scientific reports

Check for updates

OPEN

Innovative approaches in soil carbon sequestration modelling for better prediction with limited data

Mohammad Javad Davoudabadi[1,2,3,4]✉, Daniel Pagendam[4], Christopher Drovandi[1,2,3], Jeff Baldock[5] & Gentry White[1,2,3]

Soil carbon accounting and prediction play a key role in building decision support systems for land managers selling carbon credits, in the spirit of the Paris and Kyoto protocol agreements. Land managers typically rely on computationally complex models fit using sparse datasets to make these accounts and predictions. The model complexity and sparsity of the data can lead to over-fitting, leading to inaccurate results when making predictions with new data. Modellers address over-fitting by simplifying their models and reducing the number of parameters, and in the current context this could involve neglecting some soil organic carbon (SOC) components. In this study, we introduce two novel SOC models and a new RothC-like model and investigate how the SOC components and complexity of the SOC models affect the SOC prediction in the presence of small and sparse time series data. We develop model selection methods that can identify the soil carbon model with the best predictive performance, in light of the available data. Through this analysis we reveal that commonly used complex soil carbon models can over-fit in the presence of sparse time series data, and our simpler models can produce more accurate predictions.

Large-scale carbon emission from soil, one of the planet's major carbon reservoirs, into the atmosphere has deleterious impacts on global climate change, soil quality, and crop productivity[1,2]. Soil organic carbon (SOC) could be used as a significant global sink for atmospheric carbon through land-management practices, helping to reduce the atmospheric concentration of greenhouse gases and improving agricultural productivity.

International bodies and agreements such as the Intergovernmental Panel on Climate Change (IPCC) and the Paris and Kyoto Protocol agreements mitigate global warming by assessing the science related to climate change and reduce greenhouse gas emissions, especially CO_2. These agreements adopted systems of carbon accounting and trading markets. A part of these carbon markets (tracking and trading) is related to selling carbon credits by farmers, organisations certifying the credits, or providing government support for the scheme, and land-holders who apply land-management practices to sequester carbon and track the change of soil carbon sequestration in their farmlands. They usually have small datasets for tracking the changes in soil carbon as SOC sampling is time-consuming and costly.

Models can quantify changes in soil carbon stocks where there is accurate understanding of processes that govern soil carbon turnover and sequestration. Such models can also help develop a deeper understanding of the sequestration process and forecast future changes and trends in SOC. Researchers have developed computer-simulation models such as RothC[3,4], and Century[5] to help make inferences about trends in carbon stocks using time series of measurements collected over many years. For example, to improve the accounting of field emissions in the carbon footprint of agricultural products, Peter et al.[6] assess the change of SOC based on simulations with the RothC model in one of the IPCC methodological approaches (Tier 3) and compare it with other default IPCC methods. Clifford et al.[15] developed a statistical soil carbon model to estimate and forecast the amount of carbon sequestered on farmland.

[1]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia. [2]Australian Research Council Centre of Excellence for Mathematical & Statistical Frontiers (ACEMS), Victoria, Australia. [3]QUT Centre for Data Science, Queensland University of Technology, Brisbane, Australia. [4]CSIRO Data61, GPO Box 2583, Brisbane, QLD 4001, Australia. [5]CSIRO Agriculture and Food, Glen Osmond, SA, Australia. ✉email: mohammadjavad.davoudabadi@hdr.qut.edu.au

All models have their limitations and it is commonplace for modellers to make modifications that better suit specific scenarios of interest. For instance, Farina et al.[7] modified the RothC model with the aim of improving the prediction of soil carbon dynamics in semi-arid regions. At their core, models such as RothC partition the total SOC mass into specific pools. These pools are decomposable plant material (DPM), resistant plant matter (RPM), humified organic matter (HUM), microbial biomass (BIO), and inert organic matter (IOM)[1,8]. Modellers are, however, free to explore alternative means of partitioning soil carbon to suit different objectives.

The vast majority of SOC models are deterministic, yielding a single possible trajectory of soil carbon dynamics for a given set of parameters and an initial condition. On the other hand, statistical SOC models can yield ensembles of possible soil carbon trajectories. One of the main advantages of a statistical SOC model over deterministic SOC models such as RothC is introducing this randomness and providing a probabilistic method for quantifying uncertainty around model outputs. Uncertainties in SOC models arise in many ways such as around the parameters, model inputs, dynamics, and subsequently model predictions. Statistical models help to quantify uncertainties in a SOC model by modelling the different sources of randomness. Research using statistical models and sensitivity analysis (running models for different sets of parameter values) attempts to quantify uncertainties in soil carbon model outputs[9–14]. Clifford et al.[15] quantified uncertainties in model inputs, dynamics, and uncertainties in model parameters for a one pool soil carbon in a comprehensive manner using a physical-statistical model for carbon dynamics within a framework known as Bayesian hierarchical modelling (BHM). The statistical methods used by Clifford et al.[15] can be computationally burdensome, especially for more complex models such as some of the models we consider in this study. In addition, differences between the various soil carbon pools (DPM, RPM, HUM, BIO and IOM) are ignored in Clifford et al.[15]. Gurung et al.[16] identify the most important DayCent model parameters through a global sensitivity analysis for parameterization and implement a Bayesian approach using the sampling importance resampling method to calibrate the model and produce posterior distributions for the most sensitive parameters.

Microbial biomass carbon (MBC) is an important labile soil carbon fraction and the most active component of the SOC, regulating bio-geochemical processes in terrestrial ecosystems[17]. Consequently, this has drawn the attention of modellers when considering how the MBC should be treated and how it should interact with other pools of carbon. The importance of MBC in soil carbon decomposition has led to the development of a number of microbially-explicit SOC models in recent years[18–22]. Several microbial models with a similar basic structure and key bio-geochemical processes have been developed to simulate warming effects on soil organic matter (SOM) decomposition[23–25]. These models differ in model complexity and reference temperature and there have been few efforts to compare model structures. For example, Li et al.[26] have compared these models to address this question of how microbial model predictions change with increasing model complexity, and whether these predictions differ fundamentally from models with a conventional structure. More recent studies consider the interactions of microbes in a microbially-based SOC model (SOMic version 1.0)[27]. Other studies compare the fit of linear and non-linear soil bio-geochemical models (SBMs) using data assimilation with soil respiration data sourced from a meta-analysis of soil warming studies[28].

In this study, we explore the effect of relaxing some of the bio-geochemical realism of models such as RothC with respect to predicting soil carbon stocks. Bio-geochemical refers to the degree to which a model accurately represents the biological, geological, and chemical processes that govern the cycling of carbon in soil ecosystems. Our focus is using these models with the temporally sparse datasets typically available for assessing trends in soil carbon on farms, making use of two datasets from Tarlee in South Australia and Brigalow in Queensland, Australia[15,29]. These two sites are in different climatic regions, and it shows we can apply our approaches to a range of climatic regions. A pertinent scientific question is whether multi-pool models such as RothC are too complex relative to the limited data that is often available to fit them on a specific parcel of land. Therefore, we attempt to understand how model predictive performance varies when we amalgamate some of these conceptual pools in the underlying process dynamics. Specifically, we consider: (i) a single pool model considering soil carbon as a homogeneous pool that can decay and release carbon into the atmosphere[15]; (ii) a two-pool model in which we consider a single homogeneous pool of decomposable SOC and an IOM pool that does not decompose; (iii) a three-pool model which considers two pools of decomposable SOC (one of them represents the biological pool) and the IOM pool; and (iv) a five-pool model considering all pools mentioned above that are present in RothC. The two and three-pool models are novel soil carbon models that we introduce in this study. Also, the five-pool model used herein is somewhat novel in terms of the statistical modelling framework it is embedded in and its simplification in terms of time-step and reduced set of parameters compared to RothC.

