Model evaluation in relation to soil N2O emissions: An algorithmic method which accounts for variability in measurements and possible time lags

Environmental Modelling & Software 84 (2016) 251–262

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Model evaluation in relation to soil N_2O emissions: An algorithmic method which accounts for variability in measurements and possible time lags

Vasileios Myrgiotis [a, b, *], Mathew Williams [b], Robert M. Rees [a], Kate E. Smith [c], Rachel E. Thorman [c], Cairistiona F.E. Topp [a]

[a] SRUC, Edinburgh, EH9 3JG, UK
[b] School of GeoSciences, University of Edinburgh, Edinburgh, EH9 3JN, UK
[c] ADAS, Boxworth, CB33 4NN, UK

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form
30 June 2016
Accepted 1 July 2016

Keywords:
Agro-ecosystems
Soil modelling
Model evaluation
Nitrous oxide
Time lag

ABSTRACT

The loss of nitrogen from fertilised soils in the form of nitrous oxide (N_2O) is a side effect of modern agriculture and the focus of many model-based studies. Due to the spatial and temporal heterogeneity of soil N_2O emissions, the measured data can introduce limitations to the use of those statistical methods that are most commonly employed in the evaluation of model performance. In this paper, we describe these limitations and present an algorithm developed to address them. We implement the algorithm using simulated and measured N_2O data from two UK arable sites. We show that possible time lags between the measured and simulated data can affect model evaluation and that their consideration in the evaluation process can reduce measures such as the Mean Squared Error (MSE) by 30%. We also analyse the algorithm's results to identify patterns in the estimated lags and to narrow down their possible causes.

1. Introduction

Process-based agro-ecosystem models are mathematical tools that use existing knowledge of the physical, chemical and biological processes to simulate ecosystem flows of energy, nutrients and water. They provide a holistic picture of an agro-ecosystem's biogeochemical and biophysical structure and are used to communicate what is known about the system and its processes. They can also identify areas where further research is needed, and make predictions of an agro-ecosystem's behaviour under different environmental conditions and management practices (Holzworth et al., 2014). These tools are important especially since climate change and food security are crucial global issues, which are increasingly attracting the interest of citizens and governments (Godfray et al., 2010).

Evaluation is an important part of any model's application and development cycle. As a process, its aim is to examine a model's ability to capture the patterns in measured data and to assist the identification of possible reasons for a model's failure to predict the observed data (Oreskes et al., 1994; Tedeschi, 2006; Bennett et al., 2013; Bellocchi et al., 2015). Some of the most commonly scrutinised points in agro-ecosystem model evaluation concern the model's ability to predict: 1) changes in soil organic matter and soil mineral nitrogen; 2) crop yields in arable systems and cut or grazed biomass in grasslands; 3) changes in soil moisture; 4) loss of nutrients through leaching and 5) fluxes of greenhouse gases.

The statistical methods which are used to evaluate agro-ecosystem models are common to various scientific fields that work with sequences of data points (time series) (Willems, 2009; Anfossi and Castelli, 2014). These methods can be divided into a) deviation-based methods, which use the differences between the measured and simulated values (residuals) in order to provide insights into model performance; b) regression-based methods, which also use the residuals but in order to quantify the level of association between the measured and simulated data and c) probability-based methods, which use the available data to estimate the probability of statistically significant difference between the measured and modelled data.

Regression-based methods can produce their results in

* Corresponding author. SRUC, Edinburgh, EH9 3JG, UK.
E-mail address: Vasileios.Myrgiotis@sruc.ac.uk (V. Myrgiotis).

http://dx.doi.org/10.1016/j.envsoft.2016.07.002
1364-8152/

dimensional form (e.g. in units of measured/modelled data) or have their formulas adapted in such ways so as to produce results in dimensionless form (e.g. percentage). Also, distribution-based tests can be applied as part of regression-based methods such as by conducting Student's *t*-tests on the slope and the intercept (or on both in unison by using the F-test) and examining the significance of their difference from those of the 1:1 line (Bellocchi et al., 2010). Probability-based methods include comparisons between measured and simulated data (e.g. Student's *t*-test, F-test), their respective ranking (e.g. Wilcoxon-signed rank test) or cumulative distributions (e.g. Kolmogorov-Smirnov's D test) (Daniel and Cross, 2012; Stephens, 1974).

For the most frequently examined model outputs, the measured datasets are time series of the variables of interest. Because of the costs associated with setting up and conducting the measurements in agricultural ecosystems, it is common that these time series consist of data which are measured at non-uniform time intervals. Non-uniformity is a major source of complexity not only for the analysis of the data themselves but also for the evaluation of models (Gu et al., 2014; Giltrap et al., 2010; Bellocchi et al., 2010). In addition, the spatial heterogeneity of agricultural soils introduces a considerable level of uncertainty to a model's inputs and outputs as well as a high variability to the measured data, which are used to evaluate the model. The variability in measured data differs depending on the variable considered, and the impacts of the uncertainties in model input data can be unevenly shared among the main variables of interest. This paper focuses on the evaluation of a model's performance in relation to its ability to predict fluxes of nitrous oxide (N_2O) from cultivated soils. N_2O is among the main variables of interest in agro-ecosystem modelling and one on which the variability in the measured data can be particularly large. N_2O is a greenhouse gas with high global warming potential as well as an ozone depleting gas (Marschner and Rengel, 2007). To a large extent, it is produced in cultivated soils through the processes of nitrification and denitrification, which are controlled by microbes and driven by the use of nitrogenous fertilisers and by environmental conditions (Galloway et al., 2003). Nevertheless, some aspects of N_2O production in soils are not fully understood due to the complex role of soil microbes (Butterbach-Bahl and Dannenmann, 2011).

N_2O samples are typically collected using manual or automatic chambers. Despite the limitations and weaknesses, this method is widely applied and the derived N_2O data are used to evaluate agro-ecosystem models at field scale (Chadwick et al., 2014). Because of the spatial heterogeneity of soil biochemical and physical properties and the need for measurements to be representative of the examined field, the measurements are usually repeated across the experimental field. This experimental design (i.e. replication) provides a number of daily measured values from which the respective daily means and standard errors are calculated. The evaluation of agro-ecosystem models in relation to N_2O emissions can be directly and indirectly affected by certain factors:

1. The relatively large standard errors in measured data as a result of soil heterogeneity and uneven fertiliser application.
2. The existence of negative N_2O values in the measurements either because of microbial uptake or as an artefact of the experimental procedure (Cowan et al., 2014; Chapuis-Lardy et al., 2007).
3. The existence of non-uniform time intervals in the measured data due to cost constraints, field conditions and unforeseen events during sampling.
4. The possibility of time lags between measured and simulated data due to uncertainty and gaps in model inputs as well as the model's parameterisation.

This paper is based on the concept that model evaluation can be as thorough and informative as possible when multiple methods are applied (Bellocchi et al., 2010; Tedeschi, 2006; Martorana and Bellocchi, 1999; Whitmore, 1991). It presents a new evaluation algorithm that takes into account the factors listed above. The algorithm is used in order to a) integrate the variability of measured data in the model evaluation process and b) examine the impacts that time lags between the simulated and measured data may have on model evaluation. In the following sections, we describe the proposed algorithm, which we then implement to evaluate a well known agro-ecosystem model (Landscape-DNDC (Haas et al., 2012)) using measured N_2O data from two experimental sites in the UK.

2. Materials and methods

2.1. The limitations of commonly used statistics

Through devising, enhancing and combining measures and test statistics, a collection of model evaluation methods has been compiled and is available to model developers and users. Bellocchi et al., 2010 provide an excellent account of existing methods and so do Richter et al., 2012 who also rank the different methods according to their use frequency. Both authors compiled information on suggested boundaries for different evaluation measures and their corresponding model performance level. Despite the existence of such recommendations on how to interpret the results of different model evaluation tests, there is a lack of widespread agreement.

Some recently developed model evaluation methods can be found in Sanna et al., 2015, Ali and Abustan, 2014 and Ritter and Muñoz-Carpena, 2013. Their work aimed at incorporating certain −often ignored− aspects of data comparison into their proposed methods by combining multiple measures (Sanna et al., 2015; Ritter and Muñoz-Carpena, 2013), addressing certain limitations of pre-existing methods (Ali and Abustan, 2014) and considering under-explored areas (Ritter and Muñoz-Carpena, 2013).

The methods that are used to compare measured and simulated data have limitations and can produce misleading results. Such limitations become apparent when the methods are used with data that are characterised by particularities such as considerable uncertainties and/or the presence of outliers and/or irregular temporal intervals between the data points. Various authors have discussed the strengths and weaknesses of evaluation methods in detail (see references in Table 1) and their conclusions apply to model evaluation in relation to emissions of greenhouse gases from soils. In the following list we outline the main problems that affect each group of model evaluation methods, from the perspective of soil N_2O fluxes.

