Contents lists available at ScienceDirect

Environmental Modelling & Software

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

# Using site-specific data to estimate energy crop yield

A. Laurent [a, b], C. Loyce [b, a], D. Makowski [a, b], E. Pelzer [a, b, *]

[a] INRA, UMR 211 Agronomie, F-78850, Thiverval-Grignon, France
[b] AgroParisTech, UMR 211 Agronomie, F-78850, Thiverval-Grignon, France

ARTICLE INFO

*Article history:*
Received 23 April 2015
Received in revised form
4 September 2015
Accepted 16 September 2015
Available online 27 September 2015

*Keywords:*
Bayesian model
Energy crops
Site-specific data
Yield estimation
*Miscanthus* × *giganteus*

ABSTRACT

The estimation of energy crop yields is important, to help the firms responsible for collecting them to estimate biomass production in a given area, for example. A Bayesian modelling framework for site-specific yield estimation is presented in this paper. The proposed approach is based on a hierarchical model describing between-site and within-site yield variability. Probability distributions are used to describe the uncertainty of model estimations. The model can be fitted to site-specific yield data, to obtain both average and site-specific yield estimates. Site-specific yield data may be obtained from measurements for crop species other than those for which estimations are required, or from past measurements on perennial crop species grown over a period of several years at a given site. These two options were illustrated in two case studies, in which our model was used to estimate the yields of several energy crops. In most situations, site-specific yield estimations were more accurate than average estimations.

## 1. Introduction

Biomass (agricultural crops, wood, green or organic waste), as a source of renewable energy, could help to ensure security of the energy supply while reducing net greenhouse emissions and increasing agroecosystem diversity (Heaton, 2004; Kerckhoffs and Renquist, 2012). Biomass can be converted into several types of energy, such as heat, electricity, and biofuel. Energy crops compete for land with food and feed crops, and are therefore a source of controversy. The growth of energy crops on surplus cropland and degraded land unsuis for arable production appears to be a promising alternative (Metzger and Hüttermann, 2009; Rahman et al., 2014) that could reduce the competitive pressure for land. Another way of reducing this competition for land would be to select energy crops with high yields.

Yield estimations can address different types of questions. Consider, for example, *Miscanthus* × *giganteus* (hereafter referred to as *M. giganteus*), a perennial energy crop with a high yield potential (Heaton, 2004). This crop species is typically grown during 15–20 years. During the cultivation period, yield of *M. giganteus* varies from year to year. The yield tends to increase during the first 3–5 years and then reaches a maximum value (Lesur et al., 2013; Miguez et al., 2008). Temporal predictions of yield for this crop would thus be useful, as they would help farmers and collecting firms to anticipate the future yields of recently established crops in a given area, thereby making it possible to estimate more accurately the overall profitability of the crop, or the storage capacity required.

Yield estimations can also help bioenergy firms and farmers' advisers to select the most appropriate energy crop from a list of candidate species. It is generally possible to cultivate several types of energy crop in any given area (Cadoux et al., 2014). As crop yields vary considerably between sites and between years (Miguez et al., 2012), it is not easy to identify the species likely to be the most productive species. In the absence of yield data for a given energy crop at a site of interest, available yield values for other energy crops grown at the same site could be used to estimate yield of the missing crop species. The development of models of this kind could help bioenergy firms to diversify the energy feedstock.

Several types of process-based model have been proposed for the simulation of energy crop yields, particularly for *M. giganteus*. Clifton-Brown et al. (2000) developed a mechanistic model for predicting *M. giganteus* yield in Southern Ireland. Another mechanistic model, MISCANFOR, was developed by Hastings et al. (2009) for the prediction of *M. giganteus* yields as a function of climatic and soil conditions. Miguez et al. (2009) developed a semi-mechanistic model for estimating *M. giganteus* yield as a function of thermal time. Recently, Strullu et al. (2014) developed a process-based

[*] Corresponding author. INRA, UMR 211 Agronomie, F-78850, Thiverval-Grignon, France.
*E-mail address:* Elise.Pelzer@grignon.inra.fr (E. Pelzer).

http://dx.doi.org/10.1016/j.envsoft.2015.09.008
1364-8152/© 2015 Elsevier Ltd. All rights reserved.

model for simulating biomass dynamics in *M. giganteus* shoots. As this model runs for one crop species only, it cannot be used to compare yields of different crop species in a given growing area. More generally, the calibration of parameters of process-based models is difficult and requires a lot of experimental data (Wallach, 2011). Some input variables of these models may be difficult to measure in farmers' fields (e.g., mineral N in the soil at the beginning of the crop cycle), limiting their potential applications.

Lesur et al. (2013) and Miguez et al. (2008) developed statistical models for estimating yield trends over time for *M. giganteus*. Mola-Yudego and Aronsson (2008) also proposed a statistical model for estimating yields of *Salix* (another perennial crop) over a period of three years. These statistical models include only a limited number of input variables, facilitating their large-scale use, but they may not estimate yield values accurately, due to the variability of energy crop yields across sites and years.

A Bayesian modelling framework made it possible to combine statistical models with site-specific data for the estimation of energy crop yields. The principle is to adjust a statistical model to site-specific yield data, and then to use the fitted model to estimate unobserved yield values. With this approach, site-year effects are estimated through the single or small number of site-specific yield measurements used to adjust the model. These yield measurements may be collected for the species of interest or for other species cultivated at the same site. Thus, no information about soil and climate data is required. Another important advantage of the proposed Bayesian framework is that it provides a quantitative assessment of uncertainty about yield estimation, in the form of a probability distribution (Aguilera et al., 2011; Chen and Pollino, 2012).

The two datasets used in this study are described below. Then our Bayesian framework is presented and its use is illustrated in two case studies in which *i)* the yield of *M. giganteus* is estimated, using past yield data for this species collected at a specific site (case study 1), *ii)* the yields of 36 energy crop species are estimated from yield data collected for alternative crop species (case study 2). The accuracy of yield estimations is assessed in both case studies.

## 2. Materials and methods

### 2.1. Datasets used in the two case studies

The main characteristics of the two datasets are presented in Table 1 and described in detail below.

#### 2.1.1. Case study 1: dataset used to estimate yield of M. giganteus for one extra year

This dataset was used to estimate *M. giganteus* yields in future years from past yield data. Yield data were collected from 19 farmers' fields in eastern central France (Burgundy). This region has a semi-continental climate with a mean annual rainfall of 723 mm and a mean annual temperature of 10.9 °C (averaged over 2001–2014, measured locally at Ouges, 47°15′46.3″N, 5°4′26.1″E). *M. giganteus* crops were established on nine fields in 2009 and 10 fields in 2010.

From the second growing season onwards, *M. giganteus* yields were measured in February, as described by Bazot et al. (2014). Yield was not measured during the first growing season (2009 or 2010), because biomass production levels were too low (*M. giganteus* was crushed at the end of December). The last yield measurements were made in February 2014. Three to four sets of yield data were thus available per site, depending on the year of establishment. These data are presented in Fig. 1.