Our modelling framework predicts changes in soil carbon stocks and accounts for epistemic uncertainty, uncertainty in the bio-geochemical process dynamics, in a statistically defensible manner. This is particularly important in the present context. We explore structural differences in the systems of equations used to describe soil carbon process dynamics which is one of the major differences between our statistical approach and that used in the simpler regression studies (e.g. Xie et al.[28]). We develop a state-space modelling framework used for a one-pool model by[15,30] to the two, three, and RothC-like five-pool models. We develop a Bayesian model selection method known as leave-future-out cross-validation (LFO-CV)[31] to choose, for a given dataset, the best soil carbon model in terms of its out-of-sample predictive accuracy. Our approach optimally adapts to the data at hand. Fitting overly complex soil carbon models might increase the uncertainty of predictions in the presence of sparse data, and it is important when making predictions about soil carbon stocks; otherwise, a land-owner might unwittingly enter into a contract to sequester carbon that has a higher risk than anticipated. Conversely, when data are sufficiently informative, our approach supports more complexity. In addition, we explore the effect of microbes and inert organic matter on the carbon cycle decomposition by adding microbial biomass and IOM pools in the one-pool model to answer this question that by adding these pools whether we obtain better soil carbon prediction than the one-pool model in Clifford et al.[15]. Although there are a number of studies in the literature that consider the impact of microbial biomass on soil carbon sequestration and how this affects

modelling[18,19,27,32], our process of modelling the dynamics of microbial biomass in the SOC model, along with applying advanced Bayesian methods to estimate its model parameters, are the main differences between our study and aforementioned papers.

We organise the rest of the paper as follows. The datasets used in this study are described in Section "Background and description of datasets". We introduce our model framework and the LFO-CV criterion in Section "Methods". In Section "Model structure" the structure of the models is described. In Section "Results", we compare the models based on their out-of-sample predictive accuracy and quantify the uncertainty of our estimate. Section "Discussion" presents a discussion of this study and our results.

Background and description of datasets

Our model selection method is motivated by two datasets that are collected from two locations in Australia. The details of these sites are presented in the following.

Tarlee dataset

An agricultural research experiment site known as Tarlee situated 80 km north of Adelaide, South Australia was established in 1977 to examine the impact of management practices on agricultural productivity as a long-term field experiment[33]. The soil of this site is classified as a hard-setting red-brown earth with sandy loam texture. This site has a Mediterranean climate and is dominated by winter rainfall with an average of 355 mm from April to October[15,29,34]. Soil properties of that site were monitored over a 20-year period in three fields under different management practices, and soil samples covering the entire top 30 cm of the profile were obtained for the years 1979, 1985, and 1996 from all 3 rotations. Table 1 presents the time period of management treatments that were implemented in three trial fields in Tarlee.

Brigalow dataset

Brigalow is a research station in Queensland, Australia. This site is situated in a semi-arid, and subtropical climate, and consists of three forested catchments of 12–17 ha[29]. Three monitoring sites were established within each of the catchments in recognition of three soil types (a duplex soil and two clays). One catchment was planted to wheat and occasional sorghum and the other to buffel pasture and the last one was left under native Brigalow forest. At this site, on one catchment, after clearing land under Brigalow (*Acacia harpophylla*) in 1982, continuous wheat with some sorghum was established over a 18-year period. Samples were collected from the field in two distinct categories: surface samples, acquired from a depth of 0–10 cm, and profile samples, retrieved down to a depth of 200 cm. In the profile category, samples were taken at three specific intervals within the upper layers: 0–10 cm, 10–20 cm, and 20–30 cm. Table 2 shows the duration of management practices in Brigalow.

Management treatments	Field 1	Field 2	Field 3
Wheat for grain	(1979–1987) and	–	–
	(1990–1996)		
Wheat for hay	1988 and 1989	1989	–
Fallow	1997	1997	1997
Wheat for grain and fallow	–	(1979–1988) and	–
		(1990–1996)	
Wheat and pasture	–	–	(1979–1987)
Wheat and pasture for hay	–	–	1988 and 1989
Wheat for grain and pasture	–	–	(1990–1996)

Table 1. The duration of management treatments in three fields in Tarlee.

Management treatments	Soil type 1	Soil type 2	Soil type 3
Cleared	1982	1982	1982
Wheat for grain	(1985–1992) and	(1985–1992)	(1985–1992)
	(1994, 1996, 1998)	(1994, 1996, 1998)	(1994, 1996, 1998)
Sorghum for grain	1984, 1995,	1984, 1995,	1984, 1995,
	1997 and 1999	1997 and 1999	1997 and 1999
Fallow	1983 and 1993	1983 and 1993	1983 and 1993

Table 2. The duration of management treatments in Brigalow.

Methods

Soil carbon model

We can consider uncertainties in a dynamical SOC model as arising from three sources: errors in the observations, randomness or uncertainty inherent in the underlying physical processes, and uncertainties in model parameters[15]. These uncertainties are modelled through the observation model $p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})$, the process model $p(\mathbf{X}|\boldsymbol{\theta})$, and the prior $p(\boldsymbol{\theta})$. Here $\boldsymbol{\theta}$, $\mathbf{Y}$, and $\mathbf{X}$ denote unknown parameters, observations, and unobserved state process, respectively. Furthermore, the probability density function of the enclosed random variable, and the conditional probability density function given the event E are denoted by $p(.)$, and $p(.|E)$, respectively. For example, the mass of SOC, X_C, is one of the elements of $\mathbf{X}$, or the measured value of total SOC, Y_{TOC}, is one of the elements of $\mathbf{Y}$, furthermore, the decay rate of total SOC, K_C, is an example of a model parameter in a soil carbon model.

These three models form a hierarchical framework known as a Bayesian Hierarchical Model (BHM). The top level of the hierarchy contains the observation model which includes noisy observational data that depend on the state variables. This model is followed by the process model, located at the second level. At this level, latent state variables, which cannot be measured directly but can be estimated based on measurement data that depend on the latent state variables, are modelled. These two models typically rely on some unknown parameters. The third level underneath these two levels contains the parameter model[35–37]. A BHM is represented mathematically as follows:

$$p(\mathbf{Y},\mathbf{X},\boldsymbol{\theta}) = p(\mathbf{Y},\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1}$$

Note that the joint distribution $p(\mathbf{Y},\mathbf{X},\boldsymbol{\theta})$ captures all the uncertainty in the model. The advantage of analysing a model within the BHM framework is that it incorporates prior knowledge related to the parameters into the analysis by updating the distributions of these parameters with observed data. The latent state of the SOC, $\mathbf{X}$, evolves as a dynamical process and given noisy, sparse data. Inferences about soil carbon dynamics, parameters, and functions of them can be made through the posterior distribution $p(\mathbf{X},\boldsymbol{\theta}|\mathbf{Y})$. We can write the posterior distribution based on (1) as follows:

$$p(\mathbf{X},\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y})} \tag{2}$$

where $p(\mathbf{Y})$ depends only on data and may be difficult to calculate analytically or numerically, thus the posterior itself may be difficult to evaluate. Fortunately, one can draw samples from the posterior if it is not analytically tractable.

As in other recent statistical analyses[15,30] we use a state-space modelling framework, the first and second levels of the BHM, to predict changes in soil carbon stocks. State-space models are more challenging to fit in practice than simpler regression models used in[28] because they acknowledge uncertainty in the latent process dynamics. The prior information of the parameter model in the third level of the BHM is described in the following.

Prior information

As mentioned earlier, the process model and the observation model typically depend on unknown parameters, and the parameter model captures the uncertainty around these parameters. A Bayesian approach for model fitting is applied to quantify the uncertainty in parameters and predictions. This approach places a prior distribution on the unknown parameter vector $\boldsymbol{\theta}$, which is the advantage of using the Bayesian analysis since we implement our prior knowledge of parameters as part of the inferential process.

