Soil organic carbon fractions in the Great Plains of the United States: an application of mid-infrared spectroscopy

Jonathan Sanderman · Jeffrey A. Baldock · Shree R. S. Dangal · Sarah Ludwig · Stefano Potter · Charlotte Rivard · Kathleen Savage

Received: 26 June 2020 / Accepted: 6 January 2021 / Published online: 2 February 2021 © The Author(s) 2021

Abstract Spectroscopy is a powerful means of increasing the availability of soil data necessary for understanding carbon cycling in a changing world. Here, we develop a calibration transfer methodology to appropriately apply an existing mid infrared (MIR) spectral library with analyte data on the distribution of soil organic carbon (SOC) into particulate (POC), mineral-associated (MAOC), and pyrogenic (PyC) forms to nearly 8000 soil samples collected in the Great Plains ecoregion of the United States. We then use this SOC fraction database in combination with a machine learning-based predictive soil mapping approach to explore the controls on the distribution of fractions through soil profiles and across the region.

The relative abundance of each fraction had unique depth distribution profiles with POC fraction dropping exponentially with depth, the MAOC fraction having a broad distribution with a maxima at 35–50 cm, and the PyC fraction showed a slight subsurface maxima (10–20 cm) and then a steady decline with increasing depth. Within the Great Plains ecoregion, clay content was a strong control on the total amount and relative proportion of each fraction in both the surface and subsoil horizons. Sandy soils and soils in cool semi-arid regions contained significantly more POC relative to the MAOC and PyC fractions. Cultivated soils had significantly less SOC than grassland soils with losses following a predictable pattern: POC > MAOC > PyC. This SOC fraction database and resulting maps can now form the basis for improved representation of SOC dynamics in biogeochemical models.

Keywords Particulate carbon · Pyrogenic carbon · Mineral stabilization · Chemometrics · Carbon cycling · Predictive soil mapping

Introduction

Soil organic carbon (SOC) is a large and not very well constrained pool in carbon cycle models. Earth System Models (ESM) included in the CMIP5 study estimated global SOC stocks to range from 510 to 3040 PgC

Responsible Editor: William R. Wieder.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10533-021-00755-1.
Todd-Brown et al. (2013) with estimated changes under a high climate forcing scenario ranging from a loss of 72 PgC to a gain of 253 PgC over the twenty-first century (Todd-Brown et al. 2014). While there are numerous reasons for model-data disagreement, the lack of spatially comprehensive data on soil properties relevant to carbon stabilization and loss was recently cited as one of the most pressing research needs to improve SOC dynamics in ESMs (Luo et al. 2016). In particular, Luo et al. (2016) reported that developing databases with soil carbon pool information was the most commonly cited data limitation to ESM improvement.

Given that SOC is composed of a diverse set of organic compounds with greatly differing susceptibilities to microbial decomposition (Schmidt et al. 2011; Lehmann and Kleber 2015), knowledge of only the amount of SOM is insufficient to understand SOC response to environmental and anthropogenic disturbance (Lavallee et al. 2020). Lavallee et al. (2020) proposed that the division of SOC into particulate and mineral-associated organic carbon (POC and MAOC) results in two fundamentally different pools of SOC. POC is composed primarily of recognizable fragments of plant detritus with typical residence times of < 5 years and can vary from only a few percent to upwards of 50% of total SOM depending on soil texture and management history (Cambardella and Elliott 1992; Baldock et al. 2013a). POC likely provides for much of the energy demands of the soil microbial community but its role in nutrient provision will vary depending on carbon-to-nutrient ratios (Janzen et al. 1992; Gregorich et al. 2006). Additionally, POC is considered to be a nexus of aggregate formation (Six et al. 2004), especially in sandy-textured soils, and therefore can contribute significantly, albeit indirectly, to various soil structural properties. On the other hand, MAOC typically makes up the majority of SOC (von Lützow et al. 2007; Baldock et al. 2013b) and is composed primarily of highly decomposed material, often microbial in origin (Miltner et al. 2012) and is stabilized by association with reactive mineral surfaces (Schmidt et al. 2011). MAOC with typical low carbon-to-nutrient ratios represents the major reservoir of potentially plant-available nutrients (Cambardella and Elliott 1994), contributes the majority of the negative charge attributable to SOC (Kleber and Johnson 2010), and can be an important binding agent between clay particles.

Unfortunately, methodological approaches to dividing SOC into POC and MAOC by density separation or size fractionation fail to recognize that upwards of 20–50% of the total SOC pool can be composed of combustion residues of wild fires (Bird et al. 2015; Reisser et al. 2016). This pyrogenic carbon (PyC) is found distributed across both the POC and MAOC fractions (Baldock et al. 2013a; Leifeld et al. 2015). The PyC fraction has low nutritive value (Deluca et al. 2015), low energy yield and is thus slowly degraded (Liang et al. 2008), but has two important physical properties: significant microporosity and high negative charge (Kookana et al. 2011). The microporosity of PyC helps provide important habitat and refugia for soil microbes leading to greater functional diversity (Atkinson et al. 2010; Lehmann et al. 2011); while the negative charge has been shown to contribute greatly to cation exchange capacity (Liang et al. 2006), plant-water availability (Basso et al. 2013) and to reduced nutrient leaching (Yao et al. 2012).

Measuring SOC fractions clearly provide critical information on soil biogeochemical cycles but the work is extremely laborious especially when the PyC pool is also isolated (Baldock et al. 2013b). Diffuse reflectance mid infrared (MIR) spectroscopy has been used to successfully predict the distribution of SOC into fractions (Zimmermann et al. 2007; Baldock et al. 2013a; Knox et al. 2015) thereby greatly increasing the spatial coverage of SOC fraction data. Ahmed et al. (2017) used the Baldock et al. (2013b) models to explore the variation in POC and PyC across four western US states. Viscarra Rossel et al. (2019) also took advantage of the ability to model SOC fractions from MIR spectra and applied the calibration models developed by Baldock et al. (2013b) to ca. 6000 samples with MIR spectra to develop an understanding of how SOC fractions vary across the Australian continent.

As part of a larger investigation into land use and climate impacts on carbon cycling in the Great Plains ecoregion of the United States, here we present an application of efficient knowledge transfer using MIR spectroscopy and digital soil mapping in order to explore trends in the fractional allocation of SOC into particulate, mineral-associated and pyrogenic carbon pools across this broad geographic area. First, we document the spectroscopic and statistical developments necessary for applying existing carbon fraction
models to a different soil MIR spectral library. In particular, we focus on calibration transfer between FTIR instruments and extension of the training set through spiking with locally-representative samples. Second, we apply these spectroscopy model developments to understand the variation in SOC fractions across the Great Plains ecoregion of the United States.

Methods

Site and data

The Great Plains ecoregion (Omernik and Griffith 2014) of the United States, covering almost 2.1 M km² across 11 States, was selected because of the strong climatic gradients (Table S2) but small variation in soil parent material and native vegetation and the mixture of land uses. With the exception of some localized river deposits, most soils in the Great Plains have formed on deep eolian and loess deposits or glacial till left behind as the Laurentide ice sheet retreated 12,000 years ago. Native vegetation in the Great Plains was dominated by prairie grasses trending from tallgrass to shortgrass assemblages along increasing aridity gradients from east to west. Most of the former tallgrass prairie has been converted to cropland over the past 150 years with corn, soy and wheat being the three dominant crops of the region. Much of the mixed-grass and shortgrass prairie is grazed by livestock. The USDA National Soil Survey Center-Kellogg Soil Survey Laboratory (NSSC-KSSL) has collected MIR spectra and associated laboratory data on nearly 14,000 soil samples from this region. About 8500 of these samples from 1565 soil profiles had associated geolocation data (Fig. S1).

The Australian soil samples and spectra available for SOC fraction model building were from a national survey of Australian agricultural soils reported in Baldock et al. (2013a, b) plus an additional 253 soil samples that were analyzed subsequent to those publications. These 565 soil samples had undergone a full fractionation procedure, outlined below, to determine the content of particulate, mineral associated and resistant carbon fractions.

