Chapter 11

Statistical Analysis of Large Simulated Yield Datasets for Studying Climate Effects


¹L’Institut national de la recherche agronomique, Thiverval-Grignon, France
²University of Florida, Gainesville, FL, USA
³University of Bonn, Bonn, Germany
⁴Unité de recherche pluridisciplinaire sur la prairie et les plantes fourragères, Lusignan, France
⁵Genetic, Diversity and Ecophysiology of Cereals (GDEC), Clermont-Ferrand, France
⁶Blaise Pascal University, Aubière, France
⁷CIRAD, Montpellier, France
⁸CCAFS, New Delhi, India
⁹CIRAD, Montpellier, France
¹⁰Michigan State University, East Lansing, MI, USA

279
11 L’Institut national de la recherche agronomique, Avignon, France
12 German Research Center for Environmental Health, Neuherberg, Germany
13 Centre for Geo-Information, Alterra, the Netherlands
14 University of Florida, Gainesville, FL, USA
15 University of Leeds, Leeds, UK
16 International Centre for Tropical Agriculture (CIAT), Cali, Colombia
17 Wageningen University and Research Centre, Wageningen, the Netherlands
18 MTT Agrifood Research Finland, Mikkeli, Finland
19 University of East Anglia, Norwich, UK
20 Cantabrian Agricultural Research and Training Centre (CIFA), Muriedas, Spain
21 University of Tübingen, Tübingen, Germany
22 NASA Goddard Institute for Space Studies, New York, NY, USA
23 University of Nebraska-Lincoln, Lincoln, USA
24 National Laboratory for Agriculture and Environment, Ames, IA, USA
25 International Atomic Energy Agency (IAEA), Vienna, Austria
26 University of Reading, Reading, UK
27 University of Guelph, Guelph, Ontario, Canada
28 University of Hohenheim, Stuttgart, Germany
29 Joint Global Change Research Institute, College Park, MD, USA
30 Instituto Nacional de Investigación Agropecuaria (INIA), Treinta y Tres, Uruguay
31 Leibniz Centre for Agricultural Landscape Research, Müncheberg, Germany
32 University of Washington, Seattle, WA, USA
33 Univ. Politécnica de Madrid, Madrid, Spain
34 Potsdam Institute for Climate Impact Research, Potsdam, Germany
35 Indian Agricultural Research Institute, New Delhi, India
36 Landscape & Water Sciences, Department of Environment and Primary Industries, Horsham, Australia
37 Aarhus University, Tjele, Denmark
38 University of Reading, Reading, UK
39 The Pennsylvania State University, University Park, PA, USA
40 Univ. Politécnica of Madrid, Madrid, Spain
41 Rothamsted Research, Harpenden, UK
42 Food and Agriculture Organization of the UN (FAO), Rome, Italy
43 Washington State University, Pullman, WA, USA
44 Wageningen University, Wageningen, the Netherlands
Statistical Analysis of Large Simulated Yield Datasets for Studying Climate Effects

45 Chinese Academy of Science, Beijing, China
46 The New Zealand Institute for Plant & Food Research Limited, Canterbury, New Zealand
47 Commonwealth Scientific and Industrial Research Organisation (CSIRO), Ecosystem Sciences, Queensland, Australia
48 United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Beltsville, MD, USA
49 Institute for Climate and Water, Castelar, Argentina
50 AgroSystèmes et développement territorial (AGIR), Castanet-Tolosan, France
51 Arid-Land Agricultural Research Center, Maricopa, AZ, USA
52 Texas A&M University, TX, USA

*Dr. Nadine Brisson passed away in 2011 while this work was being carried out.*
Introduction

Many studies have been carried out during the last decade to study the effect of climate change on crop yields and other key crop characteristics. In these studies, one or several crop models were used to simulate crop growth and development for different climate scenarios that correspond to different projections of atmospheric CO₂ concentration, temperature, and rainfall changes (Semenov et al., 1996; Tubiello and Ewert, 2002; White et al., 2011). The Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al., 2013) builds on these studies with the goal of using an ensemble of multiple crop models in order to assess effects of climate change scenarios for several crops in contrasting environments.

