Available online at www.sciencedirect.com
ScienceDirect
Agricultural Systems 94 (2007) 52–62
AGRICULTURAL SYSTEMS
www.elsevier.com/locate/agsy

Integrating stochastic models and *in situ* sampling for monitoring soil carbon sequestration

J.W. Jones [a,*], J. Koo [a], J.B. Naab [b], W.M. Bostick [a], S. Traore [c], W.D. Graham [a]

[a] *University of Florida, Institute of Food and Agricultural Sciences (IFAS), P.O. Box 110570, Gainesville, FL 32611, USA*
[b] *SARI, Wa, Ghana*
[c] *ICRISAT, Bamako, Mali*

Received 20 February 2005; accepted 28 June 2005

Abstract

Participation in carbon (C) markets could provide farmers in developing countries incentives for improving soil fertility. However carbon traders need assurances that contract levels of C are being achieved. Thus, methods are needed to monitor and verify soil C changes over time and space to determine whether target levels of C storage are being met. Because direct measurement over the large areas needed to sequester contract amounts of C in soil is not practical, other approaches are necessary. An integrated approach is described in which an Ensemble Kalman Filter (EnKF) is used to assimilate *in situ* soil carbon measurements into a stochastic soil C model to estimate soil C changes over time and space. The approach takes into account errors in *in situ* measurements and uncertainties in the model to estimate mean and variance of soil C for each land unit within a larger land area. The approach requires initial estimates of soil C over space along with uncertainties in these estimates. Model predictions are made to estimate soil C for the next year, *in situ* soil C measurements update these predictions using maximum likelihood methods, and the spatial pattern of soil C mean, variance, and covariance thus evolve over time. This approach can also be used to provide yearly estimates of the changes in soil C over multiple fields, the variance in those estimates, and aggregate soil carbon mean and variance values each year. In this paper, the use of the EnKF is shown for an area in Ghana with 12 fields, comparing numbers of fields sampled each year and ways of selecting which fields to sample each year. The model predicts soil C changes over time using first order decomposition of existing soil C and addition of C from plant residues. The lowest intensity sampling method (sampling only 1/4 of the fields per year) resulted in the highest level of uncertainty in aggregate soil C estimate. Rotating sample fields each year improved the performance of the EnKF. These results demonstrated a quantifiable tradeoff between field sampling intensity and uncertainty in aggregate soil C estimates. The framework could be modified to use more complex biophysical models and to assimilate remote sensing data.

Keywords: Soil carbon; Stochastic model; Uncertainty; Data assimilation

1. Introduction

An accounting system is needed if soil carbon (C) sequestration is to become an accepted mechanism for reducing atmospheric CO_2 levels (Antle and Uehara, 2002). Such an accounting system must be able to provide estimates of soil C mass over an area large enough to interest potential C buyers. Although it is impossible to measure this large mass directly, samples of soil can be collected at some spatial frequency and taken to the laboratory to measure sample C concentration (mass fraction basis). Soil C mass can then be computed by multiplying concentration by soil bulk density, depth of sampling, and field area. However, there are a number of problems associated with such measurements. There are tradeoffs between costs and accuracy. First, it is costly to collect and analyze samples from large numbers of fields. A second problem is that each measurement has a high level of uncertainty; thus it may be necessary to collect large numbers of samples in

* Corresponding author.
E-mail address: jimj@ufl.edu (J.W. Jones).

0308-521X/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.agsy.2005.06.023

order to achieve an acceptable level of uncertainty. There is considerable spatial variability in soil C levels due to spatial variability of soil characteristics and land management history. Thus, uncertainties in estimates of aggregate soil C mass may be large unless the design of the monitoring system includes a relatively large number of samples over space. Reducing the number of samples to keep costs low will inherently result in more uncertainty in estimates of soil C changes over time.

Another challenging characteristic of this problem is that yearly changes in soil C are small relative to uncertainty in soil C measurements. Standard errors of C measurements in a field may be larger than annual changes of soil C. West and Post (2002) found an average soil C sequestration rate of 570 kg ha^{-1} yr^{-1} for no-till vs. conventional tillage when they analyzed data from 67 long-term experiments from around the world. In a 10-year study in Burkina Faso, the increase in soil C averaged 116 and 377 kg ha^{-1} yr^{-1} for treatments with low and high levels of both inorganic fertilizer and manure, respectively (Pichot et al., 1981). Lal (2000) observed annual rates of soil C increase of about 400 kg ha^{-1} yr^{-1} under no-till management over a 3-year experiment aimed at restoring soil C in western Nigeria. Jones et al. (2004) computed measurement standard errors of about 1000 kg ha^{-1} yr^{-1} if measurement error is 0.04% on a mass basis, a value in the mid range of those reported by Yost et al. (2002). Additional variability could be introduced from field sampling methods and inherent spatial variability of soil C in a field. One way to reduce uncertainty would be to use geostatistical methods, but this could require a relatively large number of samples in each field being monitored. It may also be possible to use geostatistical methods to estimate aggregate soil C over large areas (Yost et al., 2002).

Biophysical models can also be used to estimate soil C and its changes under different weather, soil, and management practices (Parton et al., 1988, 1994; Jones et al., 2002). However, although these models produce precise estimates, they are imperfect and their parameters for specific fields are also uncertain. Thus, errors exist in estimates of soil C from model predictions as well as from field measurements. Techniques exist to combine models and measurements to obtain better estimates of system states and model parameters. The Kalman Filter (Maybeck, 1979; Welch and Bishop, 2002) approach first uses a model to predict the state of a system, and then uses measurements to update the estimates, taking into account errors in measurements and predictions.

Variations of the Kalman Filter, originally developed for linear models, have been developed for non-linear models (e.g., Albiol et al., 1993; Graham, 2002). One variation, the Ensemble Kalman Filter (Burgers et al., 1998; Eknes and Evensen, 2002; Margulis et al., 2002), was used by Jones et al. (2004) to estimate soil C and a decomposition rate parameter over time for a single field using a non-linear model. In that study, Jones et al. (2004) demonstrated that assimilation of *in situ* soil C measurements could reduce errors in soil C estimates at the field scale and annual changes in C. However, they did not address the use of the EnKF for a spatially variable area.

The purpose of this paper is to describe and demonstrate an integrated approach for monitoring soil C over large agricultural areas, combining *in situ* sampling and modeling. The approach could be used to design a soil C monitoring program to meet specific goals and tailored to specific regional situations by analyzing tradeoffs between costs of sampling vs. accuracy of aggregate soil C estimates. It could also be used operationally to produce aggregate estimates of soil C and uncertainty of those estimates. In this paper, we first present the relationships used in the EnKF and then demonstrate its use for 12 fields in Ghana. The simple non-linear stochastic soil C model used by Jones et al. (2004) was extended in this paper to predict soil C over space and time. The resulting spatio-temporal model, implemented in the EnKF, inherently takes into account autocorrelations in soil C among different fields; an initial estimate of the covariance matrix of state variables evolves as measurements made over space and time are assimilated. Results are shown for the area in Ghana to demonstrate tradeoffs between sampling intensity and accuracy of estimates using different sampling schemes.

