
© 2016 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

http://hdl.handle.net/11262/11071
http://dx.doi.org/10.1016/j.envsoft.2016.07.002
Model evaluation in relation to soil N$_2$O 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 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$_2$O) 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$_2$O 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$_2$O 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.

© 2016 Published by Elsevier Ltd.

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; Bellocci 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...
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 (Bellocci 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; Bellocci 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 (N₂O) from cultivated soils. N₂O 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. N₂O 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 microb mes and driven by the use of nitrogenous fertilisers and by environmental conditions (Galloway et al., 2003). Nevertheless, some aspects of N₂O production in soils are not fully understood due to the complex role of soil microbes (Butterbach-Bahl and Dannenmann, 2011).

N₂O samples are typically collected using manual or automatic chambers. Despite the limitations and weaknesses, this method is widely applied and the derived N₂O 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 N₂O 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 N₂O 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 (Bellocci et al., 2010; Tedeschi, 2006; Martorana and Bellocci, 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 N₂O 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. Bellocci et al., 2010 provide an 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 N₂O fluxes.

1. Deviation based methods:
   (a) Positive and negative residuals can cancel each other out and produce unrealistic statistical values.
   (b) Negative N₂O 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 N₂O peaks can have both a short duration and a large magnitude. If a model has missed a measured peak of N₂O 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 N₂O 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
<th>Formula</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviation based</td>
<td>Bias</td>
<td>( Bias = \frac{1}{n} \sum_{i=1}^{n} (O_i - \bar{O}) )</td>
<td>Smith and Smith (2007)</td>
</tr>
<tr>
<td></td>
<td>Relative Bias</td>
<td>( RelB = \frac{n \sum_{i=1}^{n} (O_i - \bar{O})^2}{\sum_{i=1}^{n} (O_i - \bar{O})} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Fractional Bias</td>
<td>( FB = \frac{2 \sum_{i=1}^{n} (O_i - \bar{O})}{\sum_{i=1}^{n} O_i} )</td>
<td>Sanna et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>Coefficient of Residual Mass</td>
<td>( CRM = \frac{n \sum_{i=1}^{n} (O_i - \bar{O})^2}{\sum_{i=1}^{n} (O_i - \bar{O})} )</td>
<td>Sanna et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>Percent Bias</td>
<td>( PB = 100 \cdot CRM )</td>
<td>Sanna et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>Relative Error</td>
<td>( E = \frac{\text{SDs}}{\text{RMSE}} \sqrt{\sum_{i=1}^{n} (O_i - \bar{O})} )</td>
<td>Smith and Smith (2007)</td>
</tr>
<tr>
<td></td>
<td>Mean Absolute Error</td>
<td>( MAE = \frac{1}{n} \sum_{i=1}^{n} (O_i - S_i) )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Mean Squared Error</td>
<td>( MSE = \frac{1}{n} \sum_{i=1}^{n} (S_i - O_i)^2 ) and</td>
<td>Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993); Kobayashi and Salam (2000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( MSE = \text{Bias}^2 + \text{SDS}^2 + \text{LCS} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Root Mean Squared Error</td>
<td>( RMSE = \sqrt{MSE} )</td>
<td>Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)</td>
</tr>
<tr>
<td></td>
<td>Normalised Root Mean Squared Error</td>
<td>( NRMSE = \frac{RMSE}{\bar{O}} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relative Mean Squared Error</td>
<td>( RRMSE = 100 \cdot MSE )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standard Error of Prediction Corrected for bias</td>
<td>( SEPC = \frac{\sqrt{\sum_{i=1}^{n} (O_i - MBE)^2}}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})}} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Lack of Correlation weighted by the Standard deviations</td>
<td>( LCS = 2 \cdot \text{SDs} ) ( \cdot (1-r) )</td>
<td>Kobayashi and Salam (2000)</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation of the Observations</td>
<td>( SD_O = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Kobayashi and Salam (2000)</td>
</tr>
<tr>
<td></td>
<td>Standard Deviation of the Simulations</td>
<td>( SD_S = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (S_i - \bar{S})^2} )</td>
<td>Kobayashi and Salam (2000)</td>
</tr>
<tr>
<td></td>
<td>Squared Difference between Standard Deviations</td>
<td>( SDS = (SD_O - SD_S)^2 )</td>
<td>Kobayashi and Salam (2000)</td>
</tr>
<tr>
<td></td>
<td>Ratio of Standard Deviation of observations to RMSE</td>
<td>( RDP = \frac{SD_O}{RMSE} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Nondimensional error index</td>
<td>( NDI = \frac{RMSE}{\bar{O}} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Modelling Efficiency</td>
<td>( EF = 1 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Sanna et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>Nash Sutcliffe Efficiency</td>
<td>( NSE = 1 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Agreement Coefficient</td>
<td>( AC = 1 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Willmott’s index of agreement</td>
<td>( d = 1 - \frac{\sum_{i=1}^{n} (O_i - S_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Willmott et al. (2011, 1985); Richter et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Refined index of agreement</td>
<td>( dr(1) = \frac{1}{\sum_{i=1}^{n} (O_i - S_i)^2} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \sum_{i=1}^{n} [O_i - \bar{O}] \leq C \sum_{i=1}^{n} [O_i - S_i] )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>else: ( dr(2) = \frac{1}{\sum_{i=1}^{n} (O_i - \bar{O})^2} )</td>
<td>Willmott et al. (2011, 1985); Richter et al. (2012)</td>
</tr>
<tr>
<td>Regression based</td>
<td>Pearson’s correlation coefficient</td>
<td>( r = \frac{\sum_{i=1}^{n} (O_i - \bar{O})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (O_i - \bar{O})^2 \cdot \sum_{i=1}^{n} (S_i - \bar{S})^2}} )</td>
<td>Smith and Smith (2007)</td>
</tr>
<tr>
<td></td>
<td>Coefficient of determination</td>
<td>( R^2 = r^2 )</td>
<td>Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)</td>
</tr>
<tr>
<td></td>
<td>Regression (slope/intercept)</td>
<td>( S_i = m \cdot O_i + b )</td>
<td>Richter et al. (2012); Theil (1970)</td>
</tr>
<tr>
<td></td>
<td>Theil-Sen (nonparametric)</td>
<td>( S_i = m \cdot O_i + b )</td>
<td>Smith and Smith (2007)</td>
</tr>
<tr>
<td></td>
<td>t-value</td>
<td>( t = \sqrt{n - 2 \cdot \sum_{i=1}^{n} (O_i - \bar{O})^2 / (n - 1)} )</td>
<td>Richter et al. (2012)</td>
</tr>
<tr>
<td>Probability based</td>
<td>Root Mean Squared Error at 95% CI</td>
<td>( \text{RMSE}<em>{95} = \frac{100}{\text{RRMSE}} \sqrt{\sum</em>{i=1}^{n} (S_i - \text{RMSE}_{95})^2} )</td>
<td>Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)</td>
</tr>
<tr>
<td></td>
<td>Relative Error at 95% CI</td>
<td>( E_{95} = \frac{100}{\text{RRMSE}} \sqrt{\sum_{i=1}^{n} (S_i - \text{RMSE}_{95})^2} )</td>
<td>Smith and Smith (2007); Tedeschi (2006); Martorana and Bellocchi (1999); Mayer and Butler (1993)</td>
</tr>
<tr>
<td></td>
<td>Significance of difference between measured and simulated values</td>
<td>( LQFT = \frac{\sum_{i=1}^{n} (S_i - O_i)^2}{\text{RMSE}_{95}^2} )</td>
<td>Smith and Smith (2007)</td>
</tr>
</tbody>
</table>