#### 2.1.2. Case study 2: dataset used to estimate yields of different energy crop species

In this case study, the yield of an energy crop was estimated in a given area from yield data collected in the same area, but for a different crop species. The objective was to estimate the yields of an energy crop species, hereafter denoted as species 2, for site-years for which yields were measured for a reference species, hereafter denoted as Ref, different from species 2. The idea was to estimate the yield of species 2 in a given site-year from the yield of Ref measured for the same site-year, based on the correlation between the yields of the reference species Ref and species 2 in other site-years.

A dataset of 856 observations of yield, expressed in tons of dry matter per ha and per year (Laurent et al., 2015) was used. These yield data were collected in 93 experimental site-years in 12 countries and were extracted from 28 published scientific papers. Yields of at least two species were measured for each of the experimental site-years included in the dataset. A separate subset of data was defined for each of 31 pairs of species (Table 2, Fig. 2), and was used to estimate yields of one species from yield data collected for the other species in the same site-year.

### 2.2. Statistical model for yield estimations

#### 2.2.1. General framework

A hierarchical Bayesian statistical model for crop yield estimation was defined. A within-group level, a between-group level, and a level defining prior distributions were included. Groups corresponded either to sites (case study 1) or site-years (case study 2) (Table 1).

*2.2.1.1. Within-group level.* This level describes the probability distribution of the yield data within a given group (i.e., within a given site or site-year). Let $Y_{ij}$ be the jth yield data collected in the ith group. $Y_{ij}$ is related to a set of explanatory variables $X_{ij}$ (e.g., time, crop species; Table 1) as follows:

$$Y_{ij} = f(X_{ij}, \theta_i) + \varepsilon_{ij} \tag{1}$$

$$\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$$

where *f* is a function relating $Y_{ij}$ to $X_{ij}$ and to a set of group-specific parameters $\theta_i$, and $\varepsilon_{ij}$ is a residual term. Here, all residuals are assumed independent and normally distributed with variance $\sigma_\varepsilon^2$, but Eq. (1) can be modified to deal with other types of distributions.

**Table 1**
Main characteristics of the datasets used in case studies 1 and 2.

| Case study | Group (index *i*) | | Explanatory variables ($X_{ij}$) | | Model function (f) |
|---|---|---|---|---|---|
| | Definition | Number | Definition | Number | Type |
| 1 | Site | 9 or 10 | Time | 3 or 4 | Logistic |
| 2 | Site-year | From 2 to 25 | Species | 2 | Linear |

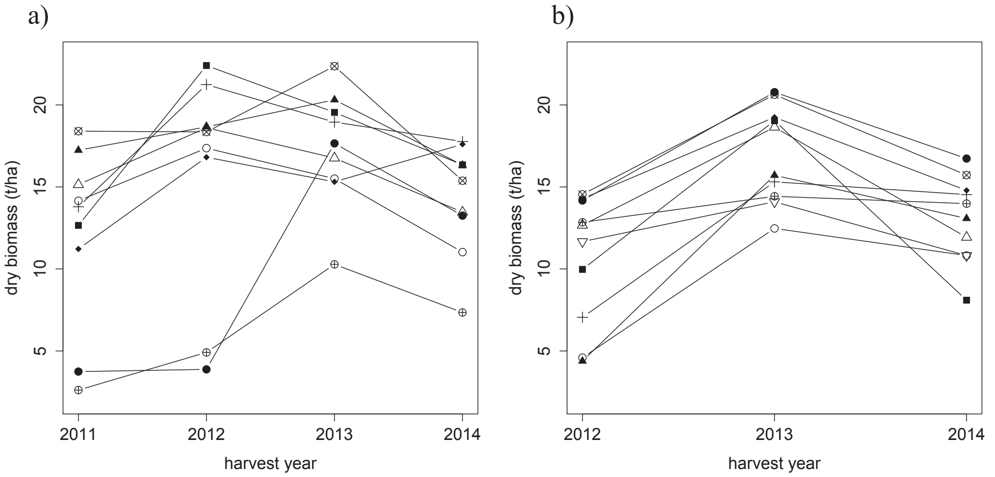

**Fig. 1.** Yield of *Miscanthus* × *giganteus* (case study 1) for the nine farmers' fields established in 2009 (subset 1) (a), and for the 10 farmers' fields established in 2010 (subset 2) (b). Points linked by a line come from the same field.

**Table 2**
Correlation coefficients (ρ) for the relationship between the yields of pairs of crop species (case study 2). Correlation coefficients were estimated from yield data obtained for 3 to 25 site-years, depending on the pair of species considered.

| Pairs of crop species | Number of site-years used for each pair of species | ρ |
|---|---|---|
| *Panicum virgatum & Sorghum bicolor* | 5 | 0.052 |
| *Saccharum* spp *& Arundo donax* | 8 | 0.057 |
| *Miscanthus* × *giganteus & Triticosecale* | 4 | 0.083 |
| *Miscanthus* × *giganteus & Pennisetum purpureum* | 6 | 0.092 |
| *Miscanthus* × *giganteus & Medicago sativa* | 4 | 0.122 |
| *Miscanthus* × *giganteus & Sorghum bicolor* | 4 | 0.129 |
| *Arundo donax & Cynara cardunculus* | 4 | 0.154 |
| *Triticosecale & Triticum aestivum* | 6 | 0.169 |
| *Triticosecale & Secale cereale* | 8 | 0.218 |
| *Triticosecale & Festuca arundinacea* | 4 | 0.219 |
| *Sorghum bicolor & Triticosecale* | 4 | 0.227 |
| *Medicago sativa & Panicum virgatum* | 4 | 0.257 |
| *Miscanthus sinensis & Panicum virgatum* | 4 | 0.263 |
| *Pennisetum purpureum & Arundo donax* | 8 | 0.268 |
| *Phalaris arundinacea & Medicago sativa* | 4 | 0.287 |
| *Triticosecale & Panicum virgatum* | 4 | 0.343 |
| *Zea mays & Medicago sativa* | 3 | 0.419 |
| *Medicago sativa & Triticosecale* | 4 | 0.593 |
| *Panicum virgatum & Festuca arundinacea* | 10 | 0.674 |
| *Miscanthus* × *giganteus & Arundo donax* | 25 | 0.708 |
| *Medicago sativa & Festuca arundinacea* | 4 | 0.728 |
| *Miscanthus* × *giganteus & Miscanthus sinensis* | 4 | 0.751 |
| *Miscanthus* × *giganteus & Phalaris arundinacea* | 4 | 0.793 |
| *Miscanthus* × *giganteus & Sida hermaphrodita* | 4 | 0.822 |
| *Miscanthus* × *giganteus & Panicum virgatum* | 10 | 0.840 |
| *Salix & Miscanthus sacchariflorus* | 4 | 0.857 |
| *Miscanthus* × *giganteus & Miscanthus sacchariflorus* | 4 | 0.859 |
| *Miscanthus* × *giganteus & Salix* | 5 | 0.876 |
| *Miscanthus* × *giganteus & s australis* | 4 | 0.923 |
| *Panicum virgatum & Panicum amarum* | 5 | 0.938 |
| *Salix & Sida hermaphrodita* | 7 | 0.966 |

*2.2.1.2. Between-group level.* This level describes the between-group variability of the parameter $\theta_i$ as follows:

$$\theta_i = \mu + b_i \quad (2)$$

with $b_i$~$N(0,\sum)$, $\mu$ is the expected value of $\theta_i$ in the population of groups, $\Sigma$ is the variance-covariance matrix of $\theta_i$. The diagonal elements of $\Sigma$ correspond to the between-group variances of $\theta_i$ and the off-diagonal elements correspond to their covariances.