1. Deviation based methods:
   (a) Positive and negative residuals can cancel each other out and produce unrealistic statistical values.
   (b) Negative N_2O measurements are used in the calculation of statistics even though, based on current understanding, models cannot predict negative fluxes.
   (c) The impact of time lags can be particularly fertiliser because N_2O peaks can have both a short duration and a large magnitude. If a model has missed a measured peak of N_2O flux by a few days the estimated Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) can be misleadingly high.
   (d) These methods use the average measured N_2O and ignore the information that replicate measurements provide.
2. *Regression based methods*:
   (a) Fail to account for model bias.

Table 1
Commonly used statistics in model evaluation.

Category	Name	Formula	Reference
Deviation based	Bias	$Bias = \frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)$	Smith and Smith (2007)
	Relative Bias	$RelB = \frac{\frac{1}{n}\sum_{i=1}^{n} O_i - S_i}{\overline{O}\cdot 100}$	Richter et al. (2012)
	Fractional Bias	$FB = 2\cdot\frac{\overline{S}-\overline{O}}{\overline{S}+\overline{O}}$	Sanna et al. (2015)
	Coefficient of Residual Mass	$CRM = \frac{\sum_{i=1}^{n} O_i - \sum_{i=1}^{n} S_i}{\sum_{i=1}^{n} O_i}$	Sanna et al. (2015)
	Percent Bias	$PB = 100\cdot CRM$	Sanna et al. (2015)
	Relative Error	$E = \frac{100}{\overline{O}}\cdot\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)$	Smith and Smith (2007)
	Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert O_i - S_i\rvert$	Richter et al. (2012)
	Mean Squared Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2$ and $MSE = Bias^2 + SDSD + LCS$	Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993); Kobayashi and Salam (2000)
	Root Mean Squared Error	$RMSE = \sqrt{MSE}$	Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)
	Normalised Root Mean Squared Error	$NRMSE = \frac{RMSE}{Range(obs)}$	Richter et al. (2012)
	Relative Root Mean Squared Error	$RRMSE = 100\frac{RMSE}{\overline{O}}$	Richter et al. (2012)
	Standard Error of Prediction Corrected for bias	$SEPC = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i - MBE)^2}$	Richter et al. (2012)
	Lack of Correlation weighted by the Standard deviations	$LCS = 2\cdot SD_S\cdot SD_O\cdot(1-r)$	Kobayashi and Salam (2000)
	Standard Deviation of the Observations	$SD_o = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - \overline{O})^2}$	Kobayashi and Salam (2000)
	Standard Deviation of the Simulations	$SD_s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - \overline{S})^2}$	Kobayashi and Salam (2000)
	Squared Difference between Standard Deviations	$SDSD = (SD_S - SD_O)^2$	Kobayashi and Salam (2000)
	Ratio of Standard Deviation of observations to RMSE	$RDP = \frac{SD_O}{RMSE}$	Richter et al. (2012)
	Nondimensional error index	$NDI = \frac{RMSE}{SD_O}$	Richter et al. (2012)
	Modelling Efficiency	$EF = 1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(S_i - \overline{S})^2}$	Sanna et al. (2015)
	Modelling percent Efficiency	$EF\% = 100\cdot(1-EF)$	Sanna et al. (2015)
	Nash Sutcliffe Efficiency	$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})}$	Richter et al. (2012)
	Agreement Coefficient	$AC = 1 - \frac{\sum_{i=1}^{n}(S_i - O_i)^2}{\sum_{i=1}^{n}(\lvert\overline{O}-\overline{S}\rvert + \lvert O_i - \overline{O}\rvert)\cdot(\lvert\overline{O}-\overline{S}\rvert + \lvert S_i - \overline{S}\rvert)}$	Richter et al. (2012)
	Willmott's index of agreement	$d = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(\lvert S_i - \overline{S}\rvert + \lvert O_i - \overline{O}\rvert)^2}$	Willmott et al. (2011, 1985); Richter et al. (2012)
	Refined index of agreement	$dr(1) = 1 - \frac{\sum_{i=1}^{n}\lvert S_i - O_i\rvert}{c\sum_{i=1}^{n}\lvert O_i - \overline{O}\rvert}$ if: $\sum_{i=1}^{n}\lvert s_i - o_i\rvert \le c\sum_{i-1}^{n}\lvert O_i - \overline{O}\rvert$ else: $dr(2) = \frac{c\sum_{i=1}^{n}\lvert O_i - \overline{O}\rvert}{\sum_{i=1}^{n}\lvert S_i - O_i\rvert} - 1$ (with $c = 2$)	Willmott et al., (2011, 1985); Richter et al. (2012)
Regression based	Pearson's correlation coefficient	$r = \frac{\sum_{i=1}^{n}(O_i - \overline{O})\cdot(S_i - \overline{S})}{n\cdot SD_S\cdot SD_O}$	Smith and Smith (2007)
	Coefficient of determination	$R^2 = r^2$	Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)
	Regression (slope/intercept)		
	Orinary Least Squares (parametric)	$S_i = m\cdot O_i + b$	
	Theil-Sen (nonparametric)	$S_i = m\cdot O_i + b$	Richter et al. (2012); Theil (1970)
Probability based	t-value	$t = \frac{Bias\cdot\sqrt{n}}{\sqrt{\sum_{i=1}^{n}(O_i - S_i - (\sum_{i=1}^{n}(O_i - S_i)/n))^2/(n-1)}}$	Smith and Smith (2007)
	Root Mean Squared Error at 95% CI	$RMSE_{95} = \frac{100}{\overline{O}}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(SE_i\cdot t_{m,95})}$	Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)
	Relative Error at 95% CI	$E_{95} = \frac{100}{\overline{O}}\cdot\frac{1}{n}\sum_{i=1}^{n}(SE_i\cdot t_{m,95})$	Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)
	Significance of difference between measured and simulated values	$LOFIT = \sum_{i=1}^{n}(O_i - S_i)^2$	
		$F = \frac{\sum_{i=1}^{n}(m_i - 1)\cdot LOFIT}{n\sum_{i=1}^{n}\sum_{j=1}^{m_i}((O_{ij} - S_i) - (O_i - S_i))^2}$	Smith and Smith (2007)

S_i: the simulated data, $\overline{S}$: the mean of the simulated data, O_i: the measured data, $\overline{O}$: the mean of the measured data.
i: the index number of the measured/simulated data, n: the number of simulated/measured data points m: the number of replicates of ith measurement O_{ij}: the jth replicate of ith measurement.
CI: the Confidence Interval, SE_i: the standard error of the ith measurement, $t_{m,95}$: the Student's *t-value* for m replicates and 95% probability (*p-value* = 0.95).

(b) Are insensitive to additive and proportional differences between simulations and measurements
(c) Can produce large coefficient of determination (R^2) values even when residuals are large.
(d) Require the measured data to be independent and normally distributed (even though data transformation and nonparametric approaches can be used e.g. Theil-Sen)

3. *Probability based methods*:
(a) The requirement of data being normally distributed is usually not met (even though data transformation and nonparametric approaches can be used e.g. Wilcoxon signed-rank test)
(b) Measured datasets are usually rather small, therefore t-tests are performed using few degrees of freedom and the null hypothesis of no difference between observed and simulated data can be difficult to reject.
(c) Large variability in measured data can lead to wide 95% confidence intervals (CI).

