



HAL
open science

Simplifying a complex computer model: Sensitivity analysis and metamodelling of an 3D individual-based crop-weed canopy model

Floriane Colas, Jean-Pierre Gauchi, Jean Villerd, Nathalie Colbach

► **To cite this version:**

Floriane Colas, Jean-Pierre Gauchi, Jean Villerd, Nathalie Colbach. Simplifying a complex computer model: Sensitivity analysis and metamodelling of an 3D individual-based crop-weed canopy model. *Ecological Modelling*, 2021, 454, pp.109607. 10.1016/j.ecolmodel.2021.109607 . hal-03461485

HAL Id: hal-03461485

<https://hal.inrae.fr/hal-03461485v1>

Submitted on 13 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Simplifying a complex computer model: sensitivity analysis and metamodelling of an 3D individual-based crop-weed canopy model

Floriane Colas¹, Jean-Pierre Gauchi², Jean Villerd^{1,3}, Nathalie Colbach¹

Corresponding author: nathalie.colbach@inrae.fr

¹ Agroécologie, AgroSup Dijon, INRAE, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France

² MaIAGE, INRAE, Université Paris-Saclay, 78350, France

³ LAE, INRAE, Univ. Lorraine, F-54500 Vandœuvre-lès-Nancy, France

* Corresponding author, address:

Nathalie Colbach

UMR Agroécologie

INRAE

17 rue Sully

BP 86510

21065 Dijon Cedex

Nathalie.Colbach@inrae.fr

Tel +33-380693033

Fax +33-380693262

Declarations of interest: none

Abstract

Complex biological models such as mechanistic research models often need to extend their current use to a broader audience. Simplification and faster simulations would increase their use. Here, a step-by-step methodology was developed and applied to partially metamodel, hence accelerate, the mechanistic model FLORSYS. This is a process-based, multiannual and multispecies model ("virtual field") which simulates crop growth and weed dynamics and allows users to assess cropping systems for crop production and biodiversity. The model is relatively slow, which makes it difficult to test numerous and diverse cropping systems needed to identify those reconciling crop production and biodiversity. Here, we (1) identified the slowest submodel of FLORSYS, i.e. the 3D voxelized light interception submodel, (2) identified and applied a relevant methodology to metamodel this submodel in the simplest situation, i.e. we predicted light interception and absorption directly at the scale of the plant instead of the voxel for a single plant in a field, and (3) extrapolated the method to more complex situations, i.e. a plant in diverse and heterogeneous crop:weed canopies, (4) replaced the original process-based FLORSYS submodel by the metamodels, which required additional equations and decision rules, (5) evaluated the metamodelled FLORSYS with independent field observations, showing

41 an adequate prediction quality combined with an increased speed at fine-grained scale since the
42 metamodelled version was 28 times faster than the process-based version. For steps 2 and 3, we used
43 the global sensitivity method based on a truncated Legendre polynomial chaos expansion (PCE)
44 whose coefficients were estimated by Partial Least Squares (PLS) regression to simultaneously
45 (i) rank inputs with respect to their polynomial and total effects on outputs via the so-called PCE-PLS
46 sensitivity indices, and (ii) provide metamodels predicting light interception and absorption at the
47 plant level. These metamodels were then shortened into parsimonious metamodels via a LASSO-PLS
48 method. The study showed that there was a trade-off between speed gain due to the metamodelled 3D
49 light submodel and the speed loss due to the additional functions for neighbourhood effects. The
50 metamodelled version is best used for testing complex systems where plant location must be modelled
51 precisely (e.g., precision agriculture, intercropping with precision sowing) whereas the voxelized
52 version with a large voxel size is better for simpler cropping systems. The present step-by-step process
53 may be helpful for investigating and speeding up other complex simulation models with interacting
54 objects/agents. It notably uses a hybrid approach, using a process-based (albeit simplified) approach
55 for the most sensitive plant stage (newly emerged tiny plants) and separate sampling plans and
56 metamodels to ensure that the more sensitive stages/components are adequately covered (small
57 plants).

