Wilkie for collating the data set and providing the map, and to Peter Schad for help with the WRB soil classes. This research was funded by the Australian Government Department of Agriculture. The data used in this study can be found in the figures and tables of this publication and at https://researchdata.ands.org.au/soil-organic-carbon-eastern-australia/631975.

Literature Cited

Archer, J. R., and P. D. Smith. 1972. The relation between bulk density, available water capacity, and air capacity of soils. *Journal of Soil Science* 23:475–480.

Arrouays, D., and P. Pelissier. 1994. Modeling carbon storage profiles in temperate forest humic loamy soils in France. *Soil Science* 157:185–192.

Australian Government Bureau of Meteorology. 2012. www.bom.gov.au.

Badgery, W. B., A. T. Simmons, B. M. Murphy, A. Rawson, K. O. Andersson, V. E. Lonergan, and R. van de Ven. 2013. Relationship between environmental and land-use variables on soil carbon levels at the regional scale in central New South Wales, Australia. *Soil Research* 51:645–656.

Blanco-Canqui, H., and R. Lal. 2008. No-tillage and soil-profile carbon sequestration: an on-farm assessment. *Soil Science Society of America Journal* 72:693–701.

Breiman, L. 2001. Random forests. *Machine Learning* 45:5–32.

Burke, I. C., C. M. Yonker, W. J. Parton, C. V. Cole, D. S. Schimel, and K. Flach. 1989. Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils. *Soil Science Society of America Journal* 53:800–805.

Charman, P. E. V., and B. W. Murphy, editors. 2007. *Soils their properties and management*, Third edition. Oxford University Press, South Melbourne, Victoria, Australia.

Chen, G., and R. Weil. 2010. Penetration of cover crop roots through compacted soils. *Plant and Soil* 331:31–43.

Don, A., J. Schumacher, and A. Freibauer. 2011. Impact of tropical land-use change on soil organic carbon stocks – a meta-analysis. *Global Change Biology* 17:1658–1670.

Don, A., C. Rödenbeck, and G. Gleixner. 2013. Unexpected control of soil carbon turnover by soil carbon concentration. *Environmental Chemistry Letters* 11:407–413.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77:802–813.

Eusterhues, K., C. Rumpel, and I. Kögel-Knabner. 2007. Composition and radiocarbon age of HF-resistant soil organic matter in a Podzol and a Cambisol. *Organic Geochemistry* 38:1356–1372.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29:1189–1232.

Gaudinski, J. B., S. E. Trumbore, E. A. Davidson, and S. Zheng. 2000. Soil carbon cycling in a temperate forest: radiocarbon-based estimates of residence times, sequestration rates and partitioning of fluxes. *Biogeochemistry* 51:33–69.

Groemping, U. 2006. Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software* 17:1–27.

Guo, L. B., and R. M. Gifford. 2002. Soil carbon stocks and land use change: a meta analysis. *Global Change Biology* 8:345–360.

Hilinski, T. E. 2001. *Implementation of exponential depth distribution of organic carbon in the CENTURY model*. Department of Soil and Crop Science, Colorado State University, Fort Collins, Colorado, USA.

Hobley, E., G. R. Willgoose, S. Frisia, and G. Jacobsen. 2013. Environmental and site factors controlling the vertical distribution and radiocarbon ages of organic carbon in a sandy soil. *Biology and Fertility of Soils* 49:1015–1026.

Hobley, E., G. R. Willgoose, S. Frisia, and G. Jacobsen. 2014. Stability and storage of soil organic carbon in a heavy-textured Karst soil from south-eastern Australia. *Soil Research* 52:476–482.

Hobley, E., B. Wilson, A. Wilkie, J. Gray, and T. Koen. 2015. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. *Plant and Soil* 390:111–127.

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics* 15:651–674.

Hudson, B. D. 1994. Soil organic matter and available water capacity. *Journal of Soil and Water Conservation* 49:189–194.

Isbell, R. 2002. *The Australian soil classification*, Revised edition. CSIRO Publishing, Collingwood, Victoria, Australia.

Janzen, H. H. 2005. Soil carbon: a measure of ecosystem response in a changing world? *Canadian Journal of Soil Science* 85:467–480.

Jobbagy, E. G., and R. B. Jackson. 2000. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecological Applications* 10:423–436.

Krull, E. S., J. A. Baldock, and J. O. Skjemstad. 2003. Importance of mechanisms and processes of the stabilisation of soil organic matter for modelling carbon turnover. *Functional Plant Biology* 30:207–222.

Lal, R. 1997. Residue management, conservation tillage and soil restoration for mitigating greenhouse effect by CO_2 enrichment. *Soil & Tillage Research* 43:81–107.

Lal, R. 2004. Soil carbon sequestration impacts on global climate change and food security. *Science* 304:1623–1627.

Luo, Z., E. Wang, and O. J. Sun. 2010. Soil carbon change and its responses to agricultural practices in Australian agro-ecosystems: a review and synthesis. *Geoderma* 155:211–223.

Lützow, M., and I. Kögel-Knabner. 2009. Temperature sensitivity of soil organic matter decomposition – what do we know? *Biology and Fertility of Soils* 46:1–15.

Michaletz, S. T., D. Cheng, A. J. Kerkhoff, and B. J. Enquist. 2014. Convergence of terrestrial plant production across global climate gradients. *Nature* http://dx.doi.org/10.1038/nature13470.