Carbon fractionation

In total 99 new samples were fractionated using the full fractionation protocol as described in Baldock et al. (2013a). These samples were selected to represent the spectral diversity expected to be encountered across the Great Plains ecoregion of the United States. Specifically, a Principal Components Analysis was conducted on baseline transformed MIR spectra of all samples from the Great Plains ecoregion contained in the NSSC-KSSL spectral library. Then, a set of 100 samples were selected to best represent the diversity in the full dataset using the Kennard-Stone algorithm (Kennard and Stone 1969). The final set of samples comes from 11 states spanning 18 degrees in latitude and 20 degrees in longitude (Fig. S1). The median mid-point sample depth was 38 cm (25–75% quartile = 10–71 cm), median SOC and inorganic carbon (IC) concentrations were 0.9% (25–75% quartile = 0.5–2.3%) and 0.5% (25–75% quartile = 0–1.8%), respectively.

The fractionation procedure (Baldock et al. 2013a) separates fine earth (< 2 mm) total organic carbon (TOC) into a particulate (POC) fraction, defined as organic material > 50 μm excluding char-like material, a mineral-associated (MAOC) fraction, defined as organic material < 50 μm excluding char-like material, and a pyrogenic (PyC) fraction defined as the organic material consisting of poly-condensed aromatic carbon determined by solid-state C-13 nuclear magnetic resonance (NMR) spectroscopy. We have adopted the MAOC terminology here while acknowledging there is no way to know if SOC in the fine size fraction is in fact mineral associated and most recent studies using the term MAOC have not distinguished between biogenic and pyrogenic C within this fraction. In brief, 10 g of the fine earth fraction was dispersed in a 5 g L⁻¹ sodium hexametaphosphate solution, then separated into a coarse and fine size fraction using an automated wet sieve shaker (Fritsch Analysette 3 Spartan). Size fractions were freeze dried and organic carbon was determined on each size fraction by elemental analysis (Elementar Vario Max Cube) after removal of carbonates using sulfurous acid as necessary. Then, the size fractionation was repeated on 10–50 g of fine earth fraction depending on carbon content of fractions. These second size fractions were prepared for solid-state C-13 NMR spectroscopy by removal of minerals. For the coarse size fraction, the
organics were floated off the sand in a large evaporating dish. For the fine size fraction, dilute (2%) hydrofluoric acid was used for demineralization and removal of paramagnetic interferences following the protocol of Skjemstad et al. (1994). NMR analyses were performed on the size fractions using a Bruker Avance 200 spectrometer equipped with a 4.7 T wide-bore superconducting magnet allowing the use of 7 mm diameter rotors which could contain up to 400 mg of sample. Cross polarization (CP) experimental parameters were optimized for each sample as described in Baldock et al. (2013a).

Not enough coarse fraction organic material was recovered for NMR on nine samples. Another 13 coarse fractions had unacceptably low signal from the CP NMR experiments so these NMR data were excluded from further analysis. For these 22 samples, we assumed that there was no resistant carbon in the coarse fraction. This is a conservative assumption and will not impact much upon the carbon mass balance since for these samples the great majority of carbon was found in the fine fraction. In fact, for these 22 samples, 91% of TOC was found, on average, in fine fraction. Good quality NMR spectra were acquired on all 99 fine fraction samples (Fig. S2). On an additional subset of fine (n = 20) and coarse (n = 7) fraction samples, direct polarization (DP) experiments were also performed to improve the relationship quantifying the underrepresentation of PyC in CP experiments (Eq. 12 in Baldock et al. 2013a). After combining these new DP data with existing Australian DP data on fractions, the new best fit for CP observability ($\phi_{CP}$) was 0.552 ($\phi_{CP} = 0.510$ in Baldock et al. 2013a) (Figure S3). This new $\phi_{CP}$ value was then applied to all US and Australian fraction data to come up with final carbon concentrations in fractions.

MIR measurement, model development and prediction

All Australian samples were finely milled and analyzed on a Thermo Nicolet 6700 FTIR spectrometer with a Pike autodiff diffuse reflectance accessory. Sixty co-added scans were acquired from 8000 to 400 cm$^{-1}$ at a resolution of 8 cm$^{-1}$. Full operating conditions can be found in Baldock et al. (2013b). The 99 samples from the Great Plains were also analyzed using this spectrometer. The Great Plains samples, after milling under similar conditions, were all scanned on a Bruker Vertex 70 with a HTS-XT high throughput diffuse reflectance accessory at the KSSL in Lincoln NE. At the KSSL, all samples were scanned in quadruplicate with spectra recorded from 6000 to 400 cm$^{-1}$. Details can be found in Dangal et al. (2019). A set of 285 Australian soils (out of the 565 available with fraction data) were also scanned at the KSSL.

A memory-based learning (MBL) model developed by Ramirez-Lopez et al. (2013) as modified by Dangal et al. (2019) was used for all chemometrics. The MBL is a local or just-in-time modeling approach where a small number of most spectrally similar samples are chosen from the training set to build a model appropriate for each sample to be predicted. This parsimonious modeling approach eliminates the need for a priori subsetting of the training data by soil type or depth or any other factor and has been shown to outperform other modeling approaches on large spectral libraries (Dangal et al. 2019). In this study, we used a local partial least square regression (L-PLSR), where a target sample is predicted by performing the local cross validation against the user specified minimum (n = 10) and maximum (n = 100) number of spectrally similar neighbors. The spectrally similar neighbors corresponding to the target sample were selected using the Mahalanobis dissimilarity computed on the principal component space. We use local L-PLSR because the only parameter that needs to be optimized is the number of principal components, which is done by performing cross-validation at each local segment. Spectra were baseline transformed prior to model fitting. All fraction data were square root transformed prior to modeling but back-transformed before reporting goodness-of-fit statistics. Prior to model building an outlier screening step (Dangal et al. 2019) was employed where a PLSR model was developed to predict each fraction using all 585 + 99 samples scanned at CSIRO and extreme outliers from the 1:1 line were removed. Five of the 99 Great Plains samples were identified as outliers in all models and thus excluded from further model development. Given the limited number of fully fractionated new Great Plains soil samples, models were evaluated using leave-one-out (LOO) cross validation instead of using an independent test set for validation.

First, we tested whether or not the original Australian training set could make accurate predictions of carbon fractions on the US soils. Next, we
tested whether or not a model from just the newly
fractionated US samples would perform well. Since
there were only 94 samples, we applied a partial least
squares regression model for each carbon fraction.
Finally, the 94 fractionated soils from the Great Plains
ecoregion were added to a database of 565 Australian
soils that have undergone the full fractionation pro-
dure. Using the MBL model with the 565 Australian
soils as the training set, we found variable success in
predicting the measured fractions on the 94 GP
samples (Table 1). After spiking the training set with
the GP samples, predictions of all fractions improved
greatly suggesting (Table 1) that the inclusion of the
local samples is necessary for accurate predictions.
Additionally, for all three fractions, the spiked model
outperformed the US-only model reducing the RMSE
by an average of 25% (Table 1).

The second technical hurdle that needed to be
addressed in order to use the Australian training set
was to account for the fact that the NSSC-KSSL
spectral library was acquired on a different instrument
than the fraction model training set. The relative
spectral response (y-axis) is known to vary slightly
between optical benches for MIR measurements and
various procedures have been proposed for trans-
ferring calibration models between instruments (Ge et al.
2011). Using the 285 AU and 94 US samples that were
scanned on both instruments, a calibration transfer
matrix was developed on first-derivative transformed
spectra using the Piecewise Direct Standardization
(PDS) approach (Bouveresse et al. 1996). Optimal
transfer was achieved with a window size of three and
retention of only one principal component. Table 2
indicates that good (i.e. high $R^2$) but slightly biased fits
could be achieved without PDS, and that PDS always
improved model fit with lower RMSE values for all
carbon fractions.

Based on these results, we moved ahead with
predictions on all GP samples that included GPS
coordinates ($n = 8500$) using the MBL model on PDS-
transformed MIR spectra. Given we are now applying
this model to a large number of samples without any
further validation data, we needed some ways of
detecting potential outliers. First, we employed an
F-ratio test (Hicks et al. 2015) that identified spectra
that fell outside of 95% of the calibration space. These
samples may predict well but we would have little
confidence in the predictions because they are unique
to our calibration set. This screening resulted in the
exclusion of 164 samples. For comparison, when the
US-only model was applied to these 8500 samples,
571 samples were identified as F-ratio outliers.