*2.2.1.3. Priors.* Prior probability distributions are given for μ and Σ. They are described separately for the two case studies in the next two sections.

### *2.2.2. Case study 1*

In this case study, each group corresponds to a specific site (i.e. a farmer's field) and *f* is a logistic function expressed as:

$$f(X_{ij},\theta_i) = \frac{Y_{\max,i}}{1+\exp\left(\frac{\alpha_i - X_{ij}}{\beta}\right)} \quad (3)$$

where $X_{ij}$ is the date (year since planting) on which the *j*th yield data are collected at the *i*th site, and $\theta_i=(\alpha_i,\beta,Y_{max,i})^T$ is the vector of

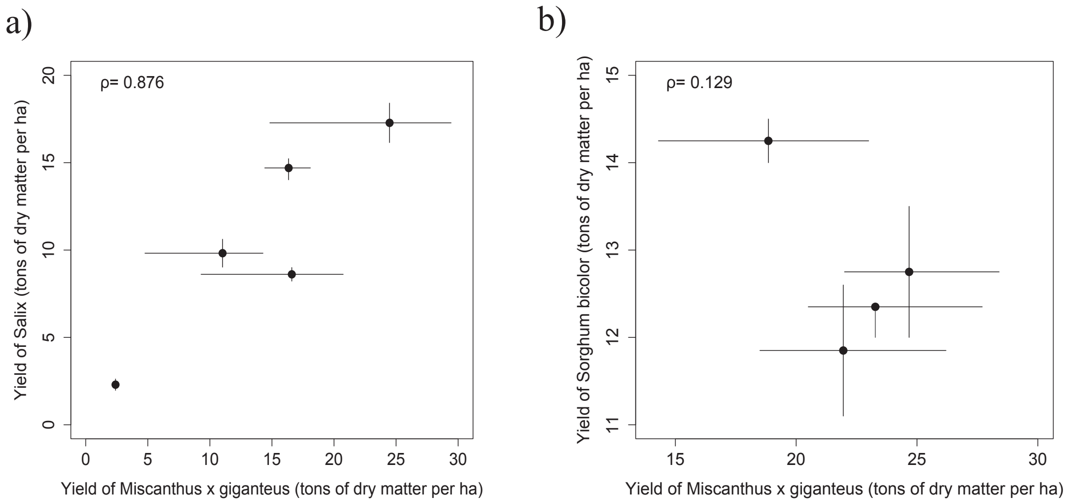

**Fig. 2.** Examples of relationships between the yields of pairs of crop species (case study 2). Black points represent yield values for *Salix* vs. *Miscanthus* × *giganteus* in five site-years (a) and yield values for *Sorghum bicolor* vs. *Miscanthus* × *giganteus* in four site-years (b). When several data are available for a given site-year, the minimum and maximum yield values are indicated by horizontal and vertical bars.

**Table 3**
Description of the two subsets of data used to evaluate model estimations for *Miscanthus* × *giganteus* yields (case study 2). The root mean square error of prediction (RMSEP) (t ha$^{-1}$) was calculated for two types of yield estimations: average estimations and site-specific estimations. The last year of yield data was estimated in both cases. RMSEP was calculated separately for the two subsets of data.

| Data subset | Number of sites (farmers' fields) | Number of years per site | Year of establishment | RMSEP average estimation (t ha$^{-1}$) | RMSEP site-specific estimation (t ha$^{-1}$) | Relative change in RMSEP (%) |
|---|---|---|---|---|---|---|
| 1 | 9 | 4 | 2009 | 5.14 | 3.16 | 62.7 |
| 2 | 10 | 3 | 2010 | 2.68 | 2.54 | 5.5 |

parameter values for the *i*th site.

Miguez et al. (2008) has already used this logistic model to estimate yield trends for *M. giganteus*. Lesur et al. (2013) compared several versions of this model, and showed that the version including two random parameters ($\alpha_i$,$Y_{max,i}$) and one fixed parameter ($\beta$) had the best performance. It was therefore assume that $\beta$ is constant and that $\alpha_i$ and $Y_{\text{max},i}$ vary across sites according to two independent Gaussian distributions (Lesur et al., 2013; Miguez et al., 2008); $\alpha_i \sim N(\mu_\alpha, \sigma^2_\alpha)$ and $Y_{\text{max},i} \sim N(\mu_{\text{Ymax}}, \sigma^2_{\text{Ymax}})$. The model output describes the yield dynamic as a function of time since planting (Lesur et al., 2013; Miguez et al., 2008). The posterior distributions of the parameters were estimated from the dataset for Burgundy described in 2.1.2 with WinBUGS software (Lunn et al., 2000) and three chains of 100,000 MCMC iterations. Convergence was evaluated with the Gelman and Rubin diagnosis (Brooks and Gelman, 1998). The first 20,000 iterations were removed.

Estimation accuracy was evaluated by calculating separate root mean square errors of prediction (RMSEP) for the average and site-specific estimations. RMSEP was estimated by leave-one-out cross-validation. For each field subset (Table 3), each of the individual yield data obtained in 2014 were removed, one at a time, and all the remaining data were used for estimating parameters with the Bayesian procedure described above. The removed data were then estimated in two different ways, using the posterior median of average estimation:

$$\frac{\mu_{\text{Ymax}}}{1+\exp\left(\frac{\mu_\alpha - X_{ij}}{\beta}\right)} \tag{4}$$

and the posterior median of site-specific estimation:

$$f(X_{ij},\theta_i) = \frac{Y_{\text{max},i}}{1+\exp\left(\frac{\alpha_i - X_{ij}}{\beta}\right)} \tag{5}$$

An average estimation represents an expected yield value in a population of sites at a given year *X*. A site-specific estimation corresponds to a yield estimate derived for a given site at a given year *X*. Average and site-specific estimations were both derived using the same yield dataset. In both cases, the yield dataset included the yield data collected at the site for which the estimation was made, but before the last date of measurement, and all the yield data collected at the other sites. However, this dataset was not used in the same way to derive average and site-specific estimations. For average estimations, yield data were used to estimate the mean parameters $\mu$ characterizing the population of sites. For site-specific estimations, yield data were used to estimate the site-specific parameters $\theta_i$.

### 2.2.3. Case study 2

A model was developed for each pair of crop species presented in Table 2. In this case study, the function *f*, defined in the general framework, is a linear function of Eq. (1), expressed as:

$$f(X_{ij},\theta_i) = X_{ij}\theta_i \tag{6}$$

where $X_{ij} = (1, X^{(2)}_{ij})$. $X^{(2)}_{ij}$ is a set of binary variables indicating which crop species corresponds to the *j*th yield data collected in the *i*th site-year. $X^{(2)}_{ij} = 1$ if $Y_{ij}$ is the log-transformed yield of species 2 at the *i*th site-year and $X^{(2)}_{ij} = 0$ if $Y_{ij}$ is the log-transformed yield of the reference species Ref. The vector $\theta_i = (\theta_i^{(ref)}, \alpha^{(2)})^T$ is the set of parameter values for the $i$th site-year, with $\theta_i^{(ref)}$ being the true log-transformed yield value in the $i$th site-year for the reference species Ref and $\alpha^{(2)}$ the true difference between the log-transformed yield of the reference species Ref and the log-transformed yield of species 2. A log transformation was used to normalize yield data as recommended in Laurent et al. (2015). The true log-transformed yield value of the reference species $\theta_i^{(ref)}$ is expressed as $\theta_i^{(ref)} = \mu_{ref} + b_i$ where $\mu_{ref}$ is the expected value of $Y_{ij}$ across site-years and $b_i$ is a random site-year effect, $b_i \sim N(0, \sigma_b^2)$.

