Environmental Modelling & Software 62 (2014) 487–494

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

The error in agricultural systems model prediction depends on the variable being predicted☆

Daniel Wallach [a, *], Peter J. Thorburn [b]

[a] INRA, UMR 1248 Agrosystèmes et développement territorial (AGIR), 31326 Castanet-Tolosan Cedex, France
[b] CSIRO Ecosystem Sciences, GPO Box 2583, Brisbane, Queensland 4001, Australia

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form
31 July 2014
Accepted 4 August 2014
Available online 29 August 2014

Keywords:
APSIM
Error
GHG emissions
Irrigation
Maize
Nitrous oxide
Uncertainty

ABSTRACT

Models are generally evaluated based on the squared error of model predictions compared with individual data. However, if major interest is in some quantity averaged over time or space it would be more pertinent to evaluate how well the model predicts this average quantity. We show that the model squared error for predictions averaged over space or time will always be smaller than average squared prediction error and how to estimate the difference between the two, using commonly available data. We illustrate with two case studies concerning irrigation management, (where major interest is in yield averaged over years) and nitrous oxide emissions (where major interest is in emissions averaged over a growing season). Squared error of the average was estimated to be only 57% and 10% of the average squared error for the irrigation and nitrous oxide emissions studies, respectively, in the limit of averaging over long times.

1. Introduction

The worldwide agricultural sector faces the significant challenge of increasing production to provide food security for a population projected to rise to 9 billion by mid-century (Godfray et al., 2010; Tilman et al., 2011) while protecting the environment and the functioning of ecosystems (Carberry et al., 2013). The worldwide agriculture sector faces these challenges while needing to both adapt to climate change (Howden et al., 2007), including adapting to increased water demand and reduced water availability (Turral et al., 2011), and to mitigate emissions of greenhouse gasses (Smith et al., 2008). Dynamic system models will play an increasingly important role in examining strategies to meet these unprecedented challenges, and there is a clear need to improve these models (Rötter et al., 2011).

Model evaluation is an important aspect of modeling in general, including integrated environmental modeling, water resources modeling and crop modeling (Laniak et al., 2013; Bennett et al., 2013; Warmink et al., 2010; Refsgaard et al., 2007). Multiple aspects of model evaluation and uncertainty have been considered (Confalonieri et al., 2010; Varella et al., 2010; Saltelli and Annoni, 2010), and specialized software developed (Wallach et al., 2011). Often however model error is not treated in detail during model development and calibration (Ahuja and Ma, 2011) although there is increasing attention being paid to understanding uncertainty of predictions from agricultural models (Rosenzweig et al., 2013; Asseng et al., 2013).

Evaluation of model performance is usually based on comparing individual model predictions with corresponding data (i.e. one predicted value with a corresponding measurement). For example, one examines how well the model predicts yield in a given year (e.g., Rötter et al., 2012; Asseng et al., 2013), or how well the model predicts nitrous oxide (N_2O) emissions from soil (Frolking et al., 1998; Del Grosso et al., 2008; Thorburn et al., 2010a) or nitrate leaching on a given day (Thorburn et al., 2011; Cichota et al., 2013). Often, however, the individual predictions do not exactly correspond to the major quantities of interest for decision-makers and other output-oriented stakeholders. Major interest might be on the probability of extreme events, or on some output averaged over space or time, or on the stability of yield over time. Here we focus on the case where major interest is in some quantity averaged over time or space. For instance, in evaluating the impact of climate change, often in practice one estimates future yield averaged over several decades (Rosenzweig and Parry, 1994; Ferrara et al., 2010; White et al., 2011). In comparing management strategies, we are again often interested in performance averaged over future

☆ Thematic Issue on Agricultural Systems Modelling & Software.
* Corresponding author.
E-mail address: wallach@toulouse.inra.fr (D. Wallach).

http://dx.doi.org/10.1016/j.envsoft.2014.08.001
1364-8152/

weather (Wallach et al., 2012; Biggs et al., 2013). In evaluating greenhouse gas emissions, we are often most interested in total emissions over some period, for example over the period of crop growth or a year (Smith et al., 2008). In studying the impact of farming systems on the use of, or effect on regional resources, particularly water, we are often interested in averaging over weather and summing over some region (Carroll et al., 2012). If major interest is in quantities averaged over space and/or time, then it would be more pertinent to evaluate how well the model predicts the average quantity of interest rather than to evaluate the model with respect to individual values. This will be especially important if the difference between prediction quality in the two cases is large.