2.2. The proposed algorithm

The development of the proposed algorithm was driven by the inability of commonly used statistics to account for possible irregular time lags between the measured and simulated time series and to consider the range of daily values that replicate measurements can provide. The main points of the proposed algorithm's concept are discussed below and its schematic diagram is presented in Fig. 1.

a) Replicate daily measurements can be used to calculate daily value ranges, which encapsulate the variability of measured N_2O. Quantifying the percentage of simulated values that fall within the respective measured ranges for each day of measurement can be a straightforward evaluation of a model's predictive accuracy. The strictness of this test depends on which method will be used to estimate the daily ranges with the standard error being the most strict method and the daily measured minimum/maximum the least. In order to perform this task we added an appropriate process to the algorithm. This process uses the upper and lower limits of daily measured data and the corresponding simulated data to return the percentage of simulated values that were inside the measured limits. Hereafter we will refer to this measure as the *accuracy* measure.

b) The correlation coefficient (r), which is one of the two most commonly used regression-based statistics, expresses the linear correlation between observed and simulated data (Duveiller et al., 2016). Because of its mathematical formulation (see Table 1), it does not reveal how successful the model was in predicting the changes in emission magnitude between successive measurements. This aspect of model performance can be examined by estimating the direction of change in the measured N_2O between two successive measurement days and comparing it with the direction of change between the simulated N_2O points that correspond to these measurement days. Repeating this process for all the data points and calculating the number of times that the simulated and measured patterns were in agreement, is an alternative way to express the correlation between observed and simulated data. In order to perform this task we added a process to the algorithm which scans the daily measured and simulated data and checks if the direction of magnitude change between two successive measured data points agrees with the respective change between the corresponding simulated data points. This check is performed in accordance to the chronological order of the data, starting from the first data point and ending at the last. When the checking process is complete, it returns a percentage value that shows how many of the direction changes between successive measured points have been predicted by the model (Fig. 2). Hereafter we will refer to this measure as the *trend prediction* measure.

c) The proposed algorithm examines the existence of possible time lags using a minimisation-of-residuals approach. Based on a user-defined range of time lags (e.g. ± 3 days) the algorithm selects, for each day of measurement, the simulated value (and corresponding lag), which has the smallest deviation from that day's measurement (Fig. 3). The time lag(s) that the algorithm predicts always refer to the position of the simulated data relative to the measured data. A lag is positive when the simulated value that is closer to the examined measured value (i.e. has the lowest residual), was simulated by the model at a day that is after the actual measurement day. A lag is negative when the simulated value that is closer to the examined measured value, was simulated by the model at a day that is before the actual measurement day.

d) A set of statistics that includes r, RMSE, the squared bias (SB), the squared difference between standard deviations (SDSD) and the lack of correlation weighted by the standard deviations (LCS) (see Table 1) along with the values of the accuracy and trend prediction measure (points *a* and *b* above) can be used to evaluate how a model performs based on a measured dataset. Calculating this set of statistics, by using the simulated time series, offers a picture of the model's performance without any kind of time lag being considered (first set of statistics). The set of simulated values that is produced with the minimisation-of-residuals process (*c*) is, in effect, a 'lagged' time series of simulated N_2O. By using this lagged simulated time series to recalculate the set of statistics and juxtaposing its results with those of the first set of statistics, we can quantify and assess the impacts of time lags on model evaluation.

e) Measurements of soil N_2O are usually more frequent around the dates when fertiliser application takes place. Therefore, a closer examination of the distribution of the estimated lags during the periods that follow these events can offer insights into the model's performance and be useful in identifying the possible causes of these lags.

2.3. Experimental data

For the model evaluation we used site information and soil N_2O measurements from two arable experiments located in the vicinity of ADAS Terrington, Cambridgeshire, eastern England (latitude 52.75, longitude 0.3, elevation 5 m a.s.l.). The sites have different soil properties and the respective measurements took place in different years; 2004–2005 (Smith et al., 2012) and 2011–2012 (Thorman et al., 2013). Winter wheat was the crop that was planted and harvested during both experiments. At both sites N_2O fluxes were monitored, using the static chamber technique (Cardenas et al., 2010; Chadwick et al., 2014) for 12 months following spring applications of manufactured nitrogen (N) fertiliser to winter wheat. At the first site, the first day of measurements was 1 March 2004 and the last was 5 March 2005. At the second site, the first day of measurements was 2 March 2011 and the last was 17 February 2012. N_2O samples were analysed in the laboratory by gas chromatography (Cardenas et al., 2010). Gravimetric topsoil moisture content was measured on every N_2O measurement occasion at the second experimental site, and periodically at the first site. Additionally at the second site, topsoil mineral N was measured concurrently with the soil moisture. The soil bulk density was used to convert the soil gravimetric moisture content to water-filled pore space (% WFPS). All experimental treatments were replicated (x3) and arranged in a randomised block design with two or five chambers per plot in the first and second experiment respectively.

In the first experiment (Smith et al., 2012), the soil texture was a silty clay loam, with a bulk density of 1.38 g/cm^3, a clay content of

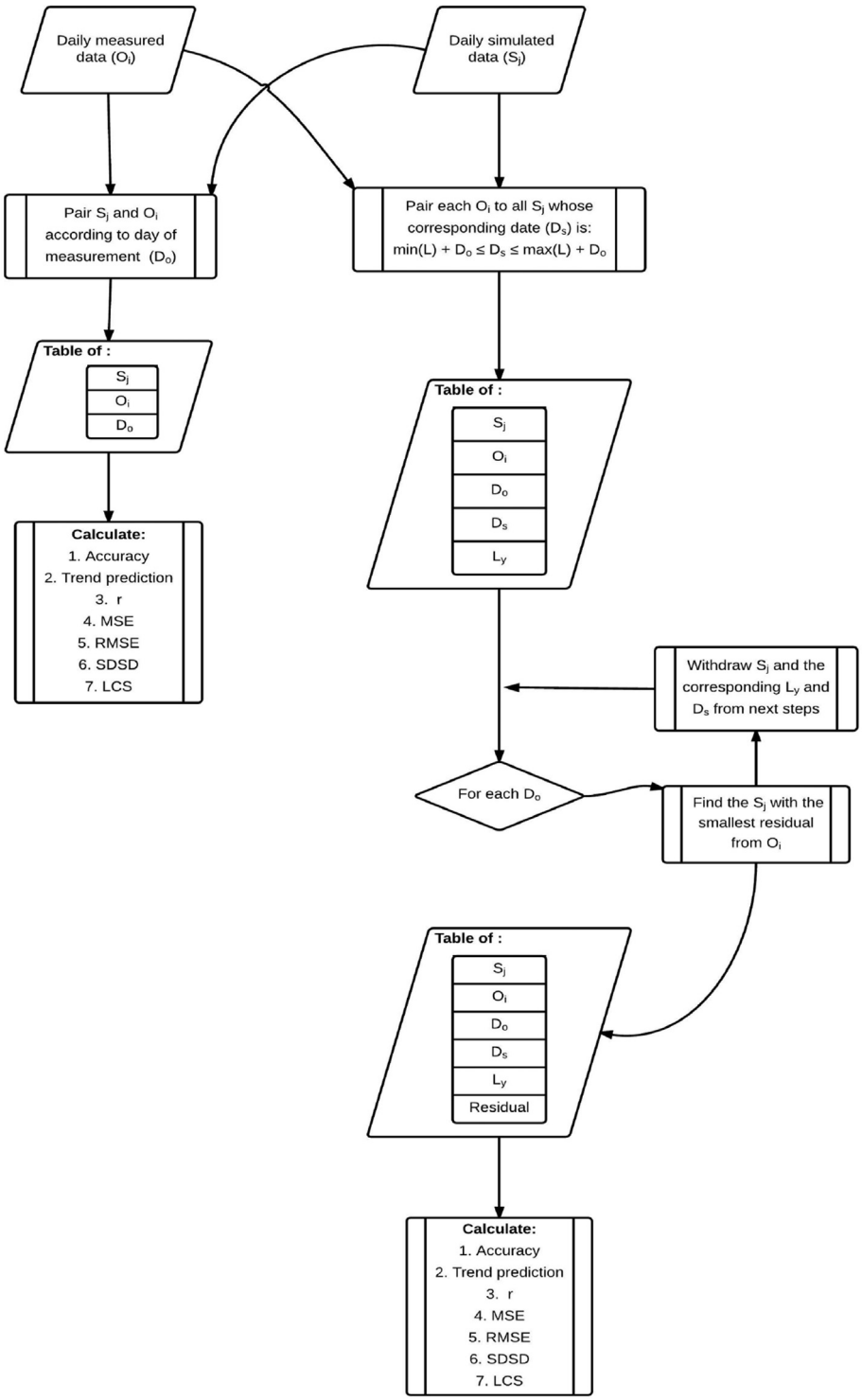

Fig. 1. Schematic description of the proposed algorithm. *S*: simulated data, *O*: measured data, *L*: time lag and *D*: measurement day, *i*: index number of measured value, *j*: index number of simulated value, *y*: index number of examined time lags.

32%, a pH of 8.1 and an organic carbon content of 1.7% (measured at 0–0.10 m depth). The total precipitation at the site during 2004 only (i.e. not the full 12 month data set) was 760 mm and the average annual temperature was 11.7 °C. The measured N_2O and soil moisture data used in this model evaluation are from 2004 and the treatment where 220 kg N ha^{-1} of urea fertiliser was applied to the soil in three doses. The measured datasets that are used in this study consist of 58 daily N_2O measurements as well as 27 daily measurements of soil moisture. The used datasets cover 2004 only and exclude 4 measurements taken during 2005 because of the large distance between the last measurement day in 2004 and the first measurement day in 2005 as well as because of the large distance between the measurement days in 2005.