With this notation, Eq. (1), for the reference species becomes:

$$Y_{ij} = \mu_{ref} + b_i + \varepsilon_{ij} \tag{7}$$

and for the species 2 becomes:

$$Y_{ij} = \mu_{ref} + \alpha^{(2)} + b_i + \varepsilon_{ij} \tag{8}$$

This model can be used to estimate the yield of crop species 2 for a site-year for which yield data were obtained for the reference crop species only. Assuming that a yield measurement for the reference crop species is available for the $i$th site-year, the probability distribution of the log-transformed yield of species 2 conditionally to the log-transformed measured yield of the reference crop species Ref is expressed as:

$$Y_{i2} \Big| Y_{i,ref} \sim N\left(\mu_{i2/ref}, \sigma^2_{i2/ref}\right)$$

with

$$\begin{aligned} \mu_{i2|ref} &= \mu_{ref} + \alpha_2 + \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2} \times \left(Y_{ref,i} - \mu_{ref}\right) \\ &= \mu_{ref} + \alpha_2 + \rho \times \left(Y_{ref,i} - \mu_{ref}\right) \end{aligned} \tag{9}$$

$$\begin{aligned} \sigma^2_{i2|ref} &= \sigma_b^2 + \sigma_\varepsilon^2 - \frac{\sigma_b^4}{\sigma_b^2 + \sigma_\varepsilon^2} \\ &= \left(\sigma_b^2 + \sigma_\varepsilon^2\right) \times \left(1 - \rho^2\right) \end{aligned} \tag{10}$$

where

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\varepsilon^2} \tag{11}$$

is the correlation between the log-transformed yield of the reference species and the log-transformed yield of species 2. Eq. (9) shows that the conditional expected value of the log-transformed yield of species 2 depends on its expected value, on the difference between the measured log-transformed yield value of the reference species and its expected value $Y_{i,ref}-\mu_{ref}$, and on the correlation between the log-transformed yields of the two species (ρ). Eq. (10) shows that the use of the measured yield value of the reference species can reduce the uncertainty concerning the yield of species 2, but the extent of the reduction depends on the correlation (ρ) between the yields of the two species. The reduction of uncertainty increases with the strength of the correlation. If the yields of the two species are not correlated, then the uncertainty is not decreased at all.

This model was fitted, by a Bayesian approach, to each pair of species in turn, using the corresponding subset of data (Table 2). Two examples of data subsets are shown in Fig. 2. The following prior distributions were used: $\mu_{ref}$~$N(0,10^5)$, $\alpha^{(2)}$~$N(0,10^5)$, $1/\sigma_b^2 \sim Gamma(10^3, 10^3)$, and $1/\sigma_\varepsilon^2 \sim Gamma(10^3, 10^3)$. The posterior distributions of the parameters were estimated with WinBUGS software (Lunn et al., 2000), with tree chains of 50,000 MCMC iterations. Convergence was evaluated with the Gelman and Rubin diagnosis (Brooks and Gelman, 1998). The first 20,000 iterations were removed.

Each fitted model was used to estimate the yield of species 2 in two different ways; by an average yield estimation equal to the posterior median of the average yield $\mu_{ref} + \alpha^{(2)}$, and by a site-specific yield estimation equal to the posterior median of $\mu_{i2/ref}$ (Eq. (9)). The average estimation represents the expected yield of species 2 in the considered population of sites. The site-specific estimation corresponds to a yield estimate derived for species 2 for a specific site. As in case study 1, the same yield dataset was used to derive the two types of yield estimation, but not in the same way. Average yield estimations were calculated using the population parameters $\mu_{ref}$ and $\alpha^{(2)}$, whereas site-specific estimations were calculated from the site-specific parameter $\mu_{i2/ref}$ (Eq. (9)). The accuracy of yield estimations was evaluated by leave-one-out cross-validation. Yield values were removed one-by-one from each subset of data, and the model was fitted to the remaining data and used to estimate the missing data, through the generation of an average estimation and a site-specific estimation. Prediction accuracy was evaluated by calculating a root mean square error of prediction (RMSEP) separately for the average and site-specific estimation. The relative change in RMSEP resulting from the use of an average estimation rather than a site-specific estimation is evaluated as follows:

$$Relative\ change\ of\ RMSEP = \frac{RMSEP\ average\ estimation - RMSEP\ site\ specific\ estimation}{RMSEP\ site\ specific\ estimation} \times 100 \tag{12}$$

## 3. Results

### 3.1. Case study 1

*M. giganteus* yields varied considerably between fields and years. In the data subset including fields of crops established in 2009, yields ranged from 2.6 to 18.4 t ha$^{-1}$ in 2011 and from 7.3 to 17.8 t ha$^{-1}$ in 2014 (Fig. 1a). In the subset of data including fields of crops established in 2010, yields ranged from 4.4 to 14.5 t ha$^{-1}$ in 2012, and from 8.1 to 16.7 t ha$^{-1}$ in 2014 (Fig. 1b).

In subset 1, the RMSEP was 3.16 t ha$^{-1}$ for site-specific estimations and 5.14 t ha$^{-1}$ for average yield estimations. For subset 2, the RMSEP was 2.54 t ha$^{-1}$ for site-specific estimations and 2.68 t ha$^{-1}$ for average yield estimations. RMSEP values were thus lower for site-specific estimations than for average estimations, for both subsets of data (Table 3). The site-specific yield estimations thus tended to be more accurate. The difference between the two types of estimation was greater when four yield values were available for yield estimation (subset 1) than when only three values were