Next, we compared the sum of the predicted
fractions (POC + MAOC + PyC) to the measured
TOC values. From this plot (Fig S6), it became clear
that our predictions broke down for samples with
greater than 100 g C kg$^{-1}$ coinciding with the highest
measured TOC value in our calibration set. These high
SOC samples ($n = 488$) were then excluded from
further analysis. A small number of samples ($n = 72$)
showed extreme mismatches between the sum of
fractions and observed TOC data. These mismatches
may have arisen from laboratory errors in analyzing
the wrong sample for TOC or FTIR analysis or were
just really poorly predicted. Either way, these 72
samples would have too much leverage on model
results and have thus been excluded. After removing
these 724 outliers from the 8500 total samples, the
predicted sum of fractions showed an excellent
 correspondence with measured TOC data (Fig. S6).

| Australian model$^a$ | Spiked model (AU + US)$^b$ | US only model$^c$ |
|----------------------|--------------------------|-----------------|
|                      | Bias | $R^2$ | RMSE | Bias | $R^2$ | RMSE | Bias | $R^2$ | RMSE |
| POC                  | 1.57 | 0.72  | 4.62 | 0.72 | 0.88  | 2.90 | 0.79 | 0.76  | 4.01  |
| MAOC                 | 0.93 | 0.79  | 3.59 | 0.47 | 0.88  | 2.65 | 0.26 | 0.82  | 3.16  |
| PyC                  | −2.64| 0.49  | 4.60 | −0.18| 0.89  | 1.37 | −0.04| 0.78  | 1.85  |

$^a$Memory based learning with test set validation (94 GP samples)

$^b$Memory based learning with leave-one-out cross validation on 94 GP samples

$^c$Partial least squares regression with leave-one-out cross validation on 94 GP samples
Predicted carbon fraction concentrations (POC, MAOC, PyC) and proportion of TOC in each fraction (fPOC, fMAOC, fPyC) calculated as fraction divided by the sum of fractions were then compared to soil and environmental data compiled on all samples. Soil data from the NSSC-KSSL Soil Characterization Database included sample depth, soil organic carbon content, clay content, pH and bulk density (BD). Analytical protocols for data generated by the NSSC-KSSL are detailed elsewhere (Soil Survey Staff 2014). All samples had measured SOC data, but gaps needed to be filled for other properties. MIR-based predictions, as developed by Dangal et al. (2019) and Sanderman et al. (2020), were used to estimate clay content for 277 samples, pH on 170 samples and BD on 3782 samples. Soil classification to the Great Group level was utilized from Hengl et al. (2017). Topographic variables including elevation, slope, topographic wetness index (TWI) and topographic position index (TPI) were extracted from CGIAR Consortium for Spatial Information (Reuter et al. 2007) and depth to bedrock was used from Pelletier et al. (2016). Climatic variables included mean monthly precipitation, temperature, solar radiation, vapor pressure deficit from WorldClim (Fick and Hijmans 2017), cloud cover from EarthEnv (Wilson and Jetz 2016), and annual snow cover NSIDC (Hall et al. 2016). MODIS-derived enhanced vegetative index (EVI) from USGS Earth Data (Didan et al. 2015). Management covariates included fertilization and irrigation area from USGS National Census of Agriculture (Falcone et al. 2016) and cattle livestock density from Gridded Livestock of the World, version 3 (Gilbert et al. 2018). Land cover classification for 2015 was used from European Space Agency’s Climate Change Initiative Land Cover version 2.0.7 product. Full list of covariate layers with links to data sources can be found in Table S1. Covariate layers were downscaled to 250 m, and transformed to a common projection prior to stacking rasters for random forest analysis. Any missing data were gap-filled using gdal_fillnodata.py.

Basic data exploration included plotting trends with depth and subsetting data into surface (0–20 cm) and subsurface (40–70 cm) groups to explore univariate relationships with covariates. Fixed depths were chosen because only 45% of the samples had information on genetic horizons. For the samples that had information on genetic horizon, 96% of the 0–20 cm samples were designated as A horizon and 92% of the 40–70 cm samples were designated as B horizon. Histograms of the maximum of fPOC, fMAOC and fPyC for each soil profile were constructed on full soil profiles (n = 1398) to better visualize the depth distribution of carbon fractions. Linear, semi-logarithmic (log-lin and lin-log) and power (log–log) regression models were applied to depth trends for all fraction data and for trends with clay for fraction data. Differences between major land cover classes (crop-land and grassland) was assessed for each fraction using one-way ANOVA after square root transformation of data.

To further explore the non-linear multivariate relationships that are known to control soil carbon distribution at larger spatial scales, quantile random forest (QRF) models (Meinshausen 2006) were built for TOC and all fractions. In order to find optimal tuning parameters in the QRF models a fivefold cross validation procedure repeated 100 times for each of the separate models was implemented. For each model, 500 trees were used and a search grid for the number of variables sampled at each tree split was searched across of a grid of 5, 10, 20, 30, 40 and 50.

Table 2 Predictive performance of Great Plains samples using spectra acquired on KSSL instrument with and without calibration transfer

|                     | No calibration transfer | After calibration transfer |
|---------------------|-------------------------|---------------------------|
|                     | Bias  | R²    | RMSE | Bias  | R²    | RMSE |
| POC (g/kg)          | 0.42  | 0.70  | 4.04 | 0.45  | 0.78  | 3.58 |
| MAOC (g/kg)         | 0.30  | 0.84  | 3.09 | 0.26  | 0.85  | 2.93 |
| PyC (g/kg)          | 0.19  | 0.71  | 2.10 | 0.04  | 0.73  | 2.01 |

*aLeave-one-out cross validation on the 94 GP samples
Within this cross-validation procedure the model which resulted in the largest $R^2$ was selected. After optimal tuning parameters were selected, the final model performance was evaluated by randomly sampling 80% of the field sites used for model training, and 20% of the sites used for model validation. Lastly, across each of the eight dependent variables modeled, spatial predictions were acquired at seven different soil depths (1, 3.5, 7.5, 15, 25, 37.5, and 67.5 cm), and at each depth three quantiles were predicted to get a measure of per pixel level uncertainty corresponding to 2.5, 50 and 97.5% confidence.

**Results**

**Depth trends**

Combining all site data, there was a strong decrease in concentration of each fraction (Fig. 1a–c) with POC decreasing relatively faster than MAOC and PyC (Fig. 1d, e). Power functions best described the decrease in POC ($R^2 = 0.42$) and fPOC ($R^2 = 0.19$) with depth, while log-lin functions best described the relationship between MAOC ($R^2 = 0.32$), fMAOC ($R^2 = 0.11$) and PyC ($R^2 = 0.21$) with depth. When all sites were grouped together, there was no trend in fPyC with depth.

The maximum proportion of each fraction was found at different depths within the soil (Fig. 2). The POC fraction peaked in the topsoil (0–10 cm), whereas, the MAOC fraction peaked around 30–50 cm. While masked when all the sites are grouped together (i.e. Fig. 1), fPyC displayed, on average, a slight subsurface maxima around 10–20 cm when examining profiles individually (Fig. 2).

**Relationships with covariates**

In order to explore univariate and multivariate relationships with covariates, data were grouped into surface (0–20 cm, $n = 2457$) and subsurface (40–70 cm, $n = 1512$) categories to minimize the overall impact of depth on soil properties. Concentration of SOC and fractions of SOC were correlated with climate (Table S2) with lowest concentrations in warm, dry environments (negative correlations with temperature and solar radiation) and highest concentrations in cooler, wetter locations (strong positive association with cloud fraction and snowfall). Correlations with climatic variables were generally much stronger for the surface than the subsurface horizon. The MAOC and PyC fractions were also positively associated with clay content.

The amount (Table S2) and proportional distribution of OC into fractions (Fig. 3) were correlated with clay content. The sandiest samples, regardless of depth, had disproportionately more POC than MAOC or PyC (Fig. 3). Soil pH was also correlated with the fractional distribution of OC with fPyC increasing with increasing pH (Table S3) but this might simply be a reflection of drier climates where pH would be higher. Correlations with edaphic properties were stronger in surface than subsurface (Table S3). Clay and pH were only weakly correlated (data not shown).