These studies generate large datasets, including thousands of simulated crop yield data. They include series of yield values obtained by combining several crop models with different climate scenarios that are defined by several climatic variables (temperature, CO₂, rainfall, etc.). Such datasets potentially provide useful information on the possible effects of different climate change scenarios on crop yields. However, it is sometimes difficult to analyze these datasets and to summarize them in a useful way due to their structural complexity; simulated yield data can differ among contrasting climate scenarios, sites, and crop models. Another issue is that it is not straightforward to extrapolate the results obtained for the scenarios to alternative climate change scenarios not initially included in the simulation protocols. Additional dynamic crop model simulations for new climate change scenarios are an option but this approach is costly, especially when a large number of crop models are used to generate the simulated data, as in AgMIP.

Statistical models have been used to analyze responses of measured yield data to climate variables in past studies (Lobell et al., 2011), but the use of a statistical model to analyze yields simulated by complex process-based crop models is a rather new idea. We demonstrate herewith that statistical methods can play an important role in analyzing simulated yield data sets obtained from the ensembles of process-based crop models. Formal statistical analysis is helpful to estimate the effects of different climatic variables on yield, and to describe the between-model variability of these effects.

These statistical methods can also be used to develop meta-models, i.e., statistical models that summarize process-based crop models, enabling scientists to explore the effects of new climate change scenarios. This approach is illustrated with two simulated yield datasets obtained by AgMIP for maize and wheat that were generated by using ensembles of process-based crop models. The yield datasets were used to develop a meta-model that provides a simplified representation of the original ensemble of crop models. The proposed meta-model is a statistical regression with random coefficients that describe the variability of the simulated yield data across the
original crop models. Once the statistical regression has been fitted to the simulated yield datasets, the meta-model can predict the effects of climate changes among the crop models and can thus be used to study alternative climate change scenarios that were not initially simulated. Finally, the results obtained with the meta-model were used to assess the risk of yield loss for maize and wheat as functions of temperature and CO₂ concentration change.

Materials and Methods

Simulated yield data

Yield data were simulated with 19 maize crop models and 26 wheat crop models in four contrasting sites for each species located in France (Lusignan), USA (Ames), Brazil (Rio Verde), and Tanzania (Morogoro) for maize, and in the Netherlands (Wageningen), Argentina (Balcarce), India (New Delhi), and Australia (Wongan Hills) for wheat.

Several climate scenarios were considered. For both species, four temperature changes (+0, +3, +6, +9°C) and five atmospheric CO₂ concentration changes (+0, +90, +180, +270, +360 ppm; where the baseline CO₂ concentration was set at 360 ppm) were considered over the 1980–2010 time-period. In addition, two precipitation changes (0% and −30%) were considered for maize and three rates of nitrogen fertilization (50%, 100%, and 150% of reference rates) were considered for wheat. The other scenarios were defined by combining temperature, CO₂ concentration, and precipitation changes. Thirty years of yield data were generated with each crop model for each of the climate scenarios (Figs. 1 and 2) and the simulated yield values were averaged over the years. Details of the maize and wheat protocols can be found in Bassu et al. (2014) and Asseng et al. (2013), respectively.

Statistical model

Simulated maize and wheat yield data were analyzed using two-level statistical random-effect models (Pinheiro and Bates, 2000). The following statistical model was used to analyze the maize yield data for each site separately:

Level 1, within crop model

\[
Y_{ij} = \alpha_0 + \alpha_{ij} \Delta T_{ij} + \alpha_{2ij} \Delta T_{ij}^2 + \alpha_{3ij} \Delta C_{ij} \\
+ \alpha_{4ij} \Delta C_{ij}^2 + \alpha_{5ij} \Delta R_{ij} + \alpha_{6ij} + \epsilon_{ij}
\]

(1)

where \(Y_{ij}\) is the 30-year mean yield simulated for the \(i\)th crop model (where \(i = 1, \ldots, P\)) and for the \(j\)th scenario (where \(j = 1, \ldots, Q_i\)). The variables \(\Delta T_{ij}, \Delta C_{ij},\)
fitted to the simulated changes among the change scenarios that the meta-model were actions of temperature