In general, the prior knowledge about parameters includes three categories: informative, weakly informative, and uninformative priors. When we have a small dataset or the dataset is sparse, the prior distribution becomes more influential and informative priors can become more useful. In this study, we obtain priors from previous studies[15,29,30] and expert opinion. The model parameters and their prior probability density functions are listed in Supplementary Tables S2 and S3 (Section B of the supplementary material).

Posterior distribution inference

To estimate the changes in SOC over time as a result of the various management practices, and to estimate the parameters driving the sequestration of carbon, we sample from the posterior distribution $p(X_{TOC},\boldsymbol{\theta}|\mathbf{Y})$, where X_{TOC} is the mass of total SOC. To this end, we draw samples from the posterior distribution $p(\mathbf{X},\boldsymbol{\theta}|\mathbf{Y})$ in (2) which can be decomposed into two components $p(\mathbf{X}|\boldsymbol{\theta},\mathbf{Y})p(\boldsymbol{\theta}|\mathbf{Y})$ and we preserve the components related to the SOC process X_{TOC} and its parameters $\boldsymbol{\theta}$. Davoudabadi et al.[30] used advanced Bayesian methods, e.g. correlated pseudo-marginal (CPM) method and the Rao-Blackwellised particle filters (RBPF) for state-space models, to reduce the computational cost of estimating uncertainties in the one-pool model presented by[15]. The CPM method, one of several particle Markov chain Monte Carlo (PMCMC) methods, is applied to the model to draw samples from $p(\boldsymbol{\theta}|\mathbf{Y})$ as the resulting likelihood is not tractable[30,38]. The CPM method in Davoudabadi et al.[30] outperforms other state of the art PMCMC methods in terms of computation time. The advantage of using this method is that it reduces the computational cost of estimating intractable likelihoods by correlating the estimators of the likelihoods in the acceptance ratio of its algorithm. Algorithm S3 in Section C.3 of the supplementary material provides the CPM algorithm. This correlation can be achieved by correlating the auxiliary random numbers used to obtain these estimators; see Davoudabadi et al.[30] and Deligiannidis et al.[38] for more details. To estimate the marginal likelihood of the state variables, we use the RBPF as the SOC model combines linear and non-linear sub-models. The RBPF algorithm estimates the marginal likelihood of the non-linear sub-model through bootstrap particle filter (BPF). It computes the marginal likelihood of the linear part of the model through the Kalman Filter (KF) algorithm[30,39]. Computing the exact likelihood of the linear sub-model

makes the RBPF algorithm an attractive algorithm in these scenarios as it reduces the computational cost of the estimated likelihood dramatically. See Davoudabadi et al.[30] for more details about the RBPF, BPF and KF algorithms. In addition, the algorithm of the KF and BPF methods are provided in Sections C.1 and C.2 of Supplementary Material , Algorithms S1 and S2 , respectively. The RBPF algorithm is reused to draw a sample of the state process from the posterior distribution $p(X_{TOC}|\boldsymbol{\theta}, \mathbf{Y})$. In the CPM algorithm, it is required to generate candidate parameters from appropriate proposal distributions. More precisely, a proposal distribution is a user-specified distribution that the user is free to choose and the Markov chain will converge to the desired posterior distribution if it is run for enough iterations. However, a proposal distribution can have a significant impact on the finite-time efficiency of the MCMC and the ideal case occurs when the proposal distribution is the desired posterior distribution which is typically unknown. The proposal distributions are presented in the supplementary material Section B.

We can quantify the uncertainty of our estimate in many ways, for example, through a 95% credible interval or the estimated expected value of functionals of interest. The inference about the mass of SOC added over a period of time can be achieved through the MCMC samples of the posterior distribution. We represent the posterior distribution $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ by M^* samples $\{(X^m, \boldsymbol{\theta}^m) : m = 1, ..., M^*\}$ and the posterior expectation of any function $g^*(\boldsymbol{X}, \boldsymbol{\theta})$ can be estimated by these samples.

$$\mathbf{E}(g^*(\boldsymbol{X}, \boldsymbol{\theta})|\boldsymbol{Y}) \approx \frac{1}{M^*} \sum_{m=1}^{M^*} g^*(X^m, \boldsymbol{\theta}^m).$$

The error of the accuracy of such estimates is negligible for sufficiently large sample size M^*. The change in SOC to field i between the first year of trial, e.g. $t = 1$, and following year t in a dataset is considered as follows

$$g^*(\boldsymbol{X}, \boldsymbol{\theta}) = X^i_{TOC(t)} - X^i_{TOC(1)};$$

and can be estimated as follows

$$\hat{g^*}(\boldsymbol{X}, \boldsymbol{\theta}) = \mathbf{E}(X^i_{TOC(t)} - X^i_{TOC(1)}|\boldsymbol{Y});$$

where $X^i_{TOC(t)}$ is the summation of other pools, for example, in the three-pool model $X^i_{TOC(t)}$ is equal to the summation of $X^i_{C(t)}$, $X^i_{IOM(t)}$, and $X^i_{B(t)}$.

The posterior variance, $\text{var}(\mathrm{X}^{\mathrm{i}}_{\mathrm{TOC(t)}} - \mathrm{X}^{\mathrm{i}}_{\mathrm{TOC(1)}}|\boldsymbol{Y})$, is a measure of uncertainty associated with this Bayes estimate.

We assess the quality of the MCMC samples through an MCMC diagnostic known as the Gelman and Rubin's convergence diagnostic statistic[40]. The Gelman and Rubin's convergence diagnostic statistic, $\hat{R}$, can be used to assess whether the MCMC samples have "mixed" sufficiently, effectively sampling from the probability distribution, and have reached a stationary distribution[40]. Gelman and Rubin's convergence diagnostic compares samples from multiple chains to assess whether the output from each chain is sufficiently similar to the others. The output from each chain is indistinguishable from the others when the scale reduction factor estimated from the sampling is less than 1.2[41].

Before estimating model parameters and conducting inference with a model, it is essential to validate our model to establish its suitability for estimating changes in soil carbon stocks. In the next section, we introduce our method for selecting between competing soil carbon models, focusing on predictive accuracy.

Model evaluation

One way to evaluate a model or compare different models is to measure predictive accuracy[42]. As our models depend on time, for model comparison and selection, we apply leave-future-out cross-validation (LFO-CV) that refits a model to different subsets of the data[31]. The LFO-CV is a fully Bayesian metric in that it uses the entire posterior distribution. This method is the approach used to compare the model's predictive accuracy for the four SOC models listed in Section "Model Structure".

Let $Y_{1:T}$ be a time series of observations and let L be the minimum number of observations from the series that we will require before making predictions for future data. To make reasonable predictions for Y_{i+1} based on $Y_{1:i}$, i should be large enough so that we can learn enough about the time series to predict future observations, otherwise, it may not be possible to make reasonable predictions. The choice of L depends on the application and how informative the data are, therefore, it may be vary from one dataset to another[31]. We would like to compute the predictive densities $p(\tilde{Y}_{t+1}|Y_{1:t})$ for each $t \in \{L, ..., T-1\}$ where $\tilde{Y}_{t+1}$ is a future vector of observed data. The expected log pointwise predictive density (ELPD) can be used as a global measure of predictive accuracy, which is

$$\text{ELPD} = \log \prod_{t=L}^{T-1} \mathbf{E}_{\theta|Y_{1:t}}(p(\tilde{Y}_{t+1}|Y_{1:t}, \theta)) = \sum_{t=L}^{T-1} \log \int p(\tilde{Y}_{t+1}|Y_{1:t}, \theta) p(\theta|Y_{1:t})\, d\theta. \tag{3}$$

In practice, the integral in (3) is intractable, however we can approximate it through Monte-Carlo methods[31]. To estimate $p(\tilde{Y}_{t+1}|Y_{1:t})$, we draw samples $(\theta^1_{1:t}, ..., \theta^S_{1:t})$ from the posterior distribution $p(\theta|Y_{1:t})$ for $t \in \{1, ..., \gamma\}$ where $\gamma \in \{L, ..., T-1\}$ using the particle MCMC method described in Section "Posterior Distribution Inference" and estimate the predictive density for $\tilde{Y}_{L+1:T}$ as follows

$$p(\tilde{Y}_{t+1}|Y_{1:t}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\tilde{Y}_{t+1}|Y_{1:t}, \theta^{s}_{1:t}). \quad (4)$$

When our model is a state-space model, we need to consider the state variables as part of the parameter space and estimate them through the particle filter methods to apply the LFO-CV. The reason for selecting ELPD instead of other global measures of accuracy such as the root mean squared error (RMSE) is that it evaluates a distribution to provide a measure of out-of-sample predictive performance rather than evaluating a point estimate like the mean or median, which we see as favourable from a Bayesian perspective[31,43].