There does not seem to have been any attempt to look in detail at the differences between prediction accuracy for individual quantities and prediction accuracy for averaged quantities. Specifically, we would like to know if there is some general relationship between how well models predict averages and how well they predict individual quantities. Secondly, what determines the differences in errors when predicting averages and when predicting individual quantities? Thirdly, can we use commonly available data to estimate the difference in errors between predicting averages compared to predicting individual results?

The purpose of this paper then is to study the above three questions. We first examine these questions theoretically, and then consider two case studies, namely predicting yield averaged over years for different irrigation strategies for maize, and predicting N_2O emissions averaged over days in different cropping systems.

2. Materials and methods

2.1. Theory

We consider the situation where we have contexts and repeated measurements of some response for each context. We are interested in averages over the repeated measures, for each context. For example, in one case study below the context is a site-management combination, and the repeated measurements refer to measurements of maize yield in several years. Here we are interested in model error averaged over years, for each site-management combination. In the second case study, the context is a site-year-management combination, and the repeated measurements are measurements of N_2O emissions on several days for each context. Here we are interested in model error averaged over days, for each site-year-management combination.

Let y_{ij} be the jth measured value for context i (e.g. the jth year of yield measurement for the ith site-management combination). Let $\hat{y}_{ij}$ be the corresponding simulated value, using some model. We can always write.

$$y_{ij} = \hat{y}_{ij} + \varepsilon_{ij} \tag{1}$$

which simply says that model error, i.e. the difference between the measured and simulated values, is noted ε_{ij}.

In terms of the above notation, the average squared error for context i is.

$$\text{aveSE}_i = (1/J)\sum_{j=1}^{J} \varepsilon_{ij}^2 \tag{2}$$

where J is the total number of measurements for context i. The squared error of the average for context i (also referred to as squared bias) is

$$\text{SEave}_i = \left[(1/J)\sum_{j=1}^{J} y_{ij} - (1/J)\sum_{j=1}^{J} \hat{y}_{ij}\right]^2 = (1/J)^2\left[\sum_{j=1}^{J} \varepsilon_{ij}\right]^2 \tag{3}$$

The quantity aveSE is average model error for individual measurements, while SEave measures how well the model predicts the average over measurements for a given context.

According to the Cauchy–Swartz inequality (Cvetkovski, 2010, Theorem 4.2 but with a slight change in notation):

$$\sum_{j=1}^{J} a_j^2 \sum_{j=1}^{J} b_j^2 \geq \left(\sum_{j=1}^{J} a_j b_j\right)^2$$

where the a_j and b_j are arbitrary numbers. Setting all the $b_j = 1$ and $a_j = \varepsilon_{ij}$ with fixed i, the inequality becomes

$$J\sum_{j=1}^{J} \varepsilon_{ij}^2 \geq \left(\sum_{j=1}^{J} \varepsilon_{ij}\right)^2$$

with equality only if all the model errors for context i are equal, which is never true. In terms of the quantities defined above this says that.

$$\text{aveSE}_i - \text{SEave}_i \geq 0 \tag{4}$$

That is, it is always true that the average squared error for the individual measurements is greater than or equal to the squared error of the average of the measurements. The inequality of Eq. (4) can also be deduced from the fact that average squared error can be decomposed into 3 positive terms, one of which is squared bias (Kobayashi and Salam, 2000).

We are interested in the difference between aveSE and SEave, and in the behavior of SEave as a function of J. This will depend on the variability of ε_{ij} for given i. To describe this variability we propose to use a mixed effects model, with a bias term that can be a nonlinear function of the explanatory variables in the original model and a random context effect. The general form of the model that we propose is.

$$\varepsilon_{ij} = g(X_{ij};\theta) + \alpha_i + \tau_{ij} \tag{5}$$

The fixed effect $g(X_{ij};\theta)$ can be a function of explanatory variables X_{ij} and parameters θ. If there is some structure in the errors ε_{ij} as a function of X_{ij}, this will be embodied in the fixed effect. If no structure is apparent, then the fixed effect will just be a constant. The examples analyzed below cover both the case with and without structure. α_i is the random effect related to the context and τ_{ij} is residual error.