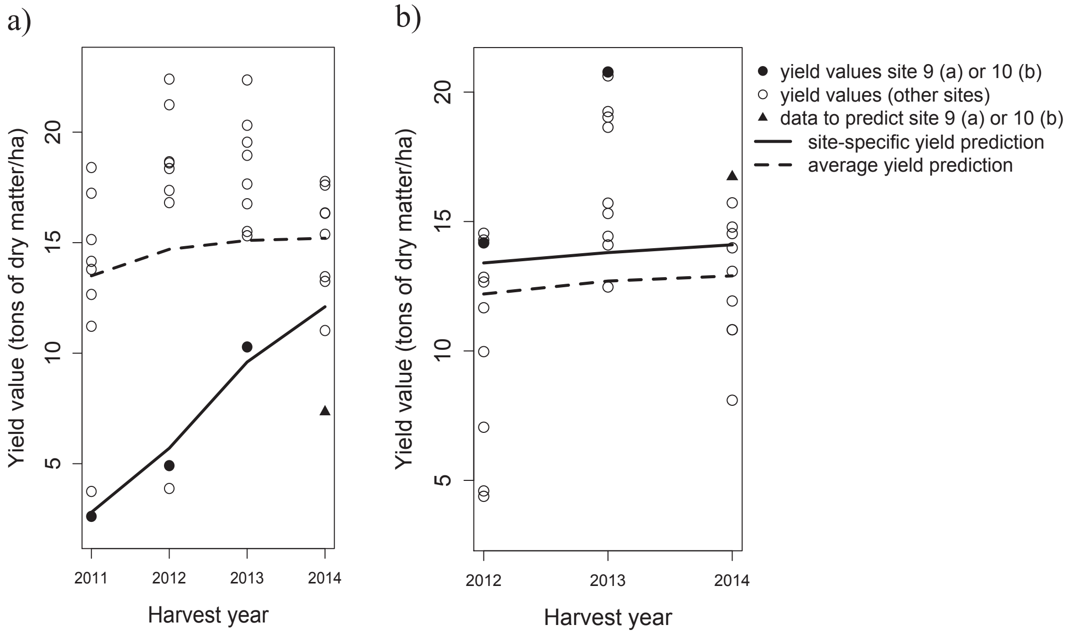

**Fig. 3.** Average and site-specific yield estimations of the logistic model (case study 1). Data collected in farmers' fields established in 2009 (a) or in 2010 (b). Site-specific estimations are presented for site 9 (a) and for site 10 (b). The first year of yield measurement was 2011 (a) or 2012 (b).

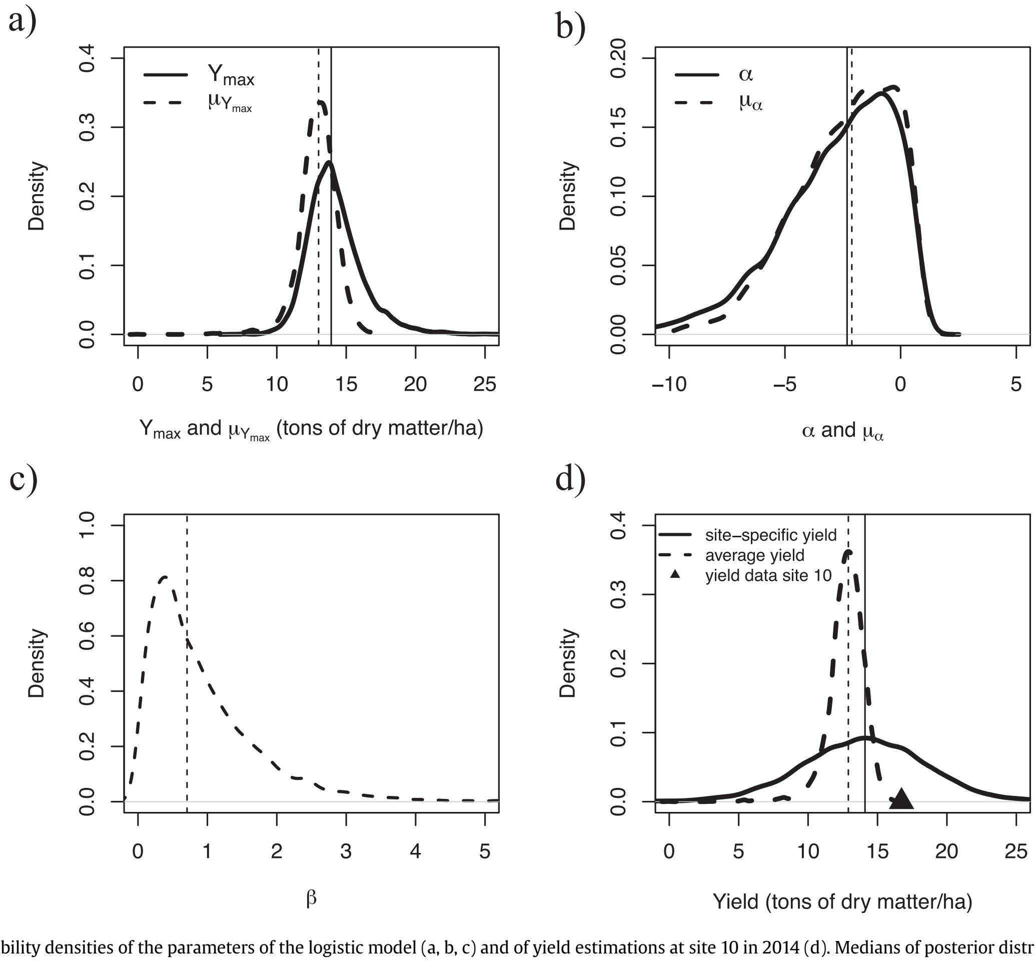

**Fig. 4.** Posterior probability densities of the parameters of the logistic model (a, b, c) and of yield estimations at site 10 in 2014 (d). Medians of posterior distributions of average and site-specific estimations are indicated by vertical dashed lines and vertical continuous lines respectively.

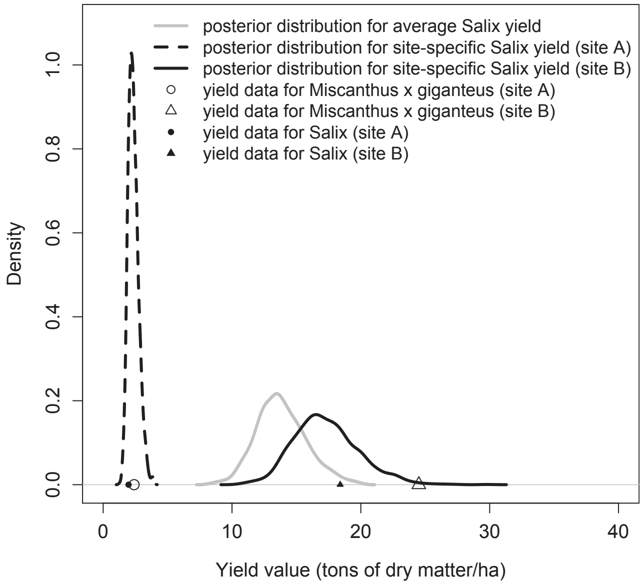

**Fig. 5.** Posterior probability densities for average and site-specific yield estimations for *Salix* in two different site-years (A and B) characterized by different *Miscanthus* × *giganteus* yields (case study 2).

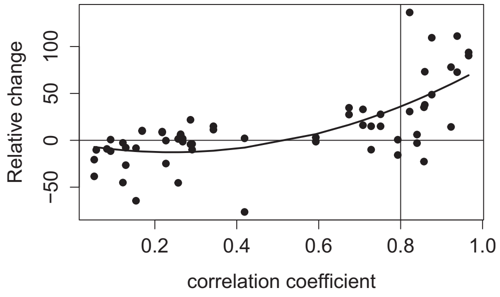

**Fig. 6.** Relative change in RMSEP resulting from the use of average rather than site-specific estimations, as a function of the coefficient of correlation between the yields of pairs of crop species (case study 2). The response curve was fitted by quadratic polynomial regression. Relative change = $\left(\frac{\text{RMSEP average estimation}-\text{RMSEP site-specific estimation}}{\text{RMSEP site-specific estimation}}\right) \times 100$.

available (subset 2) (Table 3).