Model structure

The total SOC consists of different components defined by their origin and their decay rate. These components originate from living organisms known as biotic material or non-living (abiotic) material[1,44]. Based on the RothC model, the components of the total SOC include DPM, RPM, HUM, BIO, IOM[1,8]. The one-pool model in Clifford et al.[15] considered all components mentioned above as a single pool. The process model of the one-pool model is a combination of linear and non-linear sub-models. The details of the process and the observation models of these sub-models are shown in the supplementary material Sections D.1 and D.2 , respectively. Figure 1a graphically represents the carbon emission process in the one-pool model. Based on Fig. 1a, a fraction of carbon decays is emitted into the atmosphere as CO_2 and the rest remains in the pool.

In the two-pool model, we consider the IOM pool as a second pool that is resistant to chemical and biological reactions and encompasses charcoal or charred material[8]. The IOM fraction is not subject to biological transformation and is thus constant[45]. As the IOM fraction is constant, its process model at time t is a constant value and should be estimated. The process and the observation models of the two-pool model are presented respectively in Sections E.1 and E.2. Figure 1b shows the graphical representation of the two-pool model.

The three-pool model considers the IOM and BIO as separate pools with a main pool of decomposable carbon which is an amalgamation of DPM, RPM, and HUM pools. Soil carbon decomposes from the decomposable carbon pool, and fractions are either transferred to the BIO pool or lost to the atmosphere as CO_2. Carbon present in the BIO pool that decomposes is either lost to the atmosphere as CO_2, re-assimilated as biological mass or transferred to the main soil carbon pool. Figure 1c shows the diagram of the carbon emission in the three-pool model. The process and observation models of the three-pool model are presented in detail in Sections F.1 and F.2 of the supplementary material, respectively. It is noteworthy to mention that the size of the microbial pool encompasses a small fraction of the total organic carbon, e.g. 5% of the TOC, based on expert knowledge. We implement this constraint by rejecting BIO state trajectories that exceed 5% of the TOC in the Markov chain Monte Carlo (MCMC) algorithm.

The RothC model, consisting of five conceptual pools, is the standard soil carbon used in many studies and is considered a reasonable representation of the physical sub-species of carbon in the soil. In the models presented

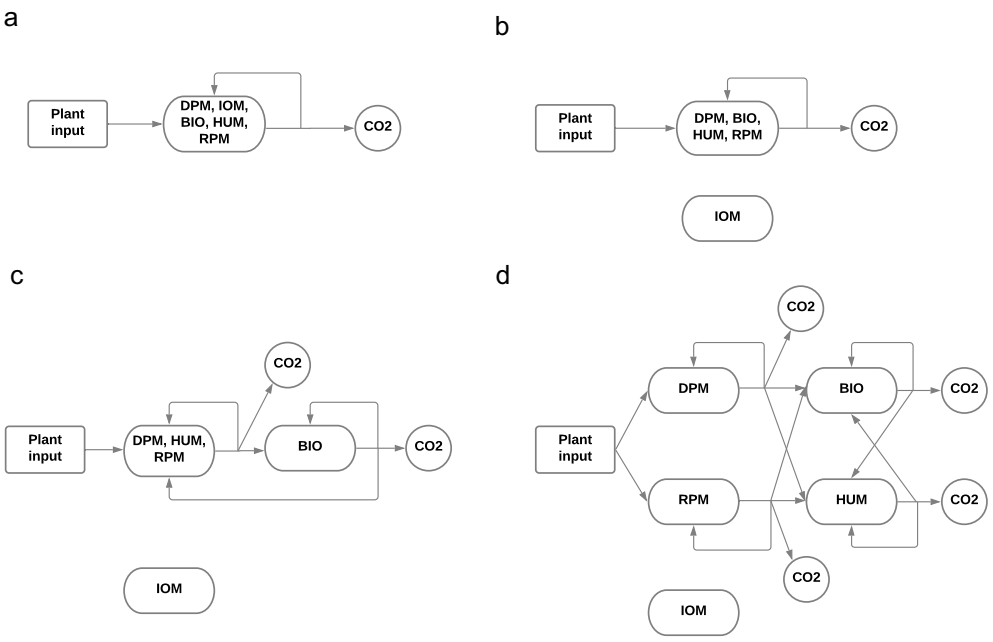

Figure 1. Graphical representation of the carbon emission in the (**a**) one-pool model, (**b**) two-pool model, (**c**) three-pool model, and (**d**) five-pool model. The five pools from RothC have been amalgamated into a single homogeneous soil carbon pool in the one-pool model. The DPM, BIO, HUM and RPM pools are amalgamated and treated as a single homogeneous pool in the two-pool model, and the DPM, HUM and RPM pools are amalgamated and treated as a single homogeneous pool in the three-pool model.

so far, we have considered the pools to be either one of the RothC pools or an amalgamation of the five RothC pools. In the five-pool model presented here, we now retain the structure presented in the RothC model without any amalgamation.

In the five-pool model, plant material is split between two conceptual pools: DPM and RPM. Decomposition of carbon from these two pools either leaves the system as CO_2 or is transformed to carbon in the BIO and HUM pools. Carbon from the BIO and HUM pools that decomposes can either be lost to the atmosphere as CO_2, or transformed to carbon in the BIO or HUM pools. The process and observation models of the carbon transfer in the five-pool model are presented mathematically in detail in Section G of the supplementary material. The five-pool model is depicted in Fig. 1d.

Results

Comparing models

We worked with four MCMC chains, each initialised with a randomly sampled parameter vector, in the Correlated Pseudo-marginal Method (CPM) method for estimating the predictive density (4). We ran each chain for 200,000 iterations discarding the first 80,000 as burn-in. We thinned these chains, choosing every 30th sample of the MCMC samples to estimate (4), therefore, S in Eq. (4) was equal to 4,000. The minimum numbers of observations, L, used for making predictions for future data in the Tarlee and Brigalow datasets were 12 and 13, respectively. The estimated expected log pointwise predictive density (ELPD) of the one, two, three, and five-pool models applied on the Tarlee dataset were −53.02, −40.55, −34.79, and −37, respectively. The estimated ELPD of those models applied on the Brigalow dataset were −36.89, −36.88, −36.48, and −49.57, respectively. Based on these results (supplementary material Tables S13 and S14) , the three-pool model outperformed the other models in the sense of yielding the best LFO predictive ability for both the Brigalow and Tarlee datasets. This three-pool model included an inert carbon pool and two decomposable pools that were conceptually equivalent to a biological pool (the decomposers) and a decomposable material pool, an amalgamation of DPM, RPM, and HUM pools. For Tarlee, the five-pool RothC-like model had the next best ELPD, but in Brigalow, the five-pool model exhibited the worst ELPD of the four models studied. The performances of the three and five-pool models in estimating the trajectories of the SOC dynamics of the Brigalow dataset are highlighted visually in Fig. 2a,b, respectively. As shown in Fig. 2b, the five-pool model increased uncertainty in the soil carbon dynamics, especially during the sparse periods, typified by wide 95% credible intervals. The significant variability in these regions stems from our practice of simulating input state values, such as the total mass of wheat dry matter (X_W), during each iteration of the particle filter algorithm and subsequently aggregating them. However, when there is no observation available for comparing these simulated values, it introduces additional variability in the trajectory of the state variables. Hence, when an observation ($Y_{(t)}$) is present, the level of uncertainty is notably lower compared to other scenarios.