2. Soil C monitoring at the aggregate scale: the Ensemble Kalman Filter

The problem is to monitor soil C in a specified soil depth (say 20 cm) over time in an area in which fields are managed to participate in a contract for sequestering C in soils. Thus, soil C in all participating fields must be aggregated to produce estimates of total C and its uncertainty in a contract project. A field in the contract area is defined as an area of land that is managed as a unit. This does not mean that the field is uniform, but instead that its boundaries are delineated and the farmer attempts to manage it uniformly. The total C in the contract is the sum of C in each of these fields. The various fields may vary in size, crops, and soil C levels. The EnKF provides the integrative framework for estimating soil C.

There are three main components in the EnKF: data, models and assimilation/estimation. For this application, data would include field measurements of soil C, but it could also include measurement of other variables using field sampling or remote sensing. Measurements may not be taken in all fields in a project in any year, and samples may not be taken every year. Accuracy will be affected by sampling design. The model in the EnKF predicts the state of the system, the mass of soil C in each field to a specified depth of soil (kg[C] ha^{-1}), as it changes with time (over years, in this case). Even if measurements are not made in each field each year, the model predicts soil C in each field every year. The assimilation component combines data and model predictions using procedures that minimize estimation errors. The predicted state of the system is updated using measurements at times when they are made;

model parameters can also be updated. Soil C sequestration is the difference between aggregate soil C at a point in time and initial soil C at the beginning of a project.

Fig. 1 shows the basic scheme of the EnKF. Each field in the project is included in the EnKF as a discrete spatial unit with its own soil C mass that varies over time due to natural processes and management activities. The EnKF includes a model that describes the dynamic changes of soil C for each field. Management may vary over time (i.e., crop rotations) and over space. Soil C and model parameters may be correlated over space. Prior knowledge of this spatial correlation helps the EnKF results converge to an optimal estimate in less time and cost than the case when there is no spatial correlation. Such correlation is taken into account explicitly in the EnKF as shown below.

The model may be simple or complex; the EnKF procedure is the same for both, although details of data assimilation calculations depend on the model and measurements. Jones et al. (2004) developed the EnKF for a single field using a simple stochastic model with one state variable (soil C) and one uncertain parameter. Koo et al. (2003) demonstrated the use of the EnKF with the DSSAT model (Jones et al., 2003; Gijsman et al., 2002) to simulate daily changes in soil C depending on daily weather data, soil properties, and crop management for each field in a hypothetical study area. This model simulates crop biomass production and soil C changes over time, and can account for crop rotations and other variations in management. There is an advantage of using this type of model in that it can be calibrated to simulate fluctuations in biomass production and changes in soil C from year to year based on weather variability. It also can be tailored to new areas due to the physical and physiological relationships in the model. But, this detailed soil-crop model requires parameters and input data that may be difficult to obtain for each field (soil properties, daily weather, and management) and a simple model may be preferred. Data that could be assimilated depend on the model. In the case of the simple models presented by Jones et al. (2004) and Bostick et al. (2003), remote sensing data could be used to estimate crop biomass, an input to those models.

The EnKF assimilation component uses inputs from the model as well as from *in situ* measurements to improve soil C estimates and model predictions over time. Periodic field measurements of soil C in all or a subset of fields improve estimates of soil C and model parameters so that future predictions are more accurate. These measurements are used to adjust predictions of soil C, not only in those fields where it is measured but also in other fields that were not sampled. The EnKF uses Monte-Carlo simulation to generate an ensemble of state variable realizations that are each propagated over time and assimilated with measurements using the Kalman update equations.

2.1. Soil carbon model

The model in the EnKF is stochastic; that is the variables in the model are random and can be characterized by a probability density function. Knowledge of the uncertainty in model predictions is necessary as is knowledge of the uncertainty in measurements of soil C and other variables. In this paper, we extend the model published by Jones et al. (2004) for multiple fields and consider only *in situ* sampling of soil C in the data assimilation process shown schematically in Fig. 1.

The model has one state variable in each field i, the mean mass of carbon per ha ($X(i, t)$, kg[C] ha^{-1}) in the top 20 cm of soil. Changes in soil C are simulated dynamically on a yearly basis (time step of one year) for each field. The model also has one unknown parameter for each field, $R(i)$, the fraction of soil C that is decomposed per year

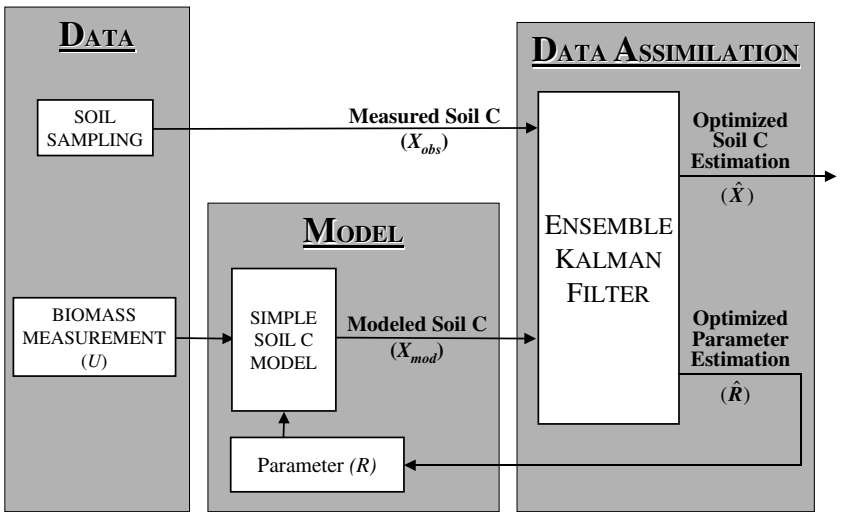

Fig. 1. Schematic diagram of the framework for integrating soil C sampling and models for monitoring soil carbon. An Ensemble Kalman Filter (EnKF) is used to combine information from the different sources to produce estimates of the mean aggregate soil C over a number of fields and to provide an estimate of the uncertainty (variance) in this estimate.