In the second experiment (Thorman et al., 2013) the soil texture was a sandy loam soil, with a bulk density of 1.35 g/cm^3, a clay content of 11%, a pH of 8.3 and an organic carbon content of 1.8%

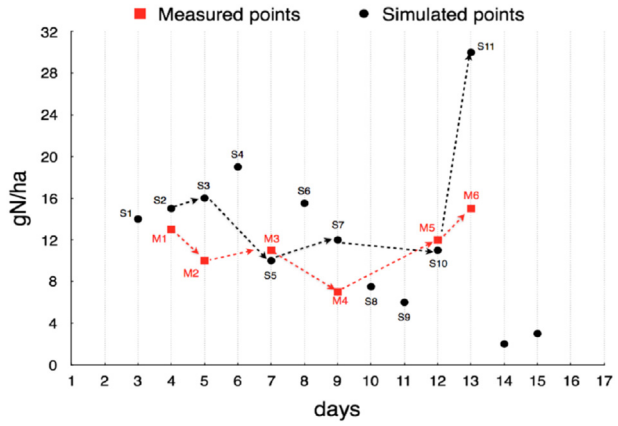

Fig. 2. Description of the concept of the trend prediction measure. The scatter plot shows simulated and measured data points during a period of 13 days from day 3 to day 15. Only six measurements were taken during this period (points M1 to M6). All the values that were simulated by the model during this period are presented (S1 to S13). The arrows show the direction of change between successive measured data points and between the corresponding simulated data points. For day 4 the model simulated value was S2 while the value M1 was measured in the field. The measured value for day 5 (M2) was smaller than that for day 4 (M1) and the direction of change between the values for these two days was negative (i.e. decreased emission). The simulated value for day 5 (S3) was larger than the simulated value for day 4 (S2) and the direction of change between them was positive (i.e. increased emission). Overall, the model predicted one direction of change correctly (between day 12 and day 13) and missed the other four (days 4–5, 5 to 7, 7 to 9 and 9 to 12). Based on the data in this figure, the algorithm's trend prediction measure is 20% (one out of five). In contrast to that, the correlation coefficient for the same data is 0.65, which suggests a moderate correlation (Bellocchi et al., 2010).

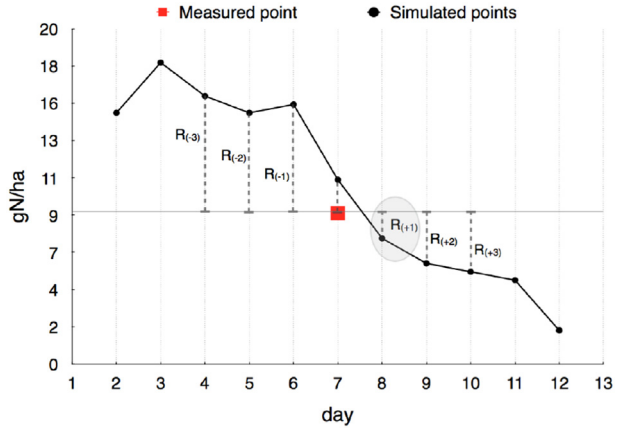

Fig. 3. Description of the concept of the minimisation of residuals. The graph shows a data point measured on day 7 (9 g N ha^{-1}). All the values that were simulated by the model at dates that are close to the measurement date are presented. For this example the time lag range is equal to ± 3 days (i.e. day 4 to day 10). The algorithm calculates the residuals of all the simulated values within this range (i.e. $R_{(-3)}$, $R_{(-2)}$, $R_{(-1)}$, $R_{(+1)}$, $R_{(+2)}$, $R_{(+3)}$) and identifies the day when the simulated value has the smallest residual (i.e. day 8). The time lag that corresponds to day 8 is + 1 day from the measurement date while the simulated value is around 7.5 g N ha^{-1}. The algorithm will: a) save the measurement date (i.e. day 7), the time lag (i.e. +1) and the simulated value (i.e. 7.5 g N ha^{-1}) in an appropriately formatted table; b) withdraw the information attached to this specific simulated point (i.e. 7.5 g N ha^{-1} at day 8) from future use and c) continue by repeating the same process for the next measured data point until it reaches the last measured data point.

(measured at 0–0.10 m depth). The total precipitation during 2011 only (i.e. not the full 12 month data set) was 470 mm and the average annual temperature was 11.9 °C. The measured N_2O, soil moisture and soil mineral N data used in this model evaluation are from 2011 and the treatment where 180 kg N ha^{-1} of ammonium nitrate (AN) fertiliser was applied in three doses. The measured data that are used in this study consist of 40 daily soil N_2O measurements, 40 daily soil moisture measurements and 40 daily soil mineral N measurements. The used datasets cover 2012 only and exclude three measurements taken during 2012 because of the large distance between the last measurement day in 2012 and the first measurement day in 2012 as well as because of the large distance between the measurement days in 2012.

2.4. Landscape-DNDC

We used the Landscape-DNDC model (version 0.23.0) to simulate the two experimental agro-ecosystems. Landscape-DNDC is a process-based ecosystem biogeochemistry model that can simulate the biogeochemistry of cropland, grassland and forest ecosystems (Haas et al., 2012). It belongs to the DNDC family of models, which includes some of the most widely-used ecosystem models (Perlman et al., 2013). The model uses information on soil properties, climatic conditions, geographic location and agricultural management as inputs. Its outputs include biomass growth, soil C and N content, emissions of C and N-based gases (e.g. ammonia, methane, carbon dioxide, nitrogen gas etc) as well as leached C and N-based compounds (e.g. nitrate, dissolved organic C etc). Hereafter, we refer to Landscape-DNDC as the *model*.

3. Results

3.1. First Terrington site

We used the measurements dataset for the urea fertiliser treatment of the first Terrington site along with the respective model outputs to implement the algorithm. We allowed the algorithm to examine the impact of the six possible time lags that constitute a ± 3-day range and used the standard deviations of the measured replicate values to define the range of measured N_2O for each day.

Fig. 4 presents a graph of the daily measured and simulated N_2O data and the results of implementing the algorithm under these instructions. The value of the accuracy measure, which was estimated with and without the use of the minimisation-of-residuals approach of the algorithm (see Fig. 3), shows that the inclusion of possible time lags in the analysis of fit leads to an accuracy that is improved by 21% (accuracy increased from 53% to 64%). Interestingly, the improvement in the trend prediction measure was larger (58% improvement from 45% to 71%).

The set of commonly used statistics (i.e. r, RMSE, MSE, SDSD, LCS) provides an insight into how time lags can influence the evaluation of the model in comparison to the field measurements. Because the MSE (presented in (g N ha^{-1})^2) is equal to the sum of SB, SDSD and LCS, we can better understand what caused the improvement in the model's prediction. We can do that because the estimated MSE value captures the role of model bias (described by SB), the role of the model prediction in relation to the patterns of fluctuations in the measured data (described by LCS) and the role of the model prediction in relation to the magnitude of fluctuations in the measured data (described by SDSD) (Kobayashi and Salam, 2000). Based on this, the observed 17% reduction in the estimated MSE (decreased by 104.35 (g N ha^{-1})^2), after the inclusion of possible time lags in the analysis, is attributed mainly to the improvement by 32% in LCS which decreased by 109.3 (g N ha^{-1})^2 and compensated for the much smaller increases in SDSD and SB (Fig. 4).

In order to provide a picture of how sensitive the algorithm's results are to the choice of the time lag window that is examined (i.e. ± 3-day) we reimplemented the algorithm after imposing a ± 1

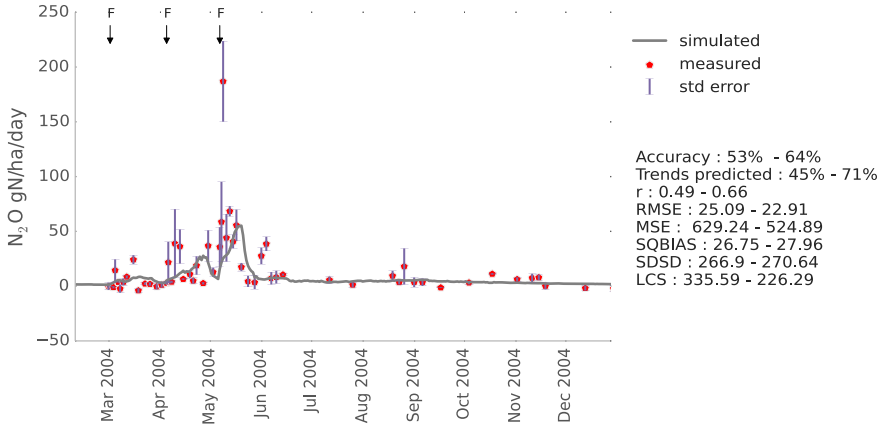

Fig. 4. Measured and simulated soil N_2O at the first Terrington site. The right sidebar presents the statistical values with (first value) and without (second value) examining the effect of the examined time lags (i.e. ± 3 days). The units of RMSE are g N ha^{-1} and the units for MSE, SB, SDSD and LCS are (g N ha^{-1})^2. 'F' indicates a date of fertiliser application with the first one representing 40 kg N ha^{-1} of urea and the latter two representing 90 kg N ha^{-1} of urea.

day deviation on the examined time lag window (i.e. set the lag window equal to ± 2 and ± 4). This ± 1 day deviation around the examined time lag window led to a relative standard deviation of 6.9% for the accuracy index, 4% for the trend prediction index, 3.8% for r and 2.1% for RMSE.

