Article:

Wallach, D, Thorburn, P, Asseng, S et al. (5 more authors) (2016) Estimating model prediction error: Should you treat predictions as fixed or random? Environmental Modelling & Software, 84. pp. 529-539. ISSN 1364-8152

https://doi.org/10.1016/j.envsoft.2016.07.010

Estimating model prediction error: Should you treat predictions as fixed or random?

running head: Estimating prediction error

authors:

Daniel Wallach[1*], Peter Thorburn[2] ,Senthold Asseng[3], Andrew J. Challinor[4,5], Frank Ewert[6], James W. Jones[3], Reimund Rotter[7], Alex Ruane[8]

[1]INRA, UMR 1248 Agrosystèmes et développement territorial (AGIR), 31326 Castanet-Tolosan Cedex, France, email: daniel.wallach@toulouse.inra.fr. tel : 33 (0)561285567.

[2]CSIRO Agriculture Flagship, Dutton Park QLD 4102, Australia, email: peter.thorburn@csiro.au,

[3]Agricultural & Biological Engineering Department, University of Florida, Gainesville, FL 32611, USA, email: sasseng@ufl.edu & jimj@ufl.edu

[4]Institute for Climate and Atmospheric Science, School of Earth and Environment, University of Leeds, Leeds LS29JT, UK, email: a.j.challinor@leeds.ac.uk

[5]CGIAR-ESSP Program on Climate Change, Agriculture and Food Security, International Centre for Tropical Agriculture (CIAT), A.A. 6713, Cali, Colombia

[6]Institute of Crop Science and Resource Conservation INRES, University of Bonn, 53115, Germany, email: fewert@uni-bonn.de & eeyshire@uni-bonn.de

[7] Natural Resources Institute Finland (Luke), FI-50100 Mikkeli, Finland, email: reimund.rotter@luke.fi

[8]NASA Goddard Institute for Space Studies, New York, NY 10025, email: alexander.c.ruane@nasa.gov

[*] corresponding author

1

2

Highlights

- It is important to estimate the uncertainty in crop model predictions.
- Two uncertainty criteria are defined, treating predictions as fixed or random
- The random criterion includes model, parameter and input uncertainty and also bias
- The random prediction criterion is specific for each prediction situation

Abstract

Crop models are important tools for impact assessment of climate change, as well as for exploring management options under current climate. It is essential to evaluate the uncertainty associated with predictions of these models. We compare two criteria of prediction error; $MSEP_{fixed}$, which evaluates mean squared error of prediction for a model with fixed structure, parameters and inputs, and $MSEP_{uncertain}(X)$, which evaluates mean squared error averaged over the distributions of model structure, inputs and parameters. Comparison of model outputs with data can be used to estimate the former. The latter has a squared bias term, which can be estimated using hindcasts, and a model variance term, which can be estimated from a simulation experiment. The separate contributions to $MSEP_{uncertain}$ (X) can be estimated using a random effects ANOVA. It is argued that $MSEP_{uncertain}$ (X) is the more informative uncertainty criterion, because it is specific to each prediction situation.

keywords: crop model; uncertainty; prediction error; parameter uncertainty; input uncertainty; model structure uncertainty

1. Introduction

Crop models are important tools in agriculture and environment, including applications in crop breeding and crop management (Boote et al., 2010). A recent major focus is in using crop models to evaluate the impact of climate change on crop production and other crop responses (Rosenzweig et al., 2013).

As for all models, it is essential to estimate the uncertainty in crop model predictions, i.e. the extent to which predicted values may differ from the true values. There is increasing recognition in the crop modeling community that more attention needs to be paid to uncertainty in the crop models (Rötter et al., 2011; Rosenzweig et al., 2013). Recently, studies have been done using both multiple climate models and multiple crop models, as a way of evaluating uncertainty arising from both types of model. Preliminary evidence indicates that in fact, the uncertainty due to the variation between crop models may be larger than that due to climate models (Asseng et al., 2013), which emphasizes the importance of estimating crop model prediction uncertainty (Koehler et al., 2013). Estimating uncertainty is of primary importance for all uses of crop models, for example for exploring crop management options under current climate (Baigorria et al., 2007). It is also of major importance for models in other fields, including climate modeling (Holzkämper et al., 2015; Tebaldi and Knutti, 2007), environmental studies (Uusitalo et al., 2015) or hydrologic modeling (Refsgaard et al., 2006; Renard et al., 2010).

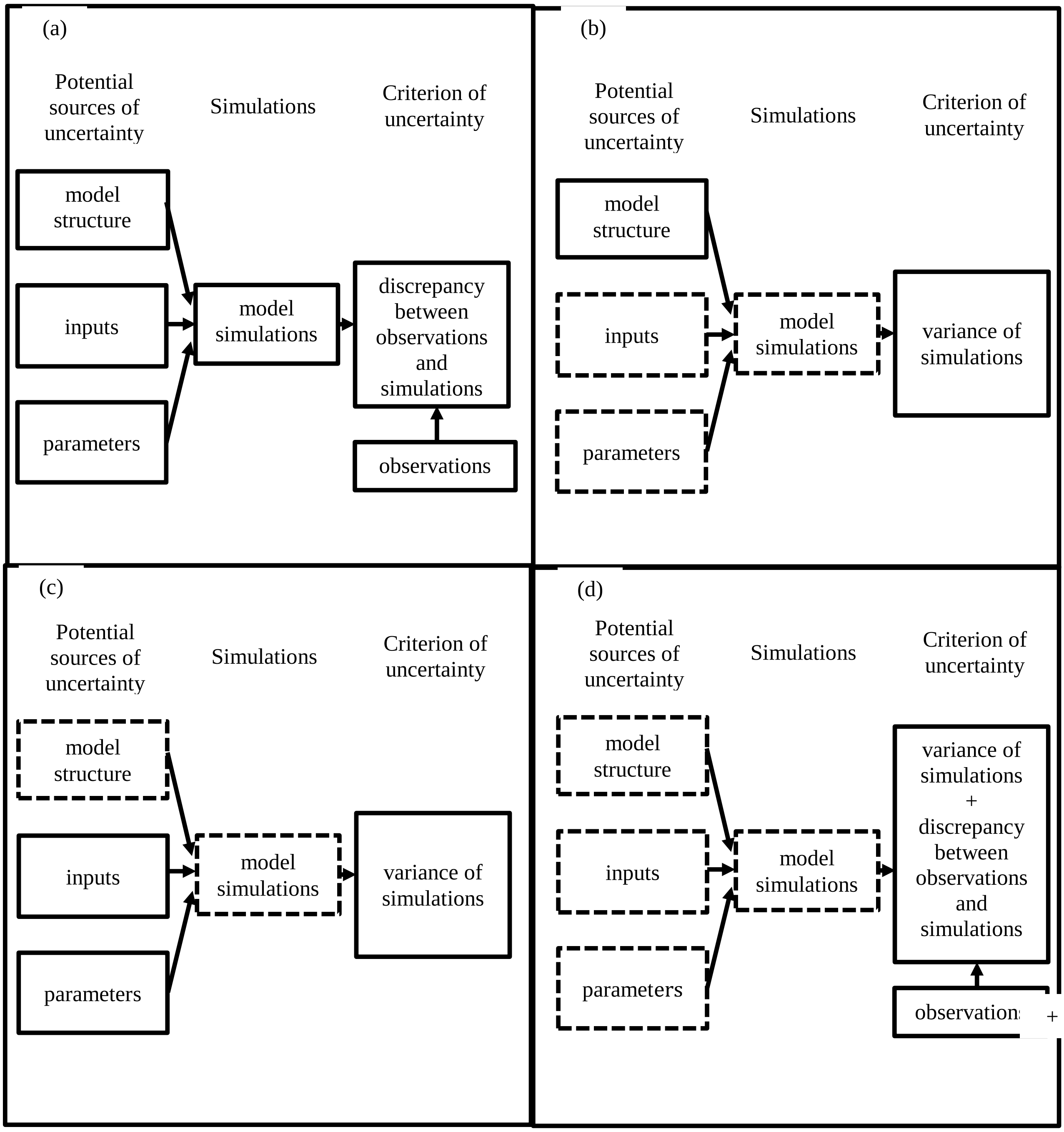

Figure 1

Schematic diagrams of different approaches to estimation of prediction uncertainty. a) Based on comparison of hindcasts with observations. b. Based on propagating input and/or parameter uncertainty through the model. c. Based on multi-model ensemble studies. d. Based on simulations with multiple model structures, multiple input vectors and multiple parameter vectors for each model. Elements that are explicitly treated as random are within a dash enclosed box.