Fig. 1. Maize yields simulated by 15 crop models in Lusignan, France for different levels of temperature change (each letter corresponds to one crop model). The reported yield data were obtained for a change of CO₂ concentration equal to +180 ppm and for the baseline (1981-2010) level of precipitation. Only 15 out of the 19 available crop models simulated this level of [CO₂] and were thus included in the figure.

\[
\begin{align*}
\varepsilon_{ij} & \sim N(0, \tau^2) \\
\alpha_{ki} & \sim N(\mu_k, \sigma_k^2), \quad k = 0, \ldots, 6
\end{align*}
\]

The distribution parameters, \( \mu_k \) (where \( k = 0, \ldots, 6 \)), are the seven mean regression coefficient values that represent the mean yield baseline (\( \mu_0 \)) and mean effects of temperature, CO₂, temperature-CO₂ interaction, and rainfall (\( \mu_1, \ldots, \mu_6 \)) over the P crop models. The seven variances, \( \sigma_k^2 \) (where \( k = 0, \ldots, 6 \)), describe the between-model variability of the random regression coefficients.
Fig. 2. Wheat yields simulated by the 26 crop models in Wageningen, the Netherlands, for different levels of temperature change (each letter corresponds to one crop model). The reported yield data were obtained for a change of CO₂ concentration equal to +180 ppm and for the baseline (1981-2010) level of fertilization.

This statistical model assumes that the ensemble of P crop models is a sample taken within a population that includes all possible crop models for a given crop while flexibly allowing for the incorporation of additional crop models in the future. The probability distributions defined by Equation (2) describe the between-crop-model variability of the yield response to climate change within the whole population of crop models. These probability distributions cover the ranges of climate effects considered in different crop models. The relationship defined by Equation (1) is assumed to be valid for all crop models, but its parameters, α_{ki}, are assumed to vary among crop models. However, this statistical model only describes 30-year mean yield responses and is not intended to describe the year-to-year variability of crop yields. Creating a statistical model that captured year-to-year variability would have required extra random terms and additional parameters, which would have overly complicated the calculated model and thus was not considered here.

The same form of model was used for wheat as was used for maize, but replaced the rainfall variable, ΔR_{ij}, with a variable that describes the level of N fertilization.
The reported yield data were for the baseline (1981–2010) crop models is a sample for a given crop while models in the future. The between-crop-model the whole population of ages of climate effects defined by Equation (1) is are assumed to vary describes 30-year mean year variability of crop variability would have such would have overly red here.

For maize, but replaced level of N fertilization in Equation (1). For the wheat pilot studies, precipitation was held constant in the wheat simulation protocol.

The population parameters of the statistical model $\mu_k$, $\sigma_k^2$, and $\tau^2$ were estimated by restricted maximum likelihood. The model-specific regression coefficients, $\alpha_{ki}$ (where $k = 0, \ldots, 6$ and $i = 1, \ldots, P$), were estimated by using the best linear unbiased predictor using the R software package “nlme” (Pinheiro and Bates, 2000) and the estimated values will be henceforth referred to as $\alpha_{ki}$. The model was fitted for each crop and each site separately, but was calculated for all crop models together.

### Assessment of the statistical model

The statistical model was compared to other statistical models, including models with fewer explanatory variables, models with fewer random coefficients, and a model that includes no random coefficients (i.e., classical linear regression). All the models were compared by using the Akaike information criterion (AIC; Pinheiro and Bates, 2000), where a lower AIC value corresponds with the better model, and we found that the model defined by Equations (1) and (2) led to lower AICs than the simpler models. The assumption that the residual errors, $\varepsilon_{ij}$, were independent was assessed by developing another statistical model that incorporated correlated residual errors. The AIC value of this model was higher and the estimated correlation coefficients were very low ($7.6 \times 10^{-4}$ and $-4.7 \times 10^{-3}$ for maize and wheat, respectively). The statistical model’s quality of fit was also assessed by using graphical analysis. Evaluating the residuals did not reveal any obvious biases due to the assumptions made on the relationships between yields and input factors.

### Estimation of the effect of climate change on yield

The statistical model described above was used to compute three different types of outputs:

- The average yield loss/gain due to climate change over the ensemble of crop models.
- The yield gain/loss estimated for individual crop models due to changes in climate variables.
- The probability that the yield loss/gain will exceed a given threshold.