\( O_i \): the measured data; \( S_i \): the mean of the measured data; \( O_i \): the measured data; \( \bar{O} \): 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 \( O_i \); 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 = 0.05).
(b) Are insensitive to additive and proportional differences between simulations and measurements
(c) Can produce large coefficient of determination (R²) 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 actually 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 N₂O. 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 direction of magnitude change between successive measurement days and comparing it with the directional correlation between observed and simulated data (Duveiller et al., 2016). Winter wheat was the crop that was planted and harvested during both experiments. At both sites N₂O 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 N₂O 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. N₂O samples were analysed in the laboratory by gas chromatography (Cardenas et al., 2010). Gravimetric topsoil moisture content was measured on every N₂O 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/cm³, a clay content of...
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 N2O 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 N₂O 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/cm³, a clay content of 11%, a pH of 8.3 and an organic carbon content of 1.8%

---

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.
nitrates (AN) fertiliser was applied in three doses. The measured data that are used in this study consist of 40 daily soil N$_2$O 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$_2$O for each day.

Fig. 4 presents a graph of the daily measured and simulated N$_2$O 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...
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 N2 O predictions (accuracy measure) is gradually improving from March to May. Most of the estimated lags in N2O prediction are positive in March and negative in April while there is a clearly positive lag in the simulated N2O values in May (Fig. 5).

Soil moisture is a major driver of N2O 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 N2O 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 N2O measurements, we used equation (1) to calculate the difference between the respective estimated lags for each day of measurement.

\[
\text{Lag Difference} = \frac{\text{Lag}_{N2O} \cdot \text{Lag}_{m}^{-1}}{|\text{Lag}_{N2O} - \text{Lag}_{m}|}
\]

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

Fig. 4. Measured and simulated soil N2O 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.

Fig. 5. The algorithm’s results for N2O for the urea treatment at the first Terrington site. The top graph shows the accuracy measure of the model’s N2O 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 N2O 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. N2O) 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 occurred 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 N2O 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 (i.e. the lag-difference becomes negative).

3.2. Second Terrington site

For the second example we implemented the algorithm using the measured N2O 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 N2O measurements were used to define the range of measured N2O 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 N2O 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. 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.)

Fig. 7. Kernel density plots of the difference between the estimated time lags in the model’s soil moisture and N2O predictions for March, April and May at the first Terrington site.
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 N₂O 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 N₂O 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 N₂O, 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 N₂O and soil moisture, N₂O 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 N₂O 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 N₂O 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...
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$_2$O 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$_2$O emissions that are driven by various interacting variables. Even more so, because the measured N$_2$O 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$_2$O. In our first example, we showed that the estimated lags in N$_2$O 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$_2$O 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
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 \( \text{N}_2\text{O} \) 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 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 \( \text{N}_2\text{O} \) 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 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 \( \text{N}_2\text{O} \)) can form a basis for focusing model development interventions and assess their impacts.

5. Conclusions

Model evaluation in relation to soil \( \text{N}_2\text{O} \) 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 \( \text{N}_2\text{O} \) 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 \( \text{N}_2\text{O} \) 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