Past crop model uncertainty studies can be grouped into three different approaches. The first is based on comparing model hindcasts to observed data (fig. 1a). A common measure of discrepancy is mean squared error, but there are many other possible measures of discrepancy, and there have been several studies devoted to examining and comparing them (Bellocchi et al., 2010; Bennett et al., 2013; Yang et al., 2014; Wallach et al., 2014). The use of hindcasts is the standard method of evaluating crop models and there have been numerous studies of this type, aiming to evaluate various models for various applications (Basso et al., 2016; Coucheney et al., 2015a). This is typically referred to by various terms, such as validation, verification and/or evaluation. The assumption is that the discrepancy between past observations and simulated values is an indication of the likely discrepancy in new predictions. That is, observed discrepancies are taken as a measure of uncertainty for predictions. There is no explicit treatment of the uncertainties in the model itself in this approach.

Short-term climate forecasts are often evaluated on the basis of skill scores, which compare some criterion of model fit with that of a naïve predictor ((Murphy 1988, Reichler and Kim 2008). This is comparable to the approach above. However, a major difference with crop models is that in general there is much more data available for testing climate models, although with remote sensing this may become less true. One result is that one can look at performance of climate models as a function of the prediction situation (geographical area, lead time), while evaluation of crop models is generally limited to estimating a single, average quality of prediction.

In a second approach (fig. 1b), the uncertainties in the model inputs or parameters are of primary concern. It is well understood that the values of the parameters in crop models are only approximations, and may have fairly large uncertainties (Dzotsi et al., 2013). Similarly, many of the input variables in crop models are difficult to estimate or measure and may have large uncertainties due to high spatial or temporal variability (Aggarwal, 1995; Bouman, 1994; Roux et al., 2014). This approach propagates the uncertainty in parameters and/or inputs through the crop model, in order to evaluate the resulting uncertainty in predictions. While these studies clearly evaluate an aspect of prediction uncertainty, the major objective is often elsewhere, namely to identify those factors (inputs or parameters) that contribute most to prediction uncertainty, using sensitivity analysis.

The third, more recent approach is based on multi-model ensembles (MMEs) (fig. 1c). For many crops multiple different crop models have been developed by different research teams. Models might for example differ in the way primary production or soil water or development rate is modeled. Model structure uncertainty is a major source of uncertainty in predictions, not only for crop models (Wintle et al., 2003) but for mechanistic models in general (Neuman, 2003). The variability between different crop models is taken as a measure of the prediction uncertainty due to uncertainty about model structure (Palosuo et al., 2011). This last approach is being used as the basis for assessing crop model prediction uncertainty in impact assessment studies (Asseng et al., 2013; Li et al., 2015).

In climate modeling, when considering climate model projections, emphasis has similarly been on uncertainty as represented by differences between models, or between different parameterizations or different initial values (Stainforth et al., 2005; Tebaldi and

Knutti, 2007). Here again however the situation is quite different than for crop models. There are no data for testing how well climate models will perform in a future, changed world, so uncertainty cannot be directly based on comparison with observations, whereas it is possible to test crop models against experiments that create conditions that may occur in the future but do not occur naturally, such as enhanced CO_2 levels (Biernath et al., 2011; Challinor and Wheeler, 2008; van Oijen and Ewert, 1999) or increased average or extreme temperatures (Asseng et al., 2004).

All of the approaches described above (comparison with hindcasts, propagation of input or parameter uncertainty, variability in multi-model ensembles) give information about crop model prediction uncertainty, but to date there have been no studies that attempt to relate them. It is important to do so, in order to obtain a better overall understanding of prediction uncertainty and how best to estimate it. We will focus on two sets of questions: (1) What are pertinent criteria of uncertainty, how can they be estimated, and how are the different approaches described above related to estimation of those criteria? (2) Given an overall criterion of uncertainty, how can one estimate the separate contributions from different sources of uncertainty?

The treatment of uncertainty here is applicable to modeling in any field. However, it should be particularly useful for crop modeling, to help interpret the multi-model ensemble studies that have become quite common recently.

2.Materials and Methods

2.1 A framework for quantitative measure of uncertainty

By "uncertainty" in model predictions we refer to the extent to which predicted values may differ from the true values. There are many ways in which this could be quantified. Most completely, one could specify the probability distribution of the difference $y - \hat{f}(\hat{X};\hat{\theta})$, where y is the true value and $\hat{f}(\hat{X};\hat{\theta})$ is the corresponding model prediction. This distribution describes, in probabilistic terms, our inability to produce perfect predictions. One must also define the "target population", which is the range of prediction situations of interest. The outcome of interest y could be any variable simulated by the model; yield is the most commonly examined variable, but outcomes related to nutrition or environmental impact are also of major interest.

As the notation shows, crop model predictions are completely determined by the triplet of model structure (i.e. the model equations) $\hat{f}$, inputs $\hat{X}$ and parameters $\hat{\theta}$. By considering uncertainty in each of these elements, we are considering the full uncertainty in crop model predictions. We assume that there is some distribution $P_{\hat{f}}$ of plausible crop model structures, and for each model structure a distribution noted $P_{\hat{\theta}|\hat{f}}$ of estimated parameter vectors. (One might decide that some parameters are part of model structure. This is not a problem. It simply means that the uncertainty of those parameters is included in structure uncertainty rather than in parameter uncertainty). We assume that for any true vector of explanatory variables X, there is a distribution noted $P_{\hat{X}|X}$ of approximations. (In general, models will share many of the same input variables. The X here refers to a set of input variables that includes the input variables of all models). The two other random variables of interest are X, the true vector of the input variables, and the output y. Both X and y have some distribution specific to the target population.

Rather than considering the full distribution of $y-\hat{f}(\hat{X};\hat{\theta})$, it is often convenient to concentrate on a summary of the distribution. A common choice is mean squared error. To emphasize the fact that we are interested in prediction error, we refer to mean squared error of prediction (MSEP). The exact definition depends on which variables are treated as random variables. To encompass the different approaches to uncertainty described in the introduction, we will be primarily concerned with two criteria based on MSEP.

The first criterion is

$$\mathrm{MSEP}_{\mathrm{fixed}} = E\left[\left(y-\hat{f}(\hat{X};\hat{\theta})\right)^2 \mid \hat{f},\hat{X},\hat{\theta}\right] \quad (1)$$

The notation is that the expectation is over all random variables, except those specified as fixed and that appear after the vertical bar. In eq. (1) the model structure $\hat{f}$, the approximation $\hat{X}$ of X and the parameter vector $\hat{\theta}$ are fixed, while X and y are treated as random variables. Crop model evaluation usually refers to this criterion, where prediction uses some specific model structure, approximation to the inputs and parameter vector.

The second criterion based on MSEP is

$$\mathrm{MSEP}_{\mathrm{uncertain}}(X) = E\left[\left(y-\hat{f}(\hat{X};\hat{\theta})\right)^2 \mid X\right]$$

Here the expectation is over $P_{\hat{f}}$, $P_{\hat{\theta}|\hat{f}}$ and $P_{\hat{X}|X}$, as well as over y for the given X. The subscript "uncertain" emphasizes the fact that in this case we explicitly treat the predictor as uncertain, by treating the model, parameters and estimated input values as random variables.

Taking the expectation over the uncertainty in the parameter vector is standard for the statistical treatment of a regression equation (Myers, 2007; Seber and Wild, 1989), and has also been proposed for process-based models (MacFarlane et al., 2000; Murphy et al., 2004; Omlin and Reichert, 1999). Other than parameter uncertainty, it has been argued that uncertainty should also take into account model structure uncertainty, which can be very important (Burnham and Anderson, 1998; Neuman, 2003). Finally, in crop models the input variables are often difficult to measure or estimate, or the measurement may be displaced in space or time relative to the required values, so it is important to include this source of uncertainty as well.

$\mathrm{MSEP}_{\mathrm{uncertain}}$ can be decomposed into two terms:

$$\begin{aligned}\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X}) &= \mathrm{E}\left\{\left[\left(\mathrm{y}-\mathrm{E}\left[\hat{\mathrm{f}}(\hat{\mathrm{X}};\hat{\theta})\,|\,\mathrm{X}\right]\right)^2\right]|\,\mathrm{X}\right\}+\mathrm{var}\left[\hat{\mathrm{f}}(\hat{\mathrm{X}};|\hat{\theta})\,|\,\mathrm{X}\right] \\ &= \text{squared bias + model variance}\end{aligned} \quad (2)$$

The first term is the squared bias when predicting using an average over model structures, approximations to X and parameter vectors, and the second term is the variance of the predictor. This second term depends only on the model, and not on the true responses. A schematic diagram of this criterion is shown in Fig. 1d. Most studies that propagate uncertainties consider only the model variance term. As eq. (2) shows, this MSEP criterion also has a squared bias term.