(yr^{-1}). It is assumed that R varies among fields, is constant over time, but is not known with certainty. The equations that describe the dynamics of this system are adapted from Jones et al. (2004), explicitly showing variable designations for each field:

$$\begin{aligned} X_{\mathrm{mod}}(i,t) &= X_{\mathrm{mod}}(i,t-1) - R(i)\cdot X_{\mathrm{mod}}(i,t-1) \\ &\quad + b\cdot U(i,t-1) + \varepsilon_{\mathrm{mod}}(i,t) \\ R(i) &= R_0(i) + \varepsilon_R(i) \end{aligned} \tag{1}$$

where $X_{\mathrm{mod}}(i, t)$ is the modeled soil C of field i at time t; i is the identifier for each field ($i = 1, 2, 3, \ldots, F$); F is the total number of fields; b is the fraction of crop residue C that was added to the soil in year $t-1$ and remains in year t; $U(i, t-1)$ is the amount of C in crop residue that is added to the soil in year $t-1$ in field i; $\varepsilon_{\mathrm{mod}}(i, t)$ is the model error term for field i and time t; $R(i)$ is the rate parameter for field i; $R_0(i)$ is the initial estimate of the parameter R for field i; $\varepsilon_R(i)$ is the error in the initial estimate for R, field i.

Model error ($\varepsilon_{\mathrm{mod}}(i, t)$) includes uncertainties in $U(i, t-1)$ and b as well as uncertainties due to the fact that the model is a simplification of reality. We assume that model errors and the parameter estimator error are normally distributed with zero-mean and are not correlated. Thus:

$$\begin{aligned} \varepsilon_{\mathrm{mod}}(i,t) &\sim N(0, \sigma^2_{\mathrm{mod}}(i)) \\ \varepsilon_R(i) &\sim N(0, \sigma^2_R(i)) \end{aligned} \tag{2}$$

where $\sigma^2_{\mathrm{mod}}(i)$ is the variance of model error for soil C in field i, $\sigma^2_R(i)$ is the variance of error for estimate of parameter R in field i.

The model error ($\varepsilon_{\mathrm{mod}}(i, t)$) is a random process that changes over time but is uncorrelated with time (i.e., white noise), whereas the decomposition rate parameter error ($\varepsilon_R(i)$) is a random variable that does not change with time.

An important characteristic of the stochastic model is that state variables may be correlated over space. The spatial correlation is expressed in the EnKF as the covariance matrix among all state variables; an estimate of the covariance matrix is given by $\boldsymbol{P}(t)$. This matrix has diagonal elements that are estimates of variance of soil C estimate at time t in each of the F fields and estimates of variance of estimates of the uncertain soil parameter at time t in each field. This matrix is written as:

$$\boldsymbol{P}(t) = \begin{bmatrix} P_{XX,t}(1,1) & \cdots & P_{XX,t}(1,F) & P_{XR,t}(1,1) & \cdots & P_{XR,t}(1,F) \\ \vdots & (XX) & \vdots & \vdots & (XR) & \vdots \\ P_{XX,t}(F,1) & \cdots & P_{XX,t}(F,F) & P_{XR,t}(F,1) & \cdots & P_{XR,t}(F,F) \\ P_{RX,t}(1,1) & \cdots & P_{RX,t}(1,F) & P_{RR,t}(1,1) & \cdots & P_{RR,t}(1,F) \\ \vdots & (RX) & \vdots & \vdots & (RR) & \vdots \\ P_{RX,t}(F,1) & \cdots & P_{RX,t}(F,F) & P_{RR,t}(F,1) & \cdots & P_{RR,t}(F,F) \end{bmatrix} \tag{3}$$

For example, the variable $P_{XX,t}(1, 1)$ is an estimate of the variance of soil C estimate in field 1 at time t, and $P_{XX,t}(1, F)$ is the estimate of covariance between soil C in field 1 and soil C in field F at time t. The variable $P_{RR,t}(1, 1)$ is the variance of the estimate of soil C decomposition rate parameter at time t in field 1 and $P_{RR,t}(1, F)$ is the covariance between estimates of the decomposition rate parameter in fields 1 and F at time t. The $P_{XR,t}(1, 1)$ is an estimate of the covariance between soil C and R in field 1 at time t.

Initial estimates are needed for all terms in this covariance matrix ($\boldsymbol{P}(0)$). The diagonal elements are more easily determined since they are variances of estimates of soil C and R. However, obtaining initial estimates of non-diagonal elements (e.g. covariance of soil C in different fields, R in different fields, or soil C and R in different fields) requires an understanding of the spatial structure of variables and the inter-variable correlations. For initializing soil C, spatial structure and correlations can be estimated in part by geostatistical analyses of spatial soil C data. In general, soil C will likely be correlated among fields. Yost et al. (1993) showed correlations of soil C over distances of 10 km or more in Hawaii. In a companion study, Bostick (personal communication) found spatial correlation of soil C up to 1 km in Mali.

Note that the distance between fields is not explicit in Eq. (3). After the covariance matrix is initialized, distance between fields is not needed to calculate the evolution of the covariance matrix. Instead, the covariance matrix can be calculated at each time step from the multiple realizations of the Monte-Carlo simulation.

2.2. Measurements

Soil C measurements may be made each year or less frequently in all or a fraction of the fields; measurements of $R(i)$ are not possible. Thus, the model has $2F$ variables that are to be estimated at each time t ($X_{\mathrm{mod}}(i, t)$ and $R(i)$). It is necessary to express the observed soil C values in terms of the true value and measurement error. This measurement equation is written as:

$$X_{\mathrm{obs}}(i,t) = X(i,t) + \varepsilon_{\mathrm{obs}}(i,t) \tag{4}$$

where $X_{\mathrm{obs}}(i, t)$ is the measured soil C of field i at time t; $\varepsilon_{\mathrm{obs}}(i, t)$ is the measurement error term for field i at time t; $X(i, t)$ is the true soil C in field i at time t.

It is assumed that the soil C measurement errors are zero-mean, normally distributed, independent in time and space, and independent from $X_{\text{mod}}(i, t)$ and $R(i)$, thus:

$$\varepsilon_{\text{obs}}(i,t) \sim N(0, \sigma^2_{\text{obs}}(i))$$

where $\sigma^2_{\text{obs}}(i)$ is the variance of soil C measurement in field i.

These equations explicitly define the relationships between model state variables and measurements, and they form the basis for including measurement error in the EnKF. One does not know $X(i, t)$, the true soil C values in each field at time t. The model, Eq. (1), estimates the true field-specific soil C values, and Eq. (4) models measurements; both have uncertainties.

Measurements (or estimates) of biomass production are also needed for each field i and for each year t ($U(i, t)$) for simulating the model for each field (Eq. (1)). Errors in measuring $U(i, t)$ are included in the model prediction error, $\varepsilon_{\text{mod}}(i, t)$ in this paper.