For maize, the average yield difference obtained between a given climate change scenario (characterized by $\Delta T$, $\Delta C$, and $\Delta R$) and the baseline scenario was expressed as:

$$\Delta Y = \mu_1 \Delta T + \mu_2 \Delta T^2 + \mu_3 \Delta C + \mu_4 \Delta C^2 + \mu_5 \Delta C \Delta T + \mu_6 \Delta R$$ (3)
The yield difference described in Equation (3) is averaged over all crop models; this difference corresponds to an average yield gain or to an average yield loss over the 19 submitted crop models. Equation (3) defines a meta-model that simulates the average output of the original ensemble of crop models. This meta-model enables the computation of the yield differences for any changes in temperature, CO\textsubscript{2}, and precipitation, \(\Delta T\), \(\Delta C\), and \(\Delta R\), at each of the four considered sites.

For a given crop model \(i\), the yield difference was expressed as:

\[
\Delta Y_i = \alpha_{e1i} \Delta T + \alpha_{e2i} \Delta T^2 + \alpha_{e3i} \Delta C + \alpha_{e4i} \Delta C^2 + \alpha_{e5i} \Delta C \Delta T + \alpha_{e6i} \Delta R \tag{4}
\]

Equation (4) defines a meta-model simulating the output of the \(i\)th crop model. The yield difference (4) is crop model-specific. It corresponds to the climate change effect on yield that would have been obtained with the \(i\)th crop model if this crop model were run for a climate change scenario that were characterized by \(\Delta T\), \(\Delta C\), and \(\Delta R\).

The statistical model defined by Equations (1) and (2) was also used to compute the probability that the yield difference will exceed a threshold and, more specifically, the probability of yield gain or of yield loss that results from a change in the temperature, CO\textsubscript{2} concentration, and/or precipitation. This probability was computed from the following Gaussian probability distribution \(N(\mu_{\Delta Y}, \sigma^2_{\Delta Y})\) as:

\[
\mu_{\Delta Y} = \mu_1 \Delta T + \mu_2 \Delta T^2 + \mu_3 \Delta C + \mu_4 \Delta C^2 + \mu_5 \Delta C \Delta T + \mu_6 \Delta R \tag{5}
\]

\[
\sigma^2_{\Delta Y} = \sigma^2_1 \Delta T^2 + \sigma^2_2 \Delta T^4 + \sigma^2_3 \Delta C^2 + \sigma^2_4 \Delta C^4 + \sigma^2_5 \Delta C^2 \Delta T^2 + \sigma^2_6 \Delta R^2 \tag{6}
\]

Similar relationships were derived for wheat by replacing precipitation change with fertilization change in Equations (3) to (6). Note that the variance defined by Equation (6) is not constant but varies as a function of the climate scenario characteristics.

Results and Discussion

Increases in temperature

Figure 3 shows the statistical meta-models’ quality of fit calculated for maize and wheat (one site per crop). The distributions of the residuals for both maize and wheat are symmetric and do not display trends. The fitted values obtained with the statistical meta-models are in most cases close to the yields simulated by the processed-based crop models. The residual standard error of the fitted statistical models is equal to 0.19 and 0.22 t/ha for maize and wheat, respectively, in the two sites considered in Fig. 3. A similar quality of fit was obtained for the other sites.

Figure 4 shows the change in the mean yield from the baseline for one maize site (Fig. 4a) and one wheat site (Fig. 4b) as affected by a CO\textsubscript{2} concentration increase
over all crop models; average yield loss over the climate change experiment with CO₂, temperature, and sites.

\[ \Delta T + \alpha_{\text{cl}} \Delta R \] (4)

for the ith crop model if the climate change model is characterized by \( \Delta T, \Delta C \), and \( \alpha_{\text{cl}} \) is also used to compare threshold and, more results from a change in the probability was defined as:

\[ \mathcal{N}(\mu_{\Delta Y}, \sigma_{\Delta Y}^2) \] as:

\[ \Delta T + \mu_{\Delta Y} \Delta R \] (5)

\[ \Delta T^2 + \sigma_{\Delta Y}^2 \Delta R^2 \] (6)