The model variance term in eq (2) can be further decomposed into contributions from the different sources of uncertainty as

$$\mathrm{var}\left[\hat{\mathrm{f}}(\hat{\mathrm{X}};\hat{\theta})\,|\,\mathrm{X}\right]=\sigma_{\hat{\mathrm{f}}}^2+\sigma_{\hat{\mathrm{X}}}^2+\sigma_{\hat{\mathrm{f}}\hat{\mathrm{X}}}^2+\sigma_{\hat{\theta}}^2 \quad (3)$$

where the terms on the right are respectively the first order effect of model structure, the first order effect of approximating X by $\hat{X}$, the interaction of model structure and approximating X by $\hat{X}$ and finally the effect of uncertainty in parameter values $\hat{\theta}$, averaged over model structures and over $\hat{X}$ (Table 1). There is no interaction term between model structure and model parameters, because the latter is nested within the former (i.e. each model has its own specific parameter vector).

Table 1

The contributions of various sources of error to $\mathrm{MSEP}_{\text{uncertain predictor}}(X)$.The random quantities over which expectations or variances are taken are shown explicitly as subscripts.

symbol	explanation	formula
$\sigma^2_{\hat{f}}$	first order contribution of model structure uncertainty	$\mathrm{var}_{\hat{f}}\left\{E_{\hat{X},\hat{\theta}}\left[\hat{f}(\hat{X};\hat{\theta})\,\vert\,X,\hat{f}\right]\right\}$
$\sigma^2_{\hat{X}}$	first order contribution of input uncertainty	$\mathrm{var}_{\hat{X}}\left\{E_{\hat{f},\hat{\theta}}\left[\hat{f}(\hat{X};\hat{\theta})\,\vert\,X,\hat{X}\right]\right\}$
$\sigma^2_{\hat{f}\hat{X}}$	interaction of model structure and input uncertainty	$\mathrm{var}_{\hat{f},\hat{X}}\left\{E_{\hat{\theta}}\left[\hat{f}(\hat{X};\hat{\theta})\,\vert\,X,\hat{f},\hat{X}\right]\right\}-\sigma^2_{\hat{f}}-\sigma^2_{\hat{X}}$
$\sigma^2_{\hat{\theta}}$	parameter uncertainty	$E_{\hat{f},\hat{X}}\left\{\mathrm{var}_{\hat{\theta}}\left[\hat{f}(\hat{X};\hat{\theta})\,\vert\,X,\hat{f},\hat{X}\right]\right\}$

Eq. 3 is the analysis of variance decomposition for an experimental design where the parameter vector is nested within model structure and input vector. This would be appropriate for simulation experiments where every model structure is crossed with every input vector, but where each combination has a different parameter vector, so that there is no relation between the parameter vectors explored for different models or different input vectors. We could also quite easily treat other experimental designs, e.g. where parameters are crossed with model structure and inputs.

2.2 Estimation of MSEP

$MSEP_{fixed}$ can be estimated based on the discrepancy between hindcasts and observed data. Suppose that we have a sample of data from the target population, (y_l , $l = 1, \ldots, L$) , unrelated to the data used to create or parameterize the model. Then the estimate is

$$\hat{MSEP}_{fixed} = (1/n)\sum_{l=1}^{L}\left(y_l - \hat{f}(\hat{X}_l; \hat{\theta})\right)^2 \quad (4)$$

If the data available for estimating MSEP have been used for parameter estimation, then the above estimator is biased. In this case one can use cross validation or a bootstrap approach (Efron, 1983). In practice, some of the parameters may vary among different hindcasts, for example if different cultivars are concerned. Then $\hat{MSEP}_{fixed}$ is estimating an average for the different parameter vectors.

In estimating $MSEP_{uncertain}(X)$, the squared bias and model variance contributions can be estimated separately. Consider first the squared bias term. Suppose that we have a sample of data for evaluation, and that for each observation in the sample, there are corresponding simulations using I model structures drawn from P_u, each combined with J input vectors drawn from $P_{\hat{X}|X}$, and for each combination K parameter vectors drawn from $P_{\hat{\theta}|\hat{f}}$. A plug-in estimate of the squared bias term, averaged over X, is

$$\hat{s}\text{quared bias} = \frac{1}{L}\sum_{l=1}^{L}\left\{\left[y_l - \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{f}_i(\hat{X}_{lj};\hat{\theta}_{ijk})\right]^2\right\} \quad (1)$$

Estimation of the model variance term does not involve observations; it involves only the model. Given simulations with multiple model structures, inputs and parameters, one simply calculates the variance of the simulated values. Thus the overall estimate of $MSEP_{uncertain}(X)$ has a term which is an average over X (the squared bias) plus a term which is specific to a particular X (the model variance).

Consider now estimation of the individual contributions to model variance (eq.(3), Table 1). Apportioning uncertainty among contributing factors is the objective of the field of sensitivity analysis (Saltelli et al., 2000). Yip et al. (2011) used an analysis of variance (ANOVA) approach to decompose the uncertainty in climate models. However, the decomposition problem has not been addressed specifically for crop models when model uncertainty is one of the sources of uncertainty. Furthermore, treatments to date use a fixed effects ANOVA, whereas in fact the factors (model structure, inputs, parameters) are better treated as samples from an infinite population. We thus develop in the appendix a random effects ANOVA, with two crossed factors (model structure and inputs) and one nested factor (parameters) (Scheffé 1959). There is no error term, because there is no measurement error. A general method for estimating the variance components is restricted maximum likelihood (REML). The R package lme4 has a function which performs this estimation (Bates et al., 2014; R Core Team, 2012).

In the case of a balanced design, we can show analytically how to estimate the variance components, which gives important insights. The ANOVA table is shown in Table 2.

Table 2

ANOVA table for experimental design with 2 crossed factors (model structure $\hat{f}$ and inputs $\hat{X}$) and one nested factor (parameters $\hat{\theta}$)

Sum of squares SS	**degrees of freedom df**	**Expected mean square E(MS)**
$SS_A = JK\sum_{i=1}^{I}\left(m_{i\cdot\cdot} - m_{\cdot\cdot\cdot}\right)^2$	I-1	$JK\left(\sigma_{\hat{f}}^2 + J^{-1}\sigma_{\hat{f}\hat{X}}^2 + J^{-1}K^{-1}\sigma_{\hat{\theta}}^2\right)$
$SS_B = IK\sum_{j=1}^{J}\left(m_{\cdot j\cdot} - m_{\cdot\cdot\cdot}\right)^2$	J-1	$IK\left(\sigma_{\hat{X}}^2 + I^{-1}\sigma_{\hat{f}\hat{X}}^2 + I^{-1}K^{-1}\sigma_{\hat{\theta}}^2\right)$
$SS_{AB} = K\sum_{i=1}^{I}\sum_{j=1}^{J}\left(m_{ij\cdot} - m_{i\cdot\cdot} - m_{\cdot j\cdot} + m_{\cdot\cdot\cdot}\right)^2$	(I-1)(J-1)	$K\left(\sigma_{\hat{f}\hat{X}}^2 + K^{-1}\sigma_{\hat{\theta}}^2\right)$
$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(m_{ijk} - m_{ij\cdot}\right)^2$	IJ(K-1)	$\sigma_{\hat{\theta}}^2$

$m_{ijk} = \hat{f}_i(\hat{X}_j;\hat{\theta}_{ijk})$ is the simulated value for model i, input vector j and parameter vector k. A dot indicates an average over the index at that position. In each row, the SS term estimates E(MS)*df.

Each sum of squares in Table 2 estimates the corresponding expected mean square (E(MS)) times the degrees of freedom df. If all df>1, then one can estimate all four E(MS) values, and from them the four variance components. If in any row df=1, that row cannot be used and then only certain combinations of the variance components can be estimated.

Some special cases are of interest. Suppose that one has a simulation experiment with multiple models (I>1), but with only a single input vector (J=1), and that each model has only a single parameter vector (K=1), different for each model. This has been the typical practice for simulations with multi crop model ensembles to date. Then only SS_A has df>1, and it estimates $\left(\sigma^2_{\hat{f}} + \sigma^2_{\hat{f}\hat{X}} + \sigma^2_{\hat{\theta}}\right)*(I-1)$. That is, the variance over simulated values of a multi-model ensemble estimates the sum of the variance components due to uncertainty in model structure, interaction between model structure and input, and parameters. It should be emphasized that the contribution of parameter uncertainty is included in the sum of squares. This occurs because by assumption each model structure is associated with a parameter vector chosen at random, independently of the other model structures. The sum of squares does not however include the effect of input uncertainty. It would be necessary to explicitly include multiple values of $\hat{X}$ in the simulations to estimate this term.

If one has multiple models (I>1) and multiple input vectors (J>1), but only a single parameter vector for each model (K=1), then one can calculate SS_A, SS_B and SS_{AB}. One can then solve for $\sigma^2_{\hat{f}}$, $\sigma^2_{\hat{X}}$ and $\sigma^2_{\hat{f}\hat{X}} + \sigma^2_{\hat{\theta}}$. Adding them gives an estimate of total model variance. However, since each model has only a single parameter value, $\sigma^2_{\hat{f}\hat{X}}$ and $\sigma^2_{\hat{\theta}}$ are confounded.

