ELSEVIER

Geoderma 81 (1997) 153–225

GEODERMA

A comparison of the performance of nine soil organic matter models using datasets from seven long-term experiments

P. Smith [a,*], J.U. Smith [a], D.S. Powlson [a], W.B. McGill [b], J.R.M. Arah [c], O.G. Chertov [d], K. Coleman [a], U. Franko [e], S. Frolking [f], D.S. Jenkinson [a], L.S. Jensen [g], R.H. Kelly [h], H. Klein-Gunnewiek [i], A.S. Komarov [d], C. Li [f], J.A.E. Molina [j], T. Mueller [g], W.J. Parton [h], J.H.M. Thornley [c], A.P. Whitmore [i]

[a] *Soil Science Department, IACR–Rothamsted, Harpenden, Herts AL5 2JQ, UK*
[b] *Department of Renewable Resources, 4-42 Earth Sciences Building, University of Alberta, Edmonton, Alberta, T6E 2E3, Canada*
[c] *Institute of Terrestrial Ecology (Edinburgh), Bush Estate, Penicuik, Midlothian EH26 0QB, UK*
[d] *Institute of Soil Science and Photosynthesis of the Russian Academy of Sciences, 142292, Pushchino, Moscow, Russian Federation*
[e] *Centre for Environmental Research, Hallesche Straße 44, D-06246 Bad Lauchstädt, Germany*
[f] *Complex Systems Research Center, Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, NH, USA*
[g] *Soil, Water and Plant Nutrition, The Royal Veterinary and Agricultural University, Thorvaldsenvej 40, DK -1871, Frederiksberg C, Copenhagen, Denmark*
[h] *Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, CO 80523, USA*
[i] *Research Institute for Agrobiology and Soil Fertility, AB-DLO Haren, PO Box 129, 9750 AC Haren, Netherlands*
[j] *Department of Soil, Water, and Climate, University of Minnesota, St. Paul, MN, USA*

Received 30 April 1996; accepted 20 February 1997

Abstract

Nine soil organic models were evaluated using twelve datasets from seven long-term experiments. Datasets represented three different land-uses (grassland, arable cropping and woodland) and a range of climatic conditions within the temperate region. Different treatments (inorganic fertilizer, organic manures and different rotations) at the same site allowed the effects of differing

* Corresponding author. Tel.: +44 1582-763133 ext. 2110; Fax: +44 1582-760 981; E-mail: pesmith@bbsrc.ac.uk

PII S0016-7061(97)00087-6

land management to be explored. Model simulations were evaluated against the measured data and the performance of the models was compared both qualitatively and quantitatively. Not all models were able to simulate all datasets; only four attempted all. No one model performed better than all others across all datasets. The performance of each model in simulating each dataset is discussed. A comparison of the overall performance of models across all datasets reveals that the model errors of one group of models (RothC, CANDY, DNDC, CENTURY, DAISY and NCSOIL) did not differ significantly from each other. Another group (SOMM, ITE and Verberne) did not differ significantly from each other but showed significantly larger model errors than did models in the first group. Possible reasons for differences in model performance are discussed in detail. © 1997 Elsevier Science B.V.

Keywords: soil organic matter; model; soil organic matter model; model evaluation; model comparison; long-term experiment; datasets; soil organic carbon

1. Introduction

There have been a number of model comparison exercises in recent years. Parton (1996a) discusses two ecosystem model comparisons, viz. VEMAP (VEMAP, 1995) and the forest model comparison (Ryan et al., 1997a,b), and two other model comparison exercises, viz. the decomposition model comparison (DECO) and the soil moisture model comparison, RICE/PILPS (Shao and Henderson-Sellers, 1997). Other recent model comparisons have included the Commission of the European Communities nitrogen model comparison (Vereecken et al., 1991) and a comparison of nitrogen turnover models (de Willigen, 1991; Otter-Nacke and Kuhlman, 1991). This paper describes a comparison of soil organic matter (SOM) models using long-term datasets.

Parton (1996a) divides recent ecosystem model comparison exercises into those used to assess the error associated with climate change predictions, those designed to advance the scientific understanding of ecosystems, and those designed to select an appropriate model for use in a particular environment. The present exercise falls most comfortably into the latter two categories.

In the future, SOM models will be used to assess the impact of global change on SOM and subsequent feedback effects. A model's simulation of future events obviously cannot be compared to measured data to verify its validity. We can, however, get some measure of performance by testing a model's ability to simulate long-term SOM changes using existing datasets. The primary purpose of this exercise, then, is to test the ability of SOM models to simulate the long-term dynamics of SOM under a variety of land-uses and in a range of climates within the temperate region as a means of identifying which SOM models are likely to be most appropriate for future global change impact assessment in different environments.

Many of the models selected for this exercise are able to predict a variety of variables other than soil organic carbon, for example, soil moisture, soil

temperature, plant biomass production, commercial crop yield, nitrate leaching and many others. This exercise, however, is restricted to the ability of models to simulate the long-term changes in soil organic matter content. Although nitrogen dynamics are intrinsically associated with soil organic matter turnover, it is the ability of a model to simulate and predict changes in soil organic carbon content that is used here to determine its performance. Shorter-term data such as surface CO_2 flux, soil water content, soil temperature, and detailed measurements of above- and below-ground C and N inputs etc. are extremely valuable for evaluating models at the level of short-term process description, thereby allowing alternative hypotheses of carbon transfer to be tested. This exercise, however, does not aim to test descriptions of short-term processes; instead it examines the ability of SOM models to simulate changes in SOM over many decades. During this exercise, we have discovered situations in which changes in SOM cannot be adequately described by our models and this has led to hypotheses of model failure and inadequate process description (see other papers in this issue). In future evaluation exercises, it will be important to examine these hypotheses in detail at the process-level using short-term data such as those described above.

The evaluation of individual models is described in other papers in this issue. In this paper we briefly present details of the nine models and the twelve treatments from seven long-term experiments as used in this exercise, and then compare the performance of the models in simulating these datasets. This body of work represents the most comprehensive evaluation and comparison to date of models for simulating long-term SOM dynamics.

2. Materials and methods

2.1. Model and dataset selection, and data provided

Nine SOM models were selected from the Global Change and Terrestrial Ecosystems Soil Organic Matter Network (GCTE SOMNET) database (Smith et al., 1996b) for participation in this exercise which began at a NATO Advanced Research Workshop held at IACR–Rothamsted, UK, in May 1995 (Powlson et al., 1996). The models are described briefly in Section 2.2 below.

Seven core long-term experimental sites were selected from GCTE SOMNET which met the minimum input requirements for running the SOM models; brief details of each site are given in Section 2.3 below. Sites were chosen to provide a broad range of land-uses, management practices and climatic conditions within the temperate climatic zone. All datasets were required to be of over 20 years duration, have full daily meteorological records, details of cropping and land-management practices (managed ecosystems only) and at least two measure-

ments of soil organic matter content (usually as total organic carbon). Some sites provided more than one treatment giving a total of twelve datasets. Most treatments were chosen to give the maximum difference in SOM response during the course of the experiment. All datasets were subject to quality control on the basis of detailed information from questionnaires (Smith et al., 1996a).

All modellers were requested to attempt to simulate as many as possible of these core datasets. The ITE and Verberne models only attempted Rothamsted Park Grass, the Prague–Ruzyně and the Waite datasets. Since ITE and Verberne were not intended for arable systems, they were forced to simulate arable crops as if they were grass. Similarly, the SOMM model which attempted all datasets, simulated arable crops as though they were grass.

Since the range of land-uses, climatic conditions and land management practices represented among the core datasets was so broad, many of the selected models had not been tested before under such conditions, and none of the models were parameterized to meet all situations. It was therefore decided at the Workshop that modellers should be allowed to tune model outputs using site-specific data by adjusting the initial distribution of carbon in the different model pools and exercising their judgement about organic carbon inputs other than those specified (e.g. root turnover, rhizodeposition etc.). Modellers were also allowed discretion to choose appropriate starting values for soil organic carbon content, though many of the modellers used common starting values (see other papers in this issue). It was decided that no modeller would be allowed to adjust run-time parameters such as SOM partitioning and turnover rate coefficients. In all cases but one, as much data as possible were made available to the modellers, including measured soil organic carbon values. In one case, for the Waite Wheat–Oats–Grass Pasture–Fallow rotation, measured soil organic carbon data were withheld but all other data were provided. Organic carbon data for the other selected Waite rotation (Wheat–Fallow) were provided to allow site-specific tuning.

When model runs had been completed by each modeller, results were collated and converted into standard units. A number of model outputs were corrected to account for different depths and soil bulk densities whilst others were converted from percentage carbon or carbon concentration values to t C ha^{-1} for the given depth. The simulated values of soil organic carbon content were then compared visually and quantitatively with measured values of soil organic carbon from the long-term experiments. The methods used for the quantitative comparison of model outputs with measured values are described in Section 2.4.

2.2. The models

The models are process-oriented multicompartment models; they are described in detail in Powlson et al. (1996) and are compared by McGill (1996). McGill (1996) noted that the models were mainly empirical in nature and all

contained a slow or inert pool of organic carbon but did not necessarily specify its nature or rate of formation. Soil texture is used in some models to modify decomposition processes. About half the models treat the soil as homogeneous with respect to depth. Concepts pertaining to the nature of litter are different from those pertaining to SOM, as indicated by inclusion of biomass as a SOM component but not a litter component. Many models in the past have specified biochemical fractions to follow C and N through litter and SOM components, but in no case were biochemical separations applied to litter or SOM. The scheme of McGill (1996) revealed convergence in the kinetic compartmentalization, growing use of soil clay content, and inclusion of an IOM component. The remainder of this section provides brief descriptions of the models for use in interpreting similarities and differences in their performance.

CANDY (CArbon-Nitrogen-DYnamics) is a modular system of simulation models and a data base system for model parameters, measurement values, initial values, weather data and soil management data (Franko et al., 1996; Franko, 1996). It simulates dynamics of soil N, temperature and water in order to provide information about N uptake by crops, leaching and water quality. CANDY uses a semi-cohort system to track litter decay, and calculates a biologically active time to allow comparisons among sites. The inert organic matter (IOM) component is calculated from the proportion of soil particles < 6 μm.

CENTURY was developed to simulate long-term (decades to centuries) SOM dynamics, plant growth and cycling of N, P and S. It was orginally developed for grasslands but has since been extended to agricultural crops, forests and savanna systems (Parton, 1996b). It uses a monthly time step with monthly average maximum and minimum temperatures and monthly precipitation data (Parton et al., 1987b; Parton and Rasmussen, 1994; Parton, 1996b). It comprises two forms of litter: Metabolic and Structural, and three SOM compartments: Active, Slow, and Passive. C leaving the Active organic matter compartment is partitioned into either CO_2 or Slow forms with the split determined by soil texture. Soil texture also regulates the rate of transfer between Slow and Passive forms. CENTURY has been used to simulate C accumulation during soil formation (Parton et al., 1987a), and changes in soil C storage following climate change scenarios (Schimel et al., 1994).

DAISY simulates crop production, and dynamics of soil water and nitrogen under diverse agricultural management systems (Hansen et al., 1991; Mueller et al., 1996). It was developed as a field management tool as well as for regional administrative purposes, and has been applied to catchment areas (Styczen and Storm, 1993), farmland areas (Jensen et al., 1994a) and specific sites (Jensen et al., 1994b). DAISY contains a hydrological model with a soil water submodel, a soil nitrogen model with a soil organic matter submodel, and a crop model with a nitrogen uptake submodel. Rate constants are modified by soil clay content, a semi-cohort accounting system is used for litter decay and soil microbial

biomass is a dependent variable although concepts of zymogenous and autochthonous microorganisms are included.

DNDC (DeNitrification and DeComposition) couples denitrification and decomposition processes as influenced by the soil environment to predict emissions of CO_2, N_2O and N_2 from agricultural soils (Li et al., 1992a,b; Li, 1996). DNDC contains four interacting submodels: soil climate, decomposition, denitrification, and plant growth (Li, 1996). The plant growth submodel includes subroutines for cropping practices such as fertilization, irrigation, tillage, crop rotation and manure addition to simulate SOM turnover in arable lands (Li et al., 1994). Clay adsorption of humads allows some soil-specificity; decomposition is first order, such that biomass formed during decomposition is a dependent variable.

The Hurley Pasture Model and the ITE (Institute of Terrestrial Ecology – Edinburgh) Forest Model share a common soil submodel, hereafter referred to as ITE. The pasture model aims to simulate N cycling in a grazed soil–plant system, and comprises three submodels (Thornley and Verberne, 1989). They are: (1) a grazing animal-intake model in which the C fluxes and N content of faeces and urine are sensitive to the N content of ingested plant material; (2) a vegetative grass growth model which responds to light, temperature and N; and (3) a soil organic matter submodel that responds to faeces, urine, and decaying plant residues. A physiologically based treatment is used for the plant component, but a functional compartment-based treatment is used for SOM (Arah, 1996). All decomposition rates are a function of the quantity of microbial biomass. Although some components are mineralized to CO_2 or NH_3 without passing through the biomass, the rate of such mineralization is a function of quantity of biomass. This model has also been combined with a transport-resistance model to describe C and N compartments and fluxes in a plantation forest soil system (Thornley and Cannell, 1992).

NCSOIL simulates N and C flow through soil microbes and organic components (Molina, 1996). It comprises four organic compartments: plant residues, microbial biomass (Pool I), humads (Pool II) and stable organic matter (Pool III) (Nicolardot et al., 1994). The original version did not include stable organic matter (Molina et al., 1983). Flows of C and N are interconnected, and increasing stability of organic matter results from metabolism and not from sorption mechanisms that would be sensitive to clay content. Microbial succession is simulated on residues, and although decomposition rate is independent of microbial biomass, microbial succession leads to more stable materials. In addition to ^{12}C and ^{14}N it simulates ^{14}C and ^{15}N dynamics. NCSOIL has been incorporated into a deterministic model (NCSWAP) of the soil–plant system that simulates interactions of N and C dynamics with crop growth and soil water (Molina and Richards, 1984; Clay et al., 1985; Lengnick and Fox, 1994). The model has been simplified to avoid having to determine too many initial variables and parameters (Molina, 1996). To do so, microbial succession was

collapsed into one microbial component with the dynamics determined by the rate of C flow through populations that consume microbes.

ROTHC is the Rothamsted C model in which the turnover of C in aerobic soil is sensitive to soil type, temperature, moisture and plant cover (Jenkinson et al., 1987; Jenkinson, 1990; Coleman and Jenkinson, 1996). Nitrogen and C dynamics are not interconnected in ROTHC, the IOM component is quantified using C-dating, and starting values are obtained by running the model to steady-state. Unlike other models examined, ROTHC has been used for calculating organic matter inputs to soil (Jenkinson and Coleman, 1994) and net primary productivity (Jenkinson et al., 1992) using soil organic matter and radiocarbon measurements. It is formulated as a discrete sums-of-exponentials that can be transformed into a continuous form which can yield analytical results that may provide additional useful insights into soil organic matter turnover (Parshotam, 1995). It uses primarily monthly input data, and shares several other basic ideas with CENTURY (Coleman and Jenkinson, 1996).

SOMM is described as the raw humus submodel of a single plant ecosystem model (SPECOM) developed for forested ecosystems (Chertov, 1990). Rates of processes are regulated by N and ash content of litter fall and it uses temperature and moisture as environmental variables (Chertov and Komarov, 1996). It treats the soil litter layers L, F, and H explicitly. C flow from L to H are governed by biological activity and progressive humification of the residual material. Activities of soil animals are implicit regulators of C flow; distinctions among humus forms such as mull in contrast to mor reflect proportions of earthworms in contrast to micro-arthropods etc. This is the only model of the group examined that simulates C accumulation in soil organic horizons explicitly.

The Verberne/Van Veen model (hereafter referred to as Verberne) aims to simulate N and water balance in a grassland soil–plant system in order to predict yield, N uptake, N leaching, N mineralization and accumulation of soil organic N (Verberne, 1992; Klein-Gunnewiek, 1996). The grass growth submodel is based on the model described by Thornley and Verberne (1989) (see ITE above). The soil is treated as a multilayer mineral soil system, and is simulated using three submodels: (1) soil water submodel, (2) soil organic matter submodel; and (3) soil N submodel. The soil organic matter submodel is based on Verberne et al. (1990). Transformations follow first-order kinetics, with physical protection caused by soil clay resulting in protected and nonprotected biomass and organic matter. Specific decomposition rates are modified by soil moisture and temperature, but microbial biomass does not influence decay rate. Clay protection reduces microbial biomass turnover by a factor of 100. Plant residues are partitioned into three compartments: 'Decomposable', 'Structural', and 'Resistant'; there is a 'Stabilized' organic matter component in addition to 'Protected' and 'Nonprotected' active organic matter. This model emphasizes the influence of clay on protection of microorganisms and soil organic components more than do other models.

2.3. The datasets

For the model evaluation and comparison exercise, twelve 'core' datasets from seven sites in the temperate regions were selected for use *by all models*, thus allowing a comparison based upon simulation of common datasets (see Powlson et al., 1996). A description of the data provided to the modellers for the evaluation exercise is given in Appendix A. The changes in soil organic matter at each core site during the course of the experiments are described in Sections 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.3.5, 2.3.6 and 2.3.7.

2.3.1. Bad Lauchstädt Static Fertilizer Experiment

The experiment, initiated in 1902 has undergone some changes since its inception but the main series is unchanged since 1906. It now consists of four fields with a sugar beet, spring barley, potatoes, winter wheat rotation. Various treatments of 30, 20 and 0 t ha^{-1} of farmyard manure are applied to different plots with further treatments consisting of a range of inorganic fertilizers, legumes every fourth year or liming. Details of soil properties, meterological factors and land-management practices for the Bad Lauchstädt Experiment are given in Appendix A and in Körschens and Müller (1996).

The two treatments used in the model evaluation exercise consist of: (a) high fertilization—30 t ha^{-1} of farmyard manure every second year plus varying rates of inorganic NPK fertilizer; (b) no fertilization. The total soil carbon content of the soil is determined every year on the main plots. The wide range of fertilizer treatments since 1902 has led to a great difference in the total carbon content of the plots, especially between the highest fertilization level and the nil treatment. Since 1956 soil carbon levels (0–30 cm) have increased from around 80 t ha^{-1} to around 95 t ha^{-1} on the high fertilization plot and have remained constant at around 65 t ha^{-1} on the nil input plot.

2.3.2. Calhoun Experimental Forest

The Calhoun Forest Experiment is an afforestation experiment with eight permanent plots and archived soil samples. It was designed to test how an acidic, formerly cultivated Ultisol meets the nutritional needs of a rapidly growing loblolly pine forest during the first four decades after planting (Richter and Markewitz, 1996).

The permanent plots have been sampled and soils archived on six occasions since seedlings were planted in 1957 (Richter and Markewitz, 1996). Following 150 years of cultivation, the old-field pine forest has rapidly accumulated carbon in forest biomass and in the soil profile over a three to four decade period, increasing in the soil from around 10 t ha^{-1} (0–15 cm) in 1962 to nearly 13 t ha^{-1} in 1990.

2.3.3. Rothamsted Park Grass Experiment

The Park Grass Experiment at Rothamsted, UK, was started in 1856 on a site which had been in grazed grassland for several centuries and was intended to

complement experiments started a few years earlier on the nutritional requirements of legumes, cereals and root crops (Poulton, 1996a).

Thirteen plots were laid out initially and a further seven added later; they range in size from 0.05 to 0.2 ha. Plots received either no treatment (unmanured), farmyard manure or N, P or K, either singly or in various combinations. The site is not grazed; in June and in autumn each year herbage is cut and made into hay or silage. The two plots used for the model evaluation exercise are (a) the organic manure treatment (N2PKNaMg until 1904 together with straw until 1897 and then farmyard manure and fishmeal on a 4-year cycle since 1905) and (b) no fertilization treatment (since 1856). Changes in total organic carbon content of the 0–23 cm soil layer (corrected for sampling depth differences) since 1856 show that on the nil inputs plot there has been little change in soil organic carbon content over the course of the experiment whilst on the organic manured plot there was (surprisingly) a decline between 1876 and 1932 and then an increase over the next 60 years. Whether the decline and subsequent increase are directly related to the change in treatment or to soil variability is unclear (Poulton, 1996a).

2.3.4. Prague–Ruzyně Plant Nutrition and Fertilization Management Experiment

The Plant Nutrition and Fertilization Experiment at Prague–Ruzyně was started in 1956 on nine fields that had been under ordinary arable cultivation for several hundred years (Klír, 1996). In 1995 five main fields with different organic and mineral fertilization regimes remain.

The results from 'field B' were selected for use in the model evaluation exercise. This field has had a two-year rotation of sugar beet and spring wheat since 1966; before that various rotations of sugar beet, spring wheat, spring barley, canola (oilseed rape) and lucerne were used. Two treatments were selected for the model evaluation: (a) the high fertilization treatments, and (b) the no fertilization treatment. Both treatments have had residues incorporated after harvest (between 1.5 and 2.0 t ha^{-1} y^{-1} since 1965) but the no fertilization treatment has received no other amendment. The high fertilization treatment has received 50 kg N ha^{-1} annually as inorganic fertilizer (to the spring wheat crop) with an additional 21 t ha^{-1} of farmyard manure every two years (after the spring wheat harvest in autumn) and 150 kg N ha^{-1} during each intermediate year (i.e. every two years to the sugar beet crop).

Total carbon has been measured every year since 1972. Changes in total carbon in the topsoil (0–20 cm) of the Prague–Ruzyně Experiment since 1972 show high inter-year variation, presumably due to site variability. Both treatments show a spread of values between about 31 t ha^{-1} to nearly 44 t ha^{-1}.

2.3.5. Tamworth Legume / Cereal Rotation on Black Earth

The Tamworth Legume/Cereal Rotation was established in 1966 on two adjacent sites with Chromic and Pellic Vertisol soils (Crocker and Holford,

1996). The trial on the Pellic Vertisol soil (i.e. black earth) was used for the model evaluation exercise. The site had been cultivated for around 100 years before the experiment began and was badly eroded by 1958. The trials were established to determine the effects of different cropping systems on the long-term sustainability of cereal (mainly wheat) production.

Two rotations were selected for the model evaluation exercise. One, referred to here as the 'lucerne/clover rotation' consisted of lucerne from 1966 to 1969, wheat from 1970 to 1978, lucerne from 1979 to 1983, sorghum from 1983 to 1987, wheat in 1987, subterranean clover from 1988–1990 and wheat from 1990 to 1994. The other, referred to here as the 'fallow rotation' consisted of long fallow from 1966 to 1969, wheat from 1970 to 1979, fallow every second year with wheat in 1981, sorghum sown in 1983 and 1985 and then wheat in 1987, 89, 91, 93 and 1994. Both rotations received various applications of superphosphate and urea. Wheat stubble was burnt in December or January up to 1976, after which it was grazed by sheep prior to being incorporated in March or April. Lucerne was grazed heavily, leaving less than 1000 kg dry matter ha^{-1}, prior to being ploughed in. It was estimated that sheep would have eaten about 75% of wheat stubble present at harvest, but less than 50% of sorghum stubble.

Soil measurements included total organic carbon and were taken yearly from 1970 to 1981, and then every second year. The site had an original organic carbon level of 1.14% in 1966 and this rose to 1.21% in 1970 following the lucerne phase of the lucerne/clover rotation. Since 1970, the soil C levels on the lucerne/clover rotation have been measured at between 23 and 29 t ha^{-1}. The initial level gradually declined during the wheat cropping phase from 1970 to 1978 except for 1976–1977. The higher level was due to ploughing in of heavy weed growth in 1976, caused by well above average rainfall during the first three months of the fallow. Soils were sampled for organic carbon measurements in June or July. The second lucerne phase from 1979–1983 had little effect on organic carbon as well below average rain occurred in three of the four years. Four sorghum crops from 1983 to 1987 lifted the organic carbon level, before it stabilised under subterranean clover. The increase in organic carbon levels under sorghum is attributed to the incorporation of sorghum residues, which were much greater than wheat residues.

In the fallow rotation, continuous fallow from 1966 to 1970 caused organic carbon content to fall to 1.05% which was significantly lower than the lucerne/subterranean clover rotation in 1970 and 1971. Since 1970, the fallow rotation has produced levels between 20 and 24 t ha^{-1} and there was no significant difference during the rest of the wheat growing phase, except for 1977. Organic carbon levels, for the fallow rotation, have continued to decline since 1976 except during the sorghum phase. Organic carbon levels have maintained a steady trend except for an increase following the sorghum phase, thus giving significant differences in organic carbon between the two rotations.