2.3. Data assimilation

The data assimilation step updates model-predicted estimates of all state variables at any time when measurements are made. This is written mathematically as:

$$\begin{pmatrix} \widehat{X}(1,t) \\ \vdots \\ \widehat{X}(F,t) \\ \widehat{R}(1) \\ \vdots \\ \widehat{R}(F) \end{pmatrix} = \begin{pmatrix} X_{\text{mod}}(1,t) \\ \vdots \\ X_{\text{mod}}(F,t) \\ R(1) \\ \vdots \\ R(F) \end{pmatrix} + \begin{bmatrix} K_{1,1}(t) \cdots K_{1,M}(t) \\ \vdots \\ K_{F,1}(t) \cdots K_{F,M}(t) \\ K_{F+1,1}(t) \cdots K_{F+1,M}(t) \\ \vdots \\ K_{2F,1}(t) \cdots K_{2F,M}(t) \end{bmatrix} \cdot \begin{pmatrix} X_{\text{obs}}(1,t) - X_{\text{mod}}(1,t) \\ \vdots \\ X_{\text{obs}}(M,t) - X_{\text{mod}}(M,t) \end{pmatrix} \tag{5}$$

where $\widehat{X}(i,t)$ is the updated (filtered) estimate of soil C in field i at time t, $\widehat{R}(i)$ is the updated (filtered) estimate of decomposition rate in field i at time t, $K_{i,j}(t)$ is the Kalman gain elements at time t and M is the number of measurements made. In this example M is less than or equal to F since the only measurement is soil C and there are F fields that could be measured.

This equation is written for the case in which soil C is measured in each of M fields. The bracket to the far right in Eq. (5) expresses the differences between observed and predicted soil C values for each sampled field. Expanding this equation for $\widehat{X}(1,t)$, one can see that the updated soil C value is the predicted value plus the sum of each Kalman gain factor multiplied by the difference between observed and predicted soil C for each field in which a measurement is made. Thus, if soil C values are spatially correlated, the Kalman gain elements, K, and measurements in each field will influence updated values of soil C and R in all fields, measured or not. When the component K values are small, updated state variable values will be near those that were predicted by the model. When the K values are large, updated state variable values will be closer to measured values.

The Kalman gain is computed from the covariance matrix, model predictions, and measurements. Written in matrix notation, the Kalman gain is:

$$K(t) = P(t)H(t)^{\text{T}}[H(t)P(t)H(t)^{\text{T}} + W(t)]^{-1} \tag{6}$$

where $K(t)$ is the Kalman gain matrix, $2F$ rows and M columns; $H(t)$ is the measurement matrix, relating observations to model state variables; $W(t)$ is the measurement error matrix.

The measurement error matrix ($W(t)$) is an $M \times M$ matrix with diagonal elements of $\sigma^2_{\text{obs}}(i)$ and all other elements equal to 0 (measurements across fields are independent from each other). The measurement matrix that relates model variables to observations, $H(t)$, is written as:

$$H(t) = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ \\ M-1 \\ M \end{array} \begin{array}{c} \overbrace{1\ 2\ 3 \cdots F-1\ F}^{X} \quad \overbrace{1\ 2\ 3 \cdots F-1\ F}^{R} \\ \begin{bmatrix} 1\,0\,0 \cdots 0\,0 & 0\,0\,0 \cdots 0\,0 \\ 0\,1\,0 \cdots 0\,0 & 0\,0\,0 \cdots 0\,0 \\ \vdots & \vdots \\ 0\,0\,0 \cdots 1\,0 & 0\,0\,0 \cdots 0\,0 \\ 0\,0\,0 \cdots 0\,1 & 0\,0\,0 \cdots 0\,0 \end{bmatrix} \end{array} \tag{7}$$

The number of rows in the $H(t)$ matrix is equal to the number of measurements. The number of columns is equal to the total number of states in the model, including all $X_{\text{mod}}(i, t)$ and $R(i)$, in this case, $2F$ columns. Note that all elements in the right half of the matrix are 0 because R is not measured. In the left half of the matrix, values could be 0 or 1. For example, the value of 1 at (2, 2) means that soil C was measured in field 2. If all diagonal values in the left half are 1, soil C in each field is measured.

3. Implementation of the ensemble Kalman Filter

In the EnKF, an ensemble of states is first created for each field using an unconditional simulation method, such as sequential Gaussian simulation (Goovaerts, 1997). This is based on the underlying spatial structure obtained from initial sampling and geostatistical analyses and on inter-variable correlations between soil C and R. Afterward, each ensemble member is updated using the model, one time step at a time. Propagation of predicted variances and covariances occurs through propagation of ensemble members for each field. The Kalman update matrix is computed (Margulis et al., 2002) and used to update each ensemble member. Thus, state estimates and their variances and covariances are automatically updated.

The EnKF calculations for this problem require several steps. First, a geostatistical analyses is performed on the measured soil C at the start of a project to obtain a semivariogram model that represents the underlying spatial structure of soil C. Then, the model and EnKF are initialized. Using the variogram model, a set of ensemble members of soil C and decomposition rate constant R are stochastically simulated at time 0 for each field. The mean and variance of initial soil C and R are needed for each field as well as the inter- and intra-variable correlations across all fields to perform this step, and the simulated ensemble should have the same spatial structure as

measured soil C. Next, estimates of inputs needed for the model are obtained. In our example, biomass produced in each field is an input in Eq. (1) and would need to be measured or estimated. Remote sensing could potentially be used, since it may not be practical to measure biomass in each field each year. In our example, we assumed that it was constant over time and space. Eq. (1) is used to simulate soil C for the next time step for each ensemble member, which are then used to compute variance and covariance of soil C and R across all fields. If soil C measurements are not made at this time step, steps 3 and 4 are repeated. Measurements, if available, are used to compute the Kalman Gain matrix, which is then used to update soil C for every ensemble member and for every field. Based on the updated ensemble, variances and covariances are computed. Finally, when aggregate estimates of soil C means and variances are needed, they are computed from ensembles of aggregate soil C values by summing mean soil C over all fields. The variance of aggregate soil C across spatially correlated fields is computed by following equation (Wackerly et al., 2002):

$$\mathrm{Var}\left(\sum_{i=1}^{F} a_i \widehat{X}_i\right) = \sum_{i=1}^{F} a_i^2 \mathrm{Var}(\widehat{X}_i) + 2\sum\sum_{i<j} a_i a_j \mathrm{Cov}(\widehat{X}_i \widehat{X}_j) \tag{8}$$

where the double sum is over all pairs of fields (i, j) with $i < j$, F is the total number of fields, a_i is the area of field i, and $\widehat{X}_i$ is the updated soil carbon estimate in the field i.

The procedure for the slightly more complex soil C model presented by Bostick et al. (2003) and for the DSSAT model are nearly the same. However, details related to measurements, the formula for the Kalman Gain matrix, and parameters to update vary among models. If the model simulates both soil C and crop biomass, the measurements of biomass can be used to improve model performance by refining one or more crop model parameters for each field as measurements are made (Koo et al., 2003).

4. Example application of the EnKF in Ghana

The purpose of this example is to demonstrate the use of the EnKF for estimating aggregate soil C over multiple fields and the uncertainty of those estimates under different field sampling intensities. The EnKF was implemented using the simple model described above in an area near the community of Wa in northwestern Ghana (lat. 10.02, long. −2.38). Measurements were made in 12 fields (each 15 m × 30 m in size) to estimate initial soil C; from 20 to 30 samples were taken from each field. These fields were part of an experiment conducted by J. Naab (personal communication) on three farms in 2003. Fig. 2 shows the 12 fields with points where soil C samples were taken. Soil samples from the top 20 cm depth of soil were analyzed using the Walkley-Black method (Jackson, 1958) to quantify concentration of C in each sample ($\mu g[C]\, g[soil]^{-1}$).