In sensitivity analysis studies, one usually has only a single model (I=1), but multiple inputs (J>1) and/or parameters (K>1). If only K>1 one can only estimate $\sigma^2_{\hat{\theta}}$. If both J>1 and K>1, and parameters are drawn independently for each input vector, then one can estimate $\sigma^2_{\hat{X}} + \sigma^2_{\hat{f}\hat{X}}$ and $\sigma^2_{\hat{\theta}}$. The case with multiple inputs all with the same parameter vector could be treated as a three way crossed design, but then the variance components are defined differently than in the nested design, and so are not directly comparable.

Above, we consider separate estimation of the bias and model variance contributions to $MSEP_{uncertain}(X)$. An alternative would be to consider the estimation of $MSEP_{uncertain}(X)$

based on observed discrepancies. Each individual observed discrepancy $\left(y-\hat{f}(\hat{X};\hat{\theta})\right)^2$ can be thought of as an estimator, based on a single model structure, a single approximation to X and a single parameter vector, of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. The estimate based on discrepancy could be compared with estimation based on separate estimation of bias and variance. .

The average of observed discrepancies over a sample from the target population is, by a similar argument, an estimator of $E\left[\mathrm{MSEP}_{\mathrm{uncertain}}(X)\right]$, where the expectation is over the X in the target population. This could be compared with estimates of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ averaged over X.

2.3 The effect of measurement error

In all of the above, we assume that the measured response y is the true response. Suppose that the response y is measured with error, as is usually the case. What is the relation between MSEP for the measured response (i.e. how well can we predict the measurement), and MSEP for the true response?

We assume that $E(\hat{y})=y$ (measurements are not biased) and $\mathrm{var}(\hat{y})=\sigma^2$ (all measurements have the same variance of measurement error). Then

$$\begin{aligned} E\left\{\left[\hat{y}-\hat{f}(\hat{X};\hat{\theta})\right]^2 \mid X\right\} &= E\left\{\left[\hat{y}-y+y-\hat{f}(\hat{X};\hat{\theta})\right]^2 \mid X\right\} \\ &= \sigma^2 + E\left\{\left[y-\hat{f}(\hat{X};\hat{\theta})\right]^2 \mid X\right\} \end{aligned} \quad (6)$$

Eq. (6) says that $\mathrm{MSEP}_{\mathrm{fixed}}(X)$ for predicting the measured response (left hand side) is larger than $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ for predicting the true response by an amount σ^2, on the average over measured values. Thus to obtain an unbiased estimate of MSEP for the true response, we should subtract σ^2 from MSEP estimated using measured responses. The same relation holds for $\text{MSEP}_{\text{fixed}}(X)$. This shows that if $\text{MSEP}_{\text{uncertain}}(X)$ is less than or comparable to σ^2, it cannot be reliably estimated because it will be confounded with measurement error.

2.4 Case study

We illustrate the above framework using previously published data from a multi-model ensemble of crop models (Asseng et al., 2013). This study also included simulations with multiple values for some of the input variables. However, each model was run with only a single parameter vector. It would be preferable to illustrate using a study which included all three of multiple model structures, multiple approximations to the input variables and multiple parameter vectors for each model, but to date there have been no such studies for crop models.

The Asseng et al. (2013) study involved running 27 different wheat models for 4 sites, in Argentina (AR), Australia (AU), India (IN) and The Netherlands (NL). Observed data were available for one year from each site. The observed yields were 5.87 t/ha (AR in 1992), 2.50 t/ha (AU in 1984), 4.18 t/ha (IN in 1985) and 7.45 t/ha (NL in 1983). As is typically the case for crop models, each model was developed and parameterized using a diversity of data, different for each model and in general not including the data from the four sites of this study. The data from the four sites here were then used to quantify the variability between models, and for model evaluation.

In the first simulation exercise, the 27 simulated yields were compared with the four observed yields. Each model was provided with the same fixed values of inputs (weather, soil characteristics, management). Times to anthesis and maturity at each site were also provided, so that the phenology parameters of each model could be adapted to the specific cultivar at each site. Then the simulated results were compared to the observed values. In a second exercise, the models were all run for the 30-year baseline period 1981-2010 for each site, using observed weather each year but the same soil and management as for the year with observed yield. In a third simulation exercise, each model was first calibrated using the observed data, and then each model was run for 1981-2010 as before, except that plant available soil moisture (PAW) was either decreased or increased by 20% compared to the value provided initially, which can be taken as a rough approximation to the uncertainty in that input. For comparison, Aggarwal (1995) assumed an uncertainty of ±15% for each of wilting point and field capacity.

The target population here is worldwide wheat fields under current climate conditions. We assume that the models in the study are a sample from plausible models. We assume that the input variables that were provided for each site-year (soil, management, weather, initial conditions) are a sample of size 1 drawn at random from the distribution of approximations to the true X for that site-year, i.e. from $P_{\hat{X}|X}$. Finally, we assume that the parameter vector used with each model is a sample of size 1 drawn at random from the distribution of parameters for that model. Each model had a different parameter vector, with phenology parameters that varied between sites while all other parameters were the same for all sites.

3. Results

First, we illustrate the differences between $MSEP_{fixed}$ and $MSEP_{uncertain}$.for the case study. We ignore measurement error, which would affect both criteria equally. $MSEP_{fixed}$ for model i can be estimated by applying eq. 4 to this case:

$$\hat{MSEP}_{fixed,i} = (1/4)\sum_{l=1}^{4}\left[y_l - \hat{f}_i(\hat{X}_l;\hat{\theta}_i) \right]^2$$

where the sum is over the 4 observed yields. The individual squared errors of the hindcasts and their average, which is the estimated value of $MSEP_{fixed}$, are shown in Table 3 for three of the models; the best model (smallest average MSEP), an intermediate model and the second worst model. (The worst model seems to be an outlier). The next to last line of the table shows the squared error per site and estimated $MSEP_{fixed}$ averaged over all 27 models. The last line concerns the case where one first averages the model predictions, and then calculates squared error and estimated $MSEP_{fixed}$. The average of predictions is a new predictor, the ensemble mean (e-mean).

Table 3

Squared errors ((t/ha)²) for hindcasts for 4 site-years and averaged over the 4 site-years, for 3 particular models, for the average over 27 models, and for the e-mean model, which is the model that predicts using the average of the 27 model predictions. The average over site-years (the last column) is the estimate of $MSEP_{fixed}$. AR is a site in Argentina, AU a site in Australia, IN a site in India and NL a site in the Netherlands.

model	**AR 1992**	**AU 1984**	**IN 1985**	**NL 1983**	**average over site-years**
			Squared error (t/ha)²		
model 1. Best	0.07	0.07	0.00	0.63	0.19
model 2. Intermediate	1.72	0.01	1.54	4.80	2.02
model 3. Second worst	12.67	2.46	2.89	0.86	4.72

Average over models	2.79	1.37	1.59	1.97	1.93
e-mean (ensemble average as predictor)	0.03	0.36	0.53	0.01	0.23

Consider now prediction mean squared error for the AR site in two other years, 2009 and 2010 (chosen arbitrarily), for which we have no data. The estimate of $\mathrm{MSEP}_{\mathrm{fixed}}$ for each model for each year (Table 4) is simply the average value obtained from the hindcasts. According to $\mathrm{MSEP}_{\mathrm{fixed}}$, the best predictor is model 1, with an average squared error of prediction of 0.19 (t/ha)². The estimated average squared error of e-mean is only slightly larger, at 0.23 (t/ha)²; the estimated average squared errors of the other models shown are quite a bit larger. However, all these estimated squared errors are averages of only 4 hindcasts, and the squared errors of the individual hindcasts are quite variable (see Table 3). This emphasizes that $\mathrm{MSEP}_{\mathrm{fixed}}$ is a very approximate estimate of prediction error.

Table 4

Values of $\mathrm{MSEP}_{\mathrm{fixed}}$ and $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ for the AR site, for 2 years without measurements. The first three rows are the estimated $\mathrm{MSEP}_{\mathrm{fixed}}$ values for the specific models in question. The last 3 rows show respectively squared bias, model variance and their sum, which is the estimate of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. Units are (t/ha)².

	AR 2009	AR 2010
model 1. $\mathrm{MSEP}_{\mathrm{fixed}}$	0.19	0.19
model 2. $\mathrm{MSEP}_{\mathrm{fixed}}$	2.02	2.02
model 3. $\mathrm{MSEP}_{\mathrm{fixed}}$	4.72	4.72
squared bias term of $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$	0.23	0.23
$\left(\sigma_{\mathrm{f}}^2+\sigma_{\mathrm{fX}}^2+\sigma_{\theta}^2\right)$ **term of** $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$	1.55	2.77
$\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$	1.78	3.00

We turn now to $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$. According to eq. (5), the estimate of the squared bias contribution is average squared error of hindcasts for the e-mean model. The value is 0.23 $(t/ha)^2$ (Table 3). This same value would be used for all predictions.