2.3.6. Rothamsted Geescroft Wilderness

Geescroft Wilderness at Rothamsted, UK, is an area of land previously in arable cropping for several centuries that was fenced off in the 1880s and left unattended; it is now a deciduous woodland (Poulton, 1996b). During the 30 years following fencing-off, the area was colonized by many damp-loving species. Although some shrubs and trees were present the area was still fairly open. By 1957 the site had reverted to woodland and most grassland species had disappeared. The site is now an oak-dominated deciduous woodland with few ground-cover plants (Poulton, 1996b).

Soil samples were taken in 1883 whilst still in clover, and the site was resampled in 1904 and 1965 and on each occasion three depths (0–22.9, 22.9–45.7, 45.7–68.6 cm) were sampled. In 1985 the top 22.9 cm only was sampled. As organic matter has accumulated, the bulk density of the sampled 0–22.9 cm layer has decreased, i.e. the soil has expanded. Since deeper soils have been sampled on Geescroft, the 'equivalent' depth can be calculated, i.e. that depth which would need to be sampled to give the same weight of fine, ignited soil as in the initial sample.

The data show a greater rate of increase in organic matter between 1965 and 1985 than in the previous 61 years. During this time the pH has declined to 4.2 from 7.1 in 1883 due to atmospheric inputs, and there has been a transition in the humus type from mull or moder to mor, i.e. a layer of largely undecomposed organic matter at the surface which is quite distinct from the mineral soil beneath. This layer varies in depth from ca. 2–6 cm. The increase in organic matter between 1965 and 1985 is probably real but it could, in part, reflect the difficulties involved with sampling an increasingly variable site. All values are corrected for the changes in bulk density (Poulton, 1996b). Since 1888, total soil organic carbon content (0–23 cm) in Geescroft Wilderness have risen from about 28 t ha^{-1} to over 60 t ha^{-1}.

2.3.7. Waite Permanent Rotation Trial

The Waite Permanent Rotation Trial was initiated in 1925 on a site previously under native grassland (Grace, 1996). There were originally ten rotations which involved wheat in various rotations and of these, seven have remained unbroken since 1925. The rotations selected for the model evaluation exercise were (a) the wheat–fallow rotation, and (b) the wheat–oats–pasture–fallow rotation. In the latter, wimmera rye grass was sown with grazing oats. The two rotations both received 209 kg ha^{-1} superphosphate until 1943, and 105 kg ha^{-1} superphosphate thereafter, applied to every cereal crop and annual pasture (no superphosphate was applied during fallow years).

Soil samples were taken from the trial in 1925, 1946, 1963, 1967, 1973, 1980, 1983 and 1993. The samples were taken at 0–10 cm and 10–22.5 cm depth increments except in 1925 (22.5 cm increments to 45 cm) and 1967 (0–22.5 cm only). Total C of stored samples from 1925, 1946 and 1963–1993

has been determined by dry combustion. Since Urrbrae red brown earth contains no calcium carbonate-carbon, this is equivalent to organic carbon content (Grace, 1996). Since 1925, total organic carbon content in the 0–22.5 cm soil layer of the wheat–fallow rotation has dropped from about 75 t ha^{-1} to about 25 t ha^{-1}, whilst in the wheat–oats–pasture–fallow rotation it has fallen to around 30 t ha^{-1}.

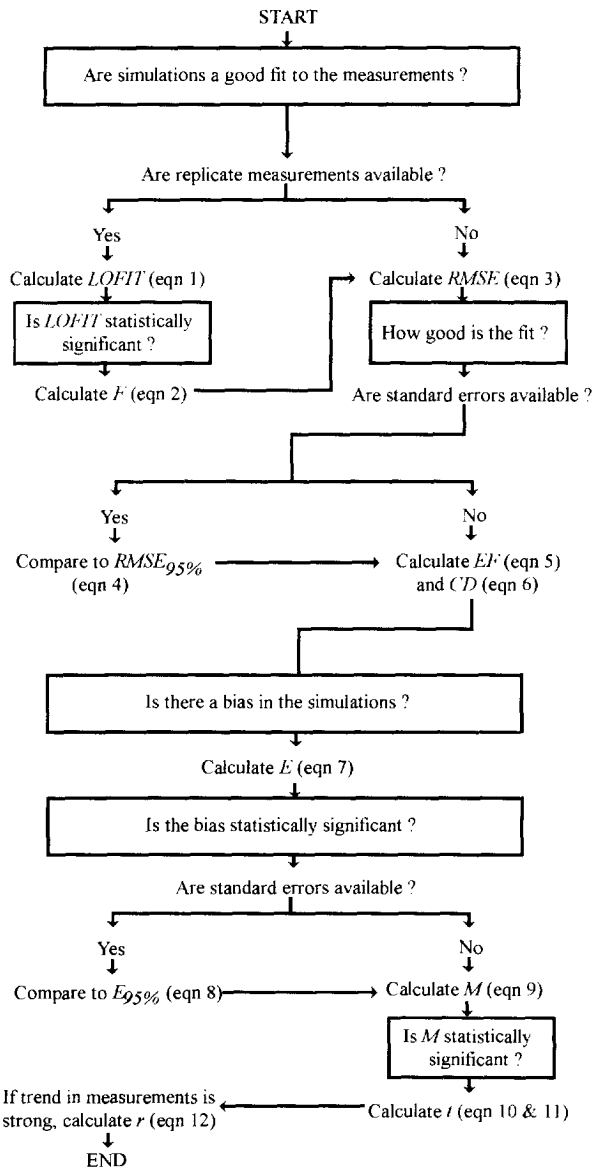


Fig. 1. A schematic representation of the statistical methods followed for the quantitative comparison between SOM model predictions and measured values.

2.4. Quantitative methods used for evaluating and comparing SOM model performance using long-term datasets

Methods for evaluating the accuracy of a simulation and determining the acceptable error are discussed in detail elsewhere (Addiscott et al., 1995; Smith et al., 1996). Each quantitative method described in this section provides information on a distinct aspect of the accuracy of the simulation.

F or t values are routinely used in statistics to derive significance levels (e.g. Chatfield, 1983). If the model output has been tuned using measured data (as is generally the case here), values of F or t cannot be used to show that measured and simulated values are significantly *related* since they are not completely independent (i.e. measured data have been used to render the measured and simulated values more similar). However, the F or t value *can* be used to show that measured and simulated values *differ* significantly even if model tuning has occurred since any parameter adjustment would be used to more closely match measurements and simulations. In this case, a significant F or t value shows that the simulations and measurements are significantly different despite attempts to render them more similar by model tuning. F and t values were used in this way for the lack-of-fit ($LOFIT$) and the mean difference (M) statistics described below.

In the following equations, O_i are the observed (measured) values, P_i are the predicted (simulated) values, $\overline{O}$ is the mean of the observed (measured) data, $\overline{P}$ is the mean of the predicted (simulated) data, and n is the number of paired values.

The method used to assess the goodness-of-fit between simulated and measured values depended on the type of measurements available (Fig. 1). If the experiments had been replicated, the lack of fit statistic, LOFIT (Whitmore, 1991), was calculated:

$$\text{LOFIT} = \sum_{j=1}^{N} m_j \bar{d}_j^2 = \sum_{j=1}^{N} m_j \left(\overline{O}_j - P_j\right)^2 \tag{1}$$

where N is the number of experiments, m_j is the number of replicates within each experiment, P_j is the simulation for the jth experiment, $\overline{O}_j$ is the mean of the measurements in the jth experiment, $\bar{d}_j$ is the mean deviation $= \overline{O}_j - P_j$.

Assuming experimental errors to be random, this statistic allows the experimental errors to be distinguished from the failure of the model. The significance of LOFIT was determined using an F-test where individual replicate measurements were available (Tamworth datasets only), i.e.:

$$F = \frac{\text{MSLOFIT}}{\text{MSE}} = \frac{\sum_{j=1}^{N} (m_j - 1) \sum_{j=1}^{N} m_j \left(\overline{O}_j - P_j\right)^2}{N \sum_{j=1}^{n} \sum_{i=1} \left((O_{i,j} - P_j) - \left(\overline{O}_j - P_j\right)\right)^2} \tag{2}$$

In line with statistical convention, a value of F greater than the critical 5% F value was taken to indicate that the total error in the simulated values was significantly greater than the error inherent in the measured values.

Where individual replicate values were not available, other tests were used. After League and Green (1991), the total difference between the simulated and the measured values were calculated as the root mean square error, RMSE:

$$\mathrm{RMSE} = \frac{100}{\overline{O}} \sqrt{\sum_{i=1}^{n} (P_i - O_i)^2 / n} \tag{3}$$

If standard errors of the measurements ($S_e(i)$) were available, the statistical significance of RMSE was assessed by comparing to the value obtained assuming a deviation corresponding to the 95% confidence interval of the measurements (corrected from Smith et al., 1996), i.e.:

$$\mathrm{RMSE}_{95\%} = \frac{100}{\overline{O}} \sqrt{\sum_{i=1} \left(t_{(n-2)95\%} \times S_e(i) \right)^2 / n} \tag{4}$$

where: $t_{(n-2)95\%}$ is Student's t distribution with $n-2$ degrees of freedom and a two-tailed P-value of 0.05.

An RMSE value less than $\mathrm{RMSE}_{95\%}$ indicates that the simulated values fall within the 95% confidence interval of the measurements.

If no estimates of the standard errors were given, the accuracy of the simulation was further assessed by calculating the modelling efficiency, EF, and the coefficient of determination, CD (League and Green, 1991). The modelling efficiency, EF, provides a comparison of the efficiency of the chosen model to the efficiency of describing the data as the mean of the observations, i.e.:

$$\mathrm{EF} = \frac{\left(\sum_{i=1}^{n} (O_i - \overline{O})^2 - \sum_{i=1}^{n} (P_i - O_i)^2 \right)}{\sum_{i-1}^{n} (O_i - \overline{O})^2} \tag{5}$$

Values for EF can be positive or negative with a maximum value of 1. A positive value indicates that the simulated values describe the trend in the measured data better than the mean of the observations. A negative value indicates that the simulated values describe the data less well than a mean of the observations.

The coefficient of determination, CD, is a measure of the proportion of the total variance in the observed data that is explained by the predicted data.

$$\mathrm{CD} = \sum_{i=1}^{n} (O_i - \overline{O})^2 / \sum_{i=1}^{n} (P_i - \overline{O})^2 \tag{6}$$

The lowest value of CD is 0. A value of 1 or above indicates that the deviation of the predictions from the mean of the measured values is less than that observed in the measurements, i.e. the model describes the measured data better than the mean of the measurements. A CD value less than 1 indicates that the deviation of the predictions from the mean of the measured values is greater than that observed in the measurements, i.e. the mean of the measurements describes the data better than did the model. Taken together, EF and CD allow RMSE to be further interpreted where standard error values of the measurements are unavailable.

The bias in the total difference between simulations and measurements was determined by calculating the relative error, E (Addiscott and Whitmore, 1987).

$$E = \frac{100}{n} \sum_{i=1}^{n} (O_i - P_i)/O_i \tag{7}$$

If the standard errors of measurements were available, the significance of E was again determined by comparing to the value obtained assuming a deviation corresponding to the 95% confidence interval of the measurements (corrected from Smith et al., 1996), i.e.:

$$E_{95\%} = \frac{100}{n} \sum_{i=1}^{n} \left(t_{(n-2)95\%} \times S_e(i) \right)/O_i \tag{8}$$

An E value greater than $E_{95\%}$ indicates that the bias in the simulation is greater than the 95% confidence interval of the measurements. The nature of the bias was further examined using the mean difference, M (Addiscott and Whitmore, 1987):

$$M = \sum_{i=1}^{n} (O_i - P_i)/n \tag{9}$$

The mean difference between measured and simulated values gives an indication of the bias in the simulation but is less informative than E because errors are not proportioned to the size of measurement. However, it is a useful statistic when standard error values are not available to derive a value for $E_{95\%}$ since M can be related directly to t as shown below:

$$t = \frac{M}{s_d/\sqrt{n}} = \frac{\sum_{i=1}^{n} (O_i - P_i)/n}{s_d/\sqrt{n}} \tag{10}$$

where:

$$s_d^2 = \frac{\sum_{i=1}^{n} (d_i - M)^2}{n-1} = \frac{\sum_{i=1}^{n} \left((O_i - P_i) - \left(\sum_{i=1}^{n} (O_i - P_i)/n \right) \right)^2}{n-1} \tag{11}$$

d_i is the difference between measured and simulated values, M is the mean difference between measured and simulated values.

The t statistic (here after Chatfield, 1983) is used to show a significant difference between simulated and measured values. In line with statistical convention, a t value greater than the critical two-tailed 2.5% t value was taken to indicate that the simulation showed a significant bias towards over- or under-estimation when compared to measured values.

To assess whether simulated values follow the same pattern as measured values, the sample correlation coefficient, r, can be calculated, i.e.:

$$r = \frac{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)\left(P_i - \overline{P}\right)}{\left(\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2\right)^{1/2}\left(\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2\right)^{1/2}} \tag{12}$$

This statistic can be useful in assessing how well the shape of the simulation matches the shape of the measured data. However, if there is no clear trend in the measured data to give a spread of paired measured and simulated data values, the correlation coefficient is of only limited use in determining how well a model simulates the measured data. Of the datasets, none showed a clear positive or negative trend in the measurements except for Geescroft (positive trend). For this reason, the correlation coefficient was used only with this dataset. The statistical procedures outlined in this section are summarized as a flow diagram in Fig. 1.

3. Results and discussion

3.1. Model performance on each dataset

Not all of the models attempted to simulate all datasets (Table 1). Reasons why certain models did not attempt certain datasets or simulated only part of them, and details of model tuning and any assumptions made during the modelling exercise are given elsewhere in this issue in papers describing the performance of individual models (Arah et al., 1997; Chertov et al., 1997; Coleman et al., 1997; Franko et al., 1997; Jensen et al., 1997; Kelly et al., 1997; Li et al., 1997; Molina et al., 1997; Whitmore et al., 1997). In this paper, we evaluate the model simulations of soil organic carbon content against the measured data at common depths using common units, and then compare the performance of the models. In this section, we initially compare individual model performance for each dataset separately (Section 3.1). In the second part (Section 3.2), models are compared with each other with respect to their performance across all datasets.

Table 1
Datasets for which models attempted to simulate the change in total organic carbon

Datasets	RothC	CANDY	DNDC	CENTURY	DAISY	SOMM	ITE	Verberne	NCSOIL
Bad Lauchstädt – no fertilization	✓	✓	✓	✓	✓	✓			✓
Bad Lauchstädt – high fertilization	✓	✓	✓	✓	✓	✓			✓
Calhoun Exp. Forest	✓	✓		✓		✓	✓		✓
Park Grass – no fertilization	✓	✓	✓	✓	✓	✓	✓	✓	✓
Park Grass – organic manure	✓	✓	✓	✓	✓	✓	✓	✓	✓
Prague–Ruzyně – no fertilization	✓	✓	✓	✓	✓	✓	✓	✓	✓
Prague–Ruzyně – high fertilization	✓	✓	✓	✓	✓	✓	✓		✓
Tamworth – fallow	✓	✓	✓	✓	✓	✓			✓
Tamworth – lucerne/clover	✓	✓	✓	✓	✓	✓			✓
Geescroft Wilderness	✓	✓		✓		✓	✓	✓	✓
Waite – Wheat/Fallow	✓	✓	✓	✓	✓	✓	✓	✓	✓
Waite – wheat/oats/pasture/fallow (blind)	✓	✓	✓	✓	✓	✓	✓	✓	✓

3.1.1. Bad Lauchstädt

The two treatments selected were the high fertilization and the no fertilization treatments. The low first measured datapoint for both treatments may well reflect changes in the analytical techniques used at different times rather than a real difference from the bulk of the measured points. Data were provided to run the models from 1956 to 1994.

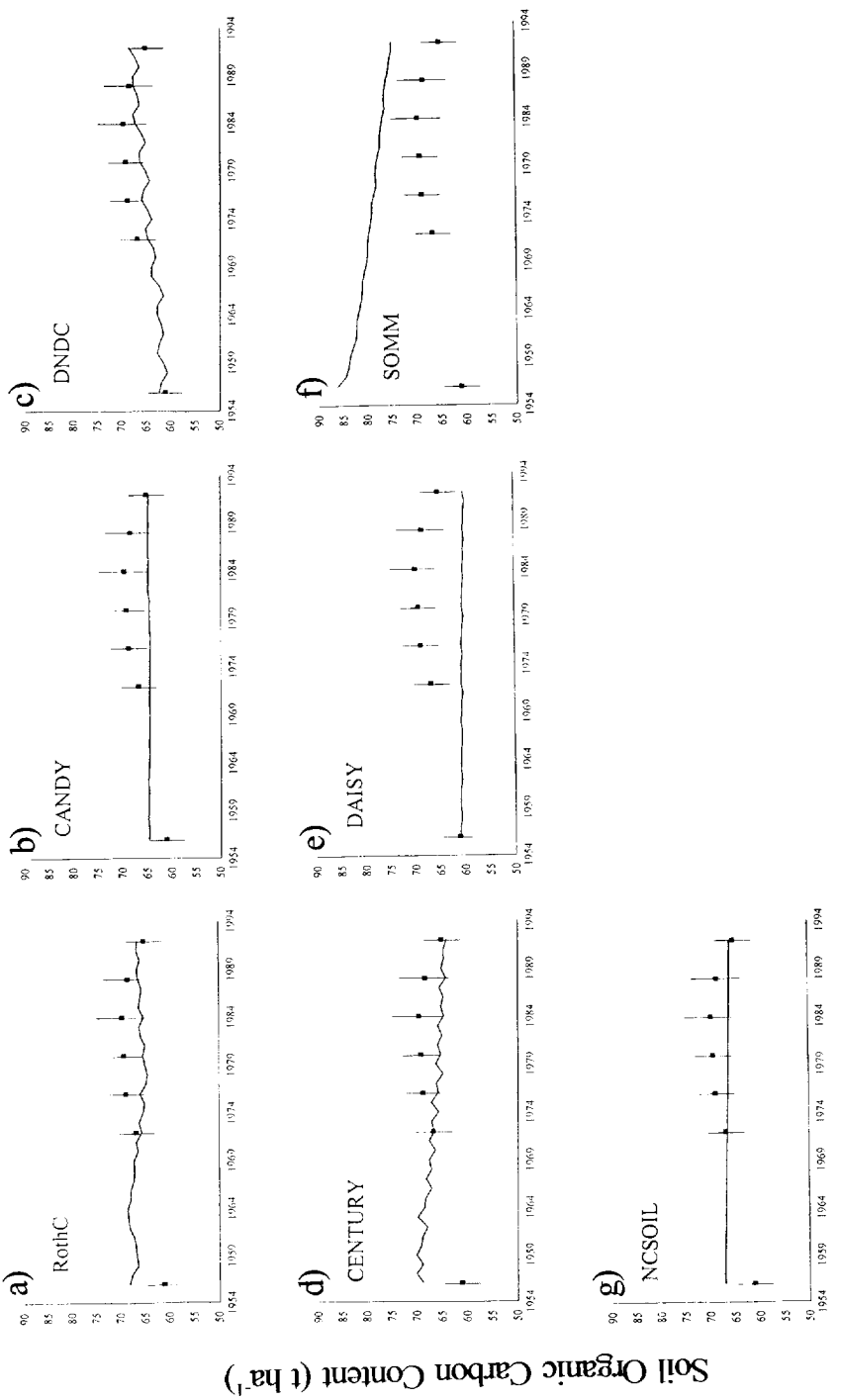
Seven of the nine models (RothC, CANDY, DNDC, CENTURY, DAISY, SOMM and NCSOIL) attempted to model the data.

3.1.1.1. Bad Lauchstädt—no fertilization. The pattern of the measured data was matched by most of the models with RothC, CANDY, DNDC, CENTURY and NCSOIL producing a simulation line that was within the standard error of most of the measured data (Fig. 2a, b, c, d and g). The close model fit for RothC, CANDY, DNDC, CENTURY and NCSOIL was reflected in the statistics (Fig. 3).

These models had RMSE values less than the $RMSE_{95\%}$ value for this data indicating that although some simulated points lay outside the standard errors of individual measured values, they fell within the 95% confidence interval for the whole dataset. DAISY showed a flat trend (Fig. 2e) which lay below most of the standard errors of the measured values but the RMSE value was within the 95% confidence interval of the dataset (Fig. 3b). SOMM showed a decreasing trend in soil organic carbon content (Fig. 2f) with all simulated points lying above the measured values and an RMSE value outside the 95% confidence interval of the measured data (Fig. 3b). A similar picture of model performance was reflected in other statistics describing total model error (LOFIT, EF and CD; Fig. 3a, c and d) but with only RothC, CANDY, CENTURY and NCSOIL adequately explaining the variance in the measured data, as shown by CD values greater than 1.

RothC, CANDY, DNDC, DAISY, CENTURY and NCSOIL all had E values within the 95% confidence interval of the data $E_{95\%}$ suggesting no bias, but M, another measure of model bias did reveal a significant bias for DAISY (Fig. 3g). The DAISY model performed less well on this dataset because it was initialised at the first measured data value. This may be anomalously low because of differences in analytical technique used at different dates; note that the first measured value of the high fertilization treatment is also low. Had DAISY used a higher initial total soil carbon value for 1956, the simulation line would be shifted closer to the measured values. The SOMM model, with a simulation line

Fig. 2. Measured and simulated values of total organic carbon in the top 30 cm of soil for the Bad Lauchstädt no fertilization treatment assuming a soil bulk density of 1.35 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM and (g) NCSOIL. ■ shows measured values with standard error bars; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Year

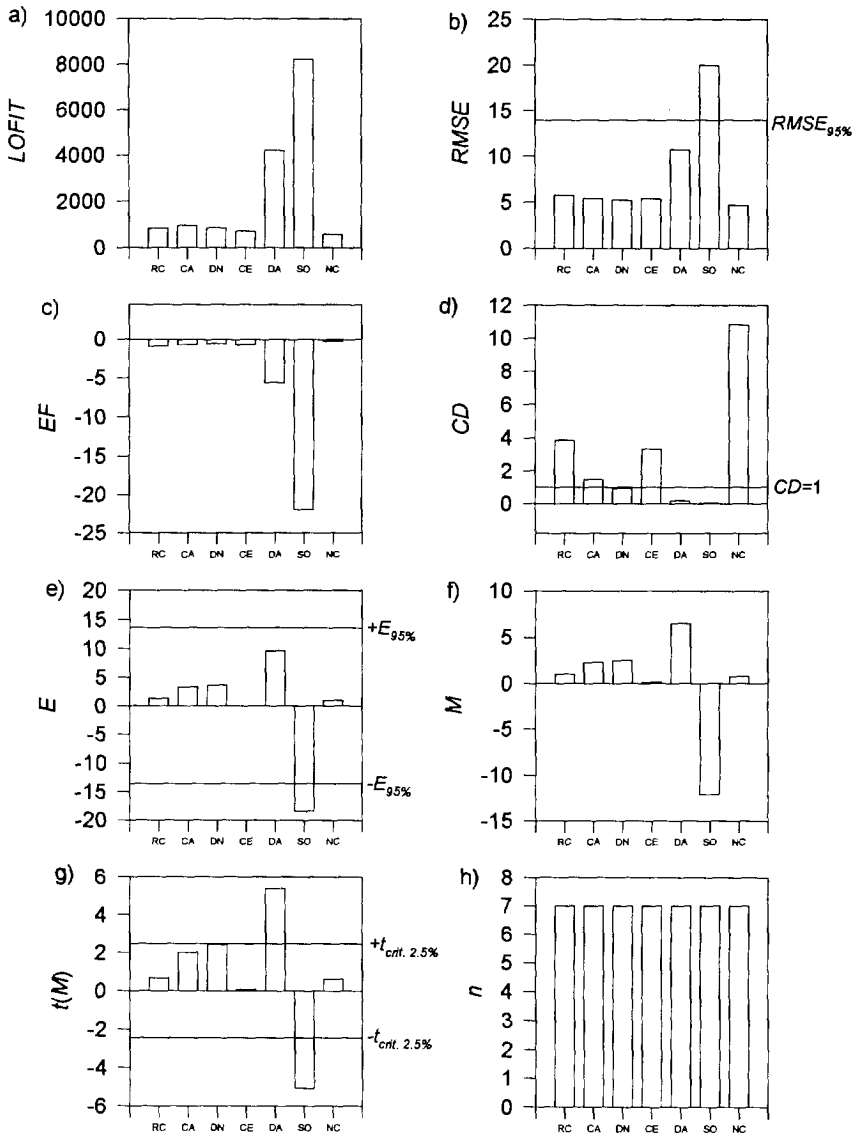


Fig. 3. Graphical representation of statistics describing the performance of models in simulating the Bad Lauchstädt no fertilization treatment. Depicted above are the following statistics: (a) lack of fit (LOFIT), (b) root mean square error (RMSE) with $RMSE_{95\%}$ value shown, (c) modelling efficiency (EF), (d) coefficient of determination (CD), (e) relative error (E) with $E_{95\%}$ values shown, (f) mean difference (M), (g) t value for M ($t(M)$) with critical 2.5% levels shown, and (h) number of paired values, n. In this figure and in all other figures the following abbreviations are used: *RC* = RothC, *CA* = CANDY, *DN* = DNDC, *CE* = CENTURY, *DA* = DAISY, *SO* = SOMM, *IT* = ITE, *Ve* = Verberne and *NC* = NCSOIL.

above the measured values and a trend opposite to that observed showed significant bias as measured by E and M.