We assume that all the α_i have identical normal distributions with expectation 0 and that all the τ_{ij} also have identical normal distributions with expectation 0. That is.

$$\begin{aligned} \alpha_i &\sim N\left(0, \sigma_\alpha^2\right) \\ \tau_{ij} &\sim N\left(0, \sigma_\tau^2\right) \end{aligned} \tag{6}$$

As discussed below, the assumption that random effects are normally distributed is often not critical. We further assume that all the random context effects α_i are independent (i.e. knowing the random effect for one context gives no information about the random effect for other contexts), that given α_i all the τ_{ij} for that context i are independent (i.e. once one has taken into account the random context effect, there is no systematic relation between the residual errors for that context), and finally that the α_i and τ_{ij} are all mutually independent (i.e. there is no systematic relation between the random context effect and the residual errors).

We now look at the expectation of aveSE and SEave, where the expectation is over contexts and over repeated measurements within a context. According to our assumptions $E(\alpha_i^2) = \sigma_\alpha^2$, $E(\tau_{ij}^2) = \sigma_\tau^2$, $E(\tau_{ij}\tau_{ij'}) = 0$ for $ij \neq i'j'$ and $E(\alpha_i\tau_{ij}) = 0$ for all ij. Then the expectations of interest are

$$E(\text{aveSE}_i) = E\left[(1/J)\sum_{j=1}^{J}\left(g(X_{ij};\theta) + \alpha_i + \tau_{ij}\right)^2\right] = (1/J)\sum_{j=1}^{J} g(X_{ij};\theta)^2 + \sigma_\alpha^2 + \sigma_\tau^2 \tag{7}$$

$$\begin{aligned} E(\text{SEave}_i) &= E\left[(1/J)^2\left[\sum_{j=1}^{J}\left(g(X_{ij};\theta) + \alpha_i + \tau_{ij}\right)\right]^2\right] \\ &= \left(1/J^2\right)\left[\sum_{j=1}^{J} g(X_{ij};\theta)\right]^2 + \sigma_\alpha^2 + \sigma_\tau^2/J \end{aligned} \tag{8}$$

If $g(X_{ij};\theta)$ is just a constant, say $g(X_{ij};\theta) = \mu$, then the above equations simplify to

$$E(\text{aveSE}_i) = \mu^2 + \sigma_\alpha^2 + \sigma_\tau^2 \tag{9}$$

and

$$E(\text{SEave}_i) = \mu^2 + \sigma_\alpha^2 + \sigma_\tau^2/J. \tag{10}$$

The first two terms on the right hand side of Eqs. (9) and (10) are identical. That is, the fixed bias (the term μ^2), and also the random context specific bias (the term σ_α^2), contribute equally to $E(\text{aveSE}_i)$ and to $E(\text{SEave}_i)$. The residual error on the other hand contributes σ_τ^2 to $E(\text{aveSE}_i)$ but only σ_τ^2/J to $E(\text{SEave}_i)$. In the more general case (Eqs. (7) and (8)), there is an additional difference between $E(\text{aveSE}_i)$ and to $E(\text{SEave}_i)$, due to the fact that $(1/J)\sum_{j=1}^{J} g(X_{ij};\theta)^2$ and $(1/J^2)[\sum_{j=1}^{J} g(X_{ij};\theta)]^2$ are in general not equal. The former is the average of squared bias for the different contexts, and the latter is the square of the overall average bias.

We can estimate θ, σ_α^2 and σ_τ^2 using standard software for estimating the parameters of a nonlinear mixed model. In the case studies here we use the lmer function which is part of the lme4 package of the R statistical software package (R Development Core Team., 2013). This function is based on the algorithm proposed in Lindstrom and Bates (1990), and uses restricted maximum likelihood (REML).