Murphy, B., A. Rawson, L. Ravenscroft, M. Rankin, and R. Millard. 2003. *Paired site sampling for soil carbon estimation – New South Wales*. Australian Greenhouse Office, Canberra, Australian Capital Territory, Australia.

Passioura, J. 1991. Soil structure and plant growth. *Australian Journal of Soil Research* 29:717–728.

Post, W. M., W. R. Emanuel, P. J. Zinke, and A. G. Stangenberger. 1982. Soil carbon pools and world life zones. *Nature* 298:156–159.

Powlson, D. S., A. P. Whitmore, and K. W. T. Goulding. 2011. Soil carbon sequestration to mitigate climate change: a critical re-examination to identify the true and the false. *European Journal of Soil Science* 62:42–55.

R Core Team. 2014 *R: a language for statistical computing, v. 3.1.2*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C., M. S. Torn, and R. J. Southard. 2005. Mineral assemblage and aggregates control carbon dynamics in a california conifer forest. *Soil Science Society of America Journal* 69:1711–1721.

Ridgeway, G. 2013. *gbm: generalized boosted regression models, v. 2.1*.

Ruehlmann, J., and M. Körschens. 2009. Calculating the effect of soil organic matter concentration on soil bulk density. *Soil Science Society of America Journal* 73:876–885.

Rumpel, C., and I. Kögel-Knabner. 2011. Deep soil organic matter – a key but poorly understood component of terrestrial C cycle. *Plant and Soil* 338:143–158.

Rumpel, C., I. Kögel-Knabner, and F. Bruhn. 2002. Vertical distribution, age, and chemical composition of organic carbon in two forest soils of different pedogenesis. *Organic Geochemistry* 33:1131–1142.

Russell, J. 1960. Soil fertility changes in the long-term experimental plots at Kybybolite, South Australia. I. Changes in pH total nitrogen, organic carbon, and bulk density. *Australian Journal of Agricultural Research* 11:902–926.

Sierra, C. A., S. E. Trumbore, E. A. Davidson, S. Vicca, and I. Janssens. 2015. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. *Journal of Advances in Modeling Earth Systems* 7:335–356. http://dx.doi.org/10.1002/2014ms000358.

Six, J., R. T. Conant, E. A. Paul, and K. Paustian. 2002. Stabilization mechanisms of soil organic matter: implications for C-saturation of soils. *Plant and Soil* 241:155–176.

Strobl, C., A. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics* 8:25.

Swift, R. S. 2001. Sequestration of carbon by soil. *Soil Science* 166:858–871.

Toosi, E. R., J. P. Schmidt, and M. J. Castellano. 2014. Soil temperature is an important regulatory control on dissolved organic carbon supply and uptake of soil solution nitrate. *European Journal of Soil Biology* 61:68–71.

Torn, M. S., S. E. Trumbore, O. A. Chadwick, P. M. Vitousek, and D. M. Hendricks. 1997. Mineral control of soil organic carbon storage and turnover. *Nature* 389:170–173.

Viscarra Rossel, R. A., C. Chen, M. Grundy, R. Searle, D. Clifford, and N. Odgers. 2014. *Soil and landscape grid Australia-Wide 3D Soil Property Maps – (3″resolution) – release 1. v2*. CSIRO, http://www.clw.csiro.au/aclep/soilandlandscapegrid/index.html.

Wang, S., M. Huang, X. Shao, R. A. Mickler, K. Li, and J. Ji. 2004. Vertical distribution of soil organic carbon in China. *Environmental Management* 33:S200–S209.

Wiesmeier, M., F. Barthold, P. Spörlein, U. Geuß, E. Hangen, A. Reischl, B. Schilling, G. Angst, M. von Lützow, and I. Kögel-Knabner. 2014a. Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). *Geoderma Regional* 1:67–78.

Wiesmeier, M., R. Hübner, P. Spörlein, U. Geuß, E. Hangen, A. Reischl, B. Schilling, M. von Lützow, and I. Kögel-Knabner. 2014b. Carbon sequestration potential of soils in southeast Germany derived from stable soil organic carbon saturation. *Global Change Biology* 20:653–665.

Wiesmeier, M., M. V. Lützow, P. Spörlein, U. Geuß, E. Hangen, A. Reischl, B. Schilling, and I. Kögel-Knabner. 2015. Land use effects on organic carbon storage in soils of Bavaria: the importance of soil types. *Soil and Tillage Research* 146(Part B):296–302.

Wilson, B. R., and V. E. Lonergan. 2013. Land-use and historical management effects on soil organic carbon in grazing systems on the Northern Tablelands of New South Wales. *Soil Research* 51:668–679.

Wilson, B. R., I. Growns, and J. Lemon. 2008. Land-use effects on soil properties on the north-western slopes of New South Wales: implications for soil condition assessment. *Soil Research* 46:359–367.

Wilson, B. R., T. B. Koen, P. Barnes, S. Ghosh, and D. King. 2011. Soil carbon and related soil properties along a soil type and land-use intensity gradient, New South Wales, Australia. *Soil Use and Management* 27:437–447.

Wynn, J. G., M. I. Bird, and V. N. L. Wong. 2004. Rayleigh distillation and the depth profile of $^{13}C/^{12}C$ ratios of soil organic carbon from soils of disparate texture in Iron Range National Park, Far North Queensland, Australia. *Geochimica et Cosmochimica Acta* 69:1961–1973.

Supporting Information

Additional Supporting Information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecs2.1214/supinfo