Setting aside the five-pool model and focusing on the one, two, and three-pool models, we see that amongst these three models, the ranking from best to worst is three-pool, two-pool, and one-pool for both study sites.

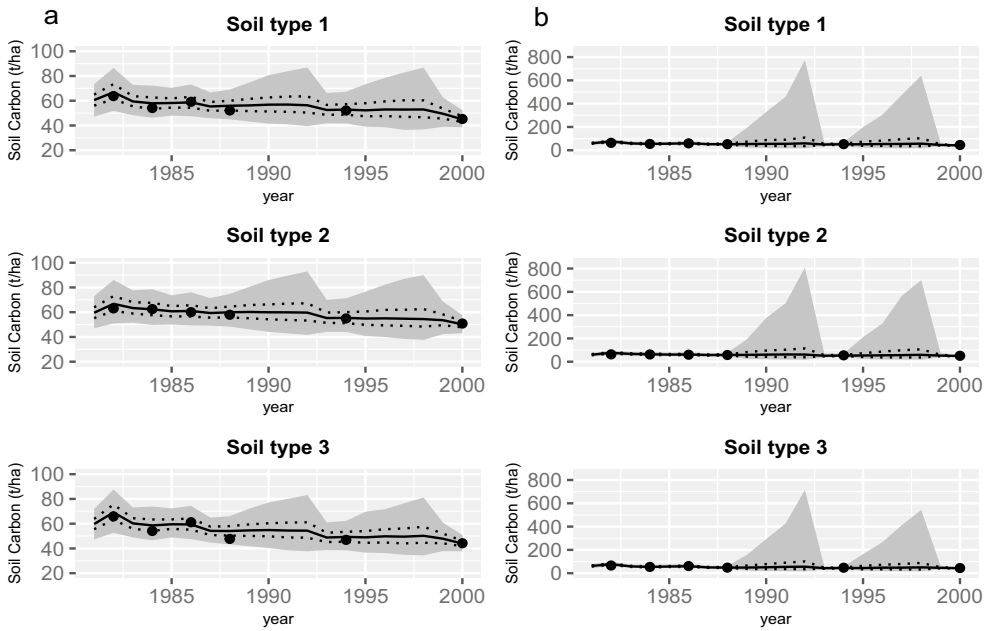

Figure 2. Soil organic carbon (SOC) dynamics of the Brigalow dataset based on (**a**) the three-pool model and (**b**) the five-pool model. The gray shaded part is the area between the 2.5th and the 97.5th percentiles for the SOC process gained by the three and five-pool models. The 25th and the 75th percentiles for the SOC process are indicated by the dashed lines. The 50th percentile is shown by the solid line and the measured SOC values are indicated by filled dots.

We cannot say with full confidence the three-pool model is the best model for the Brigalow dataset compared to the one and two-pool models as there is not much difference between their estimated ELPDs acknowledging the Monte Carlo errors. Nevertheless, we select it as the best model for the Brigalow dataset since the three-pool model has the largest ELPD.

Uncertainty quantification

The average of the SOC change between 1978 and 1997 in fields 1, 2, and 3 in the Tarlee trial based on the three-pool model were −3.81, −3.47, and 7.12, respectively (Fig. 3a). Here the negative values denote that the first two fields were expected to lose carbon over the 20-year period. The management strategies that are used in fields 1, 2, and 3 are "Wheat-Wheat", "Wheat-Fallow", and "Wheat-Pasture", respectively. This average for three soil types of the Brigalow dataset, based on the three-pool model, between 1981 and 2000 were −4.37, −0.43, and −5.13, respectively (Fig. 3b). The hardware use and computing time information are provided in Section J of the Supplementary Material.

We can find the 95% credible interval for the amount of carbon in the soil by computing the upper and lower limits of the interval which are the 97.5th and 2.5th percentiles of the posterior distribution, respectively. These percentiles for the SOC process of each soil type in the Brigalow trial and each Tarlee field are presented in Figs. 2a and 4, respectively. Due to the wide range of soil carbon stocks in Fig. 2b we also provide a separate comparison of the 50th percentiles based on three and five-pool models for Brigalow in Supplementary Figures S1a and S1b, respectively in section Supplementary Material.

As mentioned earlier in Section "Prior Information", prior knowledge plays a significant role in the presence of small and sparse datasets. We compare the prior distributions with a histogram of the samples drawn from the posteriors of some main model parameters of the three and five-pool models that are the best and the more complex models, respectively, to highlight what we have learned about those parameters. Figure 5a,b show the difference between the prior and posterior of the decomposition rate of the SOC and BIO pools of the three-pool model in Tarlee and Brigalow, respectively. Also, Figure 6a,b show the difference between the prior and posterior of the decomposition rate of each pool of the five-pool model in Tarlee and Brigalow, respectively. Based on Figures 5 and 6, it is clear that we learn quite a lot about some parameters such as K_B and K_H, and we learn little new about other parameters, namely K_C and K_D as the posterior and prior are very similar.

We calculated the Gelman and Rubin's convergence diagnostics, $\hat{R}$ for the model parameters of the three-pool model of the Tarlee dataset and the one-pool model of the Brigalow dataset. They are presented in Supplementary Tables S11 and S12, respectively, in Section H of the supplementary material. Since the values of $\hat{R}$ are less than 1.2, there is no evidence of divergence.

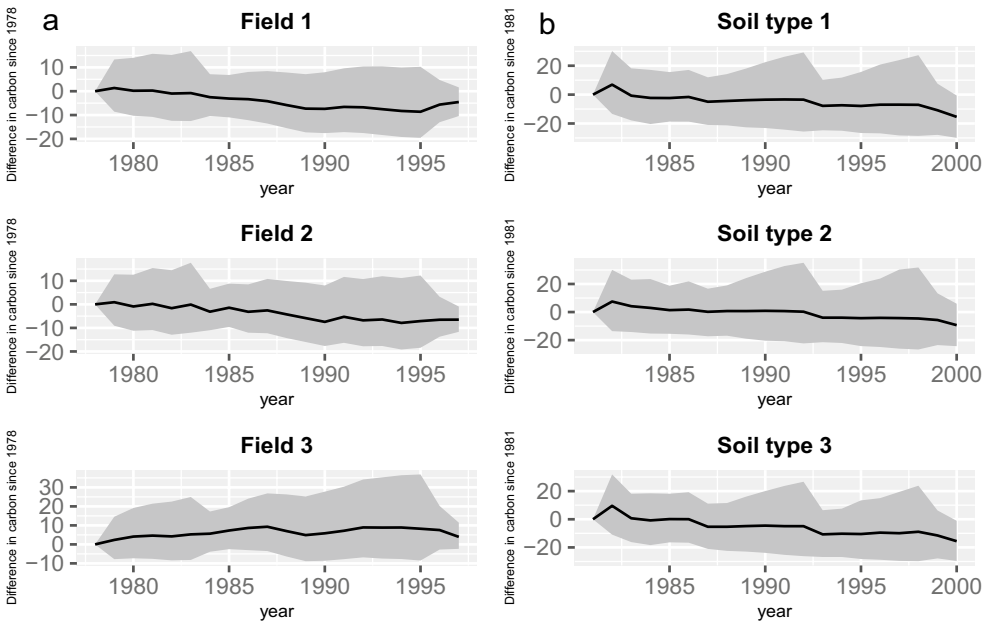

Figure 3. The expected difference of the SOC in each year from 1978 and 1981 in the (**a**) Tarlee and (**b**) Brigalow datasets, respectively, estimated based on the three-pool model. The change of the SOC stock in each field/soil type is indicated by solid line, and the gray shaded part is the area between the 2.5th and the 97.5th percentiles for the SOC process.