3.1.1.2. Bad Lauchstädt—high fertilization. All models produced simulations within the 95% confidence interval of the measured data. As with the no fertilization treatment, RothC, CANDY, DNDC, CENTURY and NCSOIL produced simulation lines within the standard error of most of the measured data (Fig. 4a, b, c, d and g).

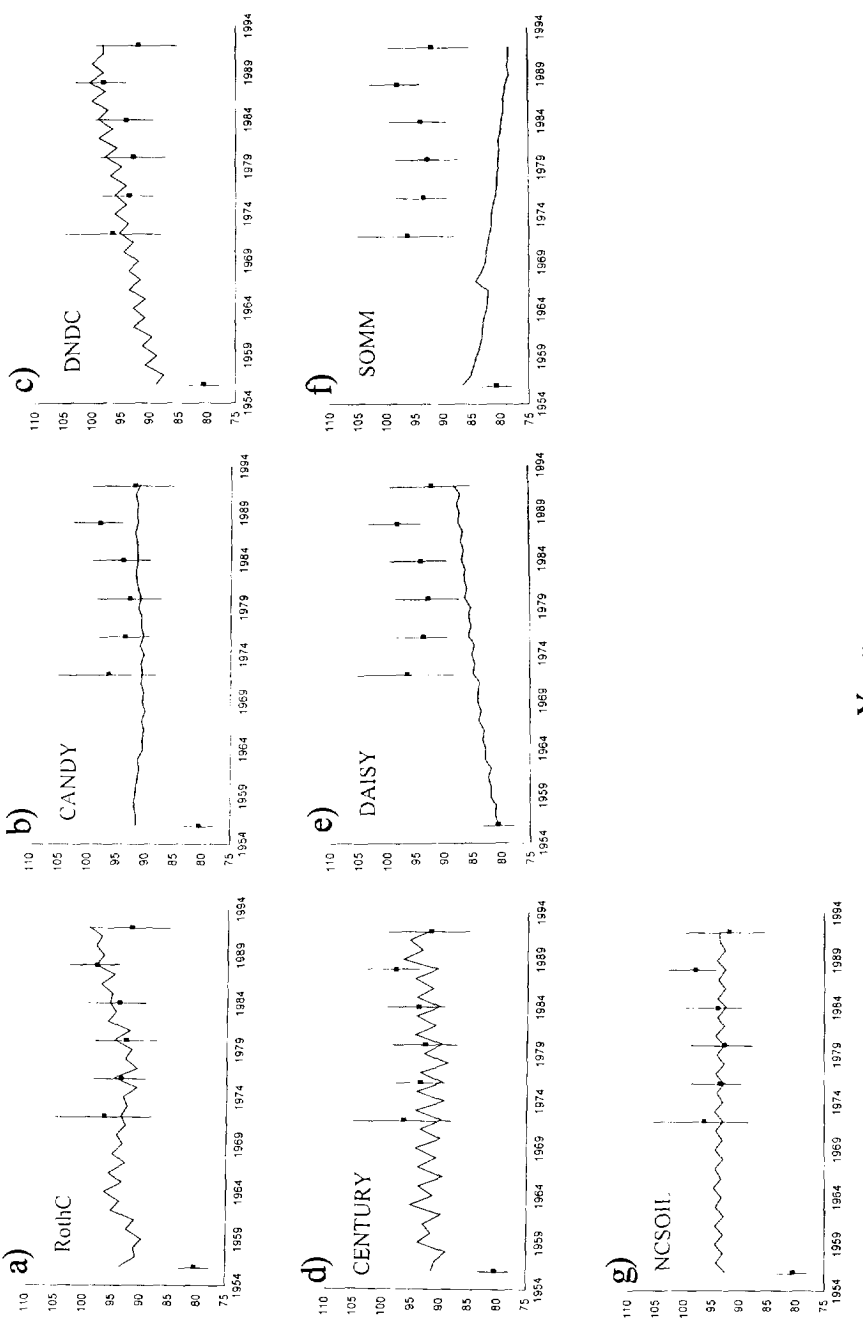
This close model fit was reflected in the statistics (Fig. 5) with all RMSE values less than $RMSE_{95\%}$ (Fig. 5b). DAISY and SOMM (Fig. 3e and f) produced simulation lines that fell below the standard errors of most of the measured values but the RMSE values were within the 95% confidence interval of the dataset (Fig. 5b). Other statistics (e.g. LOFIT; Fig. 5d) revealed a similar pattern but only CENTURY and DNDC showed positive EF values (Fig. 5c). Only RothC, CANDY, CENTURY and NCSOIL had values for CD greater than 1 (Fig. 5d).

All models had E values less than $E_{95\%}$, so none showed a bias outside the 95% confidence interval of the measured data (Fig. 5e) but values of M for DNDC, DAISY and SOMM indicates that their simulations were significantly biased (Fig. 5f) since they had t values for M that were greater than the critical two-tailed 2.5% t value (Fig. 5g). Again, the relatively poor performance of DAISY partly results from its initialisation to the anomalously low 1956 measured value (see above).

3.1.2. Calhoun Experimental Forest

Data were provided to run the models at the Calhoun Experimental Forest site from 1962 to 1993. Measured values of organic carbon for mineral soil only were used in this evaluation exercise. Six of the nine models (RothC, CANDY, CENTURY, SOMM, ITE and NCSOIL) attempted to simulate data from this experiment. Because a significant litter layer of high organic matter content has accumulated over the three decades since the experiment began (Richter and Markewitz, 1996), and because three of the models (CANDY, CENTURY and ITE) provide estimates of total organic carbon (mineral soil plus the litter layer), simulated values from only three of the models (RothC, SOMM and NCSOIL) can be compared to the measured values which are for mineral soil only. See the evaluation papers for CANDY (Franko et al., 1997), CENTURY (Kelly et al., 1997) and ITE (Arah et al., 1997) elsewhere in this issue to examine the performance of these models when simulating measured data that includes the litter layer. Since only two measured values are available, statistical comparison of the models is not possible.

As seen in Fig. 6, RothC simulated the increase in organic carbon in the top 15 cm of soil well, with simulated values within the standard error of the measured values at the beginning and end of the experiment (Fig. 6a).


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

SOMM also predicted an increase but less than that measured (Fig. 6b). By contrast, NCSOIL simulated a rise sharper than that measured, producing simulated values lower than measurements at the beginning of the experiment, and higher at the end (Fig. 6c). Based upon this qualitative visual comparison between the models it appears that RothC simulated the measured data more closely than did SOMM and NCSOIL.

3.1.3. Rothamsted Park Grass

The two treatments selected were the organic manure treatment and the no fertilization treatment. Data were provided to run the models from 1856 to 1993. All nine models attempted to model the data although NCSOIL only simulated for the period from 1958 to present.

3.1.3.1. Rothamsted Park Grass—no fertilization. Since NCSOIL simulated only the last 36 years of the experiment (Fig. 7i), it could not be compared statistically to the other models. For this run of data, however, NCSOIL's simulations had low total error and bias (Fig. 8). Visual examination of Fig. 7 shows that RothC, CANDY, DNDC, CENTURY, DAISY and Verberne were close to the measured data (Fig. 7a, b, c, d, e and h), whilst SOMM appears to dramatically underestimate organic carbon content for most of the experiment (Fig. 7f), and ITE tended to overestimate (Fig. 7g). This qualitative visual examination of the simulations is supported by the statistics (Fig. 8).

RothC, CANDY, DNDC, CENTURY, DAISY and Verberne all show low measures of total model error (RMSE and EF; Fig. 8a and b), whilst ITE and SOMM had larger errors. Since no standard errors associated with the measured data points were available, 95% confidence intervals for the measured data could not be calculated. Only DAISY had a positive EF value. Only RothC, DAISY and Verberne had CD values greater than 1 (Fig. 8c).

Model bias as revealed by E (Fig. 8d) and M (Fig. 8e) values showed a similar pattern to that seen for statistics describing total error. The t values for M in Fig. 8f indicate significant bias for SOMM and ITE. This bias was depicted in Fig. 7 (f and g, respectively) in which SOMM produced a simulation line which fell below the body of measured data, and ITE which was mostly above.

3.1.3.2. Rothamsted Park Grass—organic manure treatment. The apparent decline in soil organic carbon content of the 0–23 cm soil layer between 1876

Fig. 4. Measured and simulated values of total organic carbon in the top 30 cm of soil for the Bad Lauchstädt high fertilization treatment assuming a soil bulk density of 1.35 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM and (g) NCSOIL. ■ shows measured values with standard error bars; simulated values shown as a line.

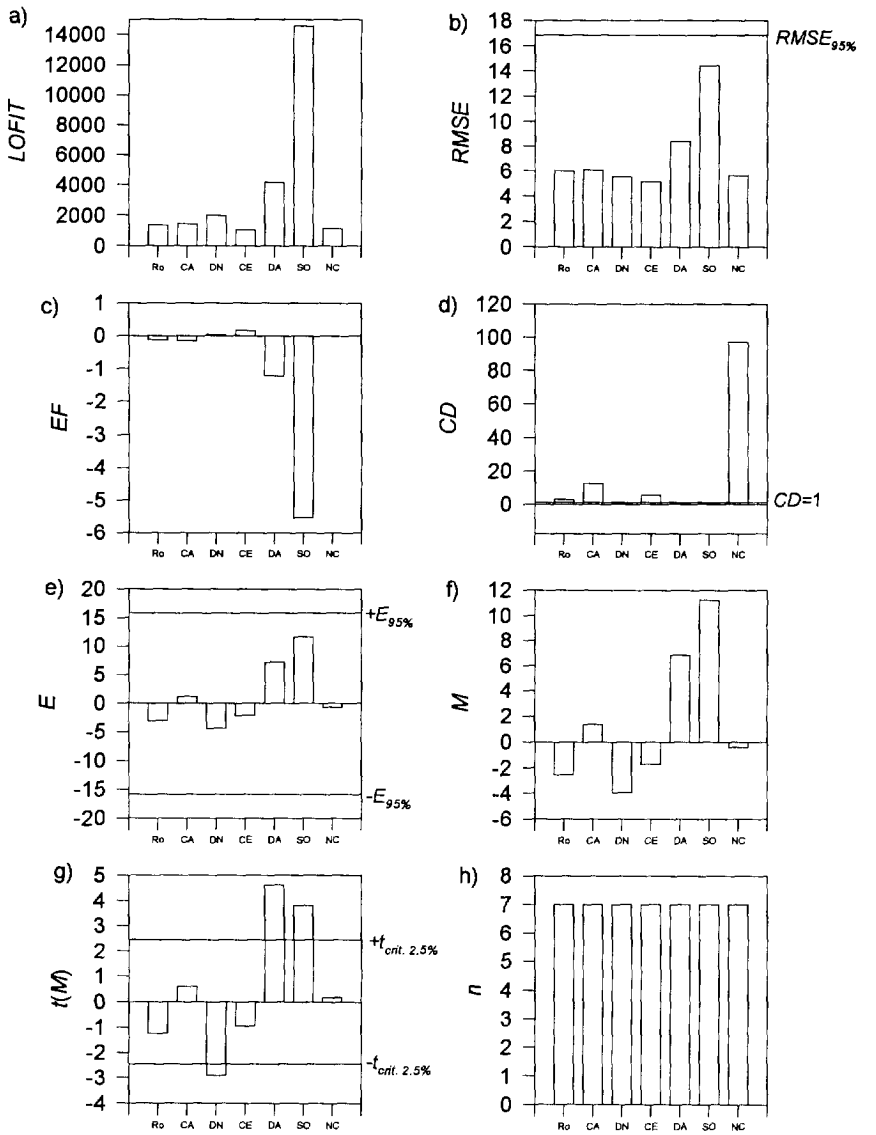
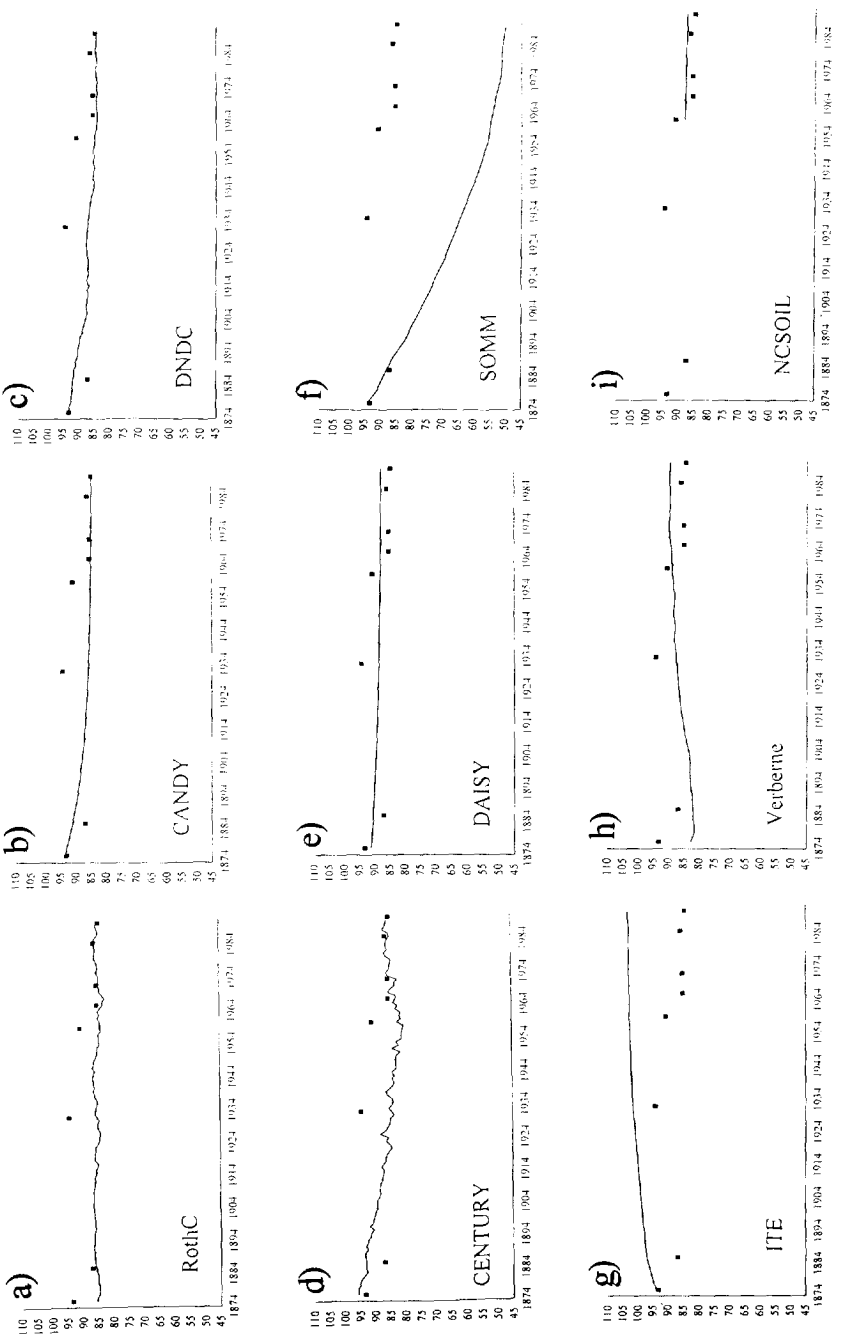


Fig. 5. Graphical representation of statistics describing the performance of models in simulating the Bad Lauchstädt high fertilization treatment. Depicted above are the following statistics: (a) lack of fit (LOFIT), (b) root mean square error (RMSE) with $RMSE_{95\%}$ value shown, (c) modelling efficiency (EF), (d) coefficient of determination (CD), (e) relative error (E) with $E_{95\%}$ values shown, (f) mean difference (M), (g) t value for M ($t(M)$) with critical 2.5% levels shown, and (h) number of paired values, n. Abbreviations as for Fig. 3.

and the next measurement in 1932 (Fig. 9), despite additions of straw (1856–1897) and then farmyard manure at 35 t ha^{-1} every four years since 1905, cannot be explained by any effect of the treatments (Poulton, 1996a). The plot

Fig. 6. Measured and simulated values of total organic carbon in the top 15 cm of soil for the Calhoun Experimental Forest assuming a soil bulk density of 1.52 g cm^{-3} for (a) RothC, (b) SOMM and (c) NCSOIL. ■ shows measured values with standard error bars; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
Verberne
i)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

was subdivided a number of times during the experiment with samples then taken from smaller subsections of the original plot (see Poulton, 1996a). It is considered most likely that the apparent decline in measured soil organic carbon content is due to spatial variability within the original plot. Given this weakness in the measured data, it is not surprising that none of the models simulated this initial decline followed by the rise in soil carbon content (Figs. 9 and 10).

NCSOIL simulated only the last 36 years of the experiment (Fig. 9i), which has only two measured points so model performance cannot be evaluated statistically or compared to the other models. A qualitative visual examination of Fig. 9 shows that the models produced very different trends, none giving a close fit. Both RothC (Fig. 9a) and ITE (Fig. 9g) simulated a rise in soil organic carbon content, as might be expected given the organic inputs. SOMM (Fig. 9f), surprisingly, simulated a steep decline. CENTURY, DAISY and Verberne (Fig. 9d, e and h, respectively) simulated only small changes during the experiment whilst CANDY (Fig. 9b) and DNDC (Fig. 9c) simulated an overall decline to about 1900 (when straw addition ended and organic amendments began) followed by a little change during this century.

RMSE values show a spread in total error among the models (Fig. 10a). Only CANDY and CENTURY had positive values EF (Fig. 10b). CANDY, DNDC, CENTURY, DAISY and Verberne all had values for CD greater than 1 (Fig. 10c). This may partly reflect the large spread in the measured data.

Model bias as indicated by E shows a similar pattern to statistics of total error (Fig. 10d) and the t values (Fig. 10f) for M (Fig. 10e) show that there is no significant bias in any of the model simulations. This may partly reflect the large scatter among the few measured values.

3.1.4. Prague–Ruzyně

The two treatments selected were the no fertilization and high fertilization treatments. Data were provided to run the models from 1972 to 1994. All nine models attempted to model data from this site. The ITE model produced simulated organic carbon estimates corresponding to only 18 of the 21 measured data points. However, as the majority of measured data was simulated this model was included in the statistical comparison with other models. The Verberne model was run only for the no fertilization treatment. As noted earlier, ITE and Verberne simulated arable crops as if they were grass.

Fig. 7. Measured and simulated values of total organic carbon in the top 23 cm of soil for the Rothamsted Park Grass no fertilization treatment assuming a soil bulk density of 1.20 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE, (h) Verberne and (i) NCSOIL. ■ show measured values; simulated values shown as a line.

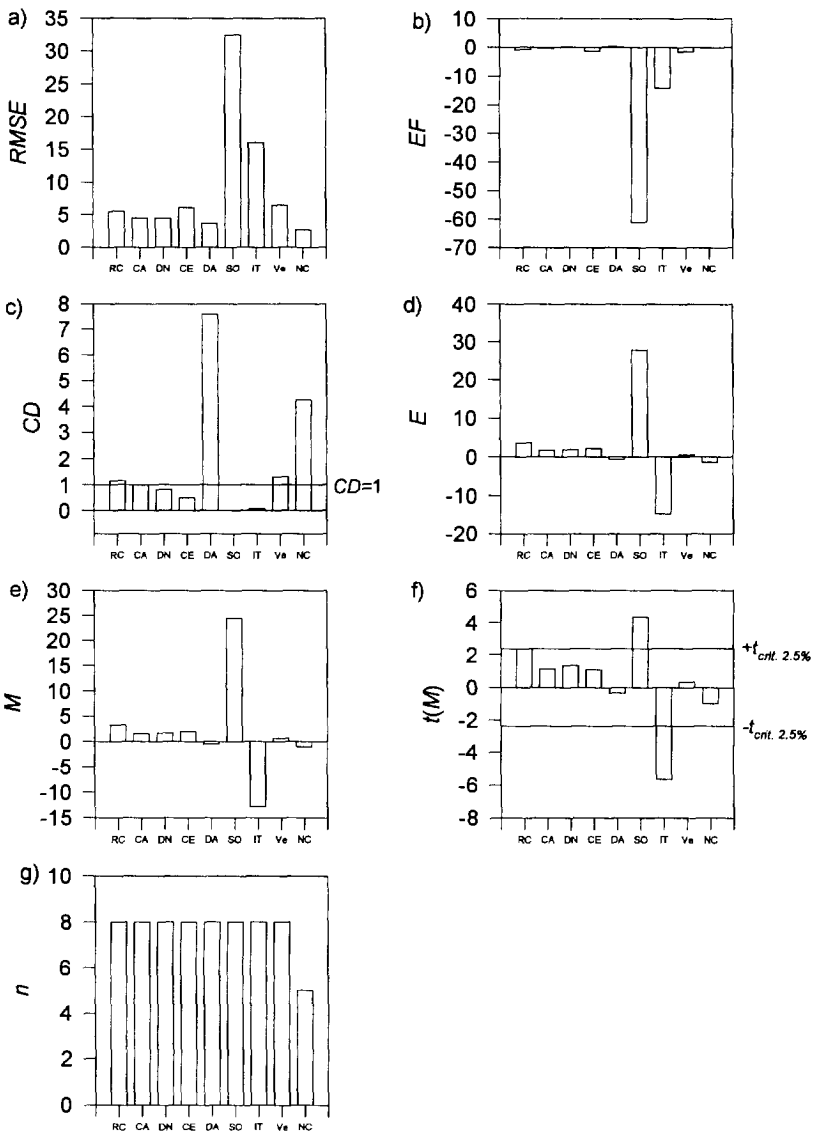


Fig. 8. Graphical representation of statistics describing the performance of models in simulating the Rothamsted Park Grass no fertilization treatment. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3. Note NCSOIL simulation only from 1959 onwards (5 values) so statistics cannot be compared directly with other models.

3.1.4.1. Prague–Ruzyně—no fertilization. A qualitative visual examination of Fig. 11 shows that RothC, CANDY, DNDC, DAISY, SOMM and NCSOIL (Fig. 11a, b, c, e, f, and i, respectively) simulated an unchanged soil organic carbon content in the top 20 cm of soil in accord with the measured data. CENTURY and Verberne (Fig. 11d and h, respectively) simulated a decline, and ITE simulated a rise (Fig. 11g).

RMSE values show that ITE and Verberne produced larger total errors than other models (Fig. 12a). No models had positive values of EF but RothC, CANDY, DNDC, DAISY and SOMM all had values greater than 1 for CD.

For measures of model bias Verberne and ITE had the highest E values (Fig. 12d). Values of t (Fig. 12f) for M (Fig. 12e) indicate a significant bias in the simulations of DNDC, NCSOIL, ITE and Verberne. The bias can be seen in Fig. 11, with the ITE simulation line above the measurements (Fig. 11g) and the Verberne simulation line below the majority of measurements (Fig. 11h). The bias in DNDC and NCSOIL (Fig. 11c and h, respectively) was less pronounced, but visible.

It is noteworthy that whilst CENTURY had a CD of less than 1 (total error) but it did not produce a significantly biased simulation. This is reflected in Fig. 11d in which a negative trend in the simulation can be seen. The figure reveals that CENTURY at first overestimated organic carbon content and later underestimated it. This led to a CD value less than 1 but a simulation that was not biased toward consistent over- or under-prediction.