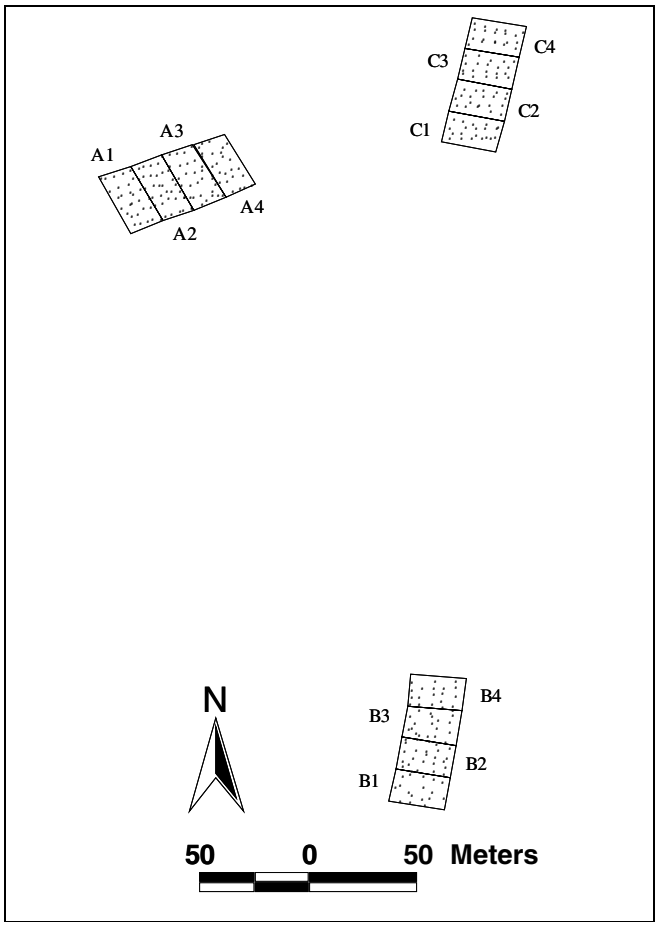

Fig. 2. Twelve fields in Wa, Ghana in which each field was sampled in 2003 for estimation of soil C (source: J. Naab, SARI, Wa, Ghana). Points within each field are where individual samples were taken.

These data were used to estimate initial conditions for use in the EnKF.

4.1. Setting initial conditions

We used GSTAT (Pebesma and Wesseling, 1998) to estimate initial soil C mean and variance for each field. The measurement data were used to create a semivariogram, which was fit to a spherical model, which had a correlation range of 53 m (Fig. 3). The nugget effect in the semivariogram was ignored because we wanted the EnKF ensemble realizations to represent the true state of the underlying spatial structure, not measurement uncertainty at a point. The spherical model was then used with block kriging to estimate mean initial soil C for each field, which ranged from 0.33% to 0.63% (Table 1). Mean and standard deviation of soil C (in kg ha^{-1}) for each of the 12 fields was computed by multiplying the percentage values in Table 1 by sample depth, bulk density, field area, and a unit conversion factor. Because bulk density was not measured, we assumed a constant bulk density value of 1.35 g cm^{-3} for the analyses, a value measured by J. Naab for similar soils in the area. More research is needed to determine the

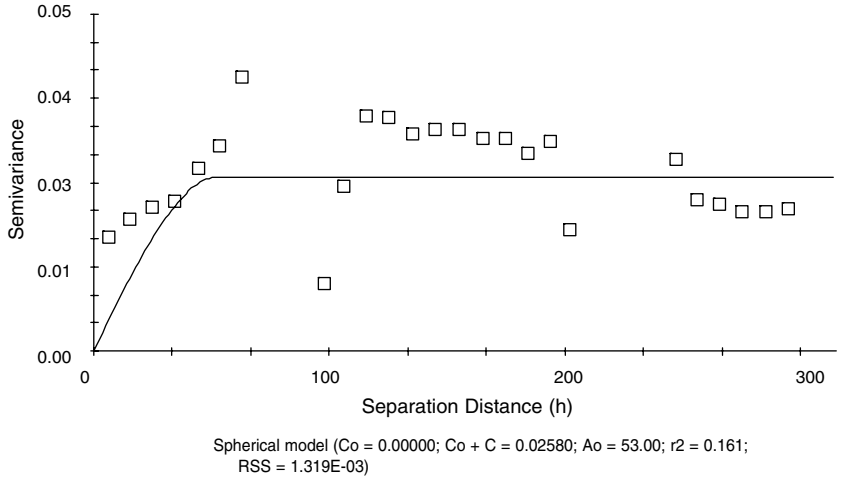

Fig. 3. Semivariogram and model calculated from the 2002 soil C measurements in Wa, Ghana.

Table 1
Soil C measurement standard deviation and block krig estimates of soil C and standard deviation in each of 12 fields in the study site

Farm	Field	Number of samples	Measurement standard deviation (%)	Initial soil C by block kriging (%)	
				Mean	Standard deviation
A	A1	30	0.06	0.58	0.19
A	A2	28	0.08	0.48	0.22
A	A3	28	0.06	0.63	0.20
A	A4	30	0.08	0.54	0.16
B	B1	20	0.06	0.54	0.22
B	B2	20	0.05	0.58	0.21
B	B3	20	0.05	0.50	0.23
B	B4	20	0.05	0.53	0.20
C	C1	20	0.04	0.38	0.18
C	C2	20	0.04	0.33	0.18
C	C3	20	0.06	0.51	0.19
C	C4	20	0.05	0.56	0.19

sensitivity to estimates of soil C changes over time when bulk density varies spatially but is assumed constant.

Realizations of C and R were generated to initialize the model variables using unconditional sequential Gaussian simulation in GSTAT (Pebesma and Wesseling, 1998).

The average semivariogram from the M realizations approximates the semivariogram (γ) obtained from the spatial analysis of field samples (Goovaerts, 1997). To generate these initial realizations in the ensemble, initial estimates of four types of underlying correlations among variables across fields were needed: correlations between (1) soil C in different fields, (2) R in different fields, (3) soil C and R in the same field, and (4) soil C and R in different fields. These correlations are used to estimate the covariance matrix $\boldsymbol{P}(0)$ (Eq. (3)) as follows. Correlations between mean soil C in different fields were obtained from a geostatistical analysis of soil C measurements (Fig. 3). Correlations between R among fields were assumed to have the same spatial structure as soil C (e.g. same spherical model and same correlation range). Changes in soil C was assumed to be perfectly correlated with changes in R in a given field. Furthermore, the correlations were assumed to be negative. This assumption was based on the nature of model behavior (Eq. (1)); a high value of R degrades soil C rapidly and consequently reduces soil C. Correlations between soil C and R in different fields were assumed to be zero.