Examples of site-specific estimations and average estimations are presented in Fig. 3, for two different fields. At site 9, which was established in 2009, the site-specific yield estimation for 2014 was 12.1 t ha$^{-1}$ and the average yield estimation for the same year was 15.2 t ha$^{-1}$ (Fig. 3a). Although not perfect, the site-specific estimation was thus closer to the actual yield (7.3 t ha$^{-1}$) than the average estimation. Similarly, at site 10, established in 2010, the site-specific yield estimation was 14.1 t ha$^{-1}$ and the average yield estimation was 12.9 t ha$^{-1}$ (Fig. 3b), whereas the actual yield was 16.7 t ha$^{-1}$. Once again, the site-specific estimation was closer to the yield data. In addition to point estimates, our model was used to calculate posterior probability distributions. Examples of distributions obtained for average and site-specific estimations are shown in Fig. 4. The posterior distribution of $\mu_{Y_{\max}}$ (average value of potential yield in the population of sites) was more peaked than the posterior distribution of $Y_{max}$ obtained for site 10 (potential yield in the considered site) (Fig. 4a). This is logical because the uncertainty about a mean value is usually lower than the uncertainty about an individual value. The posterior median obtained for $Y_{max}$ was higher than the posterior median of $\mu_{Y_{\max}}$ (Fig. 4a). This result indicates that the estimated potential yield in site 10 was higher than the average potential yield estimated for the population of sites. This is a logical result because observed yields tended to be higher in site 10 compared to the other sites (Fig. 3b). Fig. 4b did not reveal any strong difference between the posterior distributions of $\mu_a$ and $\alpha$. This result shows that the value of $\alpha$ is site 10 was close to its mean value. The posterior distribution of the mean estimate was more peaked than the posterior distribution of the site-specific yield for site 10 in 2014 (Fig. 4d). The mean yield value was thus more accurately estimated than the site-specific yield value. The posterior median of the site-specific yield value was higher than the estimated mean yield (Fig. 4d). This result is consistent with the fact that the potential yield estimated for site 10 was higher than the average potential yield in the population of sites (Fig. 4a).

### 3.2. Case study 2

The site-specific and average estimations of *Salix* yields obtained with the Bayesian linear model (*M. giganteus* as Ref and *Salix* as $S_2$) are presented in Fig. 5. One advantage of this model is that it can be used to calculate the posterior probability distributions associated with both average and site-specific estimations. These distributions described the full range of possible yield values. The median value of the posterior probability distribution can be seen as a point prediction and the whole distribution provides an indication of the uncertainty.

The posterior probability distribution for the average estimation of *Salix* yield is presented in Fig. 5 (in grey). The median of this posterior distribution was equal to 13.8 t ha$^{-1}$. The site-specific estimation was presented for two sites (sites A and B) showing contrasted *M. giganteus* yield values. The black dashed line indicates the posterior distribution for site-specific yield estimation at site A. As *M. giganteus* yield was very low at this site (2.4 t ha$^{-1}$), the median of the site-specific posterior distribution obtained for *Salix* (2.3 t ha$^{-1}$) was much lower than the posterior median obtained for average yield estimation for *Salix* (13.8 t ha$^{-1}$). The black line indicates the posterior distribution of the site-specific yield of *Salix* at site B. *M. giganteus* yield at this site was high (24.6 t ha$^{-1}$), so the median of the site-specific posterior distribution was higher (17.3 t ha$^{-1}$) than the posterior median obtained for average estimation of the *Salix* yield. For sites A and B, the medians of the site-specific posterior distributions were similar to the actual yields of *Salix* at these sites (1.97 t ha$^{-1}$ and 18.4 t ha$^{-1}$ for sites A and B, respectively). However, high levels of uncertainty remained, particularly for site B. Due to the log transformation, the size of the confidence interval was not constant and depended on yield values.

The relative changes in RMSEP resulting from the use of average estimations (medians of the average posterior distributions) instead of site-specific estimations (medians of the site-specific posterior distributions) were displayed in Fig. 6 as a function of the correlation coefficient for the yields of the pairs of species considered. This coefficient provided an indication of the strength of the relationship between the yield data for two crop species. Examples of correlations are shown in Fig. 2, for two different pairs of species, one with a high correlation coefficient (0.87) and the other with a low correlation coefficient (0.13).

It was found that the relative change in RMSEP depended on the correlation coefficient. If the correlation between the yield of the reference species and the yield of species 2 was strong, site-specific estimations for species 2 tended to be more accurate than average estimation for the same species. For correlation coefficients exceeding 0.5, the RMSEP value was reduced by the use of site-specific data for 23 of the 28 pairs of species considered (*i.e.* 2*14 couples of species). For correlation coefficients greater than 0.8, RMSEP was reduced by the use of site-specific estimations for 14 of

**Table 4**
Coefficient of correlation between the yields of the reference species and species $S_2$ (ρ) and root mean square error of prediction (RMSEP) (t $ha^{-1}$) for average estimations and for site-specific estimations of the yield of species $S_2$ (case study 2). RMSEP values were calculated separately for each pair of crop species. Only pairs of crop species with a correlation coefficient greater than 0.8 are presented here. Correlation coefficients were estimated from yield data collected in 4–10 site-years.

| Reference species | Species 2 | (ρ) | Number of site-years | RMSEP average estimation (t $ha^{-1}$) | RMSEP site-specific estimation (t $ha^{-1}$) |
|---|---|---|---|---|---|
| *Salix* | *S. hermaphrodita* | 0.966 | 7 | 6.82 | 3.52 |
| *S. hermaphrodita* | *Salix* | | | 5.29 | 2.78 |
| *P. virgatum* | *P. amarum* | 0.938 | 5 | 5.81 | 2.75 |
| *P. amarum* | *P. virgatum* | | | 4.24 | 2.46 |
| *M. giganteus* | *P. australis* | 0.923 | 4 | 1.44 | 1.26 |
| *P. australis* | *M. giganteus* | | | 10.63 | 5.97 |
| *M. giganteus* | *Salix* | 0.876 | 5 | 5.84 | 2.79 |
| *Salix* | *M. giganteus* | | | 9.39 | 6.32 |
| *M. giganteus* | *M. sacchariflorus* | 0.859 | 4 | 5.68 | 3.28 |
| *M. sacchariflorus* | *M. giganteus* | | | 10.6 | 7.69 |
| *Salix* | *M. sacchariflorus* | 0.857 | 4 | 5.65 | 4.18 |
| *M. sacchariflorus* | *Salix* | | | 4.61 | 5.96 |
| *M. giganteus* | *P. virgatum* | 0.840 | 10 | 5.66 | 5.83 |
| *P. virgatum* | *M. giganteus* | | | 9.48 | 8.93 |
| *M. giganteus* | *S. hermaphrodita* | 0.822 | 4 | 8.06 | 3.40 |
| *S. hermaphrodita* | *M. giganteus* | | | 9.34 | 6.32 |

the 16 pairs of species considered (*i.e.* 2*8 couples of species). RMSEP values are presented in Table 4 for pairs of species with correlation coefficients greater than 0.8. For these pairs, the reduction of RMSEP resulting from the use of site-specific estimations tended to be large. For *Salix* (as Ref) and *Sida hermaphrodita* (as $S_2$), the RMSEP value was 3.52 t $ha^{-1}$ for site-specific estimation, and 6.82 t $ha^{-1}$ for average estimation. For *M. giganteus* (as Ref) and *Salix* (as $S_2$) (Fig. 2a, Table 4), the RMSEP value was 2.79 for site-specific estimation, and 5.84 for average estimation method.