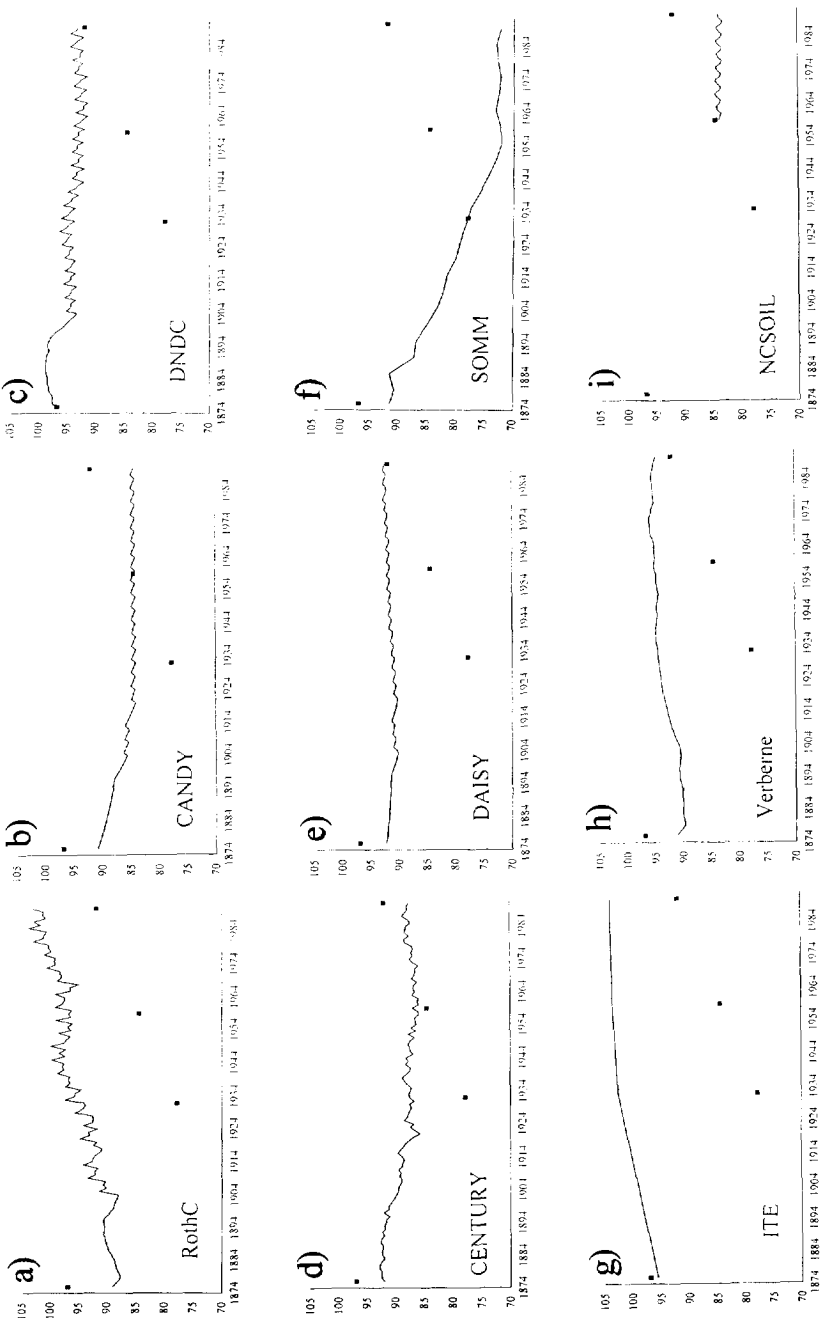
3.1.4.2. Prague–Ruzyně—high fertilization. A qualitative visual examination of Fig. 13 shows that all models simulated a gradual rise in soil organic carbon content except for CENTURY (Fig. 13d) which simulated little or no change. The measured data also appear to show a gradual rise but the wide scatter among the last three values make it difficult to assess whether the trend is real.

RMSE values show that RothC and ITE had higher total errors than other models (Fig. 14a). Four of the models (CENTURY, DAISY, SOMM and NCSOIL) had positive values of EF (Fig. 14b) and values of CD greater than 1 (Fig. 14c).

In terms of model bias RothC and ITE had the largest E values (Fig. 14d). The t values (Fig. 14 f) for M (Fig. 14e) suggest that there was no significant bias in simulations of CENTURY or DAISY, but all other models showed a significant bias.

3.1.5. Tamworth

The two Tamworth rotations selected were the lucerne/clover and the fallow rotations. Data were provided to run the models from 1970 to 1993. Seven of the nine models (RothC, CANDY, DNDC, CENTURY, DAISY, SOMM and NCSOIL) attempted to model data from the Tamworth experiment. The ITE and


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
Verberne
i)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

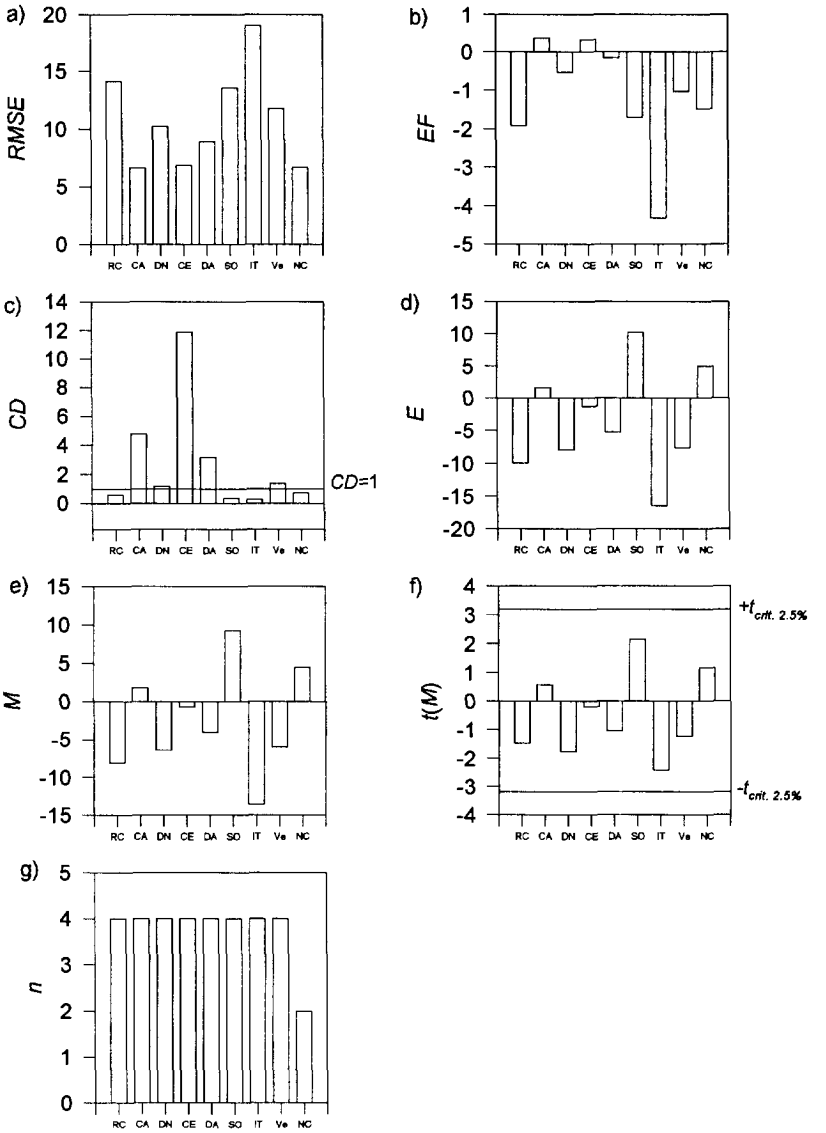
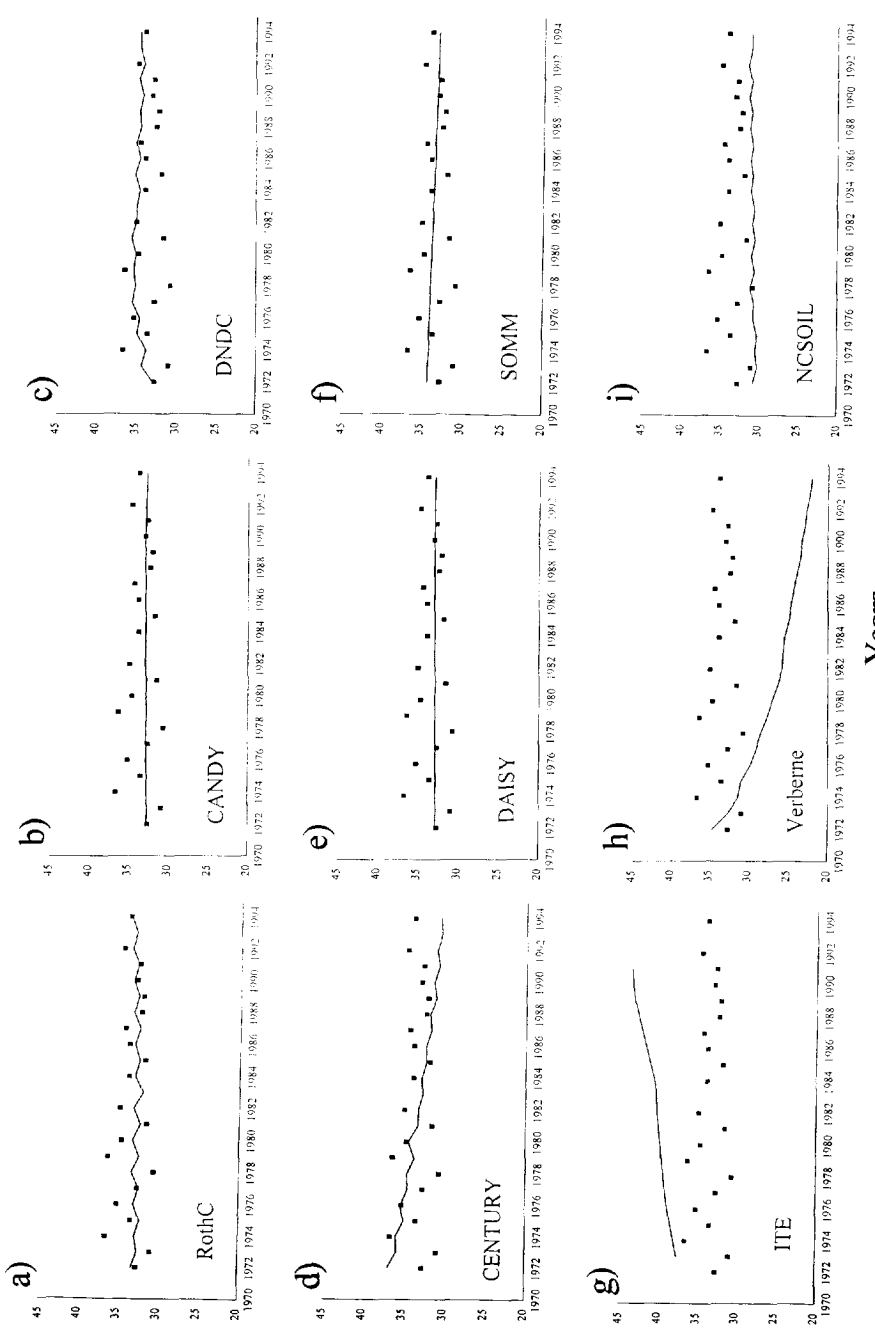


Fig. 10. Graphical representation of statistics describing the performance of models in simulating the Rothamsted Park Grass organic manure treatment. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3. Note NCSOIL simulation only from 1959 onwards (2 values) so statistics cannot be compared directly with other models.

Fig. 9. Measured and simulated values of total organic carbon in the top 23 cm of soil for the Rothamsted Park Grass organic manure treatment assuming a soil bulk density of 1.20 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE, (h) Verberne and (i) NCSOIL. ■ shows measured values; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
Verberne
i)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

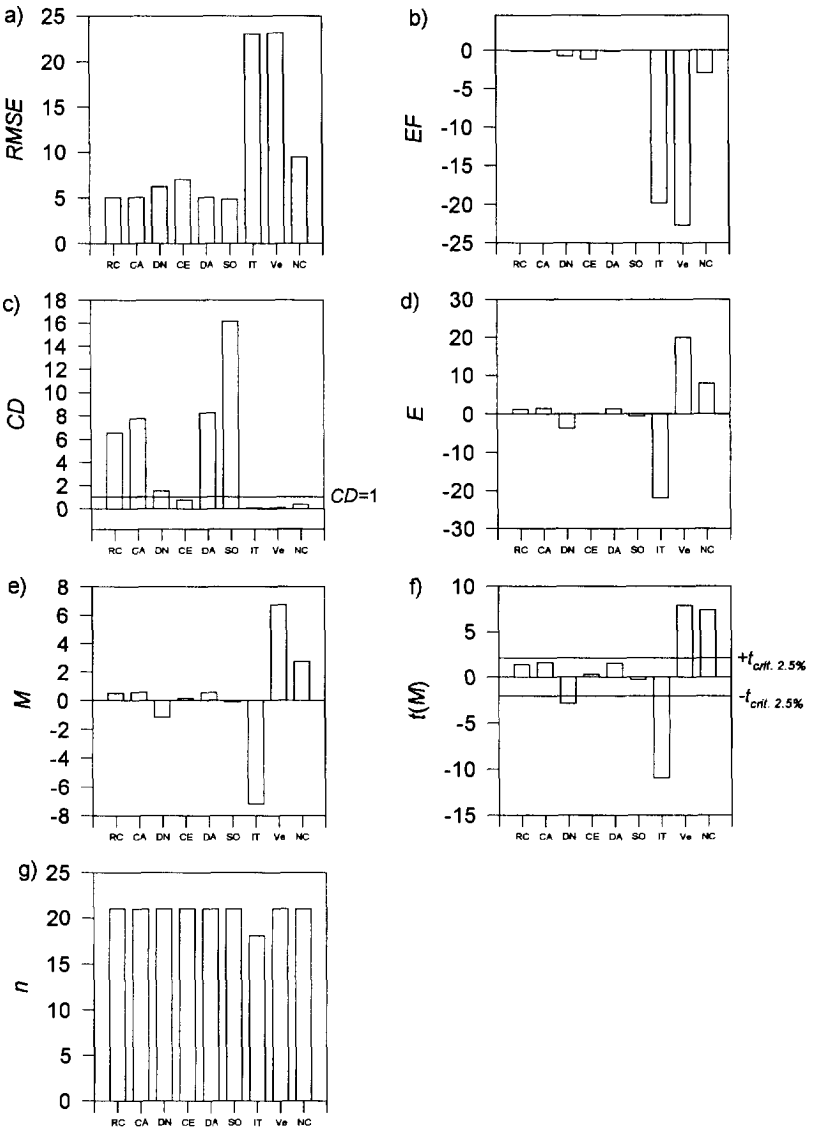
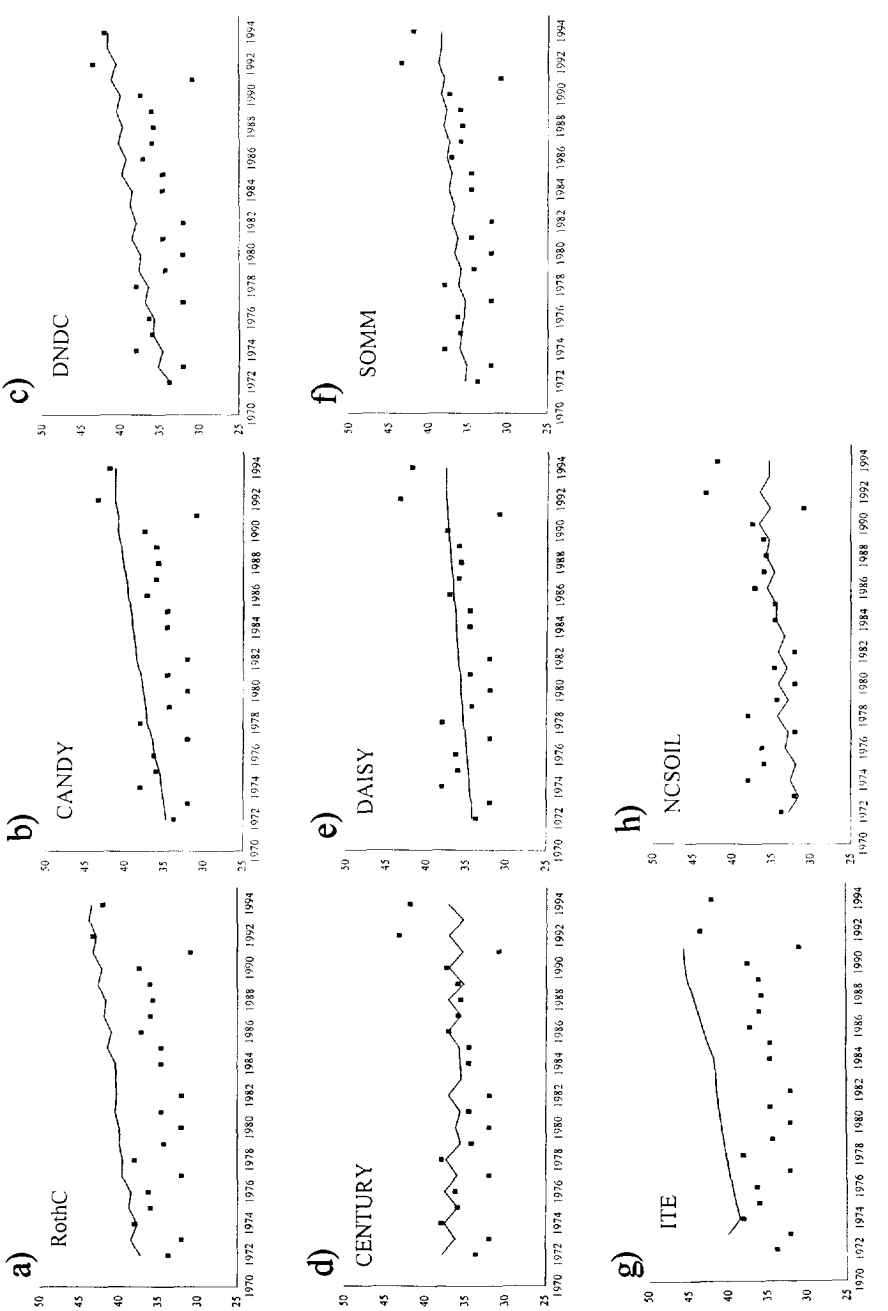


Fig. 12. Graphical representation of statistics describing the performance of models in simulating the Prague–Ruzyně no fertilization treatment. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3.

Fig. 11. Measured and simulated values of total organic carbon in the top 20 cm of soil for the Prague–Ruzyně no fertilization treatment assuming a soil bulk density of 1.42 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE, (h) Verberne and (i) NCSOIL. ■ shows measured values; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

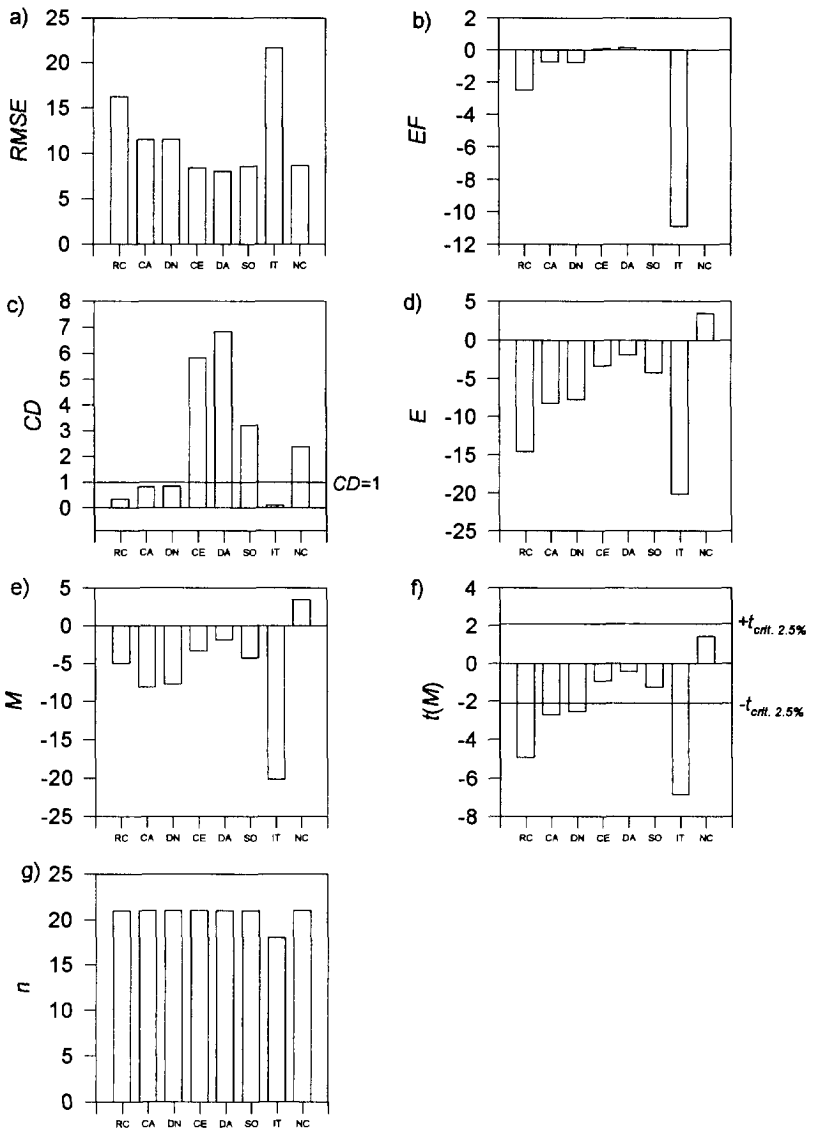


Fig. 14. Graphical representation of statistics describing the performance of models in simulating the Prague–Ruzyně no fertilization treatment. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3.

Fig. 13. Measured and simulated values of total organic carbon in the top 20 cm of soil for the Prague–Ruzyně high fertilization treatment assuming a soil bulk density of 1.42 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE and (h) NCSOIL. ■ shows measured values; simulated values shown as a line.

Verberne models did not, and NCSOIL simulated organic carbon only for the period 1970 to 1978 and so could not be compared statistically to other models.

3.1.5.1. Tamworth—fallow rotation . NCSOIL provided simulated values up to 1978 but did not simulate the latter 15 years of the experiment. For this limited simulation, the statistics for NCSOIL show low model error and bias (see Fig. 16). For the other models, visual examination of Fig. 15 shows that the models produced different trends. DNDC matches the data well including the rise and decline between 1970 and 1980 (Fig. 15c). RothC showed a similar trend but simulated values were too low (Fig. 15a). All other models, with the possible exception of CENTURY showed an approximately linear simulation with either a slight decline (CENTURY, DAISY, SOMM; Fig. 15d, e and f) or no change (CANDY; Fig. 15b) over the entire period.

DNDC, DAISY and CENTURY had the lowest LOFIT values followed by CANDY, SOMM and RothC but all models had an F value for LOFIT higher than the critical 5% value (Fig. 16b), showing that the error in the simulated values was significantly greater than the error inherent in the measured values for all models. DNDC, CENTURY and DAISY had RMSE values within the 95% confidence interval of the measured data whilst CANDY, SOMM and RothC (in increasing order of error) did not (Fig. 16c). DNDC and DAISY had positive EF values (Fig. 16d) whilst CANDY, DNDC, CENTURY and DAISY had values of 1 or above (Fig. 16e).

In terms of models bias, only SOMM and RothC showed a significant bias as estimated by E (Fig. 16f) though CANDY also showed significant bias as estimated by the t value (Fig. 16h) for M (Fig. 16g). These trends in bias can be seen in Fig. 15.

3.1.5.2. Tamworth—lucerne / clover rotation. NCSOIL performed the simulation up to 1978 only and so cannot be compared directly with other models. However, for this limited simulation, NCSOIL produced greater total error than CANDY and DAISY as measured by RMSE (outside the 95% confidence interval of the measurements) a negative EF and a CD less than 1 (see Fig. 18). Values of E outside the 95% confidence interval of the data show that the error was biased towards underprediction (see Fig. 17g). For the other models, visual examination of Fig. 17 shows that the models produced different trends with CANDY, DNDC, CENTURY and DAISY (Fig. 17b, c, d and e) simulating an overall slight rise in soil organic carbon content whilst SOMM simulated a slight decline (Fig. 17f), and RothC (Fig. 17a) simulated little or no change.

LOFIT values are shown in Fig. 18a. All models had an F value for LOFIT considerably higher than this (Fig. 18b) showing that, for all simulations, the error in the simulated values was significantly greater than the error inherent in the measured values. In terms of total error, only CANDY and DAISY had RMSE values within the 95% confidence interval of the measured data (Fig.