Warmer and wetter climate lead to relatively more MAOC and less PyC but POC tended to be independent of climate (Table S3). In contrast to the findings for carbon concentration, climate-driven trends with the proportional distribution of fraction were stronger in the subsurface than surface horizons. Perhaps due to the large regional scale of this investigation, topographic properties were poor predictors (Tables S2, S3).

The concentration of SOC and fractions in the topsoil (0–20 cm) varied by land cover classification (Fig. 4a). Grasslands had the highest concentrations of SOC and all fractions with croplands having the lowest. When all soils were pooled together, there was a difference in POC ($F = 7.5$, $P = 0.006$), a smaller difference in MAOC ($F = 2.9$, $P = 0.09$) and no difference in PyC. The relative difference between land cover classes was more dramatic when the analysis was limited by soil type with all three fractions showing highly significant differences ($P < 0.001$). For the most common Great Group (Argiustolls, $n = 650$), there was 33% less POC, 29% less MAOC and 25% less PyC in cropland than grassland samples (Fig. 4b). Given the variability across this large region, the tendency for a preferential loss of POC over other fractions did not change the average distribution of the fractions (data not shown).

**Random forest modeling**

The spatial distribution of SOC and carbon fractions were well predicted using a quantile random forest model applied to a stack of spatially continuous co-
Models produced minimal bias, a slope close to unity, $R^2$ values between 0.66 and 0.76 and RPD values around 2.3 for all properties (Table 3). Sampling depth was always the most important variable in the QRF models (Fig. S8). As a singular variable, sampling depth explained 43, 32, and 21% of the variance in POC, MAOC and PyC, respectively (Fig. 1). After accounting for sampling depth, the importance of the remaining variables varied depending on fraction. Averaged across months, temperature was the second most important variable for POC, whereas MODIS cloud fraction and solar radiation were the second and third most important variables for the MAOC and PyC fractions. Vapor pressure deficit was an important variable for POC but not MAOC or PyC. Monthly precipitation consistently ranked as some of the least important variables. Unfortunately, land cover classification and soil classification could not be included in our random forest models because not all categories present across the Great Plains were represented in the training data.

After projecting the random forest model results across the Great Plains ecoregion, large variation in the stocks of the three fractions were observed (Fig. 5). Associated uncertainty is shown in Fig. S9. In general, there was a trend of increasing amount of all fractions moving from the hot arid southwest to the cool mesic northeast with some interesting exceptions. The cool and dry northwestern edge of the Great Plains contained nearly as much POC as the mesic northeast. The Sandhills region of central Nebraska also stood out as having disproportionately low MAOC and PyC stocks relative to POC. The Sandhills region also stands out as have much greater uncertainty (Fig. S9).
Discussion

MIR spectroscopy

In this study, we took advantage of the ability to predict soil properties from easily obtained MIR spectra in order to greatly expand data availability and thus scope of inference (Nocita et al. 2015). In particular, we demonstrated that a moderately sized dataset (n = 565) of carbon fractions with appropriate processing steps could be successfully used to make predictions of carbon fractions on nearly 8000 soil samples from a different region of the world. Two different technical challenges needed to be addressed in the process of making these predictions.

Fig. 2 Depth distribution of the maximum occurrence of the proportional allocation of organic carbon to each fraction—particulate (fPOC), mineral associated (fMAOC) and pyrogenic (fPyC)—for each profile. The x-axis gives the count of the location of the maxima for each fraction in 2.5 cm bins. 1398 full profiles were included in this analysis.

Fig. 3 Relationship between clay content and proportional allocation of organic carbon to particulate (POC), mineral associated (MAOC) and pyrogenic (PyC) fractions for topsoil samples (a, c, e) and subsoil samples (b, d, f). Best fit linear, log–lin or log–log functions and goodness-of-fit (R² values) also shown.
First, the existing carbon fraction database came from Australian soils that were not necessarily representative of the types of soils encountered in the Great Plains ecoregion of the United States. In order to assess the magnitude of this challenge and provision data to resolve the problem, we fractionated 100 soil samples.

**Fig. 4** Mean concentration of particulate (POC), mineral associated (MAOC) and pyrogenic (PyC) fractions grouped by land cover classification for all topsoil (0–20 cm) samples (a) and for the most common soil type (Argiustolls) (b). Error bars represent 2 standard errors. Significance differences in fractions between land cover classes given as: ns not significant, *P < 0.1, **P < 0.01, ***P < 0.001

**Table 3** Random forest model performance for soil organic carbon (SOC) and carbon fractions

| Variable   | n  | Bias | Slope | Intercept | $R^2$ | RMSE $^b$ | RPD $^c$ |
|------------|----|------|-------|-----------|-------|-----------|----------|
| SOC (g kg$^{-1}$) | 1602 | 0.082 | 1.076 | 0.013 | 0.74 | 5.10 | 2.32 |
| POC (g kg$^{-1}$) | 1586 | 0.208 | 1.101 | 0.062 | 0.66 | 1.14 | 2.33 |
| MAOC (g kg$^{-1}$) | 1587 | 0.285 | 1.022 | 0.158 | 0.76 | 2.18 | 2.32 |
| PyC (g kg$^{-1}$) | 1586 | 0.219 | 1.059 | 0.073 | 0.73 | 1.25 | 2.21 |

*Number of samples in validation set

$^b$Root mean square error

$^c$Ratio of performance to deviation (SD/RMSE)

**Fig. 5** Predicted median 0–30 cm stocks of the particulate (POC), mineral associated (MAOC) and pyrogenic (PyC) carbon fractions for the Great Plains ecoregion of the United States.
samples from the Great Plains that represented the diversity in soil properties expected to be encountered. From the standpoint of carbon fraction content (Fig. S4) and distribution of SOC into fractions (Fig. S5), the two regions do not appear too different. Additionally, we compared the spectral data from the two regions in Principal Component space and see that there is good overlap (Fig. S7). Despite the similarities between the two sample sets, Table 1 indicated spiking of the Australia spectral dataset with the local Great Plains samples was necessary for really good unbiased predictions, particularly for the PyC fraction. Several other studies have also demonstrated that spiking with local samples is an effective method for applying spectral libraries developed outside the region of interest (Guerrero et al. 2016; Lobsey et al. 2017), although not necessarily for all soil properties (Gogé et al. 2014).

Second, the MIR spectra of the Great Plains soil database were acquired on a different FTIR spectrometer necessitating the need to evaluate calibration transfer methods. Previous work has shown that large differences in the raw spectra for the same samples acquired on the different instruments are reduced following first derivative transformation, and that PDS successfully eliminate most of the variability between the primary and secondary spectra (Dangal and Sanderman 2020). In this work, we found that calibration transfer using piecewise direct standardization did not improve $R^2$ of predictions but was necessary to remove bias in predictions resulting in up to a 50% reduction in RMSE (Table 2). Whether or not calibration transfer is always necessary is still a largely unresolved issue in soil spectroscopy and it may depend upon the specific combination of instruments and soil properties to be predicted (Dangal and Sanderman 2020).

The comparison between the spiked Australian model and the US-only model illustrated that despite any additional uncertainties that may have been introduced through the calibration transfer process the larger training set outperformed the US-only model, primarily for POC and MAOC fractions, both in terms of internal validation and applicability to a wider range of samples when applied to the 8500 samples in the Great Plains database.

Carbon fractions

Knowledge of only the total stock of SOC is likely insufficient for understanding SOC response to perturbations and for designing and reporting on SOC sequestration programs. Carbon fractions can provide a much more nuanced understanding of SOC change (Lavalle et al. 2020; Viscarra Rossel et al. 2019). In general, the faster cycling POC fraction is expected to provide for much of the energy demand of the soil microbial community while the MAOC fraction, through interactions with reactive mineral surfaces, contributes the most to longer time SOC accrual. The PyC, being of pyrogenic origin, interacts with soil microorganisms in a fundamentally different way leading to much greater persistence (Schmidt et al. 2011). The SOC fraction data generated in this study of the Great Plains ecoregion of the United States reveal a number of insights in SOC dynamics that we detail in this section.