In addition to the estimation of the statistics and model behaviour metrics, we looked into the series of irregular time lags, which the algorithm estimates and uses. We used the frequency distribution of the estimated time lags as a way to present the dominant tendency (i.e. whether positive or negative) of the lags during specific periods of time. More than 75% of all the daily measurements were conducted between March and May (Fig. 5). The accuracy of the model's N_2O predictions (accuracy measure) is gradually improving from March to May. Most of the estimated lags in N_2O prediction are positive in March and negative in April while there is a clearly positive lag in the simulated N_2O values in May (Fig. 5).

Soil moisture is a major driver of N_2O emissions and the availability of measured soil moisture data for this site offers the opportunity to examine the lags in soil moisture prediction by the model (Dobbie and Smith, 2003). We used measured soil moisture (% WFPS) data along with the corresponding simulated outputs to implement the algorithm (Fig. 6). The distributions of the estimated lags for the data-rich months show a reverse distribution to that of the lags in soil N_2O prediction (Fig. 5). It could be argued that the two sets of lags are negatively related, however, Figs. 5 and 6 do not inform us about the actual measurement dates to which each lag corresponds.

In order to better understand how the two sets of lags relate to each other throughout the period March to May, we further analysed the estimated lags. For the days on which we had both soil moisture and N_2O measurements, we used equation (1) to calculate the difference between the respective estimated lags for each day of measurement.

$$\text{Lag Difference} = \frac{Lag_{N2O} \cdot Lag_m^{-1} \cdot |Lag_{N2O} - Lag_m|}{|Lag_{N2O} \cdot Lag_m^{-1}|} \quad (1)$$

where Lag_{N2O} and Lag_m are the estimated time lag in daily N_2O and soil moisture prediction respectively. Equation (1) produces a value whose sign shows whether the soil moisture and the N_2O lag have the same or opposite direction (i.e. sign is positive or negative) and

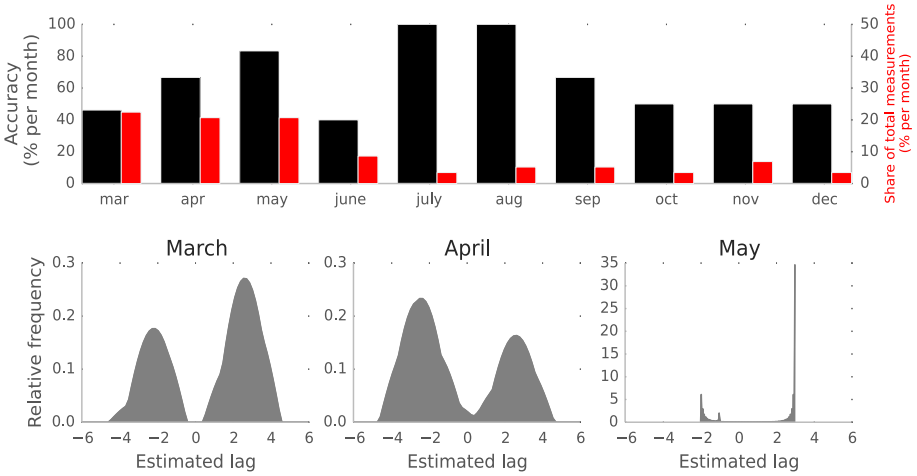

Fig. 5. The algorithm's results for N_2O for the urea treatment at the first Terrington site. The top graph shows the accuracy measure of the model's N_2O predictions for each month when time lags are considered (in black) and the percentage of total measurements that were taken in each month (in red). The three graphs in the second row show the frequency distribution of the time lags that were estimated by the algorithm for March, April and May. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

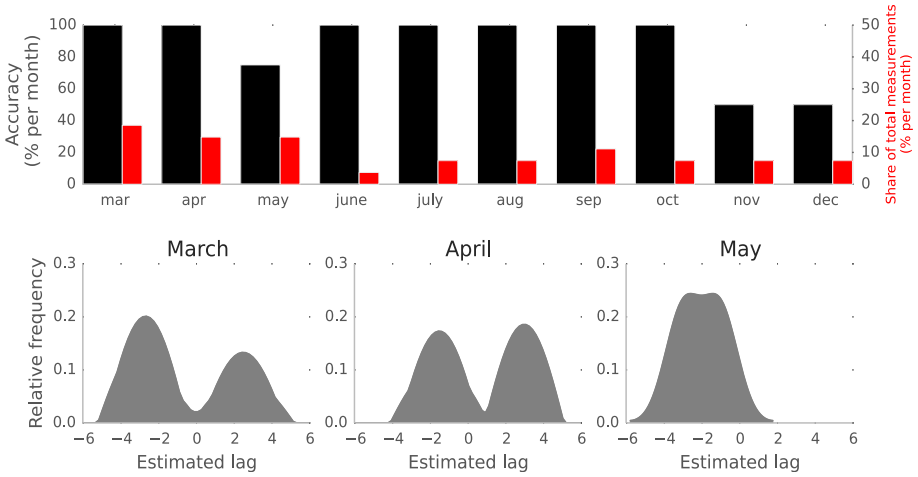

Fig. 6. The algorithm's results for soil moisture at the first Terrington site. The top graph shows the accuracy measure of the model's soil moisture predictions for each month when time lags are considered (in black) and the percentage of total measurements that were taken in each month (in red). The three graphs in the second row show the frequency distribution of the time lags that were estimated by the algorithm for March, April and May. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

whose size shows the magnitude of their difference. Because lag-difference encapsulates the date of measurement, the lag in N_2O and the lag in soil moisture prediction, it can be used to understand how the two sets of lags relate to each other and how their relationship varies through time.

The model links a driver of the simulated system (i.e. soil moisture) to one of the model's outputs (i.e. N_2O) in a way that is temporally different to that indicated by the respective measurements (Fig. 7). This difference is not constant throughout the data-rich period but changes from being rather small in March to being noticeable in May. It is possible that this increase in lag-difference is related to the increase in the amount of N added to the soil in April and May (see fertiliser applications in Fig. 4). In March the first fertiliser application occured and 40 kg N ha^{-1} of urea was added to the soil. For this month, the model produces the best-fitting simulated values for soil moisture mostly before the actual measurement date (Fig. 6). For the same month, the distribution of lag-differences (Fig. 7) shows that the lags in N_2O agree with those in soil moisture both in relation to the direction of the lags (i.e. sign of lag difference is positive) and in relation to the size of the lags (i.e. mode of lag-difference is low). During April and May two more fertiliser applications take place, each of them equal to 90 kg N ha^{-1} of urea per month. The model produces the best-fitting simulated values for soil moisture at days that are before the actual measurement day (Fig. 6) and the respective values for N_2O at days that are after the measurement day (i.e. the lag-difference becomes negative).

3.2. Second Terrington site

For the second example we implemented the algorithm using the measured N_2O dataset for the second Terrington site along with the corresponding model outputs. A ± 3-day range was used to define the six time lags that were examined and the standard deviations of the N_2O measurements were used to define the range of measured N_2O for each day.

The algorithm's results (Fig. 8) show that time lags can reduce the model's predictive accuracy by 33% (accuracy decreases from 56% to 42% if lags are not considered). As was the case for the first Terrington site, the improvement in the prediction of the trends in the measured data was large (i.e. 48% increase in the trend prediction measure) and is reflected in the similarly large increase in r (i.e. from 0.28% to 0.59%). The substantial decrease in the LCS value, when time lag is considered and relative to the size of MSE (i.e. from 3.73 to 2.23), indicates that the improvement in RMSE/MSE occurs mainly because the lagged simulated N_2O data points represent the fluctuations between the measured points far better than the respective non-lagged points.