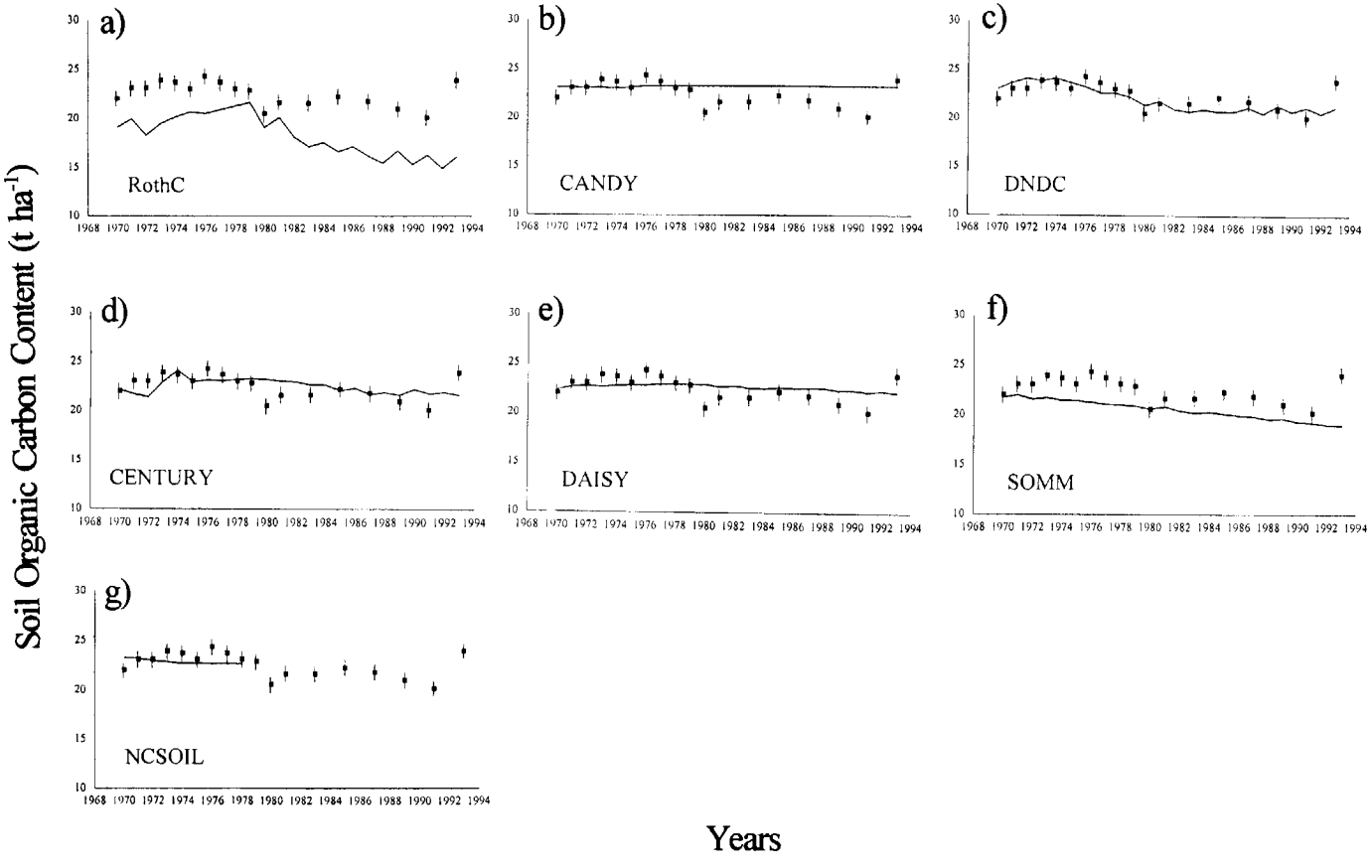


Fig. 15. Measured and simulated values of total organic carbon in the top 15 cm of soil for the Tamworth fallow rotation assuming a soil bulk density of 1.40 g cm ^{3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM and (g) NCSOIL. ■ shows measured values with standard error bars; simulated values shown as a line.

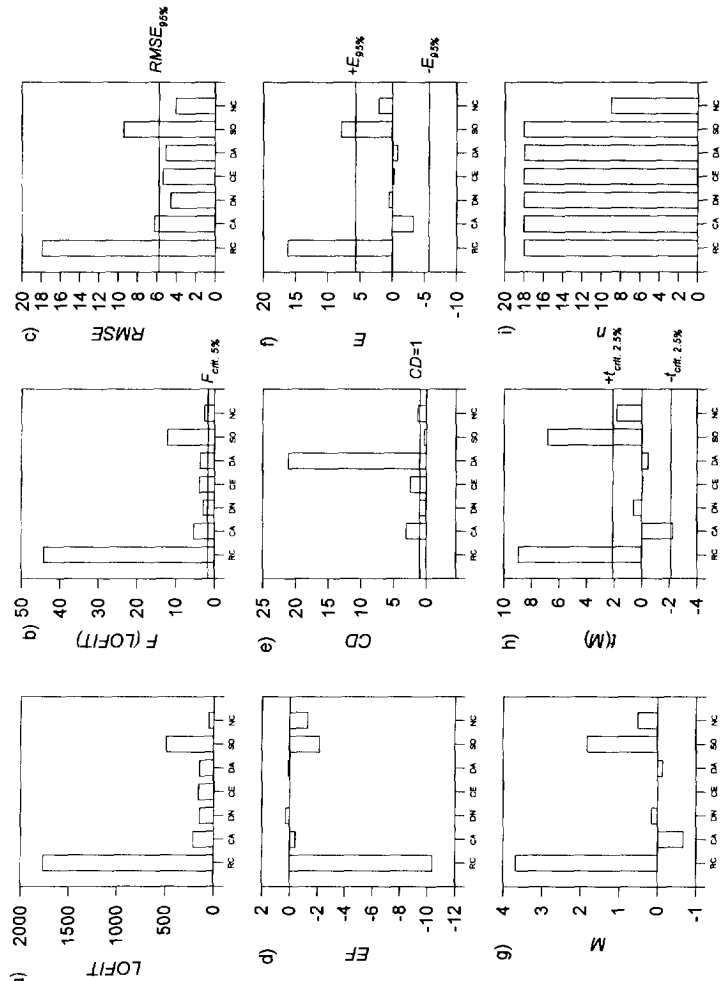


Fig. 16. Graphical representation of statistics describing the performance of models in simulating the Tamworth fallow rotation. Depicted above are the following statistics: (a) lack of fit (LOFIT), (b) F values for the LOFIT statistic with critical 5% level shown, (c) root mean square error (RMSE) with $RMSE_{95\%}$ value shown, (d) modelling efficiency (EF), (e) coefficient of determination (CD), (f) relative error (E) with $E_{95\%}$ values shown, (g) mean difference (M), (h) t value for M ($t(M)$) with critical 2.5% levels shown, and (i) number of paired values, n. Abbreviations as for Fig. 3. Note NCSOIL simulation only to 1978 (9 values) so statistics cannot be compared directly with other models.

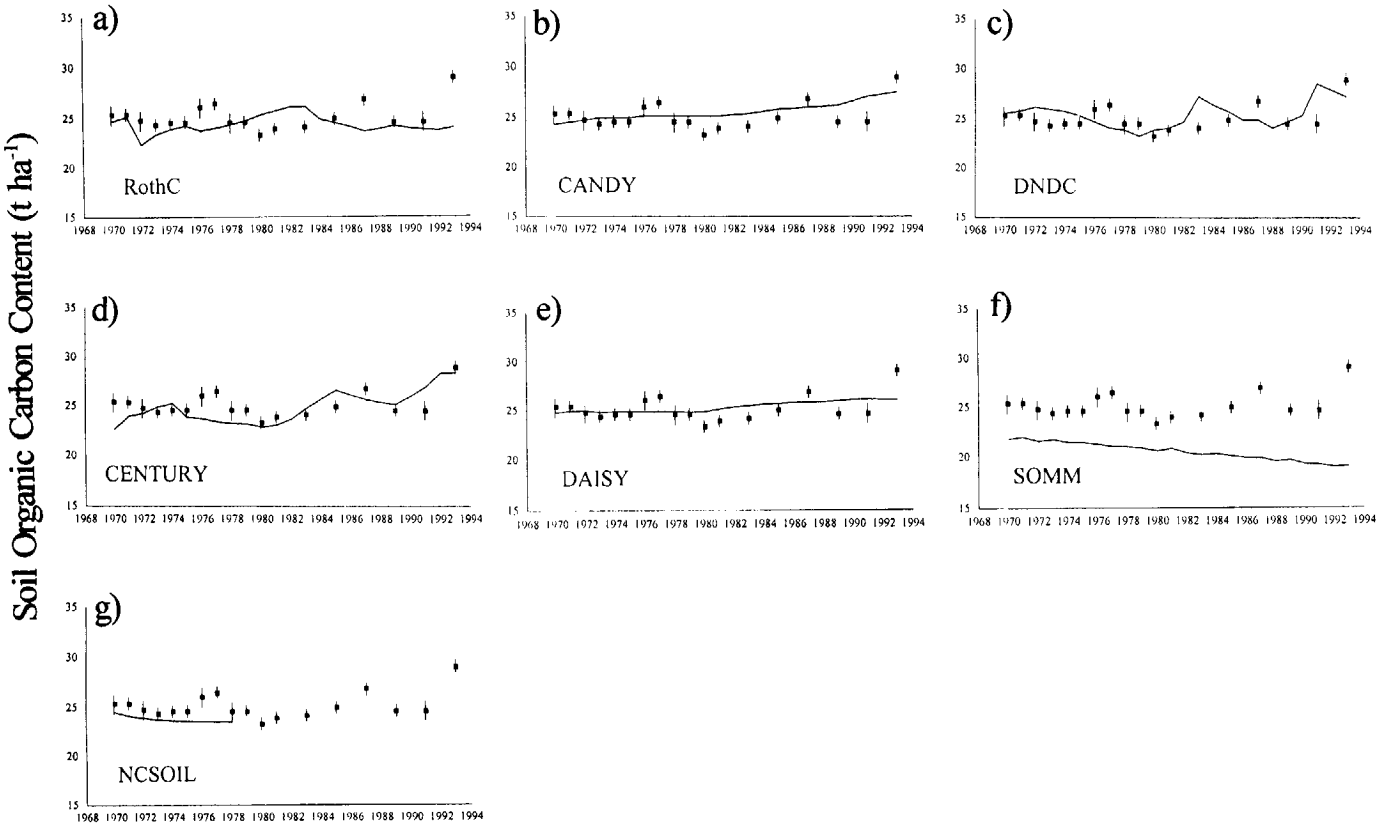


Fig. 17. Measured and simulated values of total organic carbon in the top 15 cm of soil for the Tamworth lucerne/clover rotation assuming a soil bulk density of 1.40 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM and (g) NCSOIL. ■ shows measured values with standard error bars; simulated values shown as a line.

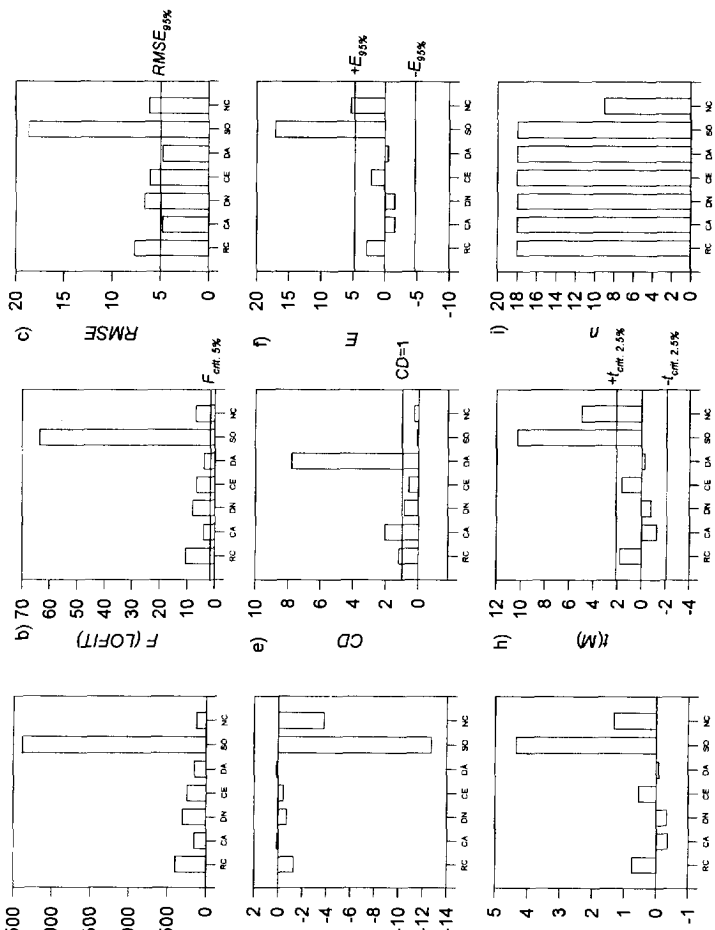


Fig. 18. Graphical representation of statistics describing the performance of models in simulating the Tamworth Lucerne/Clover rotation. Depicted above are the following statistics: (a) lack of fit (LOFIT), (b) F values for the LOFIT statistic with critical 5% level shown, (c) root mean square error (RMSE) with $RMSE_{95\%}$ value shown, (d) modelling efficiency (EF), (e) coefficient of determination (CD), (f) relative error (E) with $E_{95\%}$ values shown, (g) mean difference (M), (h) t value for M ($t(M)$) with critical 2.5% levels shown, and (i) number of paired values, n. Abbreviations as for Fig. 3. Note NCSOIL simulation only to 1978 (9 values) so statistics cannot be compared directly with other models.

18c) and positive EF values (Fig. 18d). Only DAISY, CANDY and RothC had values of CD above 1 (Fig. 18e).

Only SOMM showed significant bias as estimated by E (Fig. 18f) and M (Fig. 18g and h) with a consistent underestimation of soil organic carbon content (see Fig. 17f).

3.1.6. Rothamsted Geescroft Wilderness

Data were provided to run simulations for Geescroft Wilderness from 1883 to 1985. Six of the nine models (RothC, CENTURY, SOMM, ITE, Verberne and NCSOIL) attempted to model data from Geescroft Wilderness. CANDY provided values that included the litter layer so could not be compared with the measured data which were for mineral soil only. SOMM simulated organic carbon only for the period 1959 to present (only two measured data points) and cannot therefore be evaluated statistically or compared to other models.

Visual examination of Fig. 19 shows that all models simulate the rise in soil organic carbon content (0–23 cm layer) to approximately the final measured value, except for Verberne (Fig. 19e) which predicts little overall change. The sharp rises seen for CENTURY and Verberne (Fig. 19b and e) at 1959 and 1985 are the result of adjustment of the model output to account for changing soil bulk density (other model outputs did not require correction). The correction would ideally be continuous but could only be applied where bulk density measurements were available (at measurement dates).

Verberne had the largest total error (e.g. RMSE, Fig. 19a) and was the only model with a negative EF (Fig. 20b). Only Verberne and ITE had CD values less than 1 (Fig. 20c). In terms of model bias (E; Fig. 20d and M; Fig. 20e), Verberne showed the greatest bias but the t values for M indicate that there was no significant bias in any of the simulations, even for Verberne (Fig. 20f).

Since there is an obvious positive trend in the measured data, the correlation coefficient was used to assess how well the shape of the simulation matches the shape of the measured values. All models except Verberne showed high positive correlations between measured and simulated data (Fig. 20g). The highest values were for RothC and NCSOIL. Indeed, had the measured values not been used to tune the models, the correlations for RothC and NCSOIL would have been above the critical 5% t value (see Fig. 20h) calculated according to Smith et al. (1996).

3.1.7. Waite Rotations

Data were provided to run the models for the Waite Rotations from 1925 to 1993. However, total organic carbon values were available at 0–20 cm only for 1963, 1973, 1983 and 1993. The 1925 value given was for 0–10 cm only and was therefore omitted from the comparison exercise but was available for model tuning. Measured total organic carbon data for the wheat–oats–grass pasture–fallow rotation (blind test) were withheld from the modellers (see Section 2.1.)

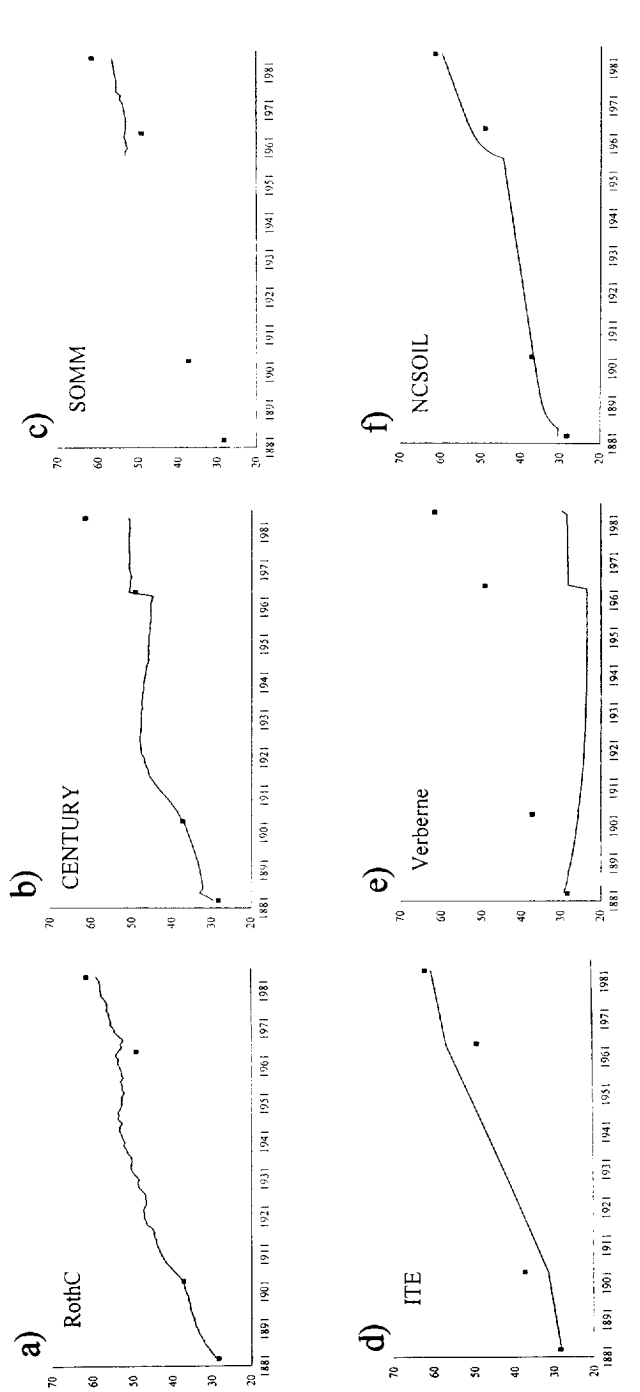


Fig. 19. Measured and simulated values of total organic carbon in the top 23 cm of soil for the Rothamsted Geescroft Wilderness assuming a soil bulk density of 1.19 g cm^{-3} for (a) RothC, (b) CENTURY, (c) SOMM, (d) ITE, (e) Verberne and (f) NCSOIL. ■ shows measured values; simulated values shown as a line.

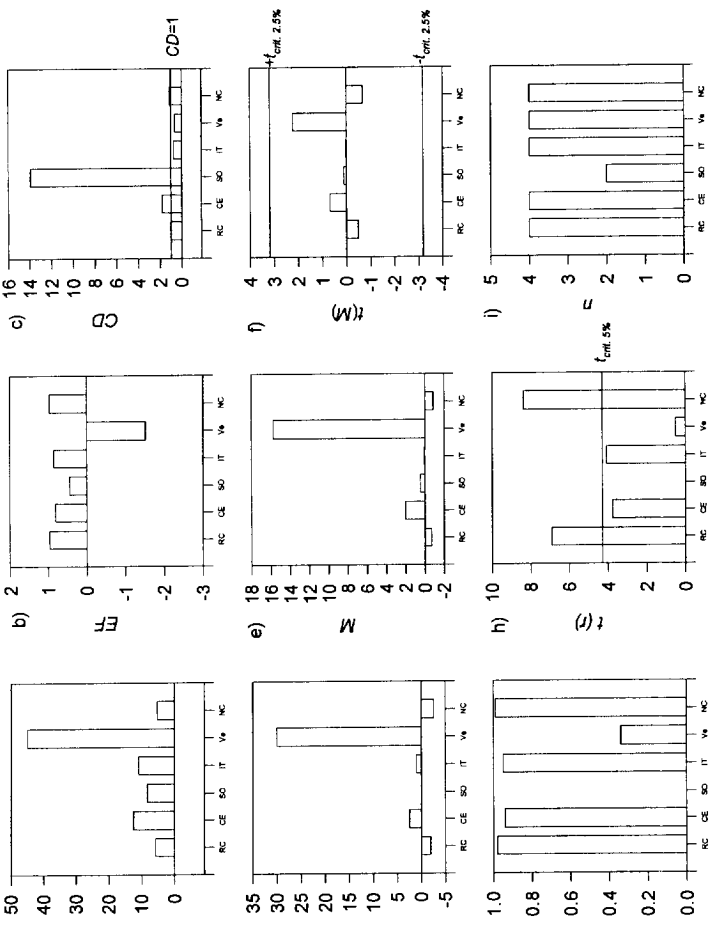


Fig. 20. Graphical representation of statistics describing the performance of models in simulating Rothamsted Geescroft Wilderness. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, (g) correlation coefficient (r), (h) t value for r with critical 5% level shown, and (i) number of paired values, n. Abbreviations as for Fig. 3. Note NCSOIL simulation only from 1959 onwards so statistics cannot be compared directly with other models.

so this was the only blind test in the evaluation. All nine models attempted to simulate data from the Waite Rotations. In simulations performed by the ITE and Verberne models, the wheat crop was modelled as though it were grass.

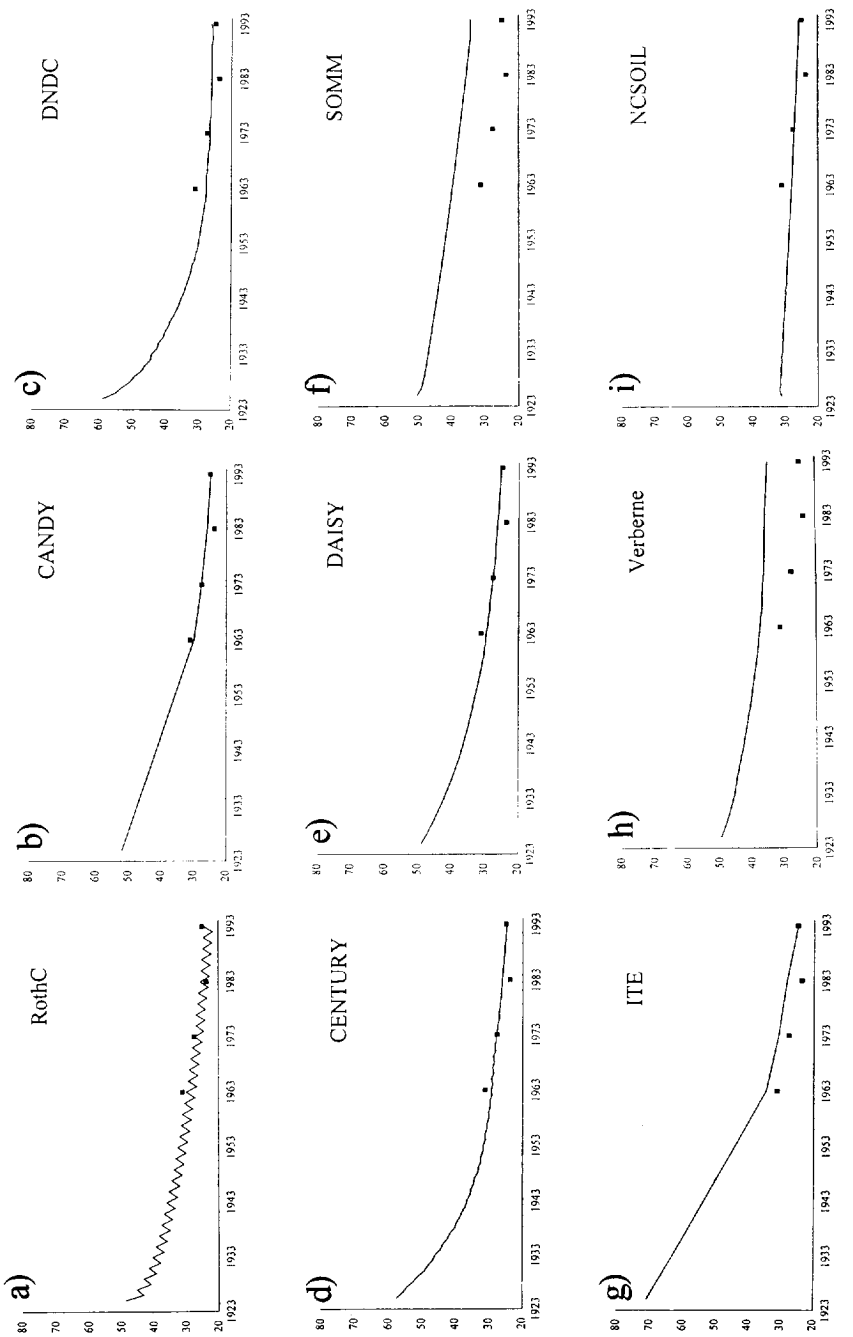
3.1.7.1. Waite Rotation—wheat / oats rotation. Visual examination of Fig. 21 shows that all models simulated a decline in soil organic carbon content. Starting values at 1925 differed widely among the simulations ranging from just above 30 to over 70 t ha^{-1}. Most simulation lines fall between the measured values between 1963 and 1993 except for those of SOMM and Verberne (Fig. 21f and h, respectively) which lie wholly above the measured data points.