Depth trends

An advantage of working with the NSSC-KSSL database is that full soil profiles were collected at nearly all locations allowing for a detailed exploration of the vertical allocation of SOC into fractions. Plotting the proportion of each fraction against depth for all samples revealed a rapid decrease in fPOC, increase in fMAOC but no trend in fPyC (Fig. 1d, e). A different picture emerged when considering each profile as a set of related data (Fig. 2). The frequency distribution of the maximum value for fPOC, fMAOC and fPyC for each profile clearly indicated POC is concentrated in the topsoil (0–10 cm), PyC peaks slightly deeper (5–20 cm), and MAOC reaches a maximum in subsoil horizons (30–50 cm).

The distribution of SOC fractions with depth is revealing and affirms the importance of including vertical transport processes when modeling SOC dynamics (Baisden et al. 2002; Koven et al. 2013) even if only considering the top 20 or 30 cm of soil. The POC fraction is directly derived from recent plant inputs, whether surface deposition of aboveground litter or root turnover, which decreases exponentially with depth, and has a relatively short mean residence time. For most soils, where vertical transport is on the order of a 0.5–2.0 mm year$^{-1}$ (Elzein and Balesdent 1995; Baisden et al. 2002; Kaste et al. 2007), this
means that most POC is mineralized or transformed into more stable MAOC by microbial cycling before moving deeper in the soil profile. The co-location of fMAOC maximum with typical clay maxima can also be explained by the relatively rapid loss of POC, vertical transport of dissolved organic carbon and subsequent retention on mineral surfaces with a greatly reduced mean residence time (Sanderman and Amundson 2008; Kaiser and Kalbitz 2012).

The slight subsurface maxima for the PyC fraction has been seen in other studies where full profiles were analyzed (Hammes et al. 2008; Hobley et al. 2016; Wang et al. 2018) and can also be interpreted in terms of relative production, transport and consumption (Major et al. 2010). The PyC fraction being pyrogenic in origin enters the soil primarily on the surface following wildfire. While a lot of PyC may be transported horizontally (Abney et al. 2017), some PyC will get incorporated into the soil profile through slow mixing processes. Importantly, the PyC fraction is not completely inert and is slowly lost to microbial processing and dissolution on timescales of centuries (Hammes et al. 2008) allowing for transport to much greater depths than POC.

While we have emphasized trends in the relative distribution of fractions in this section, it is important to emphasize that our fractionation methodology has estimated that on average 24% of TOC is consistent with a pyrogenic origin (Fig. 1). Using similar NMR spectroscopy-based methodology, Skjemstad et al. (2002) found that 10–35% of TOC was PyC across a range of US soils. However, estimates of PyC using the benzene polycarboxylic acid (BPCA) biomarker technique and hydrogen pyrolysis have fallen in the range of 4–18% for similar soils (Glaser and Amelung 2003; Lavalle et al. 2019). This discrepancy is to be expected as the NMR technique measures more of the PyC continuum than either of the other two methods (Reisser et al. 2016; Schmidt et al. 2001).

**The role of clay**

Rasmussen et al. (2018), synthesizing a broader selection of data from the NSSC-KSSL database, concluded that clay content alone does not appear to control SOC stabilization. Findings from this study where we focused on a large but relatively homogenous region, do suggest that clay content alone is still important to consider in representing SOC dynamics (Tables S2 and S3). In the topsoil (0–20 cm) and subsoil (40–70 cm), SOC was positively correlated with clay content (r = 0.28 and 0.23 for top and subsoil, respectively). However, a more nuanced picture emerges when looking at the correlations with the individual SOC fractions. Fitting with our conceptual descriptions of the formation and fate of these SOC fractions above, POC was not correlated with clay but MAOC showed a strong positive correlation for both depth increments (r = 0.43–0.44). Comparing the distribution of carbon into fractions with clay content also suggested a strong role for texture in controlling SOC dynamics (Fig. 3) with disproportionately more POC in sandy soils but MAOC and PyC dominate once clay content increases above about 8%. The role of clays and reactive mineral surfaces, generally, in controlling the stabilization of the MAOC fraction is well established (Kögel-Knabner et al. 2008; Kramer et al. 2012; Jastrow and Miller 2018). The PyC fraction was also correlated with clay content for both depths (r = 0.29–0.41); a finding that fits with the growing evidence that the PyC fraction is also retained via mineral association (Cusack et al. 2012; Soucémanianadin et al. 2014).

**Land use patterns**

The data in the NSSC-KSSL database indicate a strong impact of land use on topsoil SOC content. Soils collected from locations classified as cropland had significantly less SOC in the top 20 cm than from soils classified as grassland. When limiting the data to just Argiustolls, the most common Great Group (n = 650), there was 30% less SOC. This level of land use impact on SOC is consistent with other datasets (Guo and Gifford 2002; Sanderman et al. 2017) and is likely an underestimate of the true impact of cropping because most of the cropland sites are in the eastern wetter portions of the study region where we would expect higher SOC levels under native vegetation.

While there was relatively less C in all three fractions in cropping versus grassland sites, there was a relatively greater loss of POC than MAOC and even smaller loss of PyC (Fig. 4). These findings are consistent with observations from paired plots (Poeplau and Don 2013) and long term trials (Skjemstad et al. 2004) where the POC fraction was most sensitive to land use change. While the PyC fraction should be the least susceptible to microbially-mediated loss, our
results suggest this fraction was still susceptible to loss over the approximately 100 year history of agriculture in this region. The most likely explanation for this apparent loss of this stable SOC fraction is the well documented history of accelerated erosion in the Great Plains (DeLuca and Zabinski 2011). Water and rill erosion will remove all three fractions more or less equally (Sanderman and Chappell 2013).

Integration with predictive soil mapping

Using MIR-estimates of difficult to measure soil properties to increase data available to predictive soil mapping efforts is a valuable application of MIR technology (Hengl et al. 2015; Mirzaeitalarposhti et al. 2017; Viscarra Rossel and McBratney 2008). In this work, we have demonstrated how detailed measurements on 565 samples can be utilized to make estimates of these measurements on nearly 8000 samples using MIR spectroscopy. Estimates of carbon fractions for 7776 samples spread across 1482 geolocated soil profiles were then successfully predicted using a machine learning predictive soil mapping approach (Table 3) to arrive at 250 m resolution digital maps of carbon fractions for the Great Plains ecoregion (Fig. 5).

Several recent studies have suggested that the ratio of fast cycling to slower cycling carbon fractions can be used as an index of carbon vulnerability (Baldock et al. 2018; Viscarra Rossel et al. 2019; Gray et al. 2019). Here we have taken the data presented in Fig. 5 and calculated carbon vulnerability ($V_C$) as POC/ (MAOC + PyC), with results presented in Fig. 6. In the Great Plains, we found a broad pattern of increasing vulnerability with increasing latitude (a good proxy for annual temperature). However, in the cooler and dry regions of the Great Plains, carbon vulnerability was also high. The region of sandy soils in Nebraska also stood out as a region of exceptionally vulnerable carbon. We found that vegetation greenness was weakly correlated with total stocks and distribution of fractions (Tables S2 and S3), a finding that contrasted with the work of Ahmed et al. (2017) who found greenness (as normalized difference vegetation index) was the single strongest predictor of PyC and POC stocks for the Wyoming, Colorado, Kansas and New Mexico region. Much of the difference between this present work and that of Ahmed et al. (2017) is likely the fact that we limited our study area to non-forest and non-shrub areas whereas the study region of Ahmed et al. (2017) included cropland, grassland, shrubland and forestland.

In a similar application of mapping MIR-estimated carbon fractions across the Australian continent, Viscarra Rossel et al. (2019) found that the hot semi-arid and arid expanses that cover much of Australia had very low $V_C$ with higher vulnerability along the cooler and wetter southern and eastern rim of the continent. Mapping carbon fractions for just the state of New South Wales in Australia, Grey et al. (2019) found that lithology (siliceous versus mafic) was a strong driver of the relative abundance of fractions with soils derived from mafic parent material containing disproportionately less POC.