Similar to what was done in the first example, we reimplemented the algorithm after imposing a ± 1 day deviation on the examined time lag window in order to quantify the sensitivity of the results to the chosen time lag window. This ± 1 day deviation around the examined time lag window led to a relative standard deviation of 0.85% for the accuracy index, 4.3% for the trend prediction index, 2.1% for r and 0.9% for RMSE.

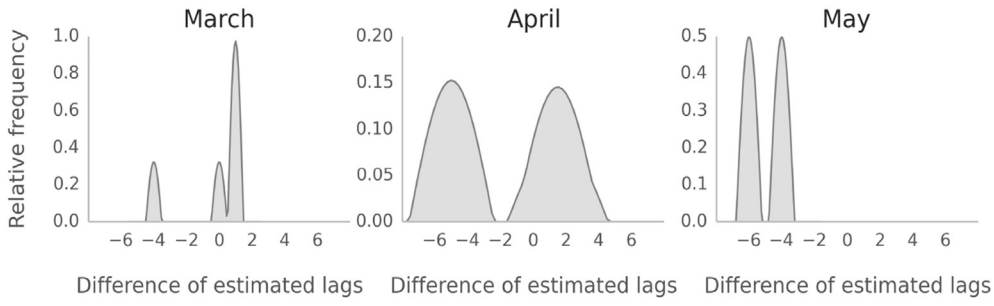

Fig. 7. Kernel density plots of the difference between the estimated time lags in the model's soil moisture and N_2O predictions for March, April and May at the first Terrington site.

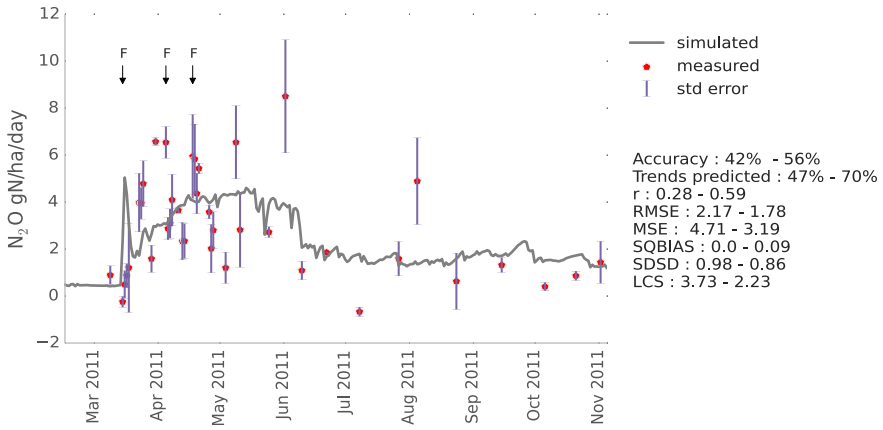

Fig. 8. Measured and simulated soil N_2O for the AN treatment at the second Terrington site. The right sidebar presents the statistical values with (first value) and without (second value) examining the effect of the examined time lags (i.e. ± 3 days). The units of RMSE are g N ha^{-1} and the units for MSE, SB, SDSD and LCS are (g N ha^{-1})^2. 'F' indicates a date of fertiliser application with each representing 60 kg N ha^{-1} of ammonium nitrate.

The analysis of the model's accuracy shows that more than half of all the measurements were taken during March and April and that the model's accuracy rises from 60% in March to 70% in April (Fig. 9). During June and October the model has produced daily outputs that were not within the respective measured limits. Most of the estimated lags for both months are positive but many negative lags have also been estimated by the algorithm (second row in Fig. 9). In order to see how the lags in the prediction of soil moisture and soil mineral N compare with those in N_2O prediction, we implemented the algorithm using measured and simulated data for soil moisture (% WFPS) and for soil mineral N (kg N ha^{-1}). The distributions of lags in soil moisture (Fig. 10) and soil mineral N (Fig. 11) during March and April look very similar. In both cases, most of the estimated lags are positive, a fact that is in line with the lags estimated for the model's soil N_2O prediction. Overall, the distribution of lags for the two data-rich months looks similar for all three variables but the similarities are more clear in soil moisture and soil mineral N.

We wanted to see how the three sets of lags (i.e. in N_2O, moisture and soil mineral N prediction) relate to each other through time. For March and April, the most data-rich months, we plotted the frequency distribution of the differences between the lags in N_2O and soil moisture, N_2O and soil mineral N as well as soil mineral N and soil moisture (Fig. 12). To estimate the differences between these sets of lags we used equation (1). In contrast to the first example, the size of the set of lag-differences was larger because all soil moisture measurement dates corresponded with those of N_2O and soil mineral N. It could be argued that the lags between the simulated and the measured values for the three variables examined are mostly positively related. The shapes of the three distribution curves reflect the fact that the estimated lags in the prediction of the three variables have the same underlying cause (Fig. 12).

4. Discussion

We presented an algorithm that compares daily measured and simulated soil N_2O data in a way that the uncertainty in the measured data can be considered and the impact of possible lags can be examined. Through this algorithm we introduced two new model evaluation measures (accuracy and trend prediction). These measures, combined with a set of commonly-used statistics, can offer a picture of the model's behaviour that is more detailed than that usually presented in modelling studies. The accuracy and the trend prediction measures can be used to quantify a model's predictive success in relation to the magnitude and the fluctuation

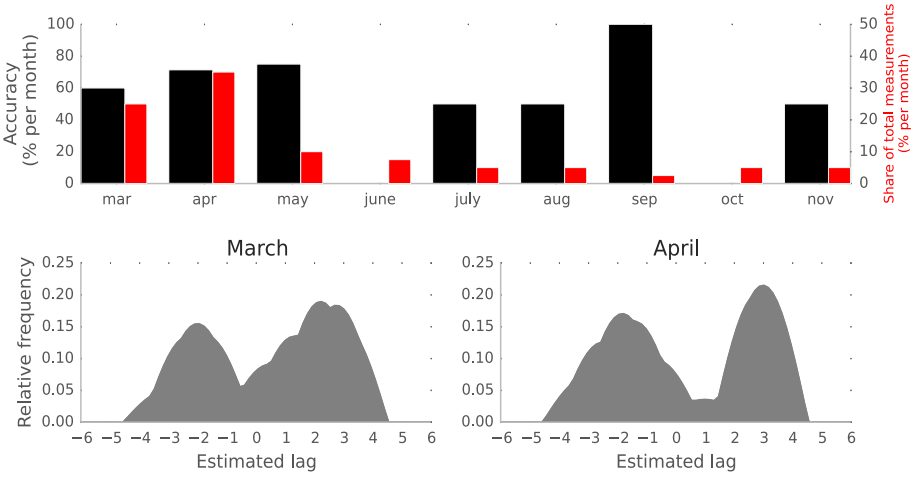

Fig. 9. The algorithm's results for N_2O prediction at the second Terrington site. The top graph shows the accuracy measure of the model's N_2O predictions for each month and when time lags are considered (in black), and the percentage of total measurements that were taken in each month (in red). The following two graphs show the frequency distribution of the time lags that were estimated by the algorithm for March and April. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

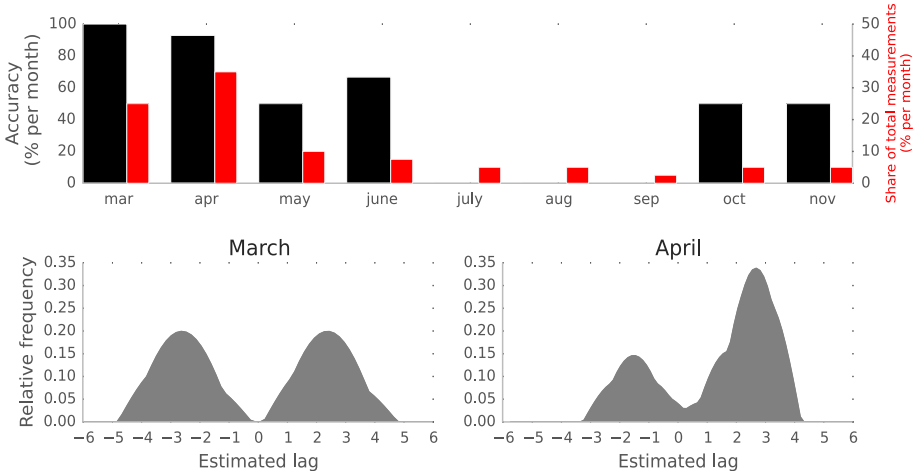

Fig. 10. The algorithm's results for soil moisture prediction at the second Terrington site. The top graph shows the accuracy measure of the model's soil moisture predictions for each month and when time lags are considered (in black), and the percentage of total measurements that were taken in each month (in red). The following two graphs show the frequency distribution of the time lags that were estimated by the algorithm for March and April. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