For weak correlations between the yields of two species, the use of site-specific yield estimations did not generally reduce RMSEP. For example, the coefficient of correlation obtained for *Sorghum bicolor* and *M. giganteus* was 0.129 (Fig. 2b). For this pair of species, the RMSEP value was 2.0 for site-specific estimation and 1.47 for average estimation (example not shown in Table 4).

## 4. Discussion

Our Bayesian framework is useful for estimating crop yields from site-specific yield data. The flexibility of this framework was illustrated in two case studies using different types of data and different types of functions. In the first case study, a nonlinear function was used to estimate yield of a perennial crop species (*M. giganteus*) grown over several years at a given location, from past site-specific yield data obtained during the first few years of crop establishment. Such estimations can help agricultural co-operatives or biomass supply companies to anticipate future storage capacity needs for *M. giganteus*. For example, about 120 $m^3$ of storage capacity is required for every hectare of a *M. giganteus* producing 12 t $ha^{-1}$ $yr^{-1}$ dry matter (El Bassam and Huisman, 2001). Such estimations could also be used as inputs in economic studies carried out to evaluate the sustainability of *M. giganteus* supply chains (De la Rua et al., 2015). In the future, similar models could also be developed for other perennial crops, such as *Panicum virgatum* (switchgrass) or *Arundo donax* (giant reed).

The second case study showed that our framework could be used to estimate the yields of different crop species from the yields obtained for other crops (provided that the yield of the crop concerned is correlated with that of the species for which estimations are required). Our approach constitutes an interesting alternative to process-based models. So far, such models were developed for a very limited number of energy crop species only. New process-based models could be developed in the future, but the development of such models is costly and takes times. Our approach could help farmers, agricultural advisers or collecting firms to select the most appropriate crop species for a given area in the absence of information about its yield in that area. The proposed statistical model can estimate the yield of a species of interest from the yield data recorded for another species grown at the same sites. Yield estimations can then be used to rank several crop species according to their expected productivity.

Both the case studies presented here showed that site-specific yield estimations were frequently, but not systematically, more accurate than average yield estimations. Our results are consistent with those of Philibert et al. (2014), who proposed a similar method for predicting $N_2O$ emissions and showed that the use of location-specific prediction rather than average prediction reduced prediction errors in most cases. Site-specific yield estimations tend to be more accurate because they are adjusted according to the characteristics of the local environment, through the use of local yield measurements. The yield data used in our study were collected under various environmental conditions, characterized by different soil and climate characteristics. Pogson et al. (2012) highlighted the importance of taking site characteristics (such as soil type and soil water content) into account when modelling crop growth. These characteristics were indirectly taken into account in our statistical model, by estimating site effects using site-specific yield data. This approach made it possible to adjust yield estimations to local characteristics without the need for input variables describing soil and climate characteristics.

Our statistical model could be used to derive average and site-specific yield estimates. In case study 1, site-specific estimations of *M. giganteus* were more accurate than average estimations if three past yield values were available for parameter estimation. When only two yield values were available per site, site-specific and average estimations performed similarly well. In case study 1, yields of *M. giganteus* were estimated using a logistic function but our framework could be implemented with other functions. In situations where longer yield time series are available, it could be useful to use an exponential function to take into account a yield decline after a certain period of time (Lesur et al., 2013). In case study 2, site-specific estimation outperformed average estimation when the yields of pairs of crop species were strongly correlated (correlation coefficient > 0.8). For weaker correlations, site-specific yield estimations were not systematically more accurate than average estimations. In the future, the approach described in case study 2 could be adapted to deal with several cultivars of the same species. In cases of strong correlation between the yields of two cultivars, the use of yield data collected for one cultivar could be used to improve yield estimations for a second cultivar grown at the same site. The parameters of equations 9 and 10 could be estimated in multisite trials comparing and ranking cultivars (Crespin and Soyer, 2002; Piepho, 1995). It would then be possible to use the fitted model to estimate the yield of a cultivar of interest at new sites, from the yield data collected for one or several other cultivars grown at this new site.

One interesting feature of the proposed Bayesian method is its description of the uncertainty associated with yield estimations through probability distributions. These distributions can be used to define ranges of plausible yield values for various energy crop species, and to explore a set of scenarios of biomass production based on realistic production levels. Yield estimates derived from our model may serve as inputs for multi-criteria assessments of cropping systems or life cycle assessment (Almeida et al., 2014; Gasol et al., 2009). The yield probability distributions computed by our models could be used to take the uncertainty associated with yield estimation into account when calculating economic indicators (e.g., crop profitability) or energy indicators (e.g., energy balance). These probability distributions may also be useful for defining agricultural scenarios in foresight studies and life cycle assessment (Hillier et al., 2009). As the outcome of such analyses is often sensitive to the yield value used (Almeida et al., 2014), it is more appropriate to use distributions of yield values rather than point values in studies assessing the environmental and economic impacts of different agricultural scenarios.

## 5. Conclusion

The proposed Bayesian method can estimate crop yields and describe the uncertainty of the estimated yield values. One interesting feature of this method is that it can adjust yield forecasts according to local characteristics, on the basis of local yield data. In practice, this method can be used to improve yield estimations from site-specific yield data for a large range of species, without the need for soil and climate input variables. When yield data are available for a given species at the site of interest, our method calculates the posterior yield probability distribution for another species grown at the same sites. For perennial crops, our method can estimate yield values from past yield data. In all cases, the proposed method provides a transparent description of the levels of uncertainty associated with yield predictions.

## Acknowledgments

This work was funded by the EU Seventh Framework Programme, through LogistEC project No. 311858: Logistics for Energy Crop Biomass. The views expressed in this work are those of the authors alone and do not necessary reflect the views of the European Commission. We thank Arnaud Gauffreteau (INRA, UMR Agronomie) for his useful and relevant comments on an earlier version of this paper.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.envsoft.2015.09.008.

## References

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environ. Model. Softw. 26, 1376–1388. http://dx.doi.org/10.1016/j.envsoft.2011.06.004.

Almeida, J., Moonen, P., Soto, I., Achten, W.M.J., Muys, B., 2014. Effect of farming system and yield in the life cycle assessment of Jatropha-based bioenergy in Mali. Energy Sustain. Dev. 23, 258–265. http://dx.doi.org/10.1016/j.esd.2014.10.001.

Bazot, M., Lesur, C., Bio-Beri, F., Lorin, M., Béjot, P., Loyce, C., 2014. Mesurer et prédire les rendements du Miscanthus (Miscanthus × giganteus) en parcelles agricoles. In: Le Cahier des Techniques de l'INRA 81.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7, 434–455. http://dx.doi.org/10.1080/10618600.1998.10474787.