Maps, such as presented in Figs. 5 and 6, can form the basis for modeling efforts to better understand climate and land use feedbacks on soil carbon dynamics. For example, Lehmann et al. (2008) demonstrated that SOC model feedback to global warming would overestimate twenty-first century CO$_2$ emissions by about 20% if the PyC pool was ignored. More generally, having accurate estimates of the distribution of SOC into carbon fractions can greatly reduce model uncertainty (Lee and Viscarra Rossel 2020). In fact, due to the greater model confidence when initialized to measurable carbon fractions, soil carbon accounting in Australia’s National Greenhouse Gas Inventory has been using a simulation model calibrated to these carbon fractions for over a decade (Skjemstad and Spouncer 2003; Skjemstad et al. 2004).

Conclusions

The dominant forms that SOC have accumulated in can matter as much as the total stock for predicting SOC response to perturbations. However, the labor and cost of applying a robust fractionation scheme severely limits the number of soil samples than can be fractionated particularly when you are interested in quantifying SOC that is pyrogenic in origin. In this work, we have extended and applied MIR spectroscopy-based predictive models for POC, MAOC and PyC fractions developed in a previous study to approximately 8000 samples from the Great Plains ecoregion of the United States. Finally, spatial covariates were able to capture much of the variance in
the distribution of each fraction, allowing us to develop spatially continuous maps of these fractions for the Great Plains ecoregion. Interpretation of the SOC fraction database and derived spatial maps provided important insights into the vertical (i.e. soil profile) and horizontal (i.e. landscape scale) distribution of SOC fractions. These insights can be used to better parameterize carbon cycle models and the maps can form the basis for initializing soil carbon models to

Fig. 6 Map of total 0–30 cm soil organic carbon stocks (a) and carbon vulnerability (b). Carbon vulnerability is defined as POC/(MAOC + PyC)
better understand land use and climate impacts on SOC cycling.

Acknowledgements The authors would like to thank Rebecca Holloway and Janine McGowan for assistance in the laboratory. We also thank Rich Ferguson and the entire staff at the Kellogg Soil Survey Laboratory for providing archived samples, data and help with the spectral library. Funding for this research was provided by USDA NIFA Award #2017-67003-26481.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Data availability All carbon fraction measurements, predictions and associated covariate data used in analyzing spatial trends and generating maps are available at Harvard Dataverse (https://doi.org/10.7910/DVN/NW8K2P). Maps of soil properties can also be downloaded from the same repository. The USDA NSSC-KSSL soil spectral library is freely available by request from the USDA (contact rich.ferguson@usda.gov). The Australia carbon fraction data and spectra can be requested from CSIRO (contact ryan.farquharson@csiro.au).

References

Abney RB, Sanderman J, Johnson D et al (2017) Post-wildfire Erosion in mountainous terrain leads to rapid and major redistribution of soil organic carbon. Front Earth Sci. https://doi.org/10.3389/feart.2017.00099

Ahmed ZU, Woodbury PB, Sanderman J et al (2017) Assessing soil carbon vulnerability in the Western USA by geospatial modeling of pyrogenic and particulate carbon stocks. J Geophys Res Biogeosciences 122:354–369. https://doi.org/10.1002/2016JG003488

Atkinson CJ, Fitzgerald JD, Hipps NA (2010) Potential mechanisms for achieving agricultural benefits from biochar application to temperate soils: a review. Plant Soil 337:1–18. https://doi.org/10.1007/s11104-010-0464-5

Baisden WT, Amundson R, Brenner DL et al (2002) A multi-isotope C and N modeling analysis of soil organic matter turnover and transport as a function of soil depth in a California annual grassland soil chronosequence. Glob Biogeochem Cycles 16:82–26. https://doi.org/10.1029/2001gb001823

Baldock JA, Hawke B, Sanderman J, MacDonald LM (2013a) Predicting contents of carbon and its component fractions in Australian soils from diffuse reflectance mid-infrared spectra. Soil Res 51:577–595. https://doi.org/10.1071/SR13077

Baldock JA, Sanderman J, Macdonald LM et al (2013b) Quantifying the allocation of soil organic carbon to biologically significant fractions. Soil Res 51:561–576. https://doi.org/10.1071/SR12374

Baldock JA, Beare MH, Curtin D, Hawke B (2018) Stocks, composition and vulnerability to loss of soil organic carbon predicted using mid-infrared spectroscopy. Soil Res 56:468–480. https://doi.org/10.1071/SR17221

Basso AS, Miguez FE, Laird DA et al (2013) Assessing potential of biochar for increasing water-holding capacity of sandy soils. GCB Bioenergy 5:132–143. https://doi.org/10.1111/gcbb.12026

Bird MI, Wynn JG, Saiz G et al (2015) The pyrogenic carbon cycle. Annu Rev Earth Planet Sci 43:273–298. https://doi.org/10.1146/annurev-earth-060614-105038

Bouveresse E, Hartmann C, Massart DL et al (1996) Standardization of near-infrared spectrometric instruments. Anal Chem 68:982–990. https://doi.org/10.1021/ac9510595

Cambardella CA, Elliott ET (1992) Particulate soil organic-matter changes across a grassland cultivation sequence. Soil Sci Soc Am J 56:777. https://doi.org/10.2136/sssaj1992.03615995005600030017x

Cambardella CA, Elliott ET (1994) Carbon and nitrogen dynamics of soil organic matter fractions from cultivated grassland soils. Soil Sci Soc Am J 58:123. https://doi.org/10.2136/sssaj1994.03615995005800010017x

Cusack DF, Chadwick OA, Hockaday WC, Vitousek PM (2012) Mineralogical controls on soil black carbon preservation. Glob Biogeochem Cycles 26:GB2019. https://doi.org/10.1029/2012GB004109

Dangal SRS, Sanderman J (2020) Is standardization necessary for sharing of a large mid-infrared spectral library? Sensors 20:6729. https://doi.org/10.3390/s20236729

Dangal SRS, Sanderman J, Wills S et al (2019) Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst 3:11. https://doi.org/10.1002/3390/soilsystems3010011

Deluca TH, Gundale MJ, Mackenzie D, Jones DL (2015) Biochar effects on soil nutrient transformations. Biochar Environ Manage 2:421–454

DeLuca TH, Zabinski CA (2011) Prairie ecosystems and the carbon problem. Front Ecol Environ 9:407–413. https://doi.org/10.1890/100063

Didan K (2015) MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m Sin Grid V006. 2015, distributed by NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13Q1.006.

Elzein A, Balesdent J (1995) Mechanistic simulation of vertical distribution of carbon concentrations and residence times in soils. Soil Sci Soc Am J 59:1328–1335. https://doi.org/10.2136/sssaj1995.03615995005900050019x

Falcone JA, LaMotte AE (2016) National 1-kilometer rasters of selected Census of Agriculture statistics allocated to land use. Biogeosciences (2021) 156:97–114 111

Bird MI, Wynn JG, Saiz G et al (2015) The pyrogenic carbon cycle. Annu Rev Earth Planet Sci 43:273–298. https://doi.org/10.1146/annurev-earth-060614-105038

Bouveresse E, Hartmann C, Massart DL et al (1996) Standardization of near-infrared spectrometric instruments. Anal Chem 68:982–990. https://doi.org/10.1021/ac9510595

Cambardella CA, Elliott ET (1992) Particulate soil organic-matter changes across a grassland cultivation sequence. Soil Sci Soc Am J 56:777. https://doi.org/10.2136/sssaj1992.03615995005600030017x

Cambardella CA, Elliott ET (1994) Carbon and nitrogen dynamics of soil organic matter fractions from cultivated grassland soils. Soil Sci Soc Am J 58:123. https://doi.org/10.2136/sssaj1994.03615995005800010017x

Cusack DF, Chadwick OA, Hockaday WC, Vitousek PM (2012) Mineralogical controls on soil black carbon preservation. Glob Biogeochem Cycles 26:GB2019. https://doi.org/10.1029/2012GB004109

Dangal SRS, Sanderman J (2020) Is standardization necessary for sharing of a large mid-infrared spectral library? Sensors 20:6729. https://doi.org/10.3390/s20236729

Dangal SRS, Sanderman J, Wills S et al (2019) Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Syst 3:11. https://doi.org/10.3390/soilsystems3010011

Deluca TH, Gundale MJ, Mackenzie D, Jones DL (2015) Biochar effects on soil nutrient transformations. Biochar Environ Manage 2:421–454

DeLuca TH, Zabinski CA (2011) Prairie ecosystems and the carbon problem. Front Ecol Environ 9:407–413. https://doi.org/10.1890/100063

Didan K (2015) MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m Sin Grid V006. 2015, distributed by NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13Q1.006.