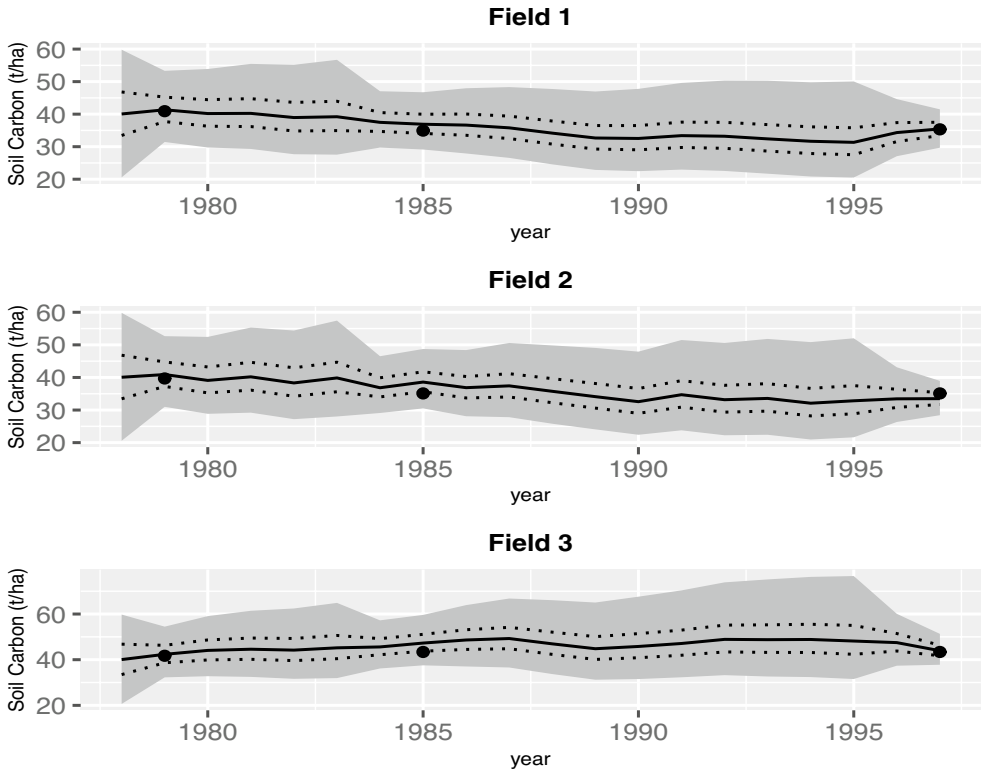

Figure 4. Soil organic carbon (SOC) dynamics in the three Tarlee fields. The gray shaded part is the area between the 2.5th and the 97.5th percentiles for the SOC process from the three-pool model. The 25th and the 75th percentiles for the SOC process are indicated by the dashed lines. The 50th percentile is shown by the solid line. The measured SOC values are indicated by filled dots.

Discussion

In this study, we have developed three new soil carbon models and compared them with the one-pool model in Clifford et al.[15] in the BHM framework, which allows us to think conditionally and critically about the parameters, the process, and the data that reside within a soil carbon model. To show these models are broadly applicable, we have implemented them for two datasets.

An important motivating question behind this study is whether multi-pool state-space models based on deterministic models such as RothC are fit for making inferences on soil carbon dynamics in commonly occurring situations where soil carbon measurements are monitored infrequently. In fitting models to two Australian datasets, we found a three-pool model (in both the cases of Tarlee and Brigalow) to have the best predictive ability of those models considered and to be better than a five-pool model, which is frequently adopted for its bio-geochemical realism. We conclude that the detail and realism included in statistical soil carbon models should consider the volume and quality of data available for making inferences. Indeed, this study has shown that some concessions in physical realism can lead to better predictive accuracy. This can be helpful for the IPCC, Paris agreement and Kyoto protocol's purposes, especially for national carbon accounting where datasets are sparse.

Furthermore, we have explored the effect of microbes and inert organic matter on the carbon cycle decomposition by adding microbial biomass and IOM pools in the Tarlee model in Clifford et al.[15]. In particular, based on the LFO-CV criterion, we have shown that the three-pool model, which includes microbial biomass and IOM pools, outperforms other models on the Tarlee and Brigalow datasets. The LFO-CV of the five-pool model is close to the three-pool model in its predictive ability for Tarlee but not for Brigalow. The reason is that the Brigalow dataset has more uninformative priors and sub-models than the Tarlee dataset. Both the Brigalow and Tarlee datasets exhibit relatively long, multi-year periods with no observation of any carbon pools, i.e. temporally sparse data. During those periods, all knowledge about the soil carbon process comes from the carbon inputs, the process dynamics and the model parameters through prior distributions. However, in the case of Brigalow, adding more pools to the model increased uncertainty in the soil carbon dynamics in each iteration of the particle filter process, causing wide variance which make it a poor predictor, typified by wide 95% credible intervals during those sparse periods. This result indicated that multi-pool models might not be as fit-for-purpose compared to some simpler models when used with sparse data over time.

In exploring soil carbon models with reduced complexity, we chose not to investigate a four-pool model. We could create such a model by combining the DPM and RPM components, for example. However, we deemed a four pool model to be too similar in structure to the five pool model, therefore not providing much additional variation in model complexity. Furthermore, our aims in this study were to explore the importance of microbe

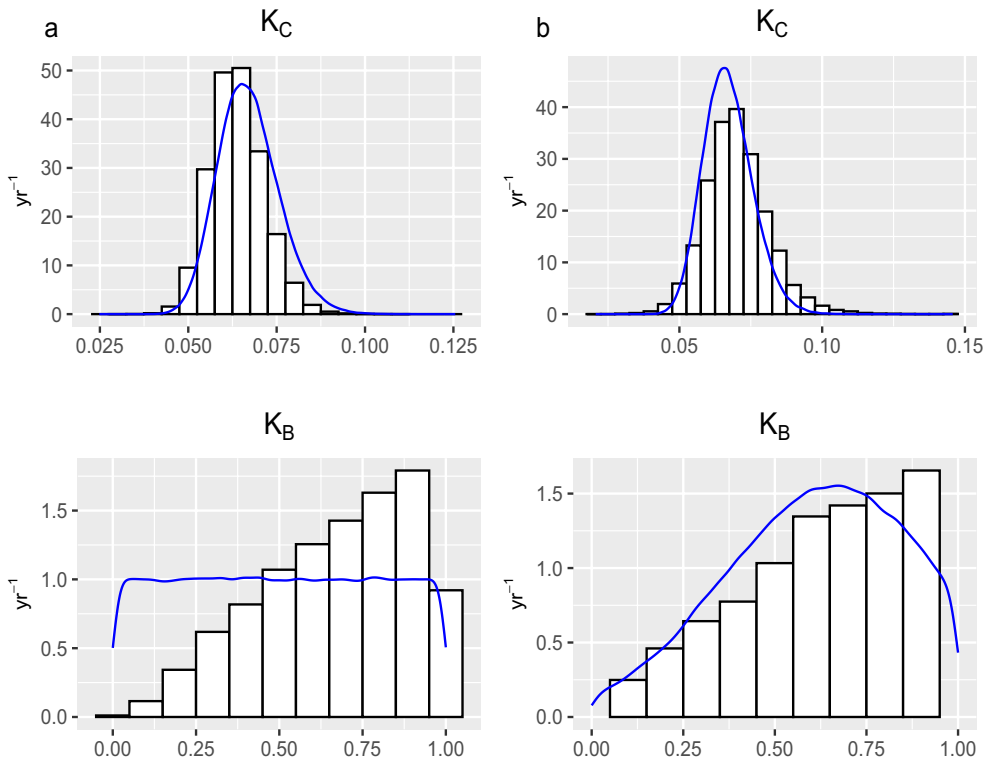

Figure 5. The marginal posterior distributions (histogram) of the SOC and BIO decomposition rates, K_C and K_B, respectively, in (**a**) Tarlee and (**b**) Brigalow. The histograms correspond to the three-pool model in both Brigalow and Tarlee. The blue densities are the prior distributions of the SOC and BIO decomposition rates.

and inert organic matter pools because they are fundamentally different from other soil carbon pools (the former being constrained in its total pool size and the latter being stable over very long time scales). The range of models used in this study provides valuable insight into whether the complexity of the RothC model is warranted when datasets are temporally sparse.