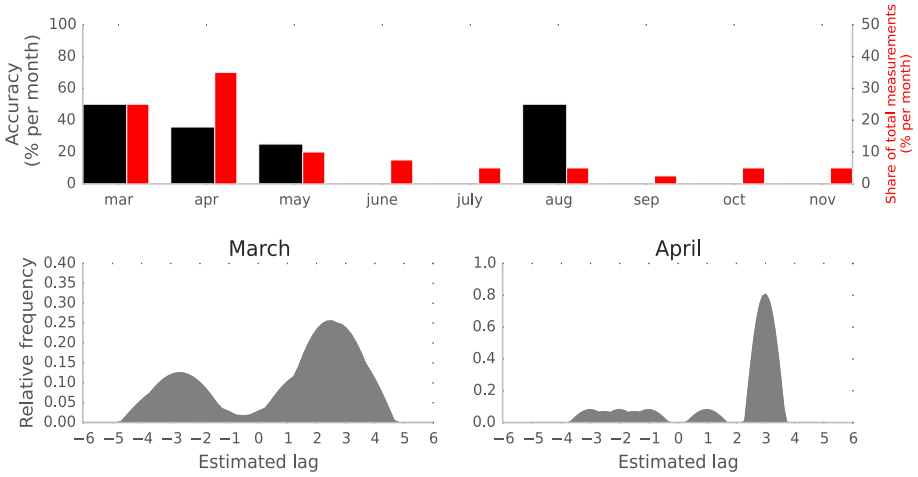

Fig. 11. The algorithm's results for soil mineral N prediction at the second Terrington site. The top graph shows the accuracy measure of the model's soil mineral N predictions for each month and when time lags are considered (in black), and the percentage of total measurements that were taken in each month (in red). The following two graphs show the frequency distribution of the time lags that were estimated by the algorithm for March and April. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

patterns in the measured data. The accuracy measure represents a strict method to assess a model's performance in relation to the measured data and, at the same time, take into account the fact that daily measured data can have significant variability. It should be noted though, that the value of the accuracy measure has to be juxtaposed with the RMSE (and MSE) when attempting to draw conclusions about a model's behaviour. This is mainly because the measured range of the daily soil N_2O data (i.e. standard error, standard deviation) can sometimes be so wide as to produce a misleadingly high value for the accuracy measure.

Using measured data from two arable sites in the UK we have shown that lags can have significant impact on model evaluation and can affect both the level of correlation between measured and simulated data and the magnitude of the sums of the residuals. Also, we used the division of MSE to three constituent statistics (SB, SDSD and LCS) to show how the level of correlation can affect the sum of residuals. By dividing the algorithm-predicted series of lag values into monthly sets and examining the frequency distribution of the lags, certain patterns in these temporally patchy series have been identified. A challenging task in relation to time lags between observed and simulated daily data, is to determine their cause. This task becomes more difficult for model outputs such as soil N_2O emissions that are driven by various interacting variables. Even more so, because the measured N_2O datasets and the measured datasets of their drivers (e.g. soil moisture, soil N content) cover small time periods, they are not continuous and can vary widely in size. In this study we implemented the algorithm using measured and simulated data for soil moisture (first and second example) and soil mineral N (second example), and compared its results with the respective results for N_2O. In our first example, we showed that the estimated lags in N_2O prediction are related to the lags in soil moisture prediction in a way that changes gradually through time. In our second example, the lags in N_2O prediction were explained by the lags in soil moisture and soil mineral N prediction, with which they had a positive relationship.

The time lags, as estimated by the algorithm, are caused by unknown emergent properties of the model. The result of these properties is that, for instance, as long as soil moisture is within the

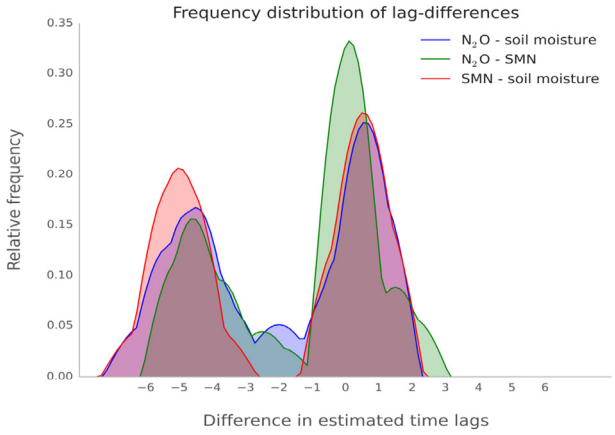

Fig. 12. Kernel density plots of the differences in estimated time lags in the model's prediction of N_2 O and soil moisture (blue), N_2O and soil mineral N (SMN) (green) and soil mineral N and soil moisture (red) during March and April at the second Terrington site. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

limits necessary for plant growth, the algorithm might be estimating a lag whose direction remains unchanged. If soil moisture for any reason (e.g. high rainfall event) falls outside these limits, the lag direction can change as a consequence of this change in an underlying process. The bimodal negatively skewed distribution, which can be seen in the results for our second example (Figs. 9–12), is typical of situations where the data are linked to two different processes. This can be an explanation of why both the lags in the prediction of the three variables and the differences between these lags have bimodal distributions. We may not be able to identify why the lags and their differences have bimodal distributions but we are able to make an important observation about the lags in the prediction of the three variables. The lags in the prediction of daily soil moisture and soil mineral N are causing similar lags to the prediction of daily N_2O emissions. Because the model explicitly describes the vertical movement of water and N in the soil, we can also argue that the lags in soil moisture prediction are causing the lags in soil mineral N prediction.

In our two examples, we have used a 3-days window around each measurement day and have implemented the algorithm in order to identify which of the six possible lags corresponds to a simulated value that is closer to the actual measured value. This choice of lag size was based on the average distance between two consecutive daily N_2O data points in the datasets that were used. Selecting a larger lag size would have complicated the interpretation of the algorithm's results significantly. On the other hand, it would have also led to more unimodal distributions of lags rather than the bimodal and bimodal negatively skewed that were presented in our results. Some of the negative values that appear in the monthly sets of lags, where for example the distribution mode is positive, are caused by the fact that the +3-days upper limit, the temporal closeness of certain measured data points and the algorithm's rule of non-duplication of simulated data can force the algorithm to select a negative lag simply because all the positive options (i.e. +1,+2,+3) have been excluded from being used during a previous step (see Figs. 1 and 3). In this paper, we quantified the sensitivity of the algorithm's results to the chosen time lag window by reimplementing the algorithm while imposing a ± 1 day deviation to the chosen time lag window (i.e. ± 3). The algorithm's results appear to be more sensitive to the choice of time lag window in the first example than in the second. This observation is in line with the fact that our analysis of estimated lags showed the existence of unstable lag patterns in the first example (Fig. 7) as opposed to the second example (Fig. 12).

It should be noted that the size and the quality of the measured data, that drive the algorithm, play a key role to the robustness of its results. Larger datasets, which include data with low variabilities, can produce results which are easier to interpret. The conclusions drawn in this study apply only to the model and the two sites used in our examples, which is suggestive of the need to implement the algorithm using additional and larger measured datasets.

The further use of the information that the algorithm provides, in ways that can lead to reductions in the estimated lags, is a process that is linked more to model development than model evaluation and was outside the scope of this study. Nevertheless, we believe that the algorithm can offer new perspectives to model development. In addition to the consideration of uncertainties in the measured data during the evaluation process, the possible existence of stable lag patterns across certain variables of interest (e.g. soil moisture and N_2O) can form a basis for focusing model development interventions on how the model links certain modelled forcing variables (e.g. modelled soil moisture) to certain modelled dependent variables (e.g. modelled N_2O). In this way, model improvement can become more targeted while remaining based on information derived from measured datasets (i.e. used to justify the interventions and assess their impacts).

5. Conclusions

Model evaluation in relation to soil N_2O emissions can be negatively affected by uncertainties in measured data and by time lags between the simulated and measured data. Time lags can be spotted through the visual assessment of a model's N_2O prediction but this is a subjective approach. In this study we presented a new model evaluation algorithm that can become part of the evaluation process in order to consider the uncertainty in measured data and quantify the impacts of time lags on different evaluation metrics. It is a well grounded and useful approach on model evaluation against soil N_2O data as well as against data for other variables which are measured sporadically (e.g. soil mineral N, soil moisture, ammonia etc).