Elzein A, Balesdent J (1995) Mechanistic simulation of vertical distribution of carbon concentrations and residence times in soils. Soil Sci Soc Am J 59:1328–1335. https://doi.org/10.2136/sssaj1995.03615995005900050019x

Falcone JA, LaMotte AE (2016) National 1-kilometer rasters of selected Census of Agriculture statistics allocated to land use. Biogeosciences (2021) 156:97–114 111
use for the time period 1950 to 2012: U.S. Geological Survey data release. https://doi.org/10.5066/F70R9MHS

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37:4302–4315. https://doi.org/10.1002/joc.5086

Ge Y, Morgan CLS, Grunwald S et al (2011) Comparison of soil reflectance spectra and calibration models obtained using multiple spectrometers. Geoderma 161:202–211. https://doi.org/10.1016/j.geoderma.2010.12.020

Gilbert M, Nicolas G, Cinardi G et al (2018) Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Sci Data 5:1–11. https://doi.org/10.1038/sdata.2018.227

Glaser B, Amelung W (2003) Pyrogenic carbon in native grassland soils along a climosequence in North America. Global Biogeochem Cycles 17:1–8. https://doi.org/10.1029/2002gb002019

Gogé F, Gomez C, Jolivet C, Joffre R (2014) Which strategy is best to predict soil properties of a local site from a national Vis-NIR database? Geoderma 213:1–9. https://doi.org/10.1016/j.geoderma.2013.07.016

Gray J, Karunaratne S, Bishop T et al (2019) Driving factors of soil organic carbon fractions over New South Wales, Australia. Geoderma 353:213–226. https://doi.org/10.1016/j.geoderma.2019.06.032

Gregorich EG, Beare MH, McKim UF, Skjemstad JO (2006) Chemical and biological characteristics of physically uncomplesed organic matter. Soil Sci Soc Am J 70:975. https://doi.org/10.2136/sssaj2005.0116

Guerrero C, Wetterlind J, Stenberg B et al (2016) Do we really need large spectral libraries for local scale SOC assessment with NIR spectroscopy? Soil Tillage Res 155:501–509. https://doi.org/10.1016/j.still.2015.07.008

Guo LB, Gifford RM (2002) Soil carbon stocks and land use change: a meta analysis. Glob Change Biol 8:345–360. https://doi.org/10.1046/j.1354-1013.2002.00486.x

Halle D, Riggs GA (2016) MODIS/Terra Snow Cover 8-Day L3 Global 500m SIN Grid, Version 6. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A2.006

Hammes K, Torn MS, Lapenias AG, Schmidt MW (2008) Centennial black carbon turnover observed in a Russian steppe soil. Biogeosciences 5:1339–1350. https://doi.org/10.5194/bg-5-1339-2008

Hengl T, Heuvelink GBM, Kempen B et al (2015) Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS ONE. https://doi.org/10.1371/journal.pone.0125814

Hengl T, de Jesus JM, Heuvelink GBM et al (2017) SoilGrids≥250m: global gridded soil information based on machine learning. PLoS ONE 12:e0169748

Hicks W, Viscarra Rossel RA, Tuomi S (2015) Developing the Australian mid-infrared spectroscopic database using data from the Australian Soil Resource Information System. Soil Res 53:922. https://doi.org/10.1071/SR15171

Hobley EU, Baldock J, Wilson B (2016) Environmental and human influences on organic carbon fractions down the soil profile. Agric Ecosyst Environ 223:152–166. https://doi.org/10.1016/j.agee.2016.03.004

Janzen HH, Campbell CA, Brandt SA et al (1992) Light-fraction organic matter in soils from long-term crop rotations. Soil Sci Soc Am J 56:1799. https://doi.org/10.2136/ssaj1992.03615995005600006025x

Jastrow JD, Miller RM (2018) Soil aggregate stabilization and carbon sequestration: feedbacks through organomineral associations. In: Follett RF, Kimble J, Stewart BA (eds) Soil processes and the carbon cycle. CRC Press, Boca Raton, pp 207–223

Kaiser K, Kalbitz K (2012) Cycling downwards - dissolved organic matter in soils. Soil Biol Biochem 52:29–32. https://doi.org/10.1016/j.soilbio.2012.04.002

Kaste JM, Heimsath AM, Bostick BC (2007) Short-term soil mixing quantified with fallout radionuclides. Geology 35:243–246. https://doi.org/10.1130/G23355A.1

Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137. https://doi.org/10.1016/10041-0

Kleber M, Johnson MG (2010) Advances in understanding the molecular structure of soil organic matter. Adv Agron 106:77–142

Knox NM, Grunwald S, McDowell ML et al (2015) Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma 239–240:229–239. https://doi.org/10.1016/j.geoderma.2014.10.019

Kögel-Knabner I, Guggenberger G, Kleber M et al (2008) Organo-mineral associations in temperate soils: Integrating biology, mineralogy, and organic matter chemistry. J Plant Nutr Soil Sci 171:61–82. https://doi.org/10.1002/jpln.200700048

Kooekana RS, Carmack AK, Van ZL et al (2011) Biochar application to soil: agronomic and environmental benefits and unintended consequences. Adv Agron 112:103–143. https://doi.org/10.1016/B978-0-12-855538-1.00003-2

Koven CD, Riley WJ, Subin ZM et al (2013) The effect of vertically resolved soil biogeochemistry and alternate soil C and N models on C dynamics of CLM4. Biogeosciences 10:7109–7131. https://doi.org/10.5194/bg-10-7109-2013

Kramer MG, Sanderman J, Chadwick OA et al (2012) Long-term carbon storage through retention of dissolved aromatic acids by reactive particles in soil. Glob Change Biol. https://doi.org/10.1111/j.1365-2486.2012.02681.x

Lavallee JM, Conant RT, Haddix ML, Follett RF, Bird MI, Paul EA (2019) Selective preservation of pyrogenic carbon across soil organic matter fractions and its influence on calculations of carbon mean residence times. Geoderma 354:113866. https://doi.org/10.1016/j.geoderma.2019.07.024

Lavallee JM, Soong JL, Cotrufo MF (2020) Conceptualizing soil organic matter into particulate and mineral-associated forms to address global change in the 21st century. Glob Change Biol 26:261–273. https://doi.org/10.1111/gcb.14859

Lee J, Viscarra Rossel RA (2020) Soil carbon simulation confined by different pool initialisation. Nutr Cycl Agroecosyst 116:245–255. https://doi.org/10.1007/s10705-019-10041-0

Lehmann J, Kleber M (2015) The contentious nature of soil organic matter. Nature 528:60
Lehmann J, Skjemstad J, Sohi S et al (2008) Australian climate–carbon cycle feedback reduced by soil black carbon. Nat Geosci 1:832–835. https://doi.org/10.1038/ngeo358

Lehmann J, Rillig MC, Thies J et al (2011) Biochar effects on soil biota? A review. Soil Biol Biochem 43:1812–1836. https://doi.org/10.1016/j.soilbiol.2011.04.022

Leifeld J, Heiling M, Hajdas I (2015) Age and thermal stability of particulate organic matter fractions indicate the presence of black carbon in soil. Radiocarbon 57:99–107. https://doi.org/10.2458/azu_rc.57.17964

Liang B, Lehmann J, Solomon D et al (2006) Black carbon increases cation exchange capacity in soils. Soil Sci Soc Am J 70:1719. https://doi.org/10.2136/sssaj2005.0383

Liang B, Lehmann J, Solomon D et al (2008) Stability of biomass-derived black carbon in soils. Geochim Cosmochim Acta 72:6069–6078. https://doi.org/10.1016/j.gca.2008.09.028

Lobsey CR, Viscarra Rossel RA, Roudier P, Hedley CB (2017) rs-local data-mines information from spectral libraries to improve local calibrations. Eur J Soil Sci 68:840–852. https://doi.org/10.1111/ejss.12490