We have shown that, the three-pool model that was found to be best suited to the Brigalow and Tarlee datasets in this study can be used to obtain good fits to observational data and can be used to estimate with uncertainty the net gain or loss of carbon overtime at each study site.

Since both datasets used in this study are not large, we have used the LFO-CV criterion for model evaluation. It is noteworthy to mention that this criterion is computationally expensive when used with a larger dataset since it requires repeating the MCMC every time a data point is introduced. Based on our experiences here, other criteria such as Pareto smoothed importance sampling LFO-CV (PSIS-LFO-CV)[31] or widely applicable information criterion (WAIC)[46] may be more relevant methods for large datasets.

We have successfully demonstrated applying advanced Bayesian methods in Davoudabadi et al.[30] to more complex SOC models. We have shown the importance of these methods in inference on soil carbon dynamics, especially in scenarios where uncertainty quantification plays a significant role in carbon sequestration accounting.

In this study, we consider the effect of the microbial biomass pool on the carbon emission decomposition rate with the limitation on the maximum size of microbes, which is 5% of the total SOC. Through this limitation, we have prevented too much carbon from entering the microbial pool and where excess, the extra amount is rejected by rejecting BIO state trajectories in the MCMC algorithm. Furthermore, the precision of the single-pool statistical model of Clifford et al.[15] has been improved upon by adding a microbial biomass and inert soil carbon pools to that model. It is possible that we could improve the growth of the population of microbes by considering a dynamic process in future studies. We could fit a model (e.g. perhaps a logistic population model with a carrying capacity) to the growth of the size of microbes. In this case, the extra amount of carbon in the BIO pool could be diverted into the other pools into which carbon could be cycled. This will be considered in future research.

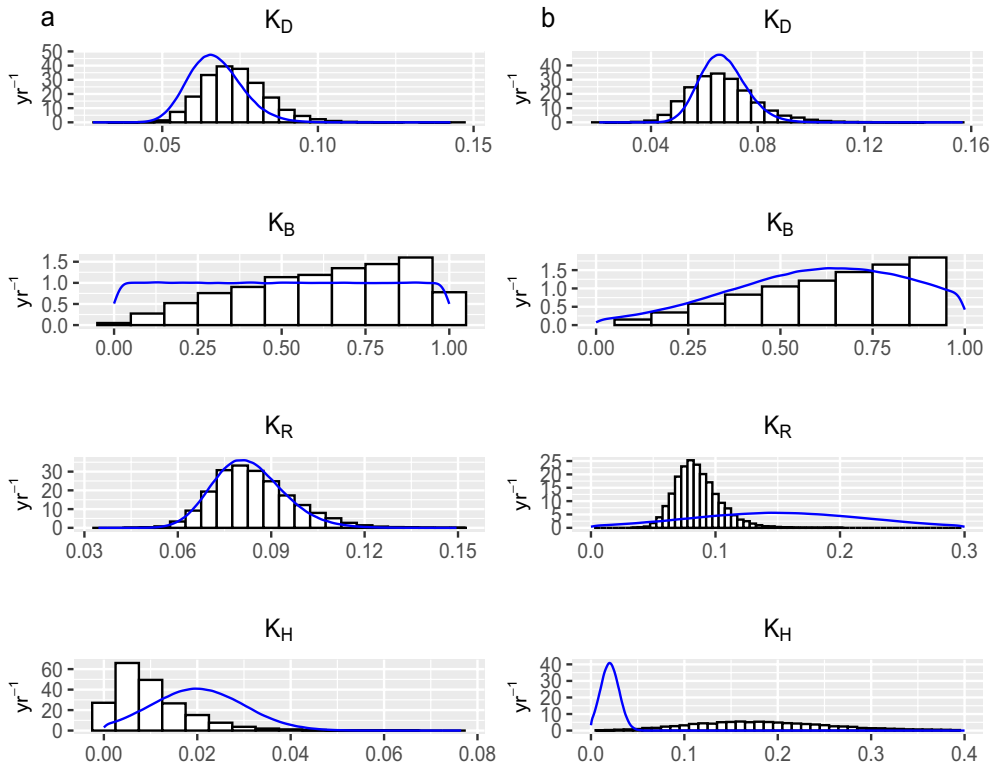

Figure 6. The marginal posterior distributions (histogram) of the DPM, BIO, RPM and HUM decomposition rates K_D, K_B, K_R, and K_H, respectively, in (**a**) Tarlee and (**b**) Brigalow. The histograms are correspond to the five-pool BIO-K model in both Brigalow and Tarlee. The blue densities are the prior distributions of the DPM, BIO, RPM and HUM decomposition rates.

Data availability

Dataset can be accessed online at:https://doi.org/10.4225/08/54F0786D6D923.

Code availability

Code for our methods and models is available at:https://github.com/MJDavoudabadi/Modelling-soil-carbon-Tarlee-and-Brigalow.

Received: 18 November 2023; Accepted: 1 February 2024
Published online: 08 February 2024

References

1. Adams, M. et al. Managing the soil-plant system to mitigate atmospheric CO2. Tech. Rep., Discussion paper for the Soil Carbon Sequestration Summit, 31 January-2 February 2011. The United States Studies Centre at the University of Sydney. (2011).
2. Shi, Z. *et al.* The age distribution of global soil carbon inferred from radiocarbon measurements. *Nat. Geosci.* **13**, 555–559 (2020).
3. Jenkinson, D. S., Hart, P. B. S., Rayner, J. H. & Parry, L. C. Modelling the turnover of organic matter in long-term experiments at Rothamsted. *INTECOL Bull.* **15**, 1–8 (1987).
4. Jenkinson, D. S. The turnover of organic carbon and nitrogen in soil. *Phil. Trans. R. Soc. Lond. B* **329**, 361–368 (1990).
5. Parton, W. J., Stewart, J. W. & Cole, C. V. Dynamics of C, N, P and S in grassland soils: A model. *Biogeochemistry* **5**, 109–131 (1988).
6. Peter, C., Fiore, A., Hagemann, U., Nendel, C. & Xiloyannis, C. Improving the accounting of field emissions in the carbon footprint of agricultural products: A comparison of default ipcc methods with readily available medium-effort modeling approaches. *Int. J. Life Cycle Assess.* **21**, 791–805 (2016).
7. Farina, R., Coleman, K. & Whitmore, A. P. Modification of the RothC model for simulations of soil organic c dynamics in dryland regions. *Geoderma* **200**, 18–30 (2013).
8. Capon, T., Harris, M. & Reeson, A. *Soil Carbon Sequestration Market Based Instruments (mbis): A Literature Review* (University of Sydney, Sydney, 2010).
9. Jones, J. W. *et al.* Integrating stochastic models and in situ sampling for monitoring soil carbon sequestration. *Agric. Syst.* **94**, 52–62 (2007).
10. Koo, J. *et al.* Estimating soil carbon in agricultural systems using ensemble Kalman filter and DSSAT-Century. *Trans. ASABE* **50**, 1851–1865 (2007).
11. Post, J., Hattermann, F. F., Krysanova, V. & Suckow, F. Parameter and input data uncertainty estimation for the assessment of long-term soil organic carbon dynamics. *Environ. Model. Softw.* **23**, 125–138 (2008).
12. Juston, J., Andrén, O., Kätterer, T. & Jansson, P. Uncertainty analyses for calibrating a soil carbon balance model to agricultural field trial data in Sweden and Kenya. *Ecol. Model.* **221**, 1880–1888 (2010).
13. Paul, K. I., Polglase, P. J. & Richards, G. P. Sensitivity analysis of predicted change in soil carbon following afforestation. *Ecol. Model.* **164**, 137–152 (2003).