In terms of total error, six of the nine models, i.e. RothC, CANDY, DNDC, CENTURY, DAISY and NCSOIL had positive values of EF (Fig. 22b) and values of 1 or above for CD (Fig. 22c). RMSE values for Verberne and SOMM were much larger than for other models (Fig. 22a). In terms of model bias (E; Fig. 22d and M; Fig. 22e) only Verberne and SOMM showed significant bias (Fig. 22f). This bias can be seen in Fig. 21 (f and h).

3.1.7.2. Waite Rotation—wheat–oats–grass-pasture–fallow rotation. The ITE model simulated only two of the measured data points and so there are too few paired values for statistical evaluation or comparison with other models. Visual examination of Fig. 23 shows that all models predict a decline in soil organic carbon content (note that ITE is based upon simulated values at 1925, 1973 and 1993 only; Fig. 23g). Again, starting values at 1925 differ widely among the simulations ranging from just above 30 to about 70 t ha^{-1}. Most models captured either the first two points or the last two points, while Verberne's simulation lay between and ITE was below all of them. The pattern of loss in the measured data suggests a possible change in analytical or sampling method, and no model simulated the more abrupt drop between points two and three.

Only two of the models, i.e. DAISY and Verberne, had positive values of EF (Fig. 24b) but all models except CENTURY and NCSOIL had CD values greater than 1 (Fig. 24c). RMSE values ranged from 9 to around 20 (Fig. 24a). There was no significant bias for any model as estimated by the t value (Fig. 24f) for M (Fig. 24e) but E values ranged from 5 and 17 (Fig. 24d). This final dataset is the only one for which models were forced to predict changes in soil organic carbon content without previous knowledge of the data. As such, it is the only part of the modelling exercise in which the models were used in a truly predictive way. The results from simulating this single dataset do not allow

Fig. 21. Measured and simulated values of total organic carbon in the top 20 cm of soil for the Waite wheat/fallow rotation assuming a soil bulk density of 1.35 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE, (h) Verberne and (i) NCSOIL. ■ shows measured values; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
Verberne
i)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

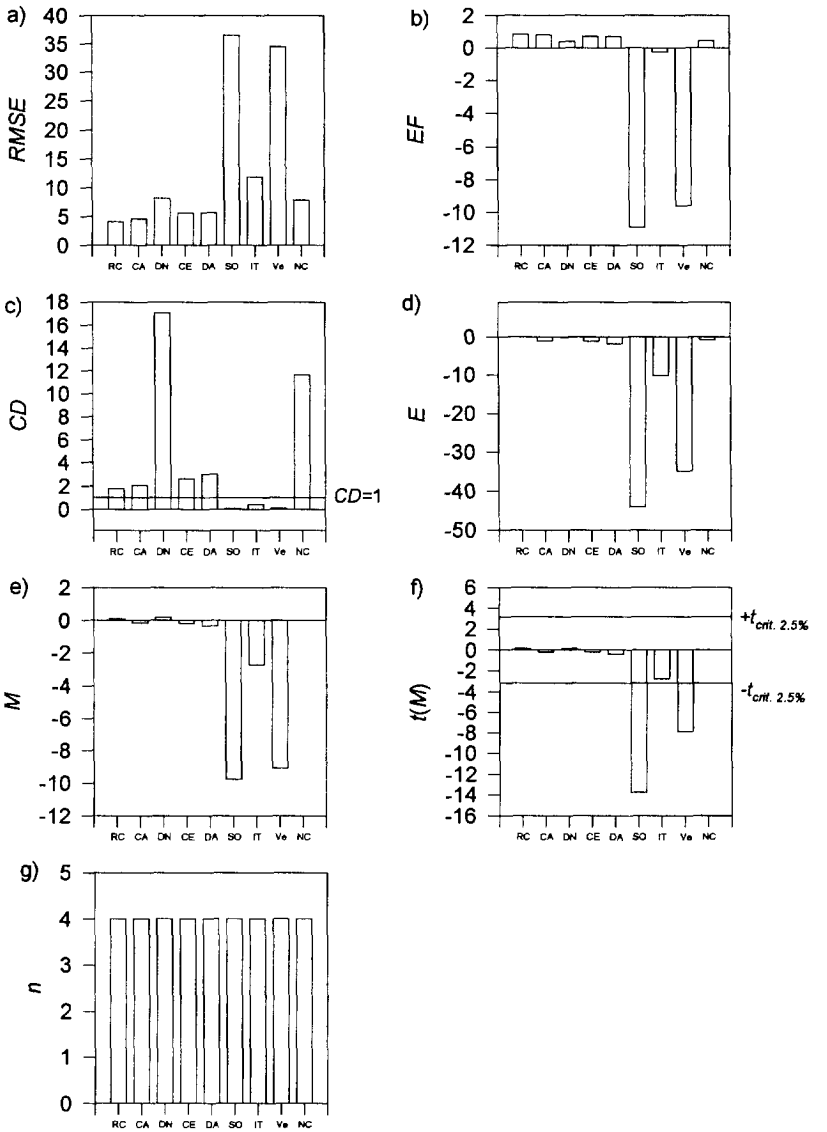
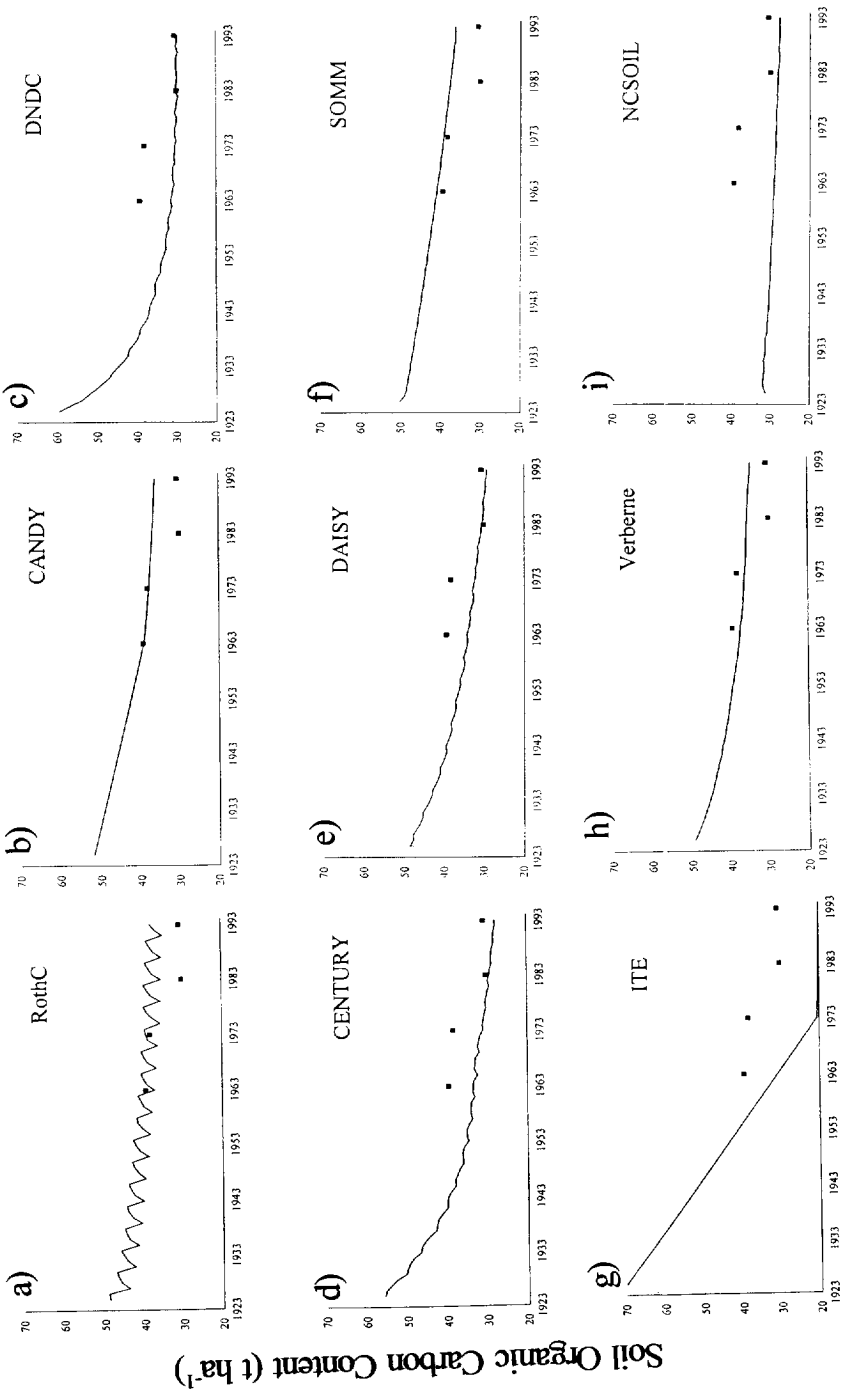


Fig. 22. Graphical representation of statistics describing the performance of models in simulating the Waite wheat/oats rotation. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3.

Fig. 23. Measured and simulated values of total organic carbon in the top 20 cm of soil for the Waite wheat–oats–grass-pasture–fallow rotation assuming a soil bulk density of 1.35 g cm^{-3} for (a) RothC, (b) CANDY, (c) DNDC, (d) CENTURY, (e) DAISY, (f) SOMM, (g) ITE (note: simulated values for 1925, 1973 and 1993 only), (h) Verberne and (i) NCSOIL. ■ shows measured values; simulated values shown as a line.


a)
RothC
b)
CANDY
c)
DNDC
d)
CENTURY
e)
DAISY
f)
SOMM
g)
ITE
h)
Verberne
i)
NCSOIL
Soil Organic Carbon Content (t ha-1)
Years

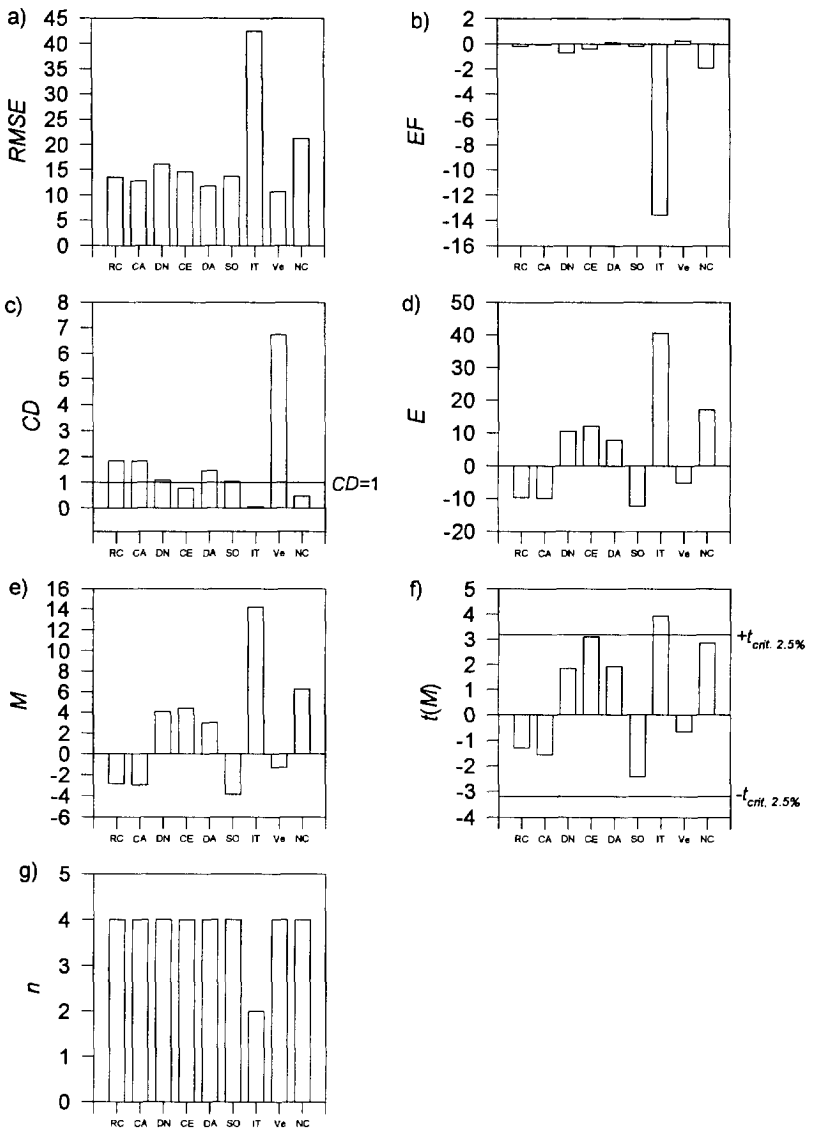


Fig. 24. Graphical representation of statistics describing the performance of models in simulating the Waite wheat–oats–grass-pasture–fallow rotation. Depicted above are the following statistics: (a) root mean square error (RMSE), (b) modelling efficiency (EF), (c) coefficient of determination (CD), (d) relative error (E), (e) mean difference (M), (f) t value for M ($t(M)$) with critical 2.5% levels shown, and (g) number of paired values, n. Abbreviations as for Fig. 3.

generalisations to be made about how well models might perform predictively, but they do provide the only information on the potential predictive ability of the models gained during this exercise.

3.2. The performance of each SOM model across all core long-term datasets

Having discussed in some detail the performance of the models for each individual dataset, in this section we attempt to summarise the performance of each model across all datasets. Table 2 summarises the frequency with which models met criteria of good model performance across all simulations.

Only four datasets allowed comparison of RMSE values with values of $RMSE_{95\%}$. Of the six models that simulated all four of these, only DAISY produced RMSE values below $RMSE_{95\%}$ for all. In terms of EF, where a positive value indicates that the simulation describes the measured data as well as the mean of the measurements, DAISY again performed best by producing a positive EF value in 60% of simulations. CD is a measure of the variance in the simulation compared to the measured mean where values of 1 or above indicate that the deviation from the measured mean among the simulated values is less than the deviation of the measured values from the measured mean. CANDY and DAISY performed best with a CD value greater than 1 on 80% of occasions.

In terms of bias in the simulations, E values could only be compared with $E_{95\%}$ values for four datasets. Of the six models that simulated all four of these, CANDY, DNDC, CENTURY and DAISY were within the 95% confidence interval of the data on all occasions. Another measure of biased error is M where low bias is shown by an M value not significantly different from zero. CENTURY showed the lowest overall bias in that it produced no M values significantly different from zero.

There are two approaches to compare an individual statistic, such as RMSE, across all datasets. One approach is to take a simple mean of the values for each dataset. This approach places equal weight on all values regardless of how many datapoints that value is based upon. A second approach is to calculate statistics that describe model fit for all datapoints simulated. With this approach, statistics generated from simulations of large datasets are weighted accordingly. CD and EF cannot be compared across datasets. Fig. 25 provides a graphical representation of the statistics of model performance across all datasets for which they attempted a simulation.

When simulated values are compared to measured values across the complete range of datapoints simulated, all models produce a very high correlation coefficient, r (Fig. 25f). Had these simulated values been obtained without tuning the model using measured data (which is not generally the case), the t values associated with these r values would have been highly significant (Fig. 25g). It should be noted, however, that the variation between sites is far greater than changes at individual sites over the course of each experiment, so even if models were provided only with initial soil carbon values, a strong positive correlation would be expected.

In terms of consistent bias in the simulations, only three models, RothC,

Table 2
Frequency that models met the criteria of good model performance across all simulations

Statistics of good model fit	RothC	CANDY	DNDC	CENTURY	DAISY	SOMM	ITE	Verberne	NCSOIL
Number of datasets	11	10	10	11	10	10	6	6	7
Number of datasets where RMSE < $RMSE_{95\%}$	2/4	3/4	3/4	3/4	4/4	1/4	0/0	0/0	2/2
Percentage of those datasets simulated where RMSE < $RMSE_{95\%}$	50	75	75	75	100	25	*	*	100
Number of datasets where EF is positive	2/11	3/10	3/10	5/11	6/10	1/10	1/6	1/6	3/7
Percentage of those datasets simulated where EF is positive	18	30	30	45	60	10	16	16	43
Number of datasets where CD > 1	8/11	8/10	5/10	7/11	8/10	3/10	0/6	3/6	5/7
Percentage of those datasets simulated where CD > 1	73	80	50	64	80	30	0	50	71
Number of datasets where $E < E_{95\%}$	3/4	4/4	4/4	4/4	4/4	1/4	0/0	0/0	2/2
Percentage of those datasets simulated where $E < E_{95\%}$	75	100	100	100	100	25	*	*	100
Number of datasets where M was not significant	9/11	8/10	5/10	11/11	8/10	3/10	3/6	4/6	5/7
Percentage of datasets simulated where M was not significant	82	80	50	100	80	30	50	67	71

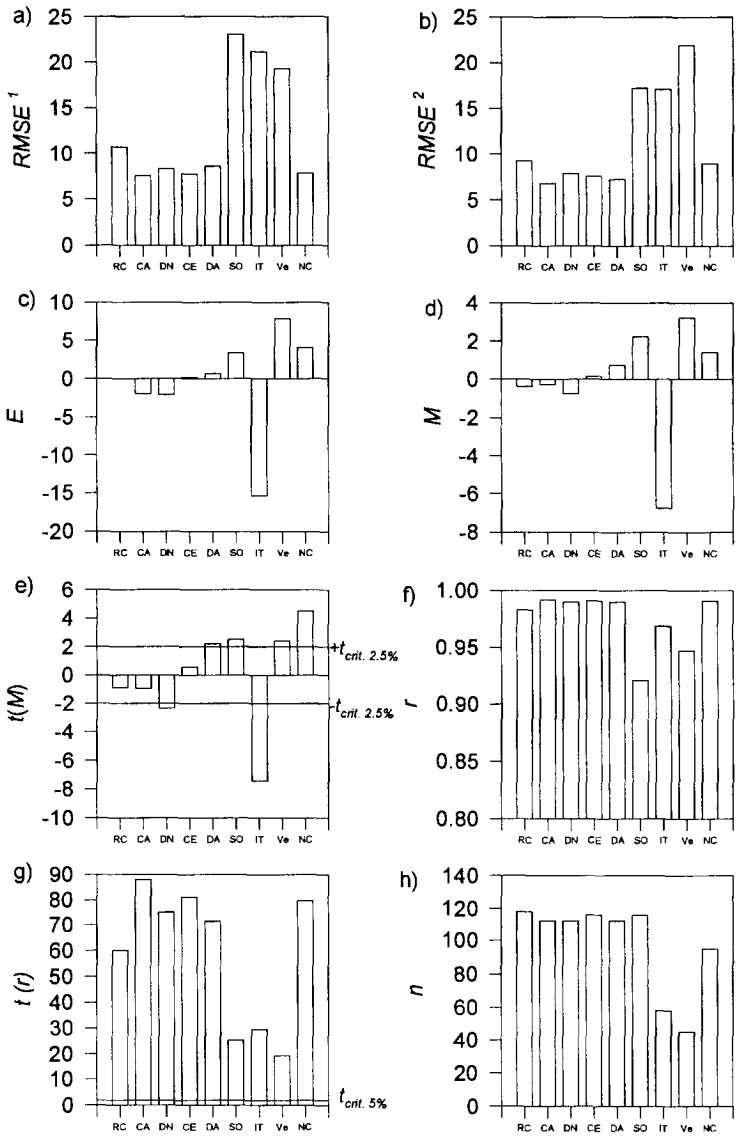


Fig. 25. Graphical representation of statistics describing the performance of all models across all datasets for which they attempted a simulation. Depicted above are the following statistics: (a) root mean square error (RMSE ^{1}) [calculated for all points simulated], (b) root mean square error (RMSE ^{2}) [the mean of all individual RMSE values produced for each dataset], (c) relative error (E), (d) mean difference (M), (e) t value for M ($t(M)$) with critical 2.5% levels shown, (f) correlation coefficient (r), (g) t value for r with critical 5% level shown, and (h) number of paired values, n. Abbreviations as for Fig. 3.

CANDY and CENTURY, do not show a significant bias over all simulations as measured by t for M (Fig. 25e). Values of E (Fig. 25c) and M (Fig. 25d) rank model bias similarly. DNDC tends toward a slight over-prediction of soil

organic carbon content over all datasets whilst DAISY, SOMM and Verberne tend to slightly under-predict soil organic carbon content. The under-prediction of NCSOIL and the over-prediction of ITE are slightly more pronounced.

For total error, RMSE values follow a similar pattern regardless of how they were derived, whether as a single value for all points simulated (Fig. 25a), or as a simple mean of all individual RMSE values (Fig. 25b). Values fall into two categories, i.e. (a) those with mean RMSE values between about 6.5 and 10 (CANDY, CENTURY, NCSOIL, DNDC, DAISY and RothC), and (b) those with mean RMSE values nearly twice as big, between 17 and 22 (Verberne, ITE and SOMM; see Fig. 25a and b). A two-sampled, two-tailed t test of simple mean RMSE values reveals that there is no significant difference between any of the models within in group (a) at the 5% level and no significant difference between any models within group (b). However, there was a significant difference between group (a) and group (b). The probabilities associated with the t distributions are given in Table 3.

Based on these figures, the group (a) models with low RMSE values can be said to more accurately simulate the measured values than the models in group (b). Attempting to further rank the models, based upon smaller differences in RMSE, e.g. between 7.15 and 7.35, is not justified since models within the two groups are not statistically significantly different from each other.

Since some datasets may be more difficult to simulate than others, i.e. mean error values for some datasets are higher than for others (e.g. mean RMSE values range between 5.65 for Bad Lauchstädt - high fertilization and 15.92 for Geescroft Wilderness; see Figs. 3 and 24) the analysis was rerun using only those datapoints for which all models simulated a value. This restricted comparison, using data drawn from simulations of Park Grass (both treatments), Prague–Ruzyně (no fertilization) and Waite (both rotations) showed that the pattern among all statistics remained unaltered. This suggests that there is no bias in the results due to different subsets of the data being modelled, i.e. no

Table 3
Probability associated with a two-sampled, two-tailed t test of root mean square error, RMSE, values produced by each model; values significant at the 5% level are shown in bold

	RothC	CANDY	DNDC	CENTURY	DAISY	SOMM	ITE	Verberne
CANDY	0.191							
DNDC	0.494	0.466						
CENTURY	0.353	0.565	0.799					
DAISY	0.269	0.737	0.390	0.326				
SOMM	**0.033**	**0.006**	**0.036**	**0.018**	**0.035**			
ITE	**0.008**	**0.000**	**0.002**	**0.000**	**0.000**	0.799		
Verberne	**0.021**	**0.008**	**0.021**	**0.007**	**0.003**	0.122	0.160	
NCSOIL	0.913	0.309	0.556	0.313	0.170	0.289	0.231	0.083

significant sampling error, so the statistics given in Fig. 25 can confidently be used to compare to overall performance of models.

4. General discussion and conclusions

4.1. Differences in model performance

In the previous sections, we have discussed the performance of each model in simulating each dataset and we have attempted to summarise the overall performance of each model across all datasets. In terms of overall performance, models fell into two groups with one group (SOMM, ITE and Verberne) performing significantly less well than the other (RothC, CANDY, DNDC, CENTURY, DAISY and NCSOIL).

One of the main differences between the two groups of models was the level of site-specific calibration used. Such model calibration was permitted in this model evaluation but was not used by all models. Differences in the level of calibration is likely to be partly responsible for the differences between the two groups of models. The other main difference between the two groups was that all those in the group with the poorer performance are less developed for application to land-uses different from those for which they were developed. ITE and SOMM were developed as forestry/grassland models and when attempting to simulate arable crops were forced to do so as if they were grass. The same is true of the Verberne model but the SOM submodel has been applied more extensively to arable systems elsewhere (Whitmore et al., 1997). SOMM and ITE performed better on forestry datasets than on others, but no better than other more generally applicable models. A problem particular to the version of the Verberne model tested in this exercise was its coupling to a complex model of dynamic water movement. Whitmore et al. (1997) have shown that when decoupled from the complex water submodel, the SOM submodel of Verberne performs significantly better than it did in this exercise.

Smaller differences between models may partly be accounted for by the adoption of different starting conditions. Not all models used the same starting levels of organic carbon. Even if the trends predicted by different models were similar, starting values might render some simulations closer to, or further from, the body of measured values.