The model variance contribution to $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$ will be different for each prediction. For each prediction we have a simulation experiment with I=26 model structures, J=1 input vectors and K=1 parameter vectors per model. We apply the ANOVA decomposition of Table 2 to the results. In the ANOVA table, since J=1 and K=1, we can only use the model structure sum of squares, which estimates $\left(\sigma_{\mathrm{f}}^2+\sigma_{\mathrm{fX}}^2+\sigma_{\theta}^2\right)*26$. We cannot estimate each component of variance separately, and the variance does not contain the first order contribution from input uncertainty. Thus we are underestimating $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$.

(Alternatively, we are assuming that there is no uncertainty in the inputs). The estimated values of squared bias, of $\left(\sigma_{\mathrm{f}}^2+\sigma_{\mathrm{fX}}^2+\sigma_{\theta}^2\right)$ and of their sum

$$\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})=\text{squared bias } +\left(\sigma_{\mathrm{f}}^2+\sigma_{\mathrm{fX}}^2+\sigma_{\theta}^2\right)$$

for AR 2009 and 2010 are shown in Table 4. Note that the squared bias term is quite small compared to the variance term.

$\mathrm{MSEP}_{\mathrm{fixed}}$ is identical for every prediction. It gives no indication as to the way prediction mean squared error varies for different prediction situations. $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$ on the other hand takes into account differences between the two prediction situations. The $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$ values are substantially larger in 2010 than in 2009, because the variability between models is larger for the 2010 conditions than for those of 2009. The conclusion is that predictions are less reliable for the 2010 conditions. A major advantage of $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$ is that it shows how prediction mean squared error varies with the prediction situation. However, this is not mean squared error for a specific model and parameter vector. It represents mean squared error averaged over plausible models and over the distribution of parameter vectors for each model.

As a second example of the difference between $\mathrm{MSEP}_{\mathrm{fixed}}$ and $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$, consider the mean squared error in predicting yield averaged over a multi-year period. This is important because climate change impact is often quantified as future yield averaged over multiple years compared to baseline yield, also averaged over multiple years. The mean squared error in predicting yield averaged over years is a different question than the average yearly mean squared error. The former will always be smaller than the latter (Wallach and

Thorburn, 2014). How much smaller however depends on exactly how much of the model discrepancy and between-model variability cancels out when averaging over years.

The hindcasts in the Asseng study only include a single observed yield at each site, and so they shed no light on how mean squared error is reduced when averaging over years. We have no reliable way of estimating $\mathrm{MSEP}_{\mathrm{fixed}}$ for a multi-year average from these data.

On the other hand, we can quite easily estimate the model variance part of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ when predicting a multi-year average yield. We first simulate for each year with each model, then take the average over years for each model and then calculate the between model variance. The results are shown in Table 5 for each of the 4 locations, for yield averaged over a 30-year baseline. The mean squared error is lower than the average mean squared error, as it must be (Wallach and Thorburn, 2014), but the differences are not very large. It seems that there is not much reduction in mean squared error due to averaging over years, i.e. the differences between models do not cancel out when averaging over the baseline years. We have not included the contribution of squared bias to $\mathrm{MSEP}_{\mathrm{uncertain}}^{\mathrm{without}\ \sigma_X^2}(X)$ because we cannot estimate it for an average over years, but in any case we know it is small relative to model variance.

Table 5

Variance of simulated values. Average of yearly variances (first line) or variance of yield averaged over the period 1981-2010. Units are $(t/ha)^2$.

	AR	AU	IN	NL

Variance of simulations each year, averaged over years (1981-2010)	2.90	1.42	1.82	2.96
Variance of simulations of average over years	2.12	1.09	1.45	2.04

Consider now the third simulation exercise of the case study. There are here I=25 models, each with J=3 input values (initial PAW or initial PAW±20%). In this case we can use SS_A, SS_B and SS_{AB} (see Table 2) to estimate σ_{f}^2, σ_{X}^2 and the sum $\sigma_{\mathrm{fX}}^2+\sigma_{\theta}^2$. The sum of those three terms is the full model variance. Results for two arbitrarily chosen prediction situations are shown in Table 6. Using the ANOVA table (Table 2) or the R function lme4 with option REML gives identical results. The model structure uncertainty contribution to prediction uncertainty is quite different for the two different sites. In both sites, the first order contribution of model structure uncertainty is much larger than the first order contribution of input uncertainty, but this is a rather artificial example in that only a single input variable was considered uncertain.

Table 6

The contributions of model structure uncertainty, input uncertainty (uncertainty in PAW) and their interaction plus parameter uncertainty, to $MSEP_{uncertain}$ at two sites. The last line shows estimated $\mathrm{MSEP}_{\mathrm{uncertain}}$. Units are $(t/ha)^2$.

site year	**AU 2009**	**NL 2009**

σ_{f}^2	0.75	2.44
σ_{X}^2	0.09	0.06
$\sigma_{\mathrm{fX}}^2 + \sigma_{\theta}^2$	0.31	0.41
squared bias	0.23	0.23
$\mathrm{MSEP}_{\mathrm{uncertain}}$	1.38	3.14

Many other specific prediction problems could be studied in a similar way. For example, one might be interested in the mean squared error in the ratio of yield under climate change to the baseline yield, considering climate as uncertain. Here the simulated quantity would be the future to baseline yield ratio and the input uncertainty would be represented by the different climate models.

4. Discussion

Each criterion $\mathrm{MSEP_{fixed}}$ and $\mathrm{MSEP_{uncertain}}(X)$ has advantages and drawbacks. One advantage of $\mathrm{MSEP_{fixed}}$ is that it automatically takes into account all sources of model uncertainty, since we compare the model predictions, which are influenced by all the sources of uncertainty, with observations. A second advantage is that it concerns a specific model and parameter vector, which is what we want if predictions will use that model and parameter vector. However, there are serious drawbacks to this criterion. First, $\mathrm{MSEP_{fixed}}$ is an average over the population from which the test data are drawn. It does not indicate how prediction mean squared error varies as a function of the specific prediction situation. As shown in the examples, $\mathrm{MSEP_{fixed}}$ is the same for all predictions. Another difficulty is that there is no mechanism for adapting the mean squared error to a population different than that which provided the data for estimating $\mathrm{MSEP_{fixed}}$. The requirement that the data for estimation of $\mathrm{MSEP_{fixed}}$ be independent of the data used for model development and calibration can be difficult to respect completely (Coucheney et al., 2015b), which could lead to underestimating prediction error. In the examples, the model phenology parameters are adjusted to data from each site-year (i.e. the data used for model evaluation have been used to some extent for parameter estimation), so $\mathrm{MSEP_{fixed}}$ estimated using those site-years probably underestimates $\mathrm{MSEP_{fixed}}$ for the target population. Finally, in practice we are often not interested in a model with specific parameters. In the examples, the predictor that is evaluated has a range of parameters for those parameters that determine phenology. Prediction may concern yet other varieties and thus use yet other parameters. This suggests that it would be more realistic to evaluate MSEP averaged over the distribution of parameters.

Consider now $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. We have shown that this criterion can be divided into two contributions, a squared bias term and a model variance term. The squared bias term involves true responses, but the model variance term does not. Estimating the squared bias term has the same disadvantages as $\mathrm{MSEP}_{\mathrm{fixed}}$. However, the limited experience with crop model ensembles suggests that the squared bias term may be substantially smaller than the model variance term when the predictions are averaged over models (Asseng et al., 2013; Bassu et al., 2014; Li et al., 2015). When this is true, it implies that when we estimate model variance, we are estimating the major contribution to $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. This is a major advantage of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$, since model variance involves simulations rather than observations. The former are very much faster and cheaper to produce than the latter, and can easily be obtained for any prediction situation, as illustrated in the examples. This is much more informative than simply evaluating an average MSEP over a target population. Another advantage is that model variance can be estimated for a target population different than the population that produced observations, if one can approximate distributions of model structure, inputs and parameters that represent the uncertainties for the new population. For example, since models include a response to CO_2, one could argue that the variability between models should represent our uncertainty about the effect of increased CO_2. A further important advantage is the possibility of estimating the separate contributions of model structure, input and parameter uncertainty. This makes it possible to target the most important causes of prediction mean squared error.

A major difficulty with $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ is that it may be difficult to define and sample from the distributions of model structure, approximations to inputs and parameters. Defining a distribution of plausible model structures is particularly difficult. In the crop model ensemble studies cited, the practice was essentially to accept all candidate models. This is meant to produce a random sample of "plausible models", but one might argue that some models are less plausible than others. Similarly, in climate modeling, there is debate over whether all models should be equally weighted (Knutti, 2010).