Luo Y, Ahlström A, Allison SD et al (2016) Toward more realistic projections of soil carbon dynamics by Earth system models. Global Biogeochem Cycles 30:40–56. https://doi.org/10.1002/2015GB005239

Major J, Lehmann J, Rondon M, Goodale C (2010) Fate of soil-applied black carbon: downward migration, leaching and soil respiration. Glob Change Biol 16:1356–1379. https://doi.org/10.1111/j.1365-2486.2009.02044.x

Meinshausen N (2006) Quantile regression forests. J Mach Learn Res 7:983–999

Miltner A, Bombach P, Schmidt-Brücken B, Kästner M (2012) SOM genesis: microbial biomass as a significant source. Biogeochemistry 111:41–55. https://doi.org/10.1007/s10533-011-9658-z

Mirzaeitalarposhti R, Demyan MS, Rasche F et al (2017) Mid-infrared spectroscopy to support regional-scale digital soil mapping on selected croplands of South-West Germany. CATENA 149:283–293. https://doi.org/10.1016/j.catena.2016.10.001

Soil Survey Staff (2014) Kellogg Soil Survey Laboratory Methods Manual; Burt, R., Soil Survey Staff, Eds.; Soil Survey Investigations Report No. 42, Version 5.0.; U.S. Department of Agriculture, Natural Resources Conservation Service: Washington, DC, USA

Nocita M, Stevens A, van Wesenmael B et al (2015) Soil spectroscopy: an opportunity to be seized. Glob Chang Biol 21:10–11. https://doi.org/10.1111/gcb.12632

Omernik JM, Griffith GE (2014) Ecoregions of the conterminous united states: evolution of a hierarchical spatial framework. Environ Manage 54:1249–1266. https://doi.org/10.1007/s00267-014-0364-1

Pelletier JD, Broxton PD, Hazenberg P, et al (2016) Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers. ORNL DAAC, Oak Ridge, Tennessee, USAhttps://doi.org/10.3334/ORNLDAAC/1304

Poepplau C, Don A (2013) Sensitivity of soil organic carbon stocks and fractions to different land-use changes across Europe. Geoderma 192:189–201. https://doi.org/10.1016/j.geoderma.2012.08.003

Ramirez-Lopez L, Behrens T, Schmidt K et al (2013) The spectrum-based learner: a new local approach for modeling soil vis-NIR spectra of complex datasets. Geoderma 195:268–279. https://doi.org/10.1016/j.geoderma.2012.12.014

Rasmussen C, Heckman K, Wieder WR et al (2018) Beyond clay: towards an improved set of variables for predicting soil organic matter content. Biogeochemistry 137:297–306. https://doi.org/10.1007/s10533-018-0424-3

Reisser M, Purves RS, Schmidt MW, Abiven S (2016) Pyrogenic carbon in soils: a literature-based inventory and a global estimation of its content in soil organic carbon and stocks. Front Earth Sci 4:80. https://doi.org/10.3389/feart.2016.00080

Reuter HI, Nelson A, Jarvis A (2007) An evaluation of void-filling interpolation methods for SRTM data. Int J Geogr Inf Sci 21:983–1008. https://doi.org/10.1080/13658810601169899

Rossel RAV, McBratney AB (2008) Diffuse reflectance spectroscopy as a tool for digital soil mapping. Digital soil mapping with limited data. Springer, Amsterdam, pp 165–172

Sanderman J, Amundson R (2008) A comparative study of dissolved organic carbon transport and stabilization in California forest and grassland soils. Biogeochemistry 89:309–327. https://doi.org/10.1007/s10533-008-9221-8

Sanderman J, Chappell A (2013) Uncertainty in soil carbon accounting due to unrecognized soil erosion. Glob Change Biol 19:264–272. https://doi.org/10.1111/gcb.12030

Sanderman J, Hengl T, Fiske GJ (2017) Soil carbon debt of 12,000 years of human land use. Proc Natl Acad Sci USA. https://doi.org/10.1073/pnas.1706103114

Sanderman J, Savage K, Dangal SRS (2020) Mid-infrared spectroscopy for prediction of soil health indicators in the United States. Soil Sci Soc Am J 84:251–261. https://doi.org/10.1002/saj2.20009

Schmidt MW, Skjemstad JO, Czimczik CI, Glaser B, Prentice KM, Gelinas Y, Kuhlbusch TA (2001) Comparative analysis of black carbon in soils. Global Biogeochem Cycles 15:163–167. https://doi.org/10.1029/2000GB001284

Schmidt MWI, Torn MS, Abiven S et al (2011) Persistence of soil organic matter as an ecosystem property. Nature 478:49–56. https://doi.org/10.1038/nature10386

Six J, Bossuyt H, Degryze S, Denef K (2004) A history of research on the link between (micro) aggregates, soil biota, and soil organic matter dynamics. Soil Tillage Res 79:7–31. https://doi.org/10.1016/j.still.2004.03.008

Skjemstad JO, Spouncer L (2003) Integrated soils modelling for the National Carbon Accounting System (estimating changes in soil carbon resulting from changes in land use). Canberra: Australian Greenhouse Office; 2003. http://hdl.handle.net/102.100.100/193053?index=1

Skjemstad JO, Clarke P, Taylor JA, Oades JM, Newman RH (1994) The removal of magnetic materials from surface soils. A solid state 13C CP/MAS NMR study. Aust J Soil Res 32:1215–1229. https://doi.org/10.1071/SR9941215

Skjemstad JO, Reicosky DC, Wilts AR, McGowan JA (2002) Charcoal carbon in US agricultural soils. Soil Sci Soc Am J 66:1249–1255. https://doi.org/10.2136/ssaj2002.1249
Skjemstad JO, Spouncer LR, Cowie B, Swift RS (2004) Calibration of the Rothamsted organic carbon turnover model (RothC ver. 26.3), using measurable soil organic carbon pools. Aust J Soil Res 42:79–88. https://doi.org/10.1071/SR03013

Soucémarianadin LN, Quideau SA, MacKenzie MD (2014) Pyrogenic carbon stocks and storage mechanisms in podzolic soils of fire-affected Quebec black spruce forests. Geoderma 217–218:118–128. https://doi.org/10.1016/j.geoderma.2013.11.010

Todd-Brown KEO, Randerson JT, Post WM et al (2013) Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. Biogeosciences 10:1717–1736. https://doi.org/10.5194/bg-10-1717-2013

Todd-Brown KEO, Randerson JT, Hopkins F et al (2014) Changes in soil organic carbon storage predicted by Earth system models during the 21st century. Biogeosciences 11:2341–2356. https://doi.org/10.5194/bg-11-2341-2014

Viscarra Rossel RA, McBratney A (2008) Diffuse reflectance spectroscopy as a tool for digital soil mapping. Digital soil mapping with limited data. Springer, Dordrecht, pp 165–172

Viscarra Rossel RA, Lee J, Behrens T et al (2019) Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat Geosci 12:547–552. https://doi.org/10.1038/s41561-019-0373-z

von Lützow M, Kögel-Knabner I, Ekschmitt K et al (2007) SOM fractionation methods: relevance to functional pools and to stabilization mechanisms. Soil Biol Biochem 39:2183–2207. https://doi.org/10.1016/j.soilbio.2007.03.007

Wang X, Sanderman J, Yoo K (2018) Climate-dependent topographic effects on pyrogenic soil carbon in southeastern Australia. Geoderma. https://doi.org/10.1016/j.geoderma.2018.02.025

Wilson AM, Jetz W (2016) Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. PLoS Biol 14:e1002415. https://doi.org/10.1371/journal.pbio.1002415

Yao Y, Gao B, Zhang M et al (2012) Effect of biochar amendment on sorption and leaching of nitrate, ammonium, and phosphate in a sandy soil. Chemosphere 89:1467–1471. https://doi.org/10.1016/j.chemosphere.2012.06.002

Zimmermann M, Leifeld J, Schmidt MWI et al (2007) Measured soil organic matter fractions can be related to pools in the RothC model. Eur J Soil Sci 58:658–667. https://doi.org/10.1111/j.1365-2389.2006.00855.x

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.