In order to elucidate further factors contributing to differences in performance of the models, we now discuss the performance of each model in turn. RothC performed simulations on all land-uses. It produced low errors for all datasets except for the Park Grass organic manure treatment, the Prague–Ruzyně high fertilization treatment and the Tamworth fallow rotation. In all three cases, the

problem was that RothC was initialised to start with the same conditions for all treatments being modelled at a particular site. This was not the case for all models, some of which used different starting criteria for each treatment. Thus the initial carbon content of the three plots being modelled on Park Grass were set to 75.1 t C ha^{-1} in 1856, to 32.2 t C ha^{-1} in the two treatments of the Prague–Ruzyně experiment in 1965 and to 23.1 t C ha^{-1} in the three rotations of the Tamworth experiment in 1966. Had this requirement been relaxed and the model allowed to run with (slightly) different *initial* carbon contents for the different treatments, the fit for the Prague–Ruzyně high fertilization treatment and the Tamworth fallow rotation could have been much improved. With the Park Grass organic manure treatment, little improvement would have resulted. As well as generating organic carbon, RothC also generates values for soil microbial biomass carbon and for the effects of the thermonuclear tests of 35 years ago on soil radiocarbon. Both soil biomass carbon and radiocarbon can be measured if the appropriate soil samples are available. Although not tested in the present paper, such measurements can provide an additional check on the workings of RothC, or for that matter, of any carbon turnover model similarly equipped.

CANDY performed simulations on all land-uses though simulations on woodland included organic matter in the litter layer and could not be compared to the measured data on mineral soil at those sites. It produced consistently low errors for all datasets it attempted.

DNDC performed simulations on grassland and arable land but did not attempt to simulate the woodland datasets. It produced consistently low errors for all datasets it attempted. Three of the sites simulated by DNDC showed rapid changes in simulated soil organic carbon content. Two were the Waite Rotations, where native grassland was converted to cultivated agriculture at the beginning of the study. Both sites lost about one-third to one-half of their soil organic carbon by the time of the first observation, 40 years after cultivation. In the DNDC simulations, most of this loss occurred within the first 20 years of cultivation; sensitivity runs with DNDC showed that transitions to lower soil organic carbon generally occurred more quickly than increases (Li et al., 1994). The other site showing rapid changes was the Park Grass Manure Treatment plot, which displayed an unexplained sharp drop in soil organic carbon over the first half of the study, and a recovery to near initial values during the second half. DNDC simulations did not capture this variation, perhaps because of improper estimation of crop (hay) residue returned to the soil each year or possibly because the measured data were unreliable. All other sites had much smaller changes in soil organic carbon, both in field observations and in the DNDC simulations. This implies that soil organic carbon pools were in rough equilibrium with annual inputs and losses. This equilibrium would be the result of historic land-use and climate.

CENTURY performed simulations on all land-uses. It produced consistently

low errors for all datasets it attempted except for Geescroft Wilderness where it failed to simulate the continuing rise in soil organic carbon between 1965 and 1985. Century model performance was best for grass and crop systems, not surprisingly as Century was conceived as a grassland model and it is most widely tested for grass and crop systems. Due to an explicit linkage between C and N dynamics in the model, Century was able to simulate low- and high-N treatments within sites. At sites with low errors in both treatments (RMSE < 10), modelling efficiency was generally better in high-N treatments. This suggests that Century is better able to capture N-induced extremes than those caused by other factors, perhaps because low-level variations are caused by small-scale variation in soil texture or soil water, or weather patterns occurring on a shorter time-scale than simulated by Century. Though Century yielded a relatively high error (RMSE > 10) in simulating Geescroft Wilderness and failed to simulate the continuing rise in soil organic carbon between 1965 and 1985, modelling efficiency was positive (meaning that Century's simulation explained the soil organic carbon dynamics better than an arithmetic mean) since it captured at least part of the trend of increase. The difficulty in capturing the continuing rise is soil organic carbon may have been because the model does not automatically account for an effect of declining pH leading to a decreasing rate of decomposition. Century's forest submodel is less well tested than the grass and crop submodel, and changes are currently being made to improve the ability to capture litter layer dynamics (see Kelly et al., 1997 for a more complete discussion of the structural limitations of the Century forest submodel).

DAISY performed simulations on arable and grassland sites. However, no attempt was made to simulate the woodland sites, because currently DAISY cannot simulate soil water and temperature dynamics in a forest system and does not incorporate leaching of dissolved organic substances between soil layers, faunal litter incorporation or low pH effects on decomposition. The DAISY model produced consistently low errors for all datasets it attempted, except for slight underprediction of soil organic carbon content in the Bad Lauchstädt no fertilization and high fertilization treatments. The latter, however, was due to the initialization of the simulation using the first measured datapoint which may be anomalously low; like RothC, the initial carbon content of each treatment at a given site was made identical when DAISY was initialised (except for Bad Lauchstädt where the two treatments are very different in soil organic carbon content). Currently, DAISY does not incorporate any mechanism for rhizodeposition or root turnover. Through the DAISY simulations it became evident that the lack of sufficient experimental data on below-ground, plant-derived carbon inputs, especially for perennial crops, may be the key problem in improving simulations of long-term changes in soil organic carbon. The DAISY model was developed for agroecosystems and the good performance with the arable datasets was thus achieved with only few assumptions about such carbon inputs. However, with the perennial grasses, especially for the Park Grass datasets,

reasonably good simulations could only be achieved by making theoretical assumptions about large below-ground carbon input from the grass. Furthermore it was evident that carbon inputs must differ greatly between the different treatments in the Park Grass experiment. Future research in this area should thus have high priority, because such data are a prerequisite if soil organic matter models are to incorporate and improve the simulation of this very important mechanism of soil carbon input.

NCSOIL performed simulations on all land-uses but some simulations were for limited parts of the datasets. It produced consistently low errors for all datasets it attempted except for the Waite wheat–oats–pasture–fallow blind test where it under-predicted soil organic carbon content. Choice of the years selected for simulation (Rothamsted Park Grass and Tamworth datasets), and simulated results for the Waite wheat–oats–pasture–fallow blind test are discussed elsewhere in this issue (Molina et al., 1997).

SOMM attempted to simulate all land-uses. SOMM was developed as a forest soil model and it performed relatively well on the Geescroft Wilderness and Calhoun (woodland) datasets. It also performed reasonably on the Prague–Ruzyně arable dataset and the Waite wheat–oats–pasture–fallow blind comparison but performed less well for other datasets. Like RothC, the initial carbon content of each treatment at a given site was made identical, thus constraining it in a similar way to that described above for RothC. Unlike most other models (with the exception of ITE and Verberne), SOMM was not calibrated to each site during the evaluation.

ITE attempted to simulate all land-uses although it was designed for grassland and forest. It performed least well on the two grassland treatments and the Prague–Ruzyně arable treatments. Its errors were smaller for the Geescroft Wilderness natural woodland regeneration dataset than for other datasets. The performance of the ITE model(s) is discussed at length elsewhere (Arah et al., 1997) and their stretching to accommodate arable systems (through simulating arable crops as grass) was mentioned earlier. The only calibration used to run the ITE simulations was the setting-up of the initial conditions. Once this was done, the model ran and simulated plants grew and died, leaf and root litter was returned to the soil and transformed from one form of SOM to another, and respiration, mineralisation, immobilisation, nitrification, denitrification and leaching all took place, responding to input weather data and nothing else. The ITE models are thus predictive in the sense outlined earlier in Section 2.4 though they do not of course simulate the weather. This should be borne in mind when comparing the results of the ITE models with others reported here, in which one or more input parameters are adjusted to optimise the overall fit to the data.

Verberne attempted to simulate all land-uses. It performed relatively well for the Park Grass grassland treatments but poorly for Geescroft Wilderness. Despite a good performance on the Waite wheat–oats–pasture–fallow rotation

blind test, it performed poorly on the Waite wheat–fallow dataset. Its only other attempt to simulate an arable dataset (Prague–Ruzyně) was relatively poor. The Verberne model did not attempt a number of the datasets because its organic matter module had recently been coupled with a sophisticated model of the dynamics of water movement in place of the simple, capacity model to which it was originally tied (Verberne et al., 1990). Since the user of the model for these exercises was not one of the authors of the model, de-coupling and a return to the original system within the time-frame of the workshop was not an option. The demands of this moisture model were such that many of the data-sets could not supply the necessary hydraulic properties of the soil nor the boundary conditions necessary to solve the equations of water flow. Sadly, therefore, the Verberne SOM turnover model could not be run against these datasets. Even so, the combined model fared relatively poorly with some of the datasets which did provide all of the necessary data. The reasons for this may also be the experimental link-up between water module and SOM dynamics. The relationship between soil moisture and decomposition rate has been established by many authors and the Verberne model uses relationships published by Van Keulen and Seligman (1987) who also simulated soil moisture content with a capacity model. Paradoxically, it may well be that the more realistic soil moisture model employed in these simulations revealed a flaw in the rates of decomposition in the Verberne model that compensated for a flaw in the original, simple moisture model. Simulations carried out by Whitmore et al. (1997) suggest that the Verberne model coupled with a capacity model could produce improved simulations. Indeed, since the structure of the Verberne model in terms of pools and flows of organic matter is not dissimilar from others in this series it would be strange if it could not. It is likely that the rates of decomposition of some of the organic matter pools in the Verberne model may need to be revised before it can be widely used with its newly coupled moisture module.

4.2. Conclusions and priorities for future research

This exercise has provided useful information for assessing which models are most suitable for simulating soil organic carbon dynamics in a given environment. Furthermore, it has identified a number of areas that require further research. One such area is the linking of SOM models to other, more complex, submodels. If models are to be used to examine the effects of global change on whole ecosystems in the future, such couplings to soil water and nitrogen models and to plant growth models will be necessary, but this coupling can result in problems. The case of the Verberne model demonstrates this point in that a SOM model which can simulate experimental data accurately (see Whitmore et al., 1997) failed to perform well in this exercise when linked to a sophisticated experimental water submodel. In this case, the performance of the

SOM submodel was constrained by the lack of adequate data on soil water parameters.

The coupling of SOM models to other submodels also led to less accurate simulation of the soils data. With simple SOM models such as RothC, total carbon inputs (e.g. from plant debris, rhizodeposition, and root turnover) are estimated as a single annual figure and are fed into the model as monthly inputs (Coleman et al., 1997). CANDY is similar in this respect and carbon inputs can be adjusted to more accurately simulate the measured data. For other SOM models with a similar structure but which have physiologically based plant growth submodels attached (e.g. ITE), carbon returns to the soil are calculated in the model and, in this exercise, were not adjusted. The relatively poor performance of the ITE models, even in ecosystems for which they were developed, suggests that the coupling of SOM and physiologically based plant growth submodels introduces a greater degree of error and uncertainty into the models. It is noteworthy that the coupling of simpler plant growth submodels to SOM models in DAISY (arable crops and grassland; Mueller et al., 1996), DNDC (arable crops and grassland; Li, 1996) and CENTURY (grasslands, arable crops, forest and savannah; Parton, 1996b) did not result in poor model performance: there appears to be some benefit in simplified approaches. Some form of coupling between SOM and plant growth modules will be important in models designed to simulate whole ecosystems. When developing such models in the future, the level of detail at which plant growth is simulated will need careful consideration.

Related to the problems of decreased accuracy when coupling SOM and physiologically based plant growth models is the problem of greater ecosystem specificity that is introduced. Simple, generic SOM models such as CANDY, RothC and NCSOIL showed good performance under a range of land-uses whereas those models with physiologically based plant growth sub-models were more limited to the landuses for which the plant growth submodels were parameterised. This is demonstrated by the application of models outside the ecosystem for which they were developed, e.g. application of the grassland/forestry models (ITE, Arah, 1996; Verbenrne, Klein-Gunnewiek, 1996) to arable ecosystems in which crops were simulated as if they were grass; a situation that inevitably introduced error into simulations. The problem extends to models with simple descriptions of plant growth in that DNDC and DAISY could not be applied to woodland ecosystems which are very different to the environments for which they were developed.

Many current SOM models do not explicitly account for changes in pH. Though some models were able to simulate long-term changes in SOM occurring concurrently with changes in pH, CENTURY's failure to simulate the final measured datapoint on the Geescroft Wilderness site was attributed to the model not automatically accounting for the effect of declining pH on decomposition rate despite some studies showing that pH affects only short-term decomposi-

tion. Since changes in pH may well occur in some ecosystems under global environmental change, further research and an explicit description linking pH and decomposition rate may be required in future models.

The scarcity of measurements of below-ground carbon inputs, either from root turnover or from rhizodeposition, is an important finding of this exercise. Models in which inputs to the soil are entered as total carbon (e.g. RothC and CANDY) are less affected by this lack of data than are models in which these inputs are entered separately from above-ground inputs (e.g. DAISY), but in either case, a better knowledge of below ground inputs would improve model performance. If SOM models are to be coupled to plant growth and other submodels in the future, more research into, and measurements of, below-ground carbon inputs will be needed to parameterise and run these models. Similarly, soil microbial biomass carbon and radiocarbon can also be measured and such measurements can provide an additional check on the workings of SOM models.

This exercise has provided a vital first step in evaluating the potential of various models for use in predicting the effects of global environmental change. Only limited information has been gained, however, about the ability of the models to predict the effects of changes in land-use (e.g. the Geescroft Wilderness Experiment) or of their sensitivity to temperature (e.g. a limited range of climatic conditions from the cool-temperate European datasets to the warm-temperate/sub-tropical American and Australian datasets). No information was gained about the sensitivity of models to changes in atmospheric carbon dioxide concentration. Before models can confidently be used to predict the effects of global environmental change, further evaluation is required to elucidate how models deal with these factors.

The exercise described here was able to evaluate how well models simulate long-term SOM dynamics in a range of ecosystems. It will be necessary in future evaluation exercises to determine the sensitivity of models to changing temperatures and atmospheric carbon dioxide concentrations. This will entail a different kind of evaluation, using a different kind of data, in order to elucidate the performance of models at the level of individual short-term processes. For these purposes, an evaluation using data such as surface CO_2 flux, soil water content, soil temperature, and detailed measurements of above- and below-ground C and N inputs etc. would be necessary as this would allow alternative hypotheses of carbon transfer to be distinguished.

Another key area that must be dealt with before models can be used in a truly predictive manner is their ability to predict soil organic carbon dynamics without site specific calibration. In this exercise, for only one dataset were models forced to run without allowing model calibration. Model calibration for other datasets was permitted because many of the models were to be used in situations (land-uses, treatments and climatic conditions) which they had not previously encountered. For use at larger scales, however, and for predicting future changes, such calibration data will not exist. A major obstacle to the confident

predictive use of models in the science of global environmental change is the need for such model calibration.

Acknowledgements

We are very grateful to the dataholders at the core sites who provided the data used in this model evaluation and comparison exercise, specifically Prof. Dr. Martin Körschens, Dr. Annett Müller, Dr. Dan Richter, Mr. Paul Poulton, Dr. Jan Klír, Mr. Graham Crocker, Dr. Ian Holford and Dr. Peter Grace. We are also grateful to Dr. Moira Mugglestone of the Statistics Department of IACR–Rothamsted for her help, and to Dr. Mac Post, Dr. Ted Elliott and Dr. Thomas Barnwell for comments on the manuscript. We thank NATO for support for the Advanced Research Workshop at which this exercise began. P.S. was funded by the Terrestrial Initiative in Global Environmental Research (TIGER) programme of the U.K. Natural Environment Research Council (NERC). IACR receives grant-aided support from the U.K. Biotechnology and Biological Science Research Council.

Appendix A

A.1. Contents of Appendix

Tables A1–A3: General properties of each core experimental site.

Tables A4–A8: Soil details for acg core site.

Tables A9–A13: Plant cover, system inputs and land management details for each site.

Table A14: Meteorological data at each site.

Table A1
The core datasets and treatments within each site used for the model evaluation exercise

Experiment	Country	Land-use	Duration	Crop/plant cover and treatments
Bad Lauchstädt Static Fertilizer Experiment (Bad Lauchstädt)	Germany	Arable	93 years	Sugar beet–spring barley–potatoes–winter wheat with: (1) organic manure plus NPK fertilizer, (2) no fertilizer
Calhoun Experimental Forest (Calhoun)	USA	Forestry	38 years	Planted loblolly pine with no fertilization
Rothamsted Park Grass (Park Grass)	UK	Grassland	139 years	Permanent grassland with: (1) no fertilizer, (2) organic manure (1905 onwards)
Prague–Ruzyně Plant Nutrition and Fertilization Management Experiment (Prague)	Czech Republic	Arable	40 years	Sugar beet, spring wheat since 1966 with: (1) organic manure plus inorganic fertilizer, (2) no fertilizer
Tamworth Legume Cereal Rotation on Black Earth (Tamworth)	Australia	Arable	29 years	(1) Lucerne/clover and cereal with urea or superphosphate, (2) fallow /cereal with urea or superphosphate
Rothamsted Geescroft Wilderness (Geescroft)	UK	Woodland	112 years	Naturally regenerated woodland with no fertilization
Waite Permanent Rotation Trial (Waite)	Australia	Arable	70 years	(1) Wheat–fallow with superphosphate, (2) wheat–oats–pasture–fallow with superphosphate

Table A2
Further details from each core site

Experiment	Latitude, Longitude	Climate region[a]	Mean annual rainfall (mm)	Mean annual air temperature (°C)	Estimated annual atmospheric nitrogen deposition (kg ha^{-1})
Bad Lauchstädt	51°24′N, 11°53′E	CT	484	8.7	50 (wet + dry)
Calhoun	34°30′N, 81°30′W	WTST	1250	17.0	10 (wet only—dry unknown)
Park Grass	51°49′N, 0°21′W	CT	728	9.1	40 (wet + dry) since 1900; 20 (wet + dry) before 1900.
Prague	50°05′N, 14°20′E	CT/CTB	523	8.0	50 (wet + dry)
Tamworth	31°06′S, 150°56′E	WTST	676	17.5	10 (wet + dry)
Geescroft	51°49′N, 0°21′W	CT	728	9.1	40 (wet + dry) since 1900; 20 (wet + dry) before 1900.
Waite	34°58′S, 138°38′E	WTST	604	16.8	2 (wet + dry)

[a] Note: CT = cool temperate, CTB = cool temperate boreal, WTST = warm-temperate–subtropical.

Table A3
Previous history, plot size, experimental design and slope of the core sites

Experiment	Previous history of the site	Size of plots (m × m)	Design: plot distribution	Slope of the site
Bad Lauchstädt	Originally grassland but site had been under arable crops for a number of years before the experiment began	26.5 × 10	Systematic	None
Calhoun	Before 1800: Eastern North American Deciduous Forest. Between 1800 and 1957, arable land under cotton, maize, wheat and pasture. 1957 planted with loblolly pine.	4 of 0.03 ha and 4 of 0.06 ha	Randomized block design	0–2% and uniform
Park Grass	Probably grazed grassland for several centuries before experiment began	Between 11.3 × 6.6 and 19.3 × 25.2	Systematic	None
Prague	Several centuries under cultivation	12 × 12	Randomized block design	< 1% and uniform
Tamworth	Originally dry sclerophyll forest with *Eucalyptus albans* dominant. Cultivated for around 100 years. Before experiment began used for arable crops and grazing. Badly eroded by 1958.	30.5 × 11.3	Randomized block design	1% and uniform
Geescroft	Arable in 1623. Arable crops 1847–1878. Fallow 1879–1882. Clover 1883–1885. Last cultivated 1883, last cut in 1885. Fenced off and regeneration allowed.	No plots, total area = 1.3 ha	No plots	< 1% and uniform
Waite	Grassland before 1925, arable rotations since then.	60 × 8.5	Randomized block design	Concave slope; % unknown

Table A4
Details of soil type and texture at each core site

Experiment	Soil type	Soil horizons (all cm)	Particle size distribution (μm)
Bad Lauchstädt	Haplic phaeozem	Ap: 0–25; Ah: 25–40; $Ah/C_{(k)}$: 40–60; $C_{(k)}$: 60–125	Top 30 cm: < 2 = 21%; 2–63 = 67.8%; > 63 = 11.2%
Calhoun	Appling series:	O: 10–0; A: 0–7.5;	A and E: < 2 = 15.4%; 2–60 = 67.6%; > 60 = 17%
	thermic, koalinitic, clayey Typic Kanhapludult	E: 7.5–34; BE: 34–58	BE: < 2 = 39.3%; 2–60 = 22.5%; > 60 = 38.2%
		B: 58–135;	B: < 2 = 57%; 2–60 = 11%; > 60 = 32%
		BC/C: 135– > 800	
Park Grass	Batcombe series;	Ah: 0–20; Eb: 20–45	Top 20 cm:
	Aquic Palendalf.	2Bt(g)1; 45–69	< 2 = 23%; 2–60 = 58%
	Chromic Luvisol	2Bt(g)2: 69–80	> 60 = 19%
Prague	Medudalf	A – Umbric: 28–33	A: < 1 = 27%; 1–10 = 31%; 10–50 = 28%; 50–100 = 7%; > 100 = 7%
		B1 – Cambic: 8–10 cm	B1: < 1 = 40%; 1–10 = 34%; 10–50 = 21%; 50–100 = 4%; > 100 = 1%
		B2 – Cambic: 8–10 cm	B2: < 1 = 49%; 1–10 = 29%; 10–50 = 15%; 50–100 = 5%; > 100 = 2%
		C – Highly weathered chalk	C: < 1 = 50%; 1–10 = 25%; 10–50 = 15%; 50–100 = 6%; > 100 = 4%
Tamworth	Pellic Vertisol (black earth)	1: 0–35; 2: 35–75; 3: 75–105 4: 105–160	Unavailable
Geescroft	Batcombe series;	L (litter): 2–0; A: 0–4; Eb:	< 2 = 21%; 2–20 = 23%; 20–2000 = 49%; > 2000 = 4%
	Aquic Palendalf	4–28; Blt: 28–46	
	Chromic Luvisol	B2t(g): 46–60 +	
Waite	Rhodoxeralf	A1: 0–15; A2: 15–26;	A1: < 2 = 18%; 2–60 = 34%; > 60 = 48%
		B1: 26–35; B2: 35–80	A2: < 2 = 30%;2–60 = 30%; > 60 = 40%
		B2ca: 80–110; C: 110 +	B1: < 2 = 47%; 2–60 = 23%; > 60 = 30%
			B2: < 2 = 60%; 2–60 = 19%; > 60 = 21%

Table A5
Soil bulk density, root depth, pH and annual erosion losses at each core site

Experiment	Root depth (m)	Bulk density ($g\ cm^{-3}$)	pH	Annual erosion losses
Bad Lauchstädt	2	1.35	6.6	Zero
Calhoun	> 3	A and E: 1.52; BE: 1.44; B: 1.40–1.44; BC: 1.42	Unavailable	< 5 kg ha^{-1}
Park Grass	> 1	1.20	4.6 on FYM plot 4.8 on nil inputs plot	Zero
Prague	> 2	1.30–1.45	6.5–7.1	No data
Tamworth	2	1.4	8.4	No data
Geescroft	> 3	1.19 in 1883; 1.18 in 1904; 0.95 in 1965; 0.88 in 1985	4.2 now (7.0 in 1883)	Zero
Waite	> 2	1.35	6.2	No data

Table A6
Soil water characteristics at each site

Experiment	Plant available water (% v/v)	Cation exchange capacity (mmol kg^{-1})	Soil water content (%)	Wilting point (%)	Hydraulic conductivity (10^{-4} cm s^{-1})
Bad Lauchstädt	32.8 (0–30 cm); 27 (30–60 cm); 27 (60–200 cm)	150–300	8–24	12.4	4.5 (0–10 cm) 1.0 (10–26 cm)
Calhoun	25–35	100–500	23 in top 3 m	5–15	No data
Park Grass	13.5–15	200	2–40	No data	No data
Prague	30	220–330	20–30	No data	No data
Tamworth	15	No data	21–35	21	No data
Geescroft	13.5–15	150	2–40	No data	No data
Waite	29 (0–15 cm); 32 (15–26 cm); 49 (36–35 cm); 55 (35–80 cm)	80 (0–15 cm)	23	19 (0–15 cm) 23 (15–26 cm)	5 (profile)

Table A7
Soil organic matter characteristics at each site

Experiment	Current organic carbon content (%)	Initial % C in profile when experiment began	Soil C:N ratio	δC^{13} and δC^{14}	Exchangeable P (mg/kg)
Bad Lauchstädt	2.07 (0–30 cm)	< 2.0	12.2	No data	210
Calhoun	50.0 (10–0 layer) 1.40 (0–7.5 cm) 0.72 (7.5–34 cm) 0.68 (34–58 cm) 0.48 (58–135 cm)	No 10–0 cm layer 0.49 (0–7.5 cm) 0.32 (7.5–34 cm) 0.34 (34–58 cm)	No data	See [a]	No data
Park Grass	All for 0–23 cm:	3.38 (0–23 cm)	12	$\delta C^{13} = -27$ (all)	'Available P': FYM = 35; nil inputs = 3.4;
	FYM plot = 3.3	0.76 (23–46 cm)		δC^{14} (see Jenkinson et al. (1992))	
	Nil inputs = 3.0	0.37 (46–69 cm) 0.29 (69–91 cm) 0.23 (91–114 cm)			
Prague	1.2–1.4 (A) 0.33 (B1) 0.28 (B2) 0.15 (C)	1.17 (0–30 cm) 0.35 (30–50 cm)	10	No data	No data
Tamworth	2 (0–10 cm) 1.4 (35–45 cm) 0.6 (105–115 cm) 0.2 (160–170 cm)	1.12 (0–10 cm) 0.8 (35–45 cm) 0.32 (105–115 cm) 0.11(160–170 cm)	10	No data	10
Geescroft	2.7 (0–23 cm)	1.04 (0–23 cm)	9–15	$C^{13} = -27$ (all)	For 0–23 cm: tot. P = 460; organic P = 159; inorg. P = 301
		0.58 (23–46 cm)		δC^{14} (see Jenkinson et al. (1992))	
		0.49 (46–69 cm)			
Waite	1.47 (0–15 cm) 0.82 (15–26 cm) 0.70 (26–35 cm) 0.23 (35–80 cm)	2.75 (0–10 cm) 0.89 (10–20 cm)	11	No data	No data

Notes: [a] At Calhoun no δC^{13} values but δC^{14} values as follows (all 1989–1992): 10–0 cm (O1, O2, O3) = 152.2, 247.3, 309.8, respectively; 0–7.5 cm = 229.6; 7.5–15 cm = 97.2; 15–35 cm = 11.2; 35–60 cm = −92.8; 60–100 cm = −301.0; 110–135 cm = −365.5; 135–165 cm = −434.9; 200–300 cm = −434.0; 300–400 cm = −630.3; 500–600 cm = −448.9.