14. Stamati, F. E., Nikolaidis, N. P. & Schnoor, J. L. Modeling topsoil carbon sequestration in two contrasting crop production to set-aside conversions with RothC-calibration issues and uncertainty analysis. *Agric. Ecosyst. Environ.* **165**, 190–200 (2013).
15. Clifford, D. *et al.* Rethinking soil carbon modelling: A stochastic approach to quantify uncertainties. *Environmetrics* **25**, 265–278 (2014).
16. Gurung, R. B., Ogle, S. M., Breidt, F. J., Williams, S. A. & Parton, W. J. Bayesian calibration of the daycent ecosystem model to simulate soil organic carbon dynamics and reduce model uncertainty. *Geoderma* **376**, 114529 (2020).
17. Paul, E. & Clark, F. Soil microbiology and biochemistry academic press. New York, USA (1996).
18. Luo, Y. *et al.* Toward more realistic projections of soil carbon dynamics by earth system models. *Global Biogeochem. Cycles* **30**, 40–56 (2016).
19. Blagodatsky, S., Blagodatskaya, E., Yuyukina, T. & Kuzyakov, Y. Model of apparent and real priming effects: Linking microbial activity with soil organic matter decomposition. *Soil Biol. Biochem.* **42**, 1275–1283 (2010).
20. Frey, S. D., Lee, J., Melillo, J. M. & Six, J. The temperature response of soil microbial efficiency and its feedback to climate. *Nat. Clim. Chang.* **3**, 395–398 (2013).
21. Moorhead, D. L. & Sinsabaugh, R. L. A theoretical model of litter decay and microbial interaction. *Ecol. Monogr.* **76**, 151–174 (2006).
22. Riley, W. *et al.* Long residence times of rapidly decomposable soil organic matter: Application of a multi-phase, multi-component, and vertically resolved model (bams1) to soil carbon dynamics. *Geosci. Model Dev.* **7**, 1335–1355 (2014).
23. Allison, S. D., Wallenstein, M. D. & Bradford, M. A. Soil-carbon response to warming dependent on microbial physiology. *Nat. Geosci.* **3**, 336–340 (2010).
24. German, D. P., Marcelo, K. R., Stone, M. M. & Allison, S. D. The m ichaelis-m enten kinetics of soil extracellular enzymes in response to temperature: A cross-latitudinal study. *Glob. Change Biol.* **18**, 1468–1479 (2012).
25. Wang, G., Post, W. M. & Mayes, M. A. Development of microbial-enzyme-mediated decomposition model parameters through steady-state and dynamic analyses. *Ecol. Appl.* **23**, 255–272 (2013).
26. Li, J., Wang, G., Allison, S. D., Mayes, M. A. & Luo, Y. Soil carbon sensitivity to temperature and carbon use efficiency compared across microbial-ecosystem models of varying complexity. *Biogeochemistry* **119**, 67–84 (2014).
27. Woolf, D. & Lehmann, J. Microbial models with minimal mineral protection can explain long-term soil organic carbon persistence. *Sci. Rep.* **9**, 1–8 (2019).
28. Xie, H. W., Romero-Olivares, A. L., Guindani, M. & Allison, S. D. A Bayesian approach to evaluation of soil biogeochemical models. *Biogeosciences* **17**, 4043–4057. https://doi.org/10.5194/bg-17-4043-2020 (2020).
29. Skjemstad, J. O., Spouncer, L. R., Cowie, B. & Swift, R. S. Calibration of the Rothamsted organic carbon turnover model (RothC ver. 26.3), using measurable soil organic carbon pools. *Soil Res.* **42**, 79–88 (2004).
30. Davoudabadi, M. J., Pagendam, D., Drovandi, C., Baldock, J. & White, G. Advanced Bayesian approaches for state-space models with a case study on soil carbon sequestration. *Environ. Model. Softw.* **136**, 104919 (2020).
31. Bürkner, P.-C., Gabry, J. & Vehtari, A. Approximate leave-future-out cross-validation for Bayesian time series models. J. Stat. Comput. Simul. 1–25 (2020).
32. Huang, Y., Liang, C., Duan, X., Chen, H. & Li, D. Variation of microbial residue contribution to soil organic carbon sequestration following land use change in a subtropical karst region. *Geoderma* **353**, 340–346 (2019).
33. Skjemstad, T. J. & Spouncer, L. NCAS calibration and verification data v1.. *CSIRO Data Collect.*https://doi.org/10.4225/08/54F0786D6D923 *(2003).*
34. Luo, Z., Wang, E. & Sun, O. J. Soil carbon change and its responses to agricultural practices in Australian agro-ecosystems: A review and synthesis. *Geoderma* **155**, 211–223 (2010).
35. Allenby, G. M. & Rossi, P. E. Hierarchical Bayes models. The Handbook of Marketing Research: Uses, Misuses, and Future Advances 418–440 (2006).
36. Berliner, L. M. Hierarchical Bayesian time series models. In Maximum entropy and Bayesian methods, 15–22 (Springer, 1996).
37. Cressie, N. & Wikle, C. K. Statistics for Spatio-Temporal Data (John Wiley & Sons, 2015).
38. Deligiannidis, G., Doucet, A. & Pitt, M. K. The correlated pseudomarginal method. *J. Royal Stat. Soc. Seri. B (Stat. Methodol.)* **80**, 839–870 (2018).
39. Doucet, A., De Freitas, N., Murphy, K. & Russell, S. Rao-Blackwellised particle filtering for dynamic Bayesian networks. In Proceedings of the Sixteenth conference on Uncertainty in Artificial Intelligence, 176–183 (Morgan Kaufmann Publishers Inc., 2000).
40. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
41. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434–455 (1998).
42. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016 (2014).
43. Vehtari, A. *et al.* A survey of Bayesian predictive methods for model assessment, selection and comparison. *Stat. Surv.* **6**, 142–228 (2012).
44. Lal, R. Managing soils and ecosystems for mitigating anthropogenic carbon emissions and advancing global food security. *Bioscience* **60**, 708–721 (2010).
45. Falloon, P., Smith, P., Coleman, K. & Marshall, S. How important is inert organic matter for predictive soil carbon modelling using the Rothamsted carbon model?. *Soil Biol. Biochem.* **32**, 433–436 (2000).
46. Watanabe, S. & Opper, M. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. **11** (2010).

Acknowledgements

We would like to thank CSIRO for providing the data used in this study. MJD was supported by QUT-CSIRO Digital Agriculture Scholarship and a CSIRO Digital Agriculture Top-Up Scholarship. CD was supported by an Australian Research Council Discovery Project (DP200102101). We gratefully acknowledge the computational resources provided by QUT's High Performance Computing (HPC) and Research Support Group.

Author contributions

M.J.D.: Conceptualization, Methodology, Formal analysis, Validation, Writing - Original draft, Coding, Interpretation. D.P. and C.D.: Methodology, Formal analysis, Validation, Critical revision of the manuscript. J.B.: Methodology, Critical revision of the manuscript. G.W.: Methodology, Formal analysis, Validation, Supervision, Critical review of the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-53516-z.

Correspondence and requests for materials should be addressed to M.J.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.