4.2. Measurements

We did not have a time series of soil C measurements for these fields to use in the EnKF. Thus, we first used the model (Eq. (1)) to generate a single 20-year time series of "true" soil C values for each field, and then generated measurements assuming that measurement errors are normally distributed around true values and that true values of soil C have the spatial structure estimated by the spherical model. We randomly chose one realization out of the ensemble generated from Eq. (1) as the true soil C values. The true value of R for each field was generated from the true soil C value at time 0 and the inter-variable correlation between soil C and R (Table 2). Note that the true mean value of R across field ($\overline{R}_{\text{true}}$) was assumed as 0.020, but

Table 2
Initial conditions for variables used in the EnKF simulation case study in Ghana

Farm	Field	$X_{true}(i, 0)$ (kg ha^{-1})	$\sigma^2_{obs}(i)$ ((kg ha^{-1})^2)	$\sigma^2_{mod}(i)$ ((kg ha^{-1})^2)	$R_{true}(i)$ (yr^{-1})	$\overline{R}(i)$ (yr^{-1})	$\sigma_R(i)$ (yr^{-1})	$U(i, t)$ (kg ha^{-1})	b (yr^{-1})
A	A1	14,339	3,500,000	20,000	0.02268	0.01492	0.0066	2000	0.2
A	A2	13,333	4,900,000	20,000	0.02472	0.01494	0.0066	2000	0.2
A	A3	16,157	2,800,000	20,000	0.01900	0.01495	0.0066	2000	0.2
A	A4	18,139	6,000,000	20,000	0.01498	0.01501	0.0066	2000	0.2
B	B1	18,907	3,300,000	20,000	0.01343	0.01503	0.0066	2000	0.2
B	B2	16,294	2,400,000	20,000	0.01872	0.01497	0.0066	2000	0.2
B	B3	15,895	2,100,000	20,000	0.01953	0.01502	0.0066	2000	0.2
B	B4	16,109	2,300,000	20,000	0.01910	0.01501	0.0066	2000	0.2
C	C1	12,973	1,500,000	20,000	0.02545	0.01479	0.0066	2000	0.2
C	C2	13,313	1,400,000	20,000	0.02476	0.01470	0.0066	2000	0.2
C	C3	14,086	2,800,000	20,000	0.02320	0.01499	0.0066	2000	0.2
C	C4	14,918	2,000,000	20,000	0.02151	0.01519	0.0066	2000	0.2

Table 3
Definition of cases compared in this study

Case no.	Description	Fields sampled per year
1 (Base case)	All fields	12
2	Same fields each year	3
3	Rotating fields each year	3
4	No field sampling	0

For example, case 2 refers to sampling three fields per year and returning to the same three fields each year.

the initial mean of R estimates across fields ($\overline{R}$) was set as 0.015 to reflect our imperfect knowledge of the true value for evaluating the EnKF. Following the example of Jones et al. (2004), we set constant values for $U(i, t)$ and b of 2000 kg ha^{-1} and 0.20, respectively, for each field to generate the "true" values (Table 2). Then, we used Eq. (4) to generate the 20-year time series of observations.

4.3. Measurement intensity comparisons

A soil C measurement scheme may achieve an acceptable level of uncertainty by intensively sampling fields over space and time. However, resources (i.e. time and cost) for conducting measurements are often limited, so the measurement scheme should be optimized in a way that minimizes uncertainties at acceptable costs. To understand the impact of different measurement schemes on the uncertainty of soil C estimates, we varied the fraction of fields sampled each year: (1) all fields and (2) 1/4th of the fields (three in this case). For the latter, we used two ways of selecting which fields to measure each year: (1) return to the same three fields each year and (2) rotate fields yearly so that all fields are sampled at approximately the same frequency. A case without any measurement was also analyzed to demonstrate how the model (Eq. (1)) behaves in a stochastic simulation. Thus, four cases were analyzed (Table 3). The value of measurement error in this example was estimated based on the assumption that one composite soil sample, composed of five mixed sub samples, is taken in each sampled field to measure soil C.

4.4. Simulation runs

Parameters for the soil C model are the same as that used by Jones et al. (2004) for the single field case, except for initial estimates of soil C and the uncertainties associated with measurements and initial conditions. For each case in Table 3, 1000 realizations of an ensemble were created using initial conditions, error covariance matrix, and parameters as explained earlier. Twenty years of simulations were used. Estimates of soil C and R means, variances, and covariances were computed at each time step from the 1000 realizations. Finally, aggregate soil C over all fields and its variance were computed from the realizations using Eq. (8). Comparisons of the different sampling schemes were made using estimates of uncertainty in aggregate soil C over all fields.

4.5. Results

Fig. 4 shows estimates of aggregate soil C for the 12 fields in Ghana in each of four cases. The dark line is the ensemble mean estimate of soil C and the dotted line is the true value. The two grey lines around the ensemble mean estimate represent plus and minus one standard deviation of the ensemble estimate. The Xs represent estimates of aggregate soil C based solely on measurements. When a field was not measured in a given year, its soil C was assumed to be the mean of the other measurements for that year. In all cases, the EnKF produced estimates that were smoother and closer to true values than the measured values. This was also true for each field in the study area (not shown).

Fig. 4(a) shows the case when all 12 fields were sampled each year. Note that variance decreased over time as more data were assimilated. Uncertainties were much larger when the same three fields were sampled yearly (Fig. 4(b)). It is interesting to note that variance of the estimate of soil C in a field that was not measured behaved differently from the variance of estimates in fields that were measured. Compared to the case when all fields were