precipitation change in variance defined in the climate scenario

for maize and for both maize and values obtained with were simulated by the fitted statistical models, in the two for the other sites. The for one maize site concentration increase of 180 ppm ([CO₂] = 540 ppm) and an increase in mean seasonal temperature ranging from 0°C to 6°C. Each emulated model yield as calculated by using the crop model-specific coefficients \( \alpha_{\text{cl}} \) (where \( k = 0, \ldots, 6 \) and \( i = 1, \ldots, P \)) is plotted with a black line and thus can be seen as a proxy of a given crop model. Negative yield differences can be interpreted as mean yield losses and, conversely, positive yield differences can be interpreted as mean yield gains without the need.
Fig. 4. The effect of climate change on yield for a CO$_2$ concentration increase of 180 ppm, and for mean temperature increases ranging from 0 to 6°C. Results are reported for (a) maize, and (b) wheat for one site per crop. Yield difference = yield obtained with a climate change scenario — yield obtained with the baseline (1981–2010). Each black curve corresponds to a given crop model. Black curves were obtained with the statistical model by using the crop model-specific coefficients $a_{ki}$ (where $k = 0, \ldots, 6$ and $i = 1, \ldots, P$). The continuous red curve indicates the mean yield response as compared to the baseline scenario, i.e., the effect averaged over all crop models. The red dashed curves indicate the first and third quartiles of climate change effect computed over all crop models. The blue line indicates zero yield difference.
for re-running the original, process-based crop model. The continuous red curve indicates the mean yield response to the given climate scenario as compared to the baseline, i.e., the effect averaged over all crop models. The red dashed curves indicate the first (probability = 0.25) and third (probability = 0.75) quartiles of the climate change effect.

According to Fig. 4, most of the crop models estimate that a temperature increase negatively impacts yield projections of both maize and wheat in all locations. But this effect is highly variable among crop models, with some models predicting little response to temperature. For maize, Fig. 4a illustrates how, on average across the ensemble of crop models, the statistical meta-model emulates a yield loss when temperatures exceed +1.5°C with a CO₂ concentration increase of +180 ppm in Lusignan, France. Contrastingly, the models suggest that wheat is more resilient to heat stress and would require a temperature threshold equal to +3.6°C before experiencing yield losses below the baseline in Wageningen, the Netherlands (Fig. 4b). Based on the emulated projections seen in Fig. 4b, the models indicate that, should climate change be limited to a +3°C change and include a CO₂ concentration increase of +180 ppm, moderate temperature increases could, in fact, lead to gains in wheat production in this location.

Figure 4 also demonstrates the large variability between crop models and displays how this variability increases as a function of temperature. The differences between the first and third quartiles are much larger for higher temperature changes, at the given CO₂ concentration. This result indicates that the differences between crop models are much larger for high-temperature changes than for low-temperature changes. The level of divergence between the predictions of the crop models thus tends to increase with the temperature value considered in the climate change scenarios.

**Increases in atmospheric CO₂ concentration**

Figure 5 shows the effect of climate change at one site for both maize and wheat yields under increasing levels of CO₂ concentration, ranging from 0 to +360 ppm from the simulated baseline concentration (i.e., 360 to 720 ppm) at a constant temperature increase of +2°C. Figure 5a illustrates how all crop model simulations predict a positive effect on maize yields in Lusignan, France when exposed to increased CO₂ concentrations, but this effect is small and varies among crop models. The mean curve suggests that the effect of a temperature increase of +2°C is negative for the range of CO₂ concentration increases, thus the benefit of the increased CO₂ concentration does not outweigh the negative effect of the temperature increase. On average, the crop models predict a yield loss for maize, even at high CO₂ concentrations, due to this +2°C temperature change.
Fig. 5. The effect of climate change on yield for a temperature increase of +2°C, and for levels of CO₂ concentration increase ranging from 0 to +350 ppm. Results are reported for (a) maize and (b) wheat for one site per crop. Yield difference = yield obtained with a climate change scenario — yield obtained with the baseline (1981–2010). Each black curve corresponds to a given crop model. The black curves were obtained with the statistical model by using the crop model-specific coefficients \( \alpha_{eki} \) (where \( k = 0, \ldots, 6 \) and \( i = 1, \ldots, P \)). The continuous red curve indicates the yield response as compared to the baseline scenario, i.e., the effect averaged over all crop models. The red dashed curves indicate the first and third quartiles of climate change effect computed over all crop models. The blue line indicates zero yield difference.