Estimating the distribution of parameter vectors also poses problems. One approach has been to estimate the distribution from the range of values found in the literature. Another approach has been to do a Bayesian estimation of the parameter vector (Iizumi et al., 2009; Wallach et al., 2012). Neither approach however fully takes into account the fact that in general crop model parameters are estimated in a series of studies, using a diversity of data. The literature approach ignores the fact that the parameters have been fit to data. The Bayesian approach takes into account only one data set, which in general is used only to determine a relatively small number of model parameters. It ignores the data used previously to estimate the remaining model parameters.

Estimating the distribution of approximations to the inputs often poses fewer problems. This can for example be based on specific studies of each input (Aggarwal, 1995). For climate change impact assessment, a major input uncertainty is uncertainty in future climate, which can be represented by results from a range of climate models.

$\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ is average squared prediction error when the model is chosen at random from the probability distribution of models, when the parameters of each model are chosen at random from the probability distribution of parameters of that model, and when the inputs are chosen at random from the probability distribution of approximations to X. Of course one would prefer to have the squared prediction error for each specific model, with a specific parameter vector and specific input vector, that is $MSEP_{fixed}(X)$. But as explained above, this is not in general possible; we can only estimate an average, $MSEP_{fixed}$, and the average is specific to situations like those used for hindcasts. Thus despite the difficulties with $MSEP_{uncertain}(X)$, we suggest that $MSEP_{uncertain}(X)$ is often a more pertinent and more informative criterion of model predictive mean squared error than $MSEP_{fixed}$, not only for climate change studies but also under current conditions. The major advantage of $MSEP_{uncertain}(X)$ is the fact that the model variance component requires only simulations, not comparison with data. As a result it can be used to obtain mean squared errors adapted to each specific prediction question, whereas $MSEP_{fixed}$ is the same for all predictions. Also, $MSEP_{uncertain}(X)$ can be used to estimate mean squared error even in the absence of data, whereas $MSEP_{fixed}$ requires observations.

For example, with respect to climate change, one might want to know mean squared error in the ratio of mid-century yield to baseline yield at some location, or mean squared error in the difference between mid-century yield with some specific adaptation strategy and without adaptation. $MSEP_{uncertain}(X)$ can be used to estimate mean squared error for each of these predictions, whereas in the absence of multi-year data $MSEP_{fixed}$ cannot estimate mean squared error in multi-year averages. Also, $MSEP_{fixed}$ could be very unreliable for estimating the effect of adaptation strategies if they are not represented in observed data.

Under current conditions, one might want to know for example mean squared error in predicting change in yield or in some environmental variable due to some change in crop management. This could in principle be estimated using hindcasts, but that would require specific experiments for each new management question. To estimate the model variance contribution to $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ on the other hand would only require new simulations.

The advantage of $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ is in large part dependent on squared bias being relatively small. This suggests that more information about the squared error contribution to $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ is important, since our evidence to date is limited. In particular, for climate change impact studies, it is important to estimate the squared bias contribution for a target population with climate similar to likely future conditions (higher CO_2, temperature).

The ANOVA table (Table 2) shows explicitly how to estimate the contributions of different uncertainties to model variance. One conclusion is that simulation studies which combine multiple model structures, multiple approximations to the inputs and multiple parameter vectors would be valuable, since they would make it possible to estimate separately all the sources of uncertainty in $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$.

The difficulties involved in determining the distributions of model structure and parameter values were discussed above. It will be important to obtain better estimations of those distributions, taking into account the complex way in which crop models are developed and parametrized. .

Determining the contributions of different sources of uncertainty to overall uncertainty could have important implications for future crop modeling work, leading to better integration of uncertainty information and model improvement. For example, if parameter uncertainty is the major contribution to overall prediction mean squared error, priority should be either on obtaining better parameter values or on developing models with fewer or more easily estimated parameters. Similarly, if input uncertainty is a major contribution, then it is important either to limit model use to cases where inputs are well known, or develop models with more easily obtained inputs.

The ANOVA results show how simulation studies with a single model structure but multiple inputs and/or parameters are related to the separate contributions to $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. Such studies should be compared between models and with values obtained from ensembles which combine multiple model structures with multiple parameter vectors for each model structure. The single model results provide a valuable additional source of information about the contributions of different sources of uncertainty. We have also shown that observed discrepancies can be related to $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$, and can therefore provide additional information about that criterion.

5. Conclusions

Agro-climatic modeling is central to anticipating the impact of climate change on agriculture, and the possibilities of reducing that impact through adaptation, as well as to agricultural and environmental decisions under current climate. Policy makers need information on the reliability of model projections, in order to make informed decisions that avoid over-reliance on model results.

Previous uncertainty work has been of three types: comparison with hindcasts, propagation of uncertainties in parameters or inputs, and evaluation of the variability in multi-model ensembles. Here we show how these different approaches can be put into a unified framework. Comparison with hindcasts is a way of estimating MSEP for a fixed model, termed $\mathrm{MSEP}_{\mathrm{fixed}}$. The other approaches are related to MSEP averaged over uncertainties in model structure, model parameters and model inputs, termed $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$. We have shown that $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ has both a squared bias term and a model variance term. The former is often ignored. Preliminary evidence suggests that it may be small, but more information is needed. We show how the contributions from different sources of uncertainty to the model variance term can be estimated from model simulation experiments.

The framework proposed here can provide the basis for obtaining information on the relative contributions of model uncertainty, parameter uncertainty, input uncertainty and bias to overall prediction mean squared error. Furthermore, since the $\mathrm{MSEP}_{\mathrm{uncertain}}(X)$ criterion that we propose is specific to each prediction situation, it can help to determine, for each specific problem, how well crop models are likely to perform.

$\mathrm{MSEP}_{\mathrm{fixed}}$ and $\mathrm{MSEP}_{\mathrm{uncertain}}(\mathrm{X})$ are complementary criteria for estimating prediction accuracy of models. Both are useful because they give different types of information about prediction error. If possible, both should be evaluated.

6. Acknowledgements

The authors are grateful to the AgMIP and MACSUR projects, which have made the collaboration behind this study possible.

7. References

Aggarwal, P.K., 1995. Uncertainties in crop, soil and weather inputs used in growth models: Implications for simulated outputs and their applications. Agric. Syst. 48, 361–384. doi:10.1016/0308-521X(94)00018-M

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P.J., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Williams, J.R., Wolf, J., 2013. Uncertainty in simulating wheat yields under climate change. Nat. Clim. Chang. 3, 827–832. doi:10.1038/nclimate1916

Asseng, S., Jamieson, P.., Kimball, B., Pinter, P., Sayre, K., Bowden, J.., Howden, S.., 2004. Simulated wheat growth affected by rising temperature, increased water deficit and elevated atmospheric CO2. F. Crop. Res. 85, 85–102. doi:10.1016/S0378-4290(03)00154-0

Baigorria, G.A., Jones, J.W., Shin, D.-W., Mishra, A., O'Brien, J.J., 2007. Assessing uncertainties in crop model simulations using daily bias-corrected Regional Circulation Model outputs. Clim. Res. 34, 211–222.

Basso, B., Liu, L., Ritchie, J.R., 2016. A comprehensive review of the CERES-Wheat, -Maize and -Rice models' performances. Adv. Agron. 136. doi:10.1016/bs.agron.2015.11.004

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? Glob. Chang. Biol. 20, 2301–20. doi:10.1111/gcb.12520

Bates, D., Maechler, M., Bolker, B., Walker, S., 2014. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7 [WWW Document]. URL http://cran.r-project.org/package=lme4

Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K., 2010. Validation of biophysical models: issues and methodologies. A review. Agron. Sustain. Dev. 30, 109–130. doi:10.1051/agro/2009001

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., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. doi:10.1016/j.envsoft.2012.09.011

Biernath, C., Gayler, S., Bittner, S., Klein, C., Högy, P., Fangmeier, A., Priesack, E., 2011. Evaluating the ability of four crop models to predict different environmental impacts on spring wheat grown in open-top chambers. Eur. J. Agron. 35, 71–82. doi:10.1016/j.eja.2011.04.001

Boote, K.J., Jones, J.W., Hoogenboom, G., White, J.W., 2010. The role of crop systems simulation in agriculture and environment. Int. J. Agric. Environ. Inf. Syst. 1 , 41–54.

Bouman, B.A.M., 1994. A framework to deal with uncertainty in soil and management parameters in crop yield simulation: A case study for rice. Agric. Syst. 46, 1–17.

Burnham, K.P., Anderson, D.A., 1998. Model selection and inference. Springer-Verlag, New York.