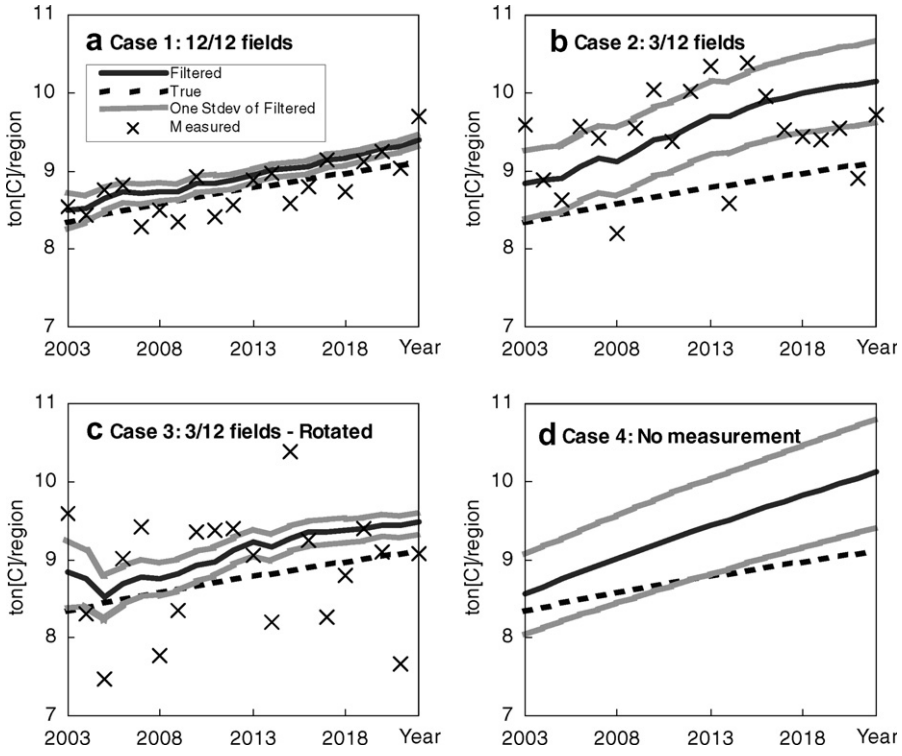

Fig. 4. Aggregated soil C sequestration in each of four cases over the study region in Ghana (0.54 ha). Aggregate measurements were based only on fields that were measured: (a) all 12 fields were measured; (b) the same three fields were measured yearly; (c) three fields were measured each year but rotated; (d) no measurements were made.

measured (Fig. 4(a)), soil C was overestimated when only 3 of the 12 fields were sampled yearly (Figs. 4(b) and (c)) and when no field was sampled (Fig. 4(d)). This was because the initially underestimated R values (Table 3) were not adjusted adequately by the EnKF. Note that underestimation of R causes slow decomposition of soil organic matter and increases the quantity of soil C remaining in the C pool. However, when the three sampled fields were rotated each year, the accuracy of estimates was dramatically improved (Fig. 4(c)) compared to the non-rotated case where only three fields were measured (Fig. 4(b)).

Fig. 5(a) depicts the ensemble variance for each field when the same three fields are measured each year. The bottom cluster shows that ensemble variances decreased for the three measured fields. The middle cluster shows fields that were near the measured fields. In those fields, ensemble variances initially decreased but later increased slightly. The top cluster shows the variances for fields that were not adjacent to measured fields. Initial variances were highest for these fields, and increased over time. When all fields were measured each year, estimated soil C variance decreased for each field, similar to results shown by Jones et al. (2004) for a single field (data not shown).

Rotating the three fields that were sampled each year had a dramatic effect on the evolution of variances for each field. Fig. 5(b) shows that variance remained low for each field over the 20 years of simulations. Variance of aggregate soil C estimate decreased when field sampling was rotated, as shown in Fig. 4(c) compared with Fig. 4(b). In both of these cases, only three fields were sampled each year. The

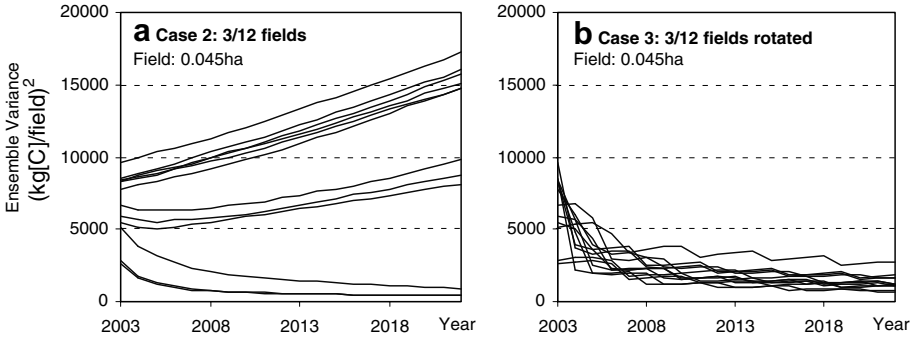

Fig. 5. Ensemble variances of soil carbon dynamics for each field comparing the case when the three measured fields were revisited each year (a) vs. when sample fields were rotated each year (b).

reason for this behavior is that the fields were spatially correlated within a relatively short correlation range. Thus, there was little information in the covariance matrix for adjusting predicted soil C for non-measured fields that were not adjacent to measured fields. By rotating the fields sampled from year to year, each was sampled at some point in time, and this information provided the basis for filtering soil C estimates and adjusting the variance estimate.

5. Discussion

Our original rationale for studying the use of the EnKF approach was that we would have remote sensing estimates of biomass that could be assimilated along with *in situ* soil samples. We also initially planned to use the DSSAT-Century model in the EnKF to take advantage of the ability of that model to account for specific crop management, soil variability and climate variability. Our original use of the simple soil C model was to gain experience with this approach. However, we now believe that the simple models (with one or two soil C state variables) may be as effective as more complex biophysical models when applied over large areas with considerable spatial variability and uncertainty in model inputs. In addition, the time required to implement simple models is much less than that for the more complex DSSAT-Century model. The amount of time required to compute the EnKF for an area with more than several hundred fields is expected to be prohibitive for practical use. One benefit of the EnKF is its ability to use a model to constrain how fast soil C can change over time and space, thus smoothing estimates when noisy measurements are expected and when measurements cannot be made for every field. Although additional research will reveal more about capabilities and limitations of this approach, we have concluded that it could be a powerful tool for operational programs in which estimates of soil C are needed over space and time.

Some observations about the use of this approach for estimating soil C are made. First, in an EnKF, every field or point must be included in the analysis if estimates of system states are needed at those points. Measurements may only be made at a fraction of those points. But, the covariance matrix structure provides information for estimating states at points that are not measured when there is spatial correlation among fields, similar to kriging using geostatistics. Secondly, in the EnKF, a stochastic spatio-temporal model describes the evolution of system states, variances, and covariances over time and space. Even when measurements are not made in a particular year, this method provides estimates of systems states and their uncertainties. One difficulty in implementing the EnKF for this type of problem is the necessity of having reliable initial estimates of all system states (for both measured and non-measured fields) and the corresponding initial covariance matrix. Although the EnKF does not necessarily require georeferenced data, the use of spatial sampling and geostatistical analysis methods are necessary to initialize the EnKF for this soil C sequestration problem.