Cadoux, S., Ferchaud, F., Demay, C., Boizard, H., Machet, J.-M., Fourdinier, E., Preudhomme, M., Chabbert, B., Gosse, G., Mary, B., 2014. Implications of productivity and nutrient requirements on greenhouse gas balance of annual and perennial bioenergy crops. GCB Bioenergy 6, 425–438. http://dx.doi.org/10.1111/gcbb.12065.

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ. Model. Softw. 37, 134–145. http://dx.doi.org/10.1016/j.envsoft.2012.03.012.

Clifton-Brown, J.C., Neilson, B., Lewandowski, I., Jones, M.B., 2000. The modelled productivity of Miscanthus × giganteus (GREEF et DEU) in Ireland. Ind. Crop. Prod. 12, 97–109. http://dx.doi.org/10.1016/S0926-6690(00)00042-X.

Crespin, C., Soyer, J., 2002. Protocole d'Expérimentation Céréales à Pailles – Essais de Valeur Agronomique et Technologique. Groupe d'Etude et de contrôle des Variétés et des Semences – Secteur Etude des Variétés, Minière, Versailles.

De la Rua, C., Lechon, Y., Morandi, F., Ostergard, H., Wohlfahrt, J., Perrin, A., Gabrielle, B., Bjorkvoll, T., Flatberg, T., Damman, S., 2015. Socio-economic effects of biomass supply chain: case studies from LogistEC project. In: 23rd European Biomass Conference and Exhibition. Vienna, Austria. http://www.researchgate.net/publication/280223279_SOCIO-ECONOMIC_EFFECTS_OF_BIOMASS_SUPPLY_CHAIN_CASE_STUDIES_FROM_LOGIST'EC_PROJECT.

El Bassam, N., Huisman, W., 2001. Harvesting and storage of Miscanthus. In: Jones, M.B., Walsh, M. (Eds.), Miscanthus for Energy and Fibre (London).

Gasol, C.M., Gabarrell, X., Anton, A., Rigola, M., Carrasco, J., Ciria, P., Rieradevall, J., 2009. LCA of poplar bioenergy system compared with *Brassica carinata* energy crop and natural gas in regional scenario. Biomass Bioenergy 33, 119–129. http://dx.doi.org/10.1016/j.biombioe.2008.04.020.

Hastings, A., Clifton-Brown, J., Wattenbach, M., Mitchell, C.P., Smith, P., 2009. The development of MISCANFOR, a new Miscanthus crop growth model: towards more robust yield predictions under different climatic and soil conditions. GCB Bioenergy 1, 154–170. http://dx.doi.org/10.1111/j.1757-1707.2009.01007.x.

Heaton, E., 2004. A quantitative review comparing the yields of two candidate C4 perennial biomass crops in relation to nitrogen, temperature and water. Biomass Bioenergy 27, 21–30. http://dx.doi.org/10.1016/j.biombioe.2003.10.005.

Hillier, J., Whittaker, C., Dailey, G., Aylott, M., Casella, E., Richter, G.M., Riche, A., Murphy, R., Taylor, G., Smith, P., 2009. Greenhouse gas emissions from four bioenergy crops in England and Wales: integrating spatial estimates of yield and soil carbon balance in life cycle analyses. GCB Bioenergy 1, 267–281. http://dx.doi.org/10.1111/j.1757-1707.2009.01021.x.

Kerckhoffs, H., Renquist, R., 2012. Biofuel from plant biomass. Agron. Sustain. Dev. 33, 1–19. http://dx.doi.org/10.1007/s13593-012-0114-9.

Laurent, A., Pelzer, E., Loyce, C., Makowski, D., 2015. Ranking yields of energy crops: a meta-analysis using direct and indirect comparisons. Renew. Sustain. Energy Rev. 46, 41–50. http://dx.doi.org/10.1016/j.rser.2015.02.023.

Lesur, C., Jeuffroy, M.-H., Makowski, D., Riche, A.B., Shield, I., Yates, N., Fritz, M., Formowitz, B., Grunert, M., Jorgensen, U., Laerke, P.E., Loyce, C., 2013. Modeling long-term yield trends of Miscanthus × giganteus using experimental data from across Europe. Field Crops Res. 149, 252–260. http://dx.doi.org/10.1016/j.fcr.2013.05.004.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10, 325–337. http://dx.doi.org/10.1023/A:1008929526011.

Metzger, J.O., Hüttermann, A., 2009. Sustainable global energy supply based on lignocellulosic biomass from afforestation of degraded areas. Naturwissenschaften 96, 279–288. http://dx.doi.org/10.1007/s00114-008-0479-4.

Miguez, F.E., Maughan, M., Bollero, G.A., Long, S.P., 2012. Modeling spatial and dynamic variation in growth, yield, and yield stability of the bioenergy crops Miscanthus × giganteus and *Panicum virgatum* across the conterminous United States. GCB Bioenergy 4, 509–520. http://dx.doi.org/10.1111/j.1757-1707.2011.01150.x.

Miguez, F.E., Villamil, M.B., Long, S.P., Bollero, G.A., 2008. Meta-analysis of the effects of management factors on Miscanthus × giganteus growth and biomass production. Agric. For. Meteorol. 148, 1280–1292. http://dx.doi.org/10.1016/j.agrformet.2008.03.010.

Miguez, F.E., Zhu, X., Humphries, S., Bollero, G.A., Long, S.P., 2009. A semimechanistic model predicting the growth and production of the bioenergy crop Miscanthus × giganteus: description, parameterization and validation. GCB Bioenergy 1, 282–296. http://dx.doi.org/10.1111/j.1757-1707.2009.01019.x.

Mola-Yudego, B., Aronsson, P., 2008. Yield models for commercial willow biomass plantations in Sweden. Biomass Bioenergy 32, 829–837. http://dx.doi.org/10.1016/j.biombioe.2008.01.002.

Philibert, A., Loyce, C., Makowski, D., 2014. Predicting nitrous oxide emissions with a random-effects model. Environ. Model. Softw. 61, 12–18. http://dx.doi.org/10.1016/j.envsoft.2014.07.002.

Piepho, H.-P., 1995. The use of multilocation trials to select cultivars that are better than a control. Plant Breed. 114, 337–340. http://dx.doi.org/10.1111/j.1439-0523.1995.tb01245.x.

Pogson, M., Hastings, A., Smith, P., 2012. Sensitivity of crop model predictions to entire meteorological and soil input datasets highlights vulnerability to drought. Environ. Model. Softw. 29, 37–43. http://dx.doi.org/10.1016/j.envsoft.2011.10.008.

Rahman, M.M., Mostafiz, S.B., Paatero, J.V., Lahdelma, R., 2014. Extension of energy crops on surplus agricultural lands: a potentially viable option in developing countries while fossil fuel reserves are diminishing. Renew. Sustain. Energy Rev. 29, 108–119. http://dx.doi.org/10.1016/j.rser.2013.08.092.

Strullu, L., Beaudoin, N., de Cortàzar Atauri, I.G., Mary, B., 2014. Simulation of biomass and nitrogen dynamics in perennial organs and shoots of Miscanthus × giganteus using the STICS model. BioEnergy Res. 7, 1253–1269. http://dx.doi.org/10.1007/s12155-014-9462-4.

Wallach, D., 2011. Crop model calibration: a statistical perspective. Agron. J. 103, 1144. http://dx.doi.org/10.2134/agronj2010.0432.