Table A8
Soil organic matter measurements that are made regularly at each site

Experiment	SOM measurements	Methods used	Regularity of measures
Bad Lauchstädt	Total carbon	Dry combustion [a]	Every year
	Soil biomass carbon	Substrate-induced respiration	Irregularly
	C in other OM fractions	Hot water soluble C – colorimetry [a]	Irregularly
	CO_2 evolution	Incubation and infra-red adsorption	Irregularly
	Total nitrogen	Conc. H_2SO_4 [b]	Every year
	N in other OM fractions	Kjeldahl [c]	Irregularly
Calhoun	Total carbon	CHN analyzer	1962, 1990
	CO_2 evolution	Incubation and infra-red adsorption	2 weekly
	Total nitrogen	CHN analyzer and Kjeldahl [c]	1962, 1968, 1972, 1977, 1982, 1990
Park Grass	Total carbon	Walkley–Black [c,d], combustion and Tinsley method [e]	1876, 1886, 1913, 1932, 1959, 1966, 1972, 1985, 1991
	Soil biomass carbon	Chloroform-fumigation-incubation method [f]	Very rarely
	CO_2 evolution	Patra et al. (1990) [g]	Very rarely
	Total nitrogen	Kjeldahl [c]	1876, 1932, 1959, 1966, 1972, 1985
	N in other OM fractions	Chloroform-fumigation-incubation method [f]	Very rarely
Prague	Total carbon	Wet combustion [c]	Every year
	CO_2 evolution	Laboratory incubation	Twice yearly
	Total nitrogen	Kjeldahl [c]	Every year
Tamworth	Oxidisable carbon	Walkley–Black [c,d]	Every 3 years
	Total nitrogen	Kjeldahl [c]	Every 2 years
Geescroft	Total carbon	Jenkinson (1965) [h]	1883, 1904, 1965, 1985
	Soil biomass carbon	Chloroform-fumigation-incubation method [f]	1970, 1985
	Total nitrogen	Kjeldahl [c]	1883, 1904, 1965, 1985
Waite	Total carbon	Dry combustion [c], Leco analyzer	1963, 1973, 1983, 1993
	Soil biomass carbon	C equivalent of Ninhydrin-reactive N	1993
	CO_2 evolution	Cores in situ and infra-red adsorption	1994
	Total nitrogen	Dry combustion	1993

Notes: [a] Deutsche Industrie-Norm, DIN H3; [b] Deutsch Industrie-Norm DIN H11; [c] see Black et al. (1965); [d] Walkley and Black (1934); [e] Tinsley (1950); [f] Jenkinson and Powlson (1976); [g] Patra et al. (1990); [h] Jenkinson (1965).

Table A9
Details of plant cover (crop) data available at each site and regularity of measurement

	Plant species present/crop	Yield	Total above ground dry matter	Total dry matter in offtake	C content of offtake	N content of offtake	Other measurements
Bad Lauchstädt	Yearly	Yearly	Irregularly	Irregularly	Irregularly	Irregularly	P, K and Mg in crop – irregularly
Calhoun	Data available	Not applicable	Every 5 years	Every 5 years	Estimated	Estimated	Above and below ground biomass; forest floor; with nutrient contents – 1990 and 1992
Park Grass	Regularly	2 × per year	2 × per year	2 × per year	No data	Every 20 years	Trace elements, heavy metals, atmospheric pollutants – recently
Prague	Yearly	Yearly	Yearly	Yearly	No data	Yearly	P, K, Ca and Mg in crop – yearly
Tamworth	Yearly	Yearly	Yearly	Yearly	No data	Yearly	P in crop – yearly
Geescroft	Data available	Not applicable	Estimates of above-ground standing trees	No data	No data	No data – once only	Heavy metals in trees and understory
Waite	Yearly	Yearly	Yearly	Yearly	No data	No data	None

Table A10
Details of plant cover and offtake required during the SOM modelling exercise

Experiment	Legumes (%)	% of yield that is dry matter	N content of offtake (%)
Bad Lauchstädt	0	Cereals 86; Sugar beet roots 25; Potatoes 24	No data
Calhoun	0	Not applicable	Not applicable
Park Grass	10 on nil inputs plot; 0 on others	100 (yield given dry)	0.57 on nil inputs plot; 0.67 on organic manured plot; 0.77 on sodium nitrate plot
Prague	22 (1956–65); 0 (1965 >)	Cereals 86; Sugar beet roots 25	Variable values
Tamworth	100 for lucerne and clover; 0 for others	Grains 90	Lucerne 3
		Pasture 100	Wheat grain 2
Geescroft	None	Not applicable	Not applicable
Waite	0	100 (yield given dry)	No data

Table A11
Details of residue returns to the system

Experiment	Annual amount of residue returned to the soil (kg ha^{-1}) or approximate % of above-ground dry matter returned	Residue parent material	N content of residue returned (kg ha^{-1})[a]
Bad Lauchstädt	No data	10 cm crop stubble and sugar beet tops ploughed in	No data
Calhoun	Carbon to O horizon: 2475 kg ha^{-1} as litterfall, 825 kg ha^{-1} as root, 85 kg ha^{-1} as dissolved organic carbon (DOC). Carbon to 0–35 m layer: 1650 kg ha^{-1} as litterfall, 290 kg ha^{-1} as DOC	Litterfall and roots from loblolly pine	26 in O horizon; 13 in 0–25 cm layer
Park Grass	No data for residue returns from ground cover; 2.24 t ha^{-1} y^{-1} chaffed wheat straw added 1856–1897 on plot receiving organic manure after 1905.	Grassland plant species[b]	No data
Prague	Estimated 1.5 t ha y^{-1} (low fertilizer) and 2 t ha y^{-1} (high fertilizer) for wheat and 0.5 t ha y^{-1} (low fertilizer) and 1 t ha y^{-1} (high fertilizer) for sugar beet.	10 cm crop stubble ploughed in	No data
Tamworth	30% of wheat and stubble returned 60% of sorghum stubble returned	Stubble burnt up to 1978; after stubble ploughed in	No data
Geescroft	No data	Leaf litter from broadleaf/ deciduous trees	No data
Waite	No data	Stubble ploughed in each year after light grazing. No straw returned deliberately	No data

Notes: [a]Ash content of residue required by one model (SOMM) – no data at any site; [b]see Jenkinson et al. (1994).

Table A12
Details of fertilizer inputs to the system

Experiment	Organic fertilizer type	Amount and frequency of organic fertilizer applied (t ha^{-1})	% dry matter of organic fertilizer applied	% C in dry matter of organic fertilizer applied	% N in dry matter of organic fertilizer applied	Amount and frequency of N in inorganic fertilizer applied (kg N ha^{-1})
Bad Lauchstädt	Cattle manure	30 (with NPK) on one plot; nil on other)	25	40	2.8	20–120 (with cattle manure) on one plot; nil on other)
Calhoun	None	Zero	Not applicable	Not applieable	Not applicable	None
Park Grass	Cattle manure and fishmeal	Organic manured plot only: 35 of cattle manure every 4 years and 0.9 of fishmeal every 4 years (between cattle manure applications)	Cattle manure 23	Cattle manure 40	Cattle manure 3	
			Fishmeal no data	Fishmeal 42	Fishmeal 6.5	
Prague	Cattle	22 (with NPK) every two years on one plot; nil on other)	23	45	2.3	12–150 (with cattle manure) on one plot; nil on the other)
Tamworth	None	Zero	Not applicable	Not applicable	Not applicable	Superphosphate only
Geescroft	None	Zero	Not applicable	Not applicable	Not applicable	None
Waite	None	Zero	Not applicable	Not applicable	Not applicable	Superphosphate only

Table A13
Tillage and irrigation information available at each site

Experiment	Tillage details	Irrigation details
Bad Lauchstädt	Ploughed to 20–30 cm yearly in September to November	Never irrigated
Calhoun	Never tilled	Never irrigated
Park Grass	Never tilled	Never irrigated
Prague	1956–85 each year: April to 7 cm with harrow; September to 10 cm with plough; November to 25 cm with plough. 1985–93: same except September ploughing replaced by 13 cm tillage in April with combinator	Never irrigated
Tamworth	Various tillage to between 5 and 20 cm throughout the year with rotary hoe, disc plough, tuned scarifier, sowing combine and mouldboard plough	Never irrigated
Geescroft	Never tilled	Never irrigated
Waite	Tilled to 8 cm every year in May with mouldboard plough until 1972; after 1972 with chisel plough	Never irrigated

Table A14
Meteorological data available at each site; frequency and year when electronic records begin

Experiment	Rainfall	Average temperature	Soil temperature	Global radiation	Sun hours	Wind-speed	Evaporation over water	Evaporation over grass	Relative humidity	Vapour pressure	Dew point
Bad Lauchstädt	Daily 1956	Daily 1956	None	Daily 1956	Daily 1956	None	Daily 1956	None	Daily 1956	Daily 1956	Daily 1956
Calhoun	Daily 1963	Daily 1963	None	Daily 1963	None	None	None	None	None	None	None
Park Grass	Daily 1856	Daily 1880	Daily 1959	Daily 1969	Daily 1891	Daily 1959	Daily 1959	Daily 1959	Daily 1959	Daily 1959	Daily 1959
Prague	Daily 1961	Daily 1961	Daily 1976	Daily 1984	Daily 1961	Daily 1961	Daily	None	None 1961	Daily	None
Tamworth	Daily	Daily	Daily	None	Daily	Daily	Daily	None	Daily	None	Daily
Geescroft	Daily 1883	Daily 1883	Daily 1959	Daily 1969	Daily 1891	Daily 1959	Daily 1959	Daily 1959	Daily 1959	Daily 1959	Daily 1959
Waite	Daily 1926	Daily 1966	None	Daily 1966	Daily 1966	Daily 1966	Daily 1966	None	None	None	None

References

Addiscott, T.M., Smith, J.U., Bradbury, N.J., 1995. Critical evaluation of models and their parameters. J. Environ. Qual. 24, 803–807.

Addiscott, T.M., Whitmore, A.P., 1987. Computer simulation of changes in soil mineral nitrogen and crop nitrogen during autumn, winter and spring. J. Agric. Sci., Camb. 109, 141–157.

Arah, J.R.M., 1996. The soil submodel of the ITE (Edinburgh) Forest and Hurley Pasture Models. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 225–230.

Arah, J.R.M., Thornley, J.H.M, Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the ITE (Edinburgh) Forest and Hurley pasture ecosystem models. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 61–74, this issue.

Black, C.A., Evans, D.D., White, J.L., Ensminger, L.E., Clark, F.E. (Eds.), 1965. Methods of Soil Analysis. Agronomy Number 9, Parts 1 and 2, American Society of Agronomy, Madison, WI.

Chatfield, C. (1983) Statistics for Technology. Chapman and Hall, London, 3rd ed., 381 pp.

Chertov, O.G., 1990. SPECOM – a single tree model of pine stand/raw humus soil ecosystem. Ecol. Modelling 50, 107–132.

Chertov, O.G., Komarov, A.S., 1996. SOMM—a model of soil organic matter and nitrogen dynamics in terrestrial ecosystems. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 231–236.

Chertov, O.G., Komarov, A.S., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the SOMM model of the humus types. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 121–135, this issue.

Clay, D.E., Clapp, C.E., Molina, J.A.E., Linden, D.R., 1985. Nitrogen-tillage-residue management, I. Simulating soil and plant behavior by the model NCSWAP. Plant Soil 84, 67–77.

Coleman, K., Jenkinson, D.S., 1996. RothC-26.3—A model for the turnover of carbon in soil. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 237–246.

Coleman, K., Jenkinson, D.S., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using RothC-26.3. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 29–44, this issue.

Crocker, G.J., Holford, I.C.R., 1996. The Tamworth legume/cereal rotation. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 313–318.

de Willigen, P., 1991. Nitrogen turnover in the soil–crop system; comparison of fourteen simulation models. Fert. Res. 27, 141–149.

Franko, U., 1996. Modelling approaches of soil organic matter turnover within the CANDY system. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 247–254.

Franko, U., Oelschlägel, B., Schenk, S., 1996. Simulation of temperature-, water-, and nitrogen-dynamics using the model CANDY. Ecol. Modelling 81, 213–222.

Franko, U., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the CANDY model. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 109–120, this issue.

Grace, P.R., 1996. The Waite Permanent Rotation Trial. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 335–340.

Hansen, S., Jensen, H.E., Nielsen, N.E., Svendsen, H., 1991. Simulation of nitrogen dynamics and biomass production in winter wheat using the Danish simulation model DAISY. Fert. Res. 27, 245–259.

Jenkinson, D.S., 1965. Studies on the decomposition of plant material in soil I. J. Soil Sci. 16, 104–115.

Jenkinson, D.S., 1990. The turnover of organic carbon and nitrogen in soil. Philos. Trans. R. Soc. London B 329, 361–368.

Jenkinson, D.S., Coleman, K., 1994. Calculating the annual input of organic matter to soil from measurements of total organic carbon and radiocarbon. Eur. J. Soil Sci. 45, 167–174.

Jenkinson, D.S., Powlson, D.S., 1976. The effects of biocidal treatments on metabolism in soil V. A method for measuring soil biomass. Soil Biol. Biochem. 8, 209–213.

Jenkinson, D.S., Hart, P.B.S., Rayner, J.H., Parry, L.C., 1987. Modelling the turnover of organic matter in long-term experiments. INTECOL Bull. 15, 1–8.

Jenkinson, D.S., Harkness, D.D., Vance, E.D., Adams, D.E., Harrison, A.F., 1992. Calculating net primary production and annual input of organic matter to soil from the amount and radiocarbon content of soil organic matter. Soil Biol. Biochem. 24, 295–308.

Jenkinson, D.S., Potts, J.M., Perry, J.N., Barnett, V., Coleman, K., Johnston, A.E., 1994. Trends in herbage yields over the last century on the Rothamsted Long-Term Continuous Hay Experiment. J. Agric. Sci., Camb. 122, 365–374.

Jensen, C., Stougaard, B., Østergaard, H.S., 1994a. Simulation of nitrogen dynamics in farmland areas of Denmark (1989–1993). Soil Use Manage. 10, 111–118.

Jensen, C., Stougaard, B., Olsen, P., 1994b. Simulation of water and nitrogen dynamics at three Danish locations by use of the DAISY model. Acta Agric. Scand., Sect. B 44, 75–83.

Jensen, L.S., Mueller, T., Nielsen, N.E., Hansen, S., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., 1997. Simulating trends in soil organic carbon in long-term experiments using the soil–plant–atmosphere model DAISY. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 5–28, this issue.

Kelly, R.H., Parton, W.J., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the century model. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 75–90, this issue.

Klein-Gunnewiek, H., 1996. Organic matter dynamics simulated by the 'Verberne'- model. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 255–262.

Klír, J., 1996. Long-term field experiment Praha–Ruzyně, Czech Republic. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 363–368.

Körschens, M., Müller, A., 1996. The Static Experiment Bad Lauchstädt, Germany. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 369–376.

Lengnick, L.L., Fox, R.H., 1994. Simulation by NCSWAP of seasonal nitrogen dynamics in corn, I. Soil nitrate. Agron. J. 86, 167–175.

Li, C., 1996. The DNDC model. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 263–268.

Li, C., Frolking, S., Frolking, T.A., 1992a. A model of nitrous oxide evolution from soil driven by rainfall events, 1. Model structure and sensitivity. J. Geophys. Res. 97, 9759–9776.

Li, C., Frolking, S., Frolking, T.A., 1992b. A model of nitrous oxide evolution from soil driven by rainfall events, 2. Model applications. J. Geophys. Res. 97, 9777–9783.

Li, C., Frolking, S., Harriss, R., 1994. Modelling carbon biogeochemistry in agricultural soils. Global Biogeochem. Cycles 8, 237–254.

Li, C., Frolking, S., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., 1997. Simulating trends in soil organic carbon in long-term experiments using the DNDC model. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 45–60, this issue.

Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. J. Contam. Hydrol. 7, 51–73.

McGill, W.B., 1996. Review and classification of ten soil organic matter (SOM) models. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 111–132.

Molina, J.A.E., 1996. Description of the model NCSOIL. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 269–274.

Molina, J.A.E., Richards, K., 1984. Simulation model of the nitrogen and carbon cycle in soil–water–plant system, NCSWAP; guide for the preparation of input data files and execution of NCSWAP. Soil Ser. 116. Dept. Soil Sci. University of Minnesota, St. Paul.

Molina, J.A.E., Clapp, C.E., Shaffer, M.J., Chichester, F.W., Larson, W.E., 1983. NCSOIL, a model of nitrogen and carbon transformations in soil: description, calibration, and behavior. Soil Sci. Soc. Am. J. 47, 85–91.

Molina, J.A.E., Crocker, G.J., Grace, P.R., Klír, J., Körschens, M., Poulton, P.R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the NCSOIL and NCSWAP models. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 91–107, this issue.

Mueller, T., Jensen, L.S., Hansen, S., Nielsen, N.E., 1996. Simulating soil carbon and nitrogen dynamics with the soil–plant–atmosphere system model DAISY. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 275–282.

Nicolardot, B., Molina, J.A.E., Allard, M.R., 1994. C and N fluxes between pools of soil organic matter: model calibration with long-term incubation data. Soil Biol. Biochem. 26, 235–243.

Otter-Nacke, S., Kuhlman, H., 1991. A comparison of the performance of N simulation models in the prediction of N min on farmers' fields in the spring. Fert. Res. 27, 341–347.

Parshotam, A., 1995. The Rothamsted soil–carbon turnover model—discrete to continuous form. Ecol. Modelling 86, 283–289.

Parton, W.J., 1996a. Ecosystem model comparison: science or fantasy world. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 133–142.

Parton, W.J., 1996b. The CENTURY model. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 283–293.

Parton, W.J., Rasmussen, P.E., 1994. Long-term effects of crop management in wheat–fallow: II. CENTURY model simulations. Soil Sci. Soc. Am. J. 58, 530–536.

Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987a. Analysis of factors controlling soil organic matter levels in great plains grasslands. Soil Sci. Soc. Am. J. 51, 1173–1179.

Parton, W.J., Stewart, J.W.B., Cole, C.V., 1987b. Dynamics of C, N, S, and P in grassland soils: a model. Biogeochemistry 5, 109–131.

Patra, D.D., Brookes, P.C., Coleman, K., Jenkinson, D.S., 1990. Seasonal changes of microbial biomass in an arable and a grassland soil which have been under uniform management for many years. Soil Biol. Biochem. 22, 739–742.

Poulton, P.R., 1996a. The Park Grass Experiment, 1856–1995. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 377–384.

Poulton, P.R., 1996b. Geescroft Wilderness, 1883–1995. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 385–390.

Powlson, D.S., Smith, P., Smith, J.U. (Eds.), 1996. Evaluation of Soil Organic Matter Models using Existing, Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, 429 pp.

Richter, D.D., Markewitz, D., 1996. Carbon changes during the growth of loblolly pine on formerly cultivated soil: the Calhoun Experimental Forest. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 397–408.

Ryan, M.G., Hunt, E.R., Jr., Ågren, G.I., Friend, A.D., Pulliam, W.M., Linder, S., McMurtrie, R.E., Aber, J.D., Rastetter, E.B., Raison, R.J., 1997a. Comparing models of ecosystem function for temperate conifer forests, I. Model description and validation. For SCOPE Volume: Effects of Climate Change on Forests and Grasslands (in press).

Ryan, M.G., Hunt, E.R. Jr, Ågren, G.I., Friend, A.D., Pulliam, W.M., Linder, S., McMurtrie, R.E., Aber, J.D., Rastetter, E.B., Raison, R.J., 1997b. Comparing models of ecosystem function for temperate conifer forests, II. Simulations of the effect of climate change. For SCOPE Volume: Effects of Climate Change on Forests and Grasslands (in press).

Schimel, D.S., Braswell, B.H. IB., Holland, E.A., McKeown, R., Ojima, D.S., Painter, T.H., Parton, W.J., Townsend, A.R., 1994. Climatic, edaphic, and biotic controls over storage and turnover of carbon in soils. Global Biogeochem. Cycles 8, 279–293.

Shao, Y., Henderson-Sellers, A., 1997. Soil moisture simulation workshop review. Global Planet. Change (in press).

Smith, J.U., Smith, P., Addiscott, T.M., 1996. Quantitative methods to evaluate and compare soil organic matter (SOM) models. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 181–200.

Smith, P., Powlson, D.S., Glendining, M.J., 1996a. Establishing a European GCTE Soil Organic Matter Network (SOMNET). In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models Using Existing Long-Term Datasets. NATO ASI Series I, Vol. 38, Springer-Verlag, Heidelberg, pp. 81–98.

Smith, P., Smith, J.U., Powlson, D.S.(Eds.), 1996b. Soil Organic Matter Network (SOMNET): 1996 Model and Experimental Metadata. GCTE Report 7, GCTE Focus 3 Office, Wallingford, 255 pp.

Styczen, M., Storm, B., 1993. Modelling of N-movement on catchment scale—a tool for analysis and decision making, I. Model description. Fert. Res. 36, 1–6.

Thornley, J.H.M., Cannell, M.G.R., 1992. Nitrogen relations in a forest plantation—soil organic matter ecosystem model. Ann. Bot. 70, 137–151.

Thornley, J.H.M., Verberne, E.L.J., 1989. A model of nitrogen flows in grassland. Plant Cell Environ. 12, 863–886.

Tinsley, J., 1950. The determination of organic carbon in soils by dichromate mixtures. Trans. 4th Int. Congr. Soil Sci., Amsterdam, Vol. 1, pp. 161–164.

Van Keulen, H., Seligman, N.G., 1987. Simulation of water use, nitrogen nutrition and growth of a spring wheat crop. Simulation Monographs, PUDOC, Wageningen, 310 pp.

VEMAP Members, 1995. Vegetation/ecosystem modeling and analysis project. Comparing biogeography and biogeochemistry models in a continental-scale study of terrestrial ecosystem responses to climate change and CO_2 doubling. Global Biogeochem. Cycles 9: 407–437.

Verberne, E.L.J., 1992. Simulation of nitrogen and water balance in a system of grassland and soil. DLO-Instituut voor Bodemvruchtbaarheid, Oosterweg 92, Postbus 30003, 9750 RA Haren, 56 pp. + Appendices.

Verberne, E.L.J., Hassink, J., de Willigen, P., Groot, J.J.R., Van Veen, J.A., 1990. Modelling soil organic matter dynamics in different soils. Neth. J. Agric. Sci. 38, 221–238.

Vereecken, H., Jansen, E.J., Hack-ten Broeke, M.J.D., Swerts, M., Engelke, R., Fabrewitz and Hansen, S., 1991. Comparison of simulation results of five nitrogen models using different datasets. In: Soil Groundwater Research Report II. Nitrate in Soils. Commission of the European Communities, pp. 321–338.

Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29–38.

Whitmore, A.P., 1991. A method for assessing the goodness of computer simulations of soil processes. J. Soil Sci. 42, 289–299.

Whitmore, A.P., Klein-Gunnewiek, H., Crocker, G.J., Klír, J., Körschens, M., Poulton, P.R., 1997. Simulating trends in soil organic carbon in long-term experiments using the Verberne/MOTOR model. In: Smith, P. Powlson, D.S., Smith, J.U., Elliott, E.T. (Eds.), Evaluation and Comparison of Soil Organic Matter Models Using Datasets From Seven Long-Term Experiments. Geoderma 81(1–2), 137–151, this issue.