If there is little correlation among system states, the power of the EnKF is limited and the design of an appropriate sampling scheme is very important. The EnKF clearly showed that if the same subset of fields is sampled each year, there is little information to improve predicted soil C in non-measured fields as the system evolves. In this case, rotating sample fields over time was superior to sampling the same fields each year. When only 3 of the 12 fields were sampled each year, rotating the sampled fields greatly reduced the estimated standard deviation of aggregate soil C. Finally, one could implement the EnKF by measuring and modeling soil C at points or by measuring and modeling average soil C in fields. In our case study, the EnKF estimated average soil C for each field, our basic unit of analysis. These values were aggregated by multiplying areas by average soil C. The rationale for our use of the field in this study is that fields are managed as units, and we assumed that management was uniform across each field. This may not be the case. However, the landscape in most agricultural settings is composed of patches of different land use and management practices, which cause discontinuities in soil C over space. Thus, assumptions that soil C varies smoothly over space and is stationary are questionable. Nevertheless, it may be possible to decompose the landscape into different land use and management types and treat soil C as a pseudo continuous variable over space for spatial analysis purposes, avoiding dissimilar land units. Data need to be collected over large areas with sufficient sampling intensity to quantify spatial variability at short and long distances (Walter et al., 2003). Additional research is needed to evaluate discrete vs. continuous state variables in the EnKF.

Acknowledgements

This research was supported by the Florida Agricultural Experiment Station and by the Soil Management Collaborative Research Program (SM CRSP) through a Grant (LAG-G-00-97-00002-00) from the US Agency of International Development and by a grant from NASA entitled "Carbon from Communities: A Satellite View"; approved for publication as Florida Agric. Exp. Stn. Journal Series No. R-10958.

References

Albiol, J., Robuste, J., Casas, C., Poch, M., 1993. Biomass estimation in plant cell cultures using an extended Kalman Filter. Biotechnology Progress 9, 174–178.

Antle, J.M., Uehara, G., 2002. Creating incentives for sustainable agriculture: defining, estimating potential and verifying compliance with carbon contracts for soil carbon projects in developing countries. In: A Soil Carbon Accounting System for Emissions Trading – Special Publication, Soil Management Collaborative Research Support Program. University of Hawaii, Honolulu, HI, pp. 1–12.

Bostick, W.M., Koo, J., Jones, J.W., Gijsman, A.J., Traore, P.S., Bado, B.V., 2003. Combining model estimates and measurements through an ensemble Kalman filter to estimate carbon sequestration. ASAE Paper #33042. ASAE, St. Joseph, MI 49085. 13 pp.

Burgers, G., van Leeuwen, P.J., Evensen, G., 1998. Analysis scheme in the Ensemble Kalman Filter. Monthly Weather Review 126, 1719–1724.

Eknes, M., Evensen, G., 2002. An Ensemble Kalman Filter with a 1-D marine ecosystem model. Journal of Marine Systems 36, 75–100.

Gijsman, A.J., Hoogenboom, G., Parton, W.J., Kerridge, P.C., 2002. Modifying DSSAT crop models for low-input agricultural systems using a soil organic matter-residue module from CENTURY. Agronomy Journal 94, 462–474.

Goovaerts, P., 1997. Geostatistics for Natural Resource Evaluation. Oxford University Press, Oxford.

Graham, W.D., 2002. Estimation and prediction of hydrogeochemical parameters using Extended Kalman Filtering. In: Govindaraju, Rao S. (Ed.), Stochastic Methods in Subsurface Contaminant Hydrology. ASCE Publications, Reston, VA, pp. 327–363.

Jackson, M.L., 1958. Chemical Analysis. Prentice Hall, Inc., Englewood Cliffs, NJ.

Jones, J.W., Gijsman, A.J., Parton, W.J., Boote, K.J., Doraiswamy, P., 2002. Predicting soil carbon accretion: the role of biophysical models in monitoring and verifying soil carbon. In: A Soil Carbon Accounting System for Emissions Trading – Special Publication, Soil Management Collaborative Research Support Program (Ed.). University of Hawaii, Honolulu, HI, pp. 41–68.

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., Ritchie, J.T., 2003. The DSSAT cropping system model. European Journal of Agronomy 18, 235–265.

Jones, J.W., Graham, W.D., Wallach, D., Bostick, W.M., Koo, J., 2004. Estimating soil carbon levels using an ensemble Kalman filter. Transactions of the ASAE 47, 331–339.

Koo, J., Bostick, W.M., Jones, J.W., Gijsman, A.J., Naab, J.B., 2003. Estimating soil carbon in agricultural using ensemble Kalman filter and DSSAT-CENTURY. ASAE Paper #33041. ASAE, St. Joseph, MI 49085, 13 pp.

Lal, R., 2000. Land use and cropping system effects on restoring soil carbon pool of degraded Alfisols in western Nigeria. In: Lal, R., Kimble, J.M., Stewart, B.A. (Eds.), Global Change and Tropical Ecosystems. Lewis Publishers, Boca Raton, pp. 157–165.

Margulis, S.A., McLaughlin, D., Entekhabi, D., Dunne, S., 2002. Land data assimilation and soil moisture estimation using measurements from the Southern Great Plains 1997 Field Experiment. Water Resources Research 38, 1299.

Maybeck, Peter S., 1979Stochastic Models, Estimation, and Control, vol. 1. Academic Press, New York.

Parton, W.J., Stewart, J.W.B., Cole, C.V., 1988. Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry 5, 109–131.

Parton, W.J., Ojima, D.S., Cole, C.V., Schimel, D.S., 1994. A general model for soil organic matter dynamics: sensitivity to litter chemistry, texture and management. In: Bryant, R.B., Arnold, R.W. (Eds.), Quantitative Modeling of Soil Forming Processes. Special Publication 39. SSSA, Madison, WI, pp. 147–167.

Pebesma, E.J., Wesseling, C.G., 1998. Gstat, a program for geostatistical modelling, prediction and simulation. Computers and Geosciences 24, 17–31.

Pichot, J., Sedogo, M.P., Poulain, J.F., Arrivets, J., 1981. Fertility evolution in a tropical ferruginous soil under the effect of organic manure and inorganic fertilizer applications. L'Agriculture Tropicale 37, 122–133.

Wackerly, D.D., Mendenhall III, W., Scheaffer, R.L., 2002. Mathematical Statistics with Applications, sixth ed. Duxbury Thompson Learning, Pacific Grove, CA, 255 pp.

Walter, C., Viscarra Rossel, R.A., McBratney, A.B., 2003. Spatio-temporal simulation of the field-scale evolution of organic carbon over the landscape. Soil Science Society of America Journal 67, 1477–1486.

Welch, G., Bishop, G., 2002. An introduction to the Kalman Filter. Department of Computer Science, University of North Carolina, Chapel Hill, NC. Report TR-95-041, 16 pp.

West, T.O., Post, W.M., 2002. Soil organic carbon sequestration rates by tillage and crop rotation: a global data analysis. Soil Science Society of America Journal 66, 1930–1946.

Yost, R., Loague, K., Green, R., 1993. Reducing variance in soil organic carbon estimates: soil classification and geostatistical approaches. Geoderma 57, 247–262.

Yost, R.S., Doraiswamy, P., Doumbia, M., 2002. Defining the contract area: using spatial variation in land, cropping systems and soil organic carbon. In: A Soil Carbon Accounting System for Emissions Trading – Special Publication, Soil Management Collaborative Research Support Program (Ed.). University of Hawaii, Honolulu, HI, pp. 13–40.