It is important to note that the algorithm's effectiveness is constrained by the size, variety and quality of the measured data. In this paper we used measured data from two UK arable sites and the results of a single model, therefore, our conclusions are specific to that model and those two sites. The further use of the algorithm with more extensive measured data from different types of agro-ecosystems as well as the use of different models, is needed. We aspire that a more widespread use of the algorithm will contribute to the refinement of its underlying concepts and increase its applicability. In order to facilitate this procedure the algorithm's code (written in python 2.7) is freely available upon request.

Acknowledgements

The authors are grateful to the Scottish Government for supporting this work. The collection of data at the first Terrington site (2004) was carried out by ADAS and sponsored by the UK Department for Environment and Rural Affairs (DEFRA) through the NT26 programme. The collection of data at the second Terrington site (2011) was carried out by ADAS and sponsored by DEFRA and the Scottish Government through Sustainable Arable LINK Project LK09128, and we acknowledge the contributions of ADAS, Agricultural Industries Confederation, Bayer CropScience, British Sugar, Country Land and Business Association, The Co-operative Group, Frontier Agriculture, GrowHow UK, AHDB-HGCA, Hill Court Farm Research, NFU, North Energy Associates, PGRO,

Renewable Energy Association, Rothamsted Research (North Wyke), Scotch Whisky Research Institute, SoilEssentials, SRUC, Vivergo fuels, Warburtons and Yara UK.

References

Ali, M.H., Abustan, I., 2014. A new novel index for evaluating model performance. J. Nat. Resour. Dev. 1–9.

Anfossi, D., Castelli, S.T., 2014. Atmospheric tracer experiment uncertainties related to model evaluation. Environ. Model. Softw. 51 (C), 166–172.

Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K., 2010. Validation of biophysical models: issues and methodologies. A review. Agron. Sustain. Dev. 30 (1), 109–130.

Bellocchi, G., Rivington, M., Matthews, K., 2015. Deliberative processes for comprehensive evaluation of agroecological models. A review. Agron. Sustain. 30 (1), 109–130.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andréassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40 (c), 1–20.

Butterbach-Bahl, K., Dannenmann, M., 2011. Denitrification and associated soil N. Curr. Opin. Environ. Sustain. 3 (5), 389–395.

Cardenas, L.M., Thorman, R., Ashlee, N., Butler, M., Chadwick, D., Chambers, B., Cuttle, S., Donovan, N., Kingston, H., Lane, S., Dhanoa, M.S., Scholefield, D., 2010. Quantifying annual N_2O emission fluxes from grazed grassland under a range of inorganic fertiliser nitrogen inputs. Agric. Ecosyst. Environ. 136 (3–4), 218–226.

Chadwick, D.R., Cardenas, L., Misselbrook, T.H., Smith, K.A., Rees, R.M., Watson, C.J., McGeough, K.L., Williams, J.R., Cloy, J.M., Thorman, R.E., Dhanoa, M.S., 2014. Optimizing chamber methods for measuring nitrous oxide emissions from plot-based agricultural experiments. Eur. J. Soil Sci. 65 (2), 295–307.

Chapuis-Lardy, L., Wrage, N., Metay, A., Chotte, J.-L., Bernoux, M., 2007. Soils, a sink for N_2 O? A review. Glob. Change Biol. 13 (1), 1–17.

Cowan, N.J., Famulari, D., Levy, P.E., Anderson, M., Bell, M.J., Rees, R.M., Reay, D.S., Skiba, U.M., 2014. An improved method for measuring soil N_2O fluxes using a quantum cascade laser with a dynamic chamber. Eur. J. Soil Sci. 65 (5), 643–652.

Daniel, W.W., Cross, C.L., 2012. Biostatistics: A Foundation for Analysis in the Health Sciences, tenth ed. A Foundation for Analysis in the Health Sciences, Wiley Global Education.

Dobbie, K.E., Smith, K.A., 2003. Nitrous oxide emission factors for agricultural soils in Great Britain: the impact of soil waterâĂŘfilled pore space and other controlling variables. Glob. Change Biol. 9 (2), 204–218.

Duveiller, G., Fasbender, D., Meroni, M., 2016. Revisiting the concept of a symmetric index of agreement for continuous datasets. Sci. Rep. 6, 19401.

Galloway, J.N., Aber, J.D., Erisman, J.W., Seitzinger, S.P., Howarth, R.W., Cowling, E.B., Cosby, B.J., 2003. The nitrogen cascade. Bioscience 53 (4), 341–356.

Giltrap, D.L., Li, C.L., Saggar, S., 2010. DNDC: a process-based model of greenhouse gas fluxes from agricultural soils. Agric. Ecosyst. Environ. 136 (3–4), 292–300.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: the challenge of feeding 9 billion people. Science 327 (5967), 812–818.

Gu, J., Loustau, D., Henault, C., Rochette, P., Cellier, P., Nicoullaud, B., Grossel, A., Richard, G., 2014. Modeling nitrous oxide emissions from tile-drained winter wheat fields in central france. Nutrient Cycl. Agroecosyst. 98 (1), 27–40. http://dx.doi.org/10.1007/s10705-013-9593-6.

Haas, E., Klatt, S., Fröhlich, A., Kraft, P., Werner, C., Kiese, R., Grote, R., Breuer, L., Butterbach-Bahl, K., 2012. LandscapeDNDC: a process model for simulation of biosphere–atmosphere–hydrosphere exchange processes at site and regional scale. Landsc. Ecol. 28 (4), 615–636.

Holzworth, D.P., Snow, V., Janssen, S., Athanasiadis, I.N., Donatelli, M., Hoogenboom, G., White, J.W., Thorburn, P., 2014. Environmental Modelling & Software. Environ. Model. Softw. 1–11.

Kobayashi, K., Salam, M.U., 2000. Comparing simulated and measured values using mean squared deviation and its components. Agron. J. 92 (2), 345–352.

Marschner, P., Rengel, Z. (Eds.), 2007. Nutrient Cycling in Terrestrial Ecosystems, vol. 35. Springer, Berlin.

Martorana, F., Bellocchi, G., 1999. A review of methodologies to evaluate agro-ecosystem simulation models. Italian J. Agron. 3 (1), 19–40.

Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecol. Model. 68 (1), 21–32.

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, validation, and confirmation of numerical models in the Earth sciences. Science 263 (5147), 641–646.

Perlman, J., Hijmans, R.J., Horwath, W.R., 2013. Modelling agricultural nitrous oxide emissions for large regions. Environ. Model. Softw. 48 (C), 183–192.

Richter, K., Atzberger, C., Hank, T.B., Mauser, W., 2012. Derivation of biophysical variables from Earth observation data: validation and statistical measures. J. Appl. Remote Sens. 6 (1), 063557–1–25.

Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45.

Sanna, M., Bellocchi, G., Fumagalli, M., Acutis, M., 2015. A new method for analysing the interrelationship between performance indicators with an application to agrometeorological models. Environ. Model. Softw. 73 (C), 286–304.

Smith, J., Smith, P., 2007. Environmental Modelling: An Introduction. OUP, Oxford. URL. http://books.google.co.uk/books?id=RIENngEACAAJ.

Smith, K.A., Dobbie, K.E., Thorman, R., Watson, C.J., Chadwick, D.R., Yamulki, S., Ball, B.C., 2012. The effect of N fertilizer forms on nitrous oxide emissions from UK arable land and grassland. Nutrient Cycl. Agroecosyst. 93 (2), 127–149.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Stat. Assoc. 69 (347), 730.

Tedeschi, L.O., 2006. Assessment of the adequacy of mathematical models. Agric. Syst. 89 (2–3), 225–247.

Theil, H., 1970. Economic Forecast and Policy, Vol. XV of Contributions to Economic Analysis, second ed. North-Holland Publishing Company, Amsterdam, The Netherlands.

Thorman, R.E., Smith, K.E., Rees, R.M., Chauhan, M., Bennett, G., Malkin, S., Munro, D.G., Sylvester-Bradley, R., 2013. Nitrous oxide emissions associated with nitrogen use on arable crops in England. Int. Fertil. Soc. Proc. 715 (3–4), 1–42.

Whitmore, A.P., 1991. Method for assessing the goodness of computer simulation of soil processes. Eur. J. Soil Sci. 42 (2), 289–299.

Willems, P., 2009. A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. Environ. Model. Softw. 24 (3), 311–321.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., Rowe, C.M., 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. Oceans (1978–2012) 90 (C5), 8995–9005.

Willmott, C.J., Robeson, S.M., Matsuura, K., 2011. A refined index of model performance. Int. J. Climatol. 32 (13), 2088–2094.