Challinor, A.J., Wheeler, T.R., 2008. Use of a crop model ensemble to quantify CO2 stimulation of water-stressed and well-watered crops. Agric. For. Meteorol. 148, 1062–1077. doi:10.1016/j.agrformet.2008.02.006

Coucheney, E., Buis, S., Launay, M., Constantin, J., Mary, B., García de Cortázar-Atauri, I., Ripoche, D., Beaudoin, N., Ruget, F., Andrianarisoa, K.S., Le Bas, C., Justes, E., Léonard, J., 2015a. Accuracy, robustness and behavior of the STICS soil–crop model for plant, water and nitrogen outputs: Evaluation over a wide range of agro-environmental conditions in France. Environ. Model. Softw. 64, 177–190. doi:http://dx.doi.org/10.1016/j.envsoft.2014.11.024

Coucheney, E., Buis, S., Launay, M., Constantin, J., Mary, B., García de Cortázar-Atauri, I., Ripoche, D., Beaudoin, N., Ruget, F., Andrianarisoa, K.S., Le Bas, C., Justes, E., Léonard, J., 2015b. Accuracy, robustness and behavior of the STICS soil–crop model for plant, water and nitrogen outputs: Evaluation over a wide range of agro-environmental conditions in France. Environ. Model. Softw. 64, 177–190. doi:10.1016/j.envsoft.2014.11.024

Dzotsi, K.A., Basso, B., Jones, J.W., 2013. Development, uncertainty and sensitivity analysis of the simple SALUS crop model in DSSAT. Ecol. Modell. 260, 62–76. doi:10.1016/j.ecolmodel.2013.03.017

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331.

Holzkämper, A., Klein, T., Seppelt, R., Fuhrer, J., 2015. Assessing the propagation of uncertainties in multi-objective optimization for agro-ecosystem adaptation to climate change. Environ. Model. Softw. 66, 27–35. doi:10.1016/j.envsoft.2014.12.012

Iizumi, T., Yokozawa, M., Nishimori, M., 2009. Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. Agric. For. Meteorol. 149, 333–348.

Knutti, R., 2010. The end of model democracy? Clim. Change 102, 395–404. doi:10.1007/s10584-010-9800-2

Koehler, A.-K., Challinor, A.J., Hawkins, E., Asseng, S., 2013. Influences of increasing temperature on Indian wheat: quantifying limits to predictability. Environ. Res. Lett. 8, 034016. doi:10.1088/1748-9326/8/3/034016

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Singh, B.-, Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B., 2015. Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob. Chang. Biol. 21, 1328–41. doi:10.1111/gcb.12758

MacFarlane, D.W., Green, E.J., Valentine, H.T., 2000. Incorporating uncertainty into the parameters of a forest process model. Ecol. Modell. 134, 27–40. doi:10.1016/S0304-3800(00)00329-X

Murphy, A.H., 1988. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. Mon. Weather Rev. 116, 2417–2424. doi:10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A.H., Epstein, E.S., 1989. Skill scores and correlation coefficients in model verification. Mon. Weather Rev. 117, 572–581.

Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., Webb, M.J., Collins, M., Stainforth, D.A., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430, 768–72. doi:10.1038/nature02771

Myers, R.H., 2007. Classical and modern regression with applications. PWS-Kent, Boston .

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Environ. Res. Risk Assess. 17, 291–305. doi:10.1007/s00477-003-0151-7

Omlin, M., Reichert, P., 1999. A comparison of techniques for the estimation of model prediction uncertainty. Ecol. Modell. 115, 45–59. doi:10.1016/S0304-3800(98)00174-4

Palosuo, T., Kersebaum, K.C., Angulo, C., Hlavinka, P., Moriondo, M., Olesen, J.E., Patil, R.H., Ruget, F., Rumbaur, C., Takáč, J., Trnka, M., Bindi, M., Çaldağ, B., Ewert, F., Ferrise, R., Mirschel, W., Şaylan, L., Šiška, B., Rötter, R., 2011. Simulation of winter wheat yield and its variability in different climates of Europe: A comparison of eight crop growth models. Eur. J. Agron. 35, 103–114. doi:10.1016/j.eja.2011.05.001

R Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austrai.

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resour. 29, 1586–1597. doi:10.1016/j.advwatres.2005.11.013

Reichler, T., Kim, J., 2008. How well do coupled models simulate today's climate? Bull. Am. Meteorol. Soc. 89, 303+. doi:10.1175/BAMS-89-3-303

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S.W., 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resour. Res. 46, n/a–n/a. doi:10.1029/2009WR008328

Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P., Antle, J.M., Nelson, G.C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., Winter, J.M., 2013. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. Agric. For. Meteorol. 170, 166–172. doi:10.1016/j.agrformet.2012.09.011

Rötter, R.P., Carter, T.R., Olesen, J.E., Porter, J.R., 2011. Crop–climate models need an overhaul. Nat. Clim. Chang. 1, 175–177. doi:10.1038/nclimate1152

Roux, S., Brun, F., Wallach, D., 2014. Combining input uncertainty and residual error in crop model predictions: A case study on vineyards. Eur. J. Agron. 52, 191–197. doi:10.1016/j.eja.2013.09.008

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity analysis. Wiley , New York.

Scheffé, H., 1959. The analysis of variance. John Wiley & Sons, New York.

Seber, G.A.F., Wild, C.J., 1989. Nonlinear regression. Wiley , New York.

Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J., Allen, M.R., 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433, 403–6. doi:10.1038/nature03301

Tebaldi, C., Knutti, R., 2007. The use of the multi-model ensemble in probabilistic climate projections. Philos. Trans. A. Math. Phys. Eng. Sci. 365, 2053–75. doi:10.1098/rsta.2007.2076

Uusitalo, L., Lehikoinen, A., Helle, I., Myrberg, K., 2015. An overview of methods to evaluate uncertainty of deterministic models in decision support. Environ. Model. Softw. 63, 24–31. doi:10.1016/j.envsoft.2014.09.017

Van Oijen, M., Ewert, F., 1999. The effects of climatic variation in Europe on the yield response of spring wheat cv. Minaret to elevated CO2 and O3: an analysis of open-top chamber experiments by means of two crop growth simulation models. Eur. J. Agron. 10, 249–264. doi:10.1016/S1161-0301(99)00014-3

Wallach, D., Keussayan, N., Brun, F., Lacroix, B., Bergez, J.-E., 2012. Assessing the uncertainty when using a model to compare irrigation strategies. Agron. J. 104, 1274–1283. doi:10.2134/agronj2012.0038

Wallach, D., Thorburn, P.J., 2014. The error in agricultural systems model prediction depends on the variable being predicted. Environ. Model. Softw. 62, 487–494. doi:10.1016/j.envsoft.2014.08.001

Wintle, B.A., McCarthy, M.A., Volinsky, C.T., Kavanagh, R.P., 2003. The Use of Bayesian Model Averaging to Better Represent Uncertainty in Ecological Models. Conserv. Biol. 17, 1579–1590. doi:10.1111/j.1523-1739.2003.00614.x

Yang, J.M., Yang, J.Y., Liu, S., Hoogenboom, G., 2014. An evaluation of the statistical methods for testing the performance of crop models with observed data. Agric. Syst. doi:10.1016/j.agsy.2014.01.008

Yip, S., Stephenson, D.B., Hawkins, E., 2011. A Simple, Coherent Framework for Partitioning Uncertainty in Climate Predictions. J. Clim. 24, 4634–4643.

Appendix

Estimation of contributions from uncertainty in model structure, input vector and parameter vector to model variance.

We derive the ANOVA table for a random effects model with three factors (model formulations, input vectors and parameter vectors) , where parameter vectors is nested within model formulations x input vectors. The treatment here follows very closely that of (Scheffé, 1959, chapter 7). However, that source does not treat the case of interest here, of two crossed factors and one nested factor. The relation to $\mathrm{MSEP}_{\mathrm{random\,predictor}}$ is original here.

The random effects

Model formulations are assumed to come from an infinite population of models. It is not specifically the models in the sample that are of interest, but rather the variability in the population of models. Let u be an index which identifies a model. The models in the sample (u_i, i=1,…,I) are assumed to be drawn at random from the distribution of models.

The values of the input vector for each prediction situation are assumed to come from an infinite population of possible input vectors. For example, one of the input variables could be soil depth. This can be measured at various points within a field, but because of spatial heterogeneity there will be variability in the measurements, and therefore uncertainty in the true average soil depth. The distribution of $\hat{X} \mid X$ represents the distribution of estimated soil depth, for a specific field with some true average soil depth X. It is assumed that all models use the same estimated values of the input vector. Let v be an index which identifies an input vector. The input vectors in the sample (v_j, j=1,…,J) are assumed to be drawn at random from the distribution of input vectors.

The values of the parameter vector are assumed to come from an infinite population of possible parameter values. Let w|uv be an index which identifies a parameter vector for model formulation u, input vector v. The parameter vectors in the sample (v_k, k=1,…,K) are assumed to be drawn at random for each combination of u,v from the distribution of parameter vectors for those u,v. Thus this factor is nested within the combination of models and input values.