The fitted curves obtained for CO₂ response of wheat in Wageningen, the Netherlands are displayed in Figure 5b. Compared to maize, the effect of a CO₂ concentration increase is stronger for wheat. This effect is highly variable among crop models; some models demonstrate positive slopes over the range of CO₂ concentrations, whereas others show slopes close to zero. When averaged across all of the crop models, the effect of the +2°C temperature increase is positive (yield gain) as soon as the CO₂ concentration increase reaches 100 ppm.

**Probabilistic thresholds**

An important advantage of the meta-model is that it handles the interpolation between temperature levels and between CO₂ concentration levels. Our meta-model can thus be used to calculate temperature and [CO₂] thresholds that lead to yield loss or yield gain. Table 1 shows the thresholds of [CO₂] increase required to obtain a probability of maize and wheat yield gain higher than 0.5 (i.e., 50% chance of yield gain). These thresholds were computed for two values of temperature increase (+2 and +4°C) and four sites per crop. It would have been very difficult to compute them without a meta-model. Results show that, for wheat, climate change has 50% chance of resulting in a yield gain if [CO₂] increases by at least +117 ppm (depending on
the site) and if temperature concurrently increases by +2°C (Table 1). Required levels of \([\text{CO}_2]\) increases are much higher for maize (Table 1). Our meta-model has another advantage, although not fully illustrated here; it can be used to quantify the effects of temperature, \([\text{CO}_2]\), and precipitation on yields, their interactions, and their variability between the four considered sites and between all the considered crop models. It thus constitutes a powerful tool for exploring complex crop model responses to climate factors.

**Conclusions**

Our study shows that yield data simulated by an ensemble of complex dynamic-process crop models can be summarized by using statistical meta-models that are based on random coefficient regressions. These statistical models describe the between-crop model variability of the simulated yield data by using probability distributions. They can be used to compute key quantities such as mean yield loss, percentiles of yield loss, and probabilities of yield loss as functions of temperature change and \(\text{CO}_2\) concentration change. These statistical meta-models can thus be helpful for analyzing the risk of yield loss due to climate change at the locations where the original simulations were conducted. The capabilities of the meta-models were illustrated by using two yield datasets generated by the
AgMIP pilot model intercomparison studies for maize and wheat. The results showed that:

- High CO₂ concentrations are likely to outweigh the negative effect of increasing temperature, leading to yield gains for wheat with climate change, but not for maize.
- Thresholds of temperature increase and CO₂ concentrations that lead to a specific risk level of maize and wheat yield loss can be computed by using the meta-model.
- The divergence between the maize and wheat crop models and therefore the uncertainty in the simulated results increases as a function of temperature (the higher the temperature change, the higher the between-crop model variability).

In the future, the proposed meta-model could be extended in two different ways. First, it could be applied to a dataset that includes simulations obtained for a higher number of sites. It will then be possible to include co-variables that describe site characteristics (e.g., soil type, agricultural practices) in the meta-model in order to explain the origin of the between-site variability. Second, our meta-model could be extended in order to describe the between-year variability of yields and to analyze the risk of extreme yield values for different climate scenarios. This could be achieved by including one or several additional random effects in the meta-model and by fitting this model to yearly yield data.

Acknowledgements

G. J. O’Leary was supported by the Victorian Department of Environment and Primary Industries and the Australian Grains Research and Development Corporation. P. Bertuzzi, J.-L. Durand, D. Makowski, P. Marte, D. Ripoche, and D. Wallach were partly supported by the INRA ACCAF meta-program. S. G. was supported by a grant from the Ministry of Science, Research and Arts of Baden-Württemberg (AZ Zu33-721.3-2) and the Helmholtz Centre for Environmental Research — UFZ, Leipzig.

References

and wheat. The results indicate the positive effect of increasing temperature change, but not for interactions that lead to specific computed by using the models and therefore the variability of temperature (the top model variability).

This could be achieved in different ways, as obtained for a higher order to make the meta-model could be used to yield and to analyze the yield variability. The meta-model achieved by using the meta-model and by using the models and therefore the variability of temperature (the top model variability).