There is no error term here, since we are interested in simulated values which have no measurement error. The experimental setup for the simulation experiment is shown in Table A 1

Table A 1

Experimental design for simulations, showing I model formulations crossed with J input vectors. For each combination, there are K parameter vectors, all drawn independently from the distribution of parameters of the model in question.

Model formulation	Input vector	Parameter vector
$\hat{f}_1$	$\hat{X}_1$	$\hat{\theta}_{111},\ldots,\hat{\theta}_{11K}$
$\vdots$	$\vdots$	$\vdots$
$\hat{f}_1$	$\hat{X}_J$	$\hat{\theta}_{1J1},\ldots,\hat{\theta}_{1JK}$
$\vdots$	$\vdots$	$\vdots$
$\hat{f}_I$	$\hat{X}_1$	$\hat{\theta}_{I11},\ldots,\hat{\theta}_{I1K}$
$\vdots$	$\vdots$	$\vdots$
$\hat{f}_I$	$\hat{X}_J$	$\hat{\theta}_{IJ1},\ldots,\hat{\theta}_{IJK}$

Decompose *m(u,v,w)*

To conform to the notation in (Scheffé, 1959) we write the predictor as m(u,v,w). In the main text of this study the predictor is written as $\hat{f}(\hat{X};\hat{\theta})$, so we have u= $\hat{f}$ is model formulation, v= $\hat{X}$ is the estimated input vector and w= $\hat{\theta}$ is the estimated parameter vector. We define

$$
\begin{aligned}
&\mu = m(\cdot,\cdot,\cdot)\\
&a(u) = m(u,\cdot,\cdot) - m(\cdot,\cdot,\cdot)\\
&b(v) = m(\cdot,v,\cdot) - m(\cdot,\cdot,\cdot)\\
&c(u,v) = m(u,v,\cdot) - m(u,\cdot,\cdot) - m(\cdot,v,\cdot) + m(\cdot,\cdot,\cdot)\\
&t(u,v,w) = m(u,v,w) - m(u,v,\cdot)
\end{aligned}
$$

where a dot means that the expectation is taken over that variable. Then :

$$m(u,v,w) = \mu + a(u) + b(v) + c(u,v) + t(u,v,w)$$

In terms of our sample, we have

$$m_{ijk} = \mu + a_i + b_j + c_{ij} + t_{ijk}$$

where

$$
\begin{aligned}
a_i &= a(u_i)\\
b_j &= b(v_j)\\
c_{ij} &= c(u_i, v_j)\\
t_{ijk} &= t(u_i, v_j, w_k)
\end{aligned}
\qquad (A1)
$$

Since the joint distribution of (u_i, v_j, w_k) is the same as that of (u,v,w), the $\{a_i\}$ are all identically distributed, and similarly for $\{b_j\}$, $\{c_{ij}\}$ and $\{t_{ijk}\}$. They all have expectation 0. The variances are (see Table 1 in main)

$$
\begin{aligned}
\sigma_A^2 &= \mathrm{var}(a(u)) = \mathrm{var}(a_i) = \mathrm{var}\left[E(m(u,v,w)\,|\,u\right] \\
\sigma_B^2 &= \mathrm{var}(b(v)) = \mathrm{var}(b_j) = \mathrm{var}\left[E(m(u,v,w)\,|\,v\right] \\
\sigma_{AB}^2 &= \mathrm{var}(c(u,v)) = \mathrm{var}(c_{ij}) = \mathrm{var}\left[E(m(u,v,w)\,|\,u,v\right] - \sigma_A^2 - \sigma_B^2 \\
\sigma_T^2 &= \mathrm{var}(t(u,v,w)) = \mathrm{var}(t_{ijk}) \\
&= \mathrm{var}\left[E(t(u,v,w)\,|\,uv\right] + E\left[\mathrm{var}(t(u,v,w)\,|\,uv)\right] \\
&= E\left[\mathrm{var}(t(u,v,w)\,|\,uv)\right]
\end{aligned}
\quad \text{(A2)}
$$

The third equation above arises from that fact that

$$m(u,v,\square) - m(\square,\square,\square) = E\left[m(u,v,w)\,|\,u,v\right] - E\left[m(u,v,w)\right] = a(u) + b(v) + c(u,v)$$

Since all the terms on the right are uncorrelated, the total variance is the sum of the variances, so $\mathrm{var}\left[E(m(u,v,w)\,|\,u,v\right] = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2$. Rearranging gives the relation shown for σ_{AB}^2. The fourth relation results from the fact that $E\left[t(u,v,w)\,|\,u,v\right] = m(u,v,\square) - m(u,v,\square) = 0$ and so $\mathrm{var}\left\{E\left[t(u,v,w)\,|\,u,v\right]\right\} = 0$.

The total variance of the predictor is

$$\mathrm{var}\left[m_{ijk}\right] = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_T^2 \quad \text{(A3)}$$

which follows from the fact that all the random effects in eq. (A1) can be shown to be uncorrelated. According to eq. A2, $\sigma_A^2 = \mathrm{var}\left[E(m(u,v,w)\,|\,u\right]$ is the variance of the conditional expectation when model formulation is fixed. This is the usual criterion for the first order effect of u, which here represents model formulation. Similarly, $\sigma_B^2 = \mathrm{var}\left[E(m(u,v,w)\,|\,v\right]$ is the usual criterion for the first order effect of factor v, here input variables, and σ_{AB}^2 is the usual definition of the interaction effect, here between model formulation and input variables. On the other hand, σ_T^2, the parameter effect, is the expectation of conditional variance. That is, the contribution of parameter uncertainty is the variance due to parameters averaged over model formulations and input variables. This is unlike usual sensitivity analysis, but is appropriate for a nested factor.

Estimation of the variance components

The sums of squares of interest can be rewritten as shown in Table A 2. Here the dot notation indicates an average over the indices replaced by a dot. To obtain the MS column in the table, note that

$$
\begin{aligned}
m_{ijk} &= \mu + a_i + b_j + c_{ij} + t_{ijk} \\
m_{i\cdot\cdot} &= \mu + a_i + b_\cdot + c_{i\cdot} + t_{i\cdot\cdot} \\
m_{\cdot j\cdot} &= \mu + a_\cdot + b_j + c_{\cdot j} + t_{\cdot j\cdot} \\
m_{ij\cdot} &= \mu + a_i + b_j + c_{ij} + t_{ij\cdot} \\
m_{\cdot\cdot\cdot} &= \mu + a_\cdot + b_\cdot + c_{\cdot\cdot} + t_{\cdot\cdot\cdot}
\end{aligned}
$$

See (Scheffé, 1959) for the demonstration that groups of terms are uncorrelated. The expected mean squares E(MS) are the same as in Table 2 in main.

Table A 2

SS	MS	E(MS)
SS_A	$JK(1/I)\sum_{i=1}^{I}\left(m_{i\cdot\cdot} - m_{\cdot\cdot\cdot}\right)^2$ $= JK(1/I)\sum_{i=1}^{I}\left(a_i - a_\cdot + c_{i\cdot} - c_{\cdot\cdot} + t_{i\cdot\cdot} - t_{\cdot\cdot\cdot}\right)^2$	$JK\left(\sigma_A^2 + J^{-1}\sigma_{AB}^2 + J^{-1}K^{-1}\sigma_T^2\right)$
SS_B	$IK(1/I)\sum_{j=1}^{J}\left(m_{\cdot j\cdot} - m_{\cdot\cdot\cdot}\right)^2$ $= IK(1/I)\sum_{j=1}^{J}\left(b_i - b_\cdot + c_{i\cdot} - c_{\cdot\cdot} + t_{i\cdot\cdot} - t_{\cdot\cdot\cdot}\right)^2$	$IK\left(\sigma_B^2 + I^{-1}\sigma_{AB}^2 + I^{-1}K^{-1}\sigma_T^2\right)$

SS_{AB}	$K(1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J}\left(m_{ij\cdot}-m_{i\cdot\cdot}-m_{\cdot j\cdot}+m_{\cdot\cdot\cdot}\right)^2$ $=K(1/IJ)\sum_{i=1}^{I}\sum_{j=1}^{J}\left(c_{ij}-c_{i\cdot}-c_{\cdot j}+c_{\cdot\cdot}+t_{ij\cdot}-t_{i\cdot\cdot}-t_{\cdot j\cdot}+t_{\cdot\cdot\cdot}\right)^2$	$K\left(\sigma_{AB}^2+K^{-1}\sigma_T^2\right)$
SS_T	$(1/IJK)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(m_{ijk}-m_{ij\cdot}\right)^2$ $=(1/IJK)\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(t_{ijk}-t_{ij\cdot}\right)^2$	σ_T^2

1