58

59 **Keywords**

60 Metamodel; light interception; photosynthetically active radiation PAR; crop:weed canopy; sensitivity
61 analysis; simulation time

62

63 **1 Introduction**

64 The study of biological problems usually requires complex mechanistic models, especially when
65 dealing with weed dynamics (Holst et al., 2007a; Colbach, 2010). Indeed, even though weeds are
66 considered to be the most harmful pest for crop production (Oerke, 2006), national and European
67 policies now increasingly focus on the role of weeds for biodiversity (Marshall et al., 2003; Petit et al.,
68 2011) and the need to reduce herbicide use because of health issues and environmental concerns
69 (Stoate et al., 2009; Waggoner et al., 2013). Unfortunately, to date, no alternative weed control
70 technique is, alone, as efficient as herbicides, and thus, several cultural techniques must be combined
71 to control weeds (Liebman and Gallandt, 1997). Many weed dynamics models exist to understand and
72 predict weed dynamics (see reviews by Colbach and Debaeke, 1998; Holst et al., 2007b; Freckleton
73 and Stephens, 2009; Bagavathiannan et al., 2020). Only a few take into account the long-term effects
74 of the weed impacts on crops, the multiplicity of weed species, the complexity of cropping systems, or
75 the impact on crop production and biodiversity. To date, FLORSYS (Gardarin et al., 2012; Munier-
76 Jolain et al., 2013; Colbach et al., 2014; Colbach et al., 2021) is the one model that answers all these
77 requirements. This is a process-based "virtual field" model which simulates the effects of cropping
78 systems on weed dynamics as well as on crop production and weed-related biodiversity, thus making

79 possible a multiobjective design of cropping (Colbach et al., 2017; Colbach et al., 2021).
80 Unfortunately, models like FLORSYS that are accurate enough to reproduce the effects of agricultural
81 practices on weed dynamics are time-consuming and complex (Colbach, 2010). In order to use
82 FLORSYS to screen numerous cropping systems and identify sustainable herbicide-sparse cropping
83 systems, the model must be accelerated and simplified. This question is common to many other
84 mechanistic models whose use is often limited by their complexity and slowness.

85 Complex mechanistic models can be simplified by decreasing their precision level as some problems
86 do not require the same high precision level (Kleijnen and Sargent, 2000; Renton, 2011). To simplify a
87 model without losing precision is more difficult and requires different methods. Global sensitivity
88 analyses can explore the model and understand its behaviour to identify which inputs change the
89 outputs the most. This allows developers to assign constant values to minor inputs and to simplify
90 equations (Cox et al., 2006). Global sensitivity analyses thus help to find the correct level of
91 complexity for a metamodel by identifying the non-influential inputs (Faivre et al., 2013). Then,
92 metamodelling aims at emulating the original model, linking inputs and outputs by less detailed but
93 faster equations, which simplifies model use for practical applications. Examples are metamodelling of
94 the noTG forest model (Marie and Simioni, 2014), phoma stem canker control (Hossard et al., 2015)
95 and the bio-geo-chemical DNDC-EUROPE model (Villa-Vialaneix et al., 2012). Sometimes,
96 metamodelling a whole model can be impractical, particularly if there are too many inputs and
97 outputs. If a model consists of several submodels, as is the case of FLORSYS, a more practical solution
98 is to perform a local metamodelling on the submodel using most of the computing time (Marie and
99 Simioni, 2014).

100 Metamodelling requires several steps (Kleijnen and Sargent, 2000) that summarize as (1) what is the
101 purpose of the metamodel (*i.e.* what goal, what is the accuracy needed), (2) what do we know about
102 the model to be metamodelled (*i.e.* which inputs, which domain of applicability, which outputs), (3)
103 what method to use (which type of metamodel to use, which experimental design) and (4) how to
104 evaluate the metamodel (*i.e.* what fitting, which validity). Many sensitivity analysis and
105 metamodelling methods exist like the widespread Sobol indices or FAST. Mahévas and Iooss (2013)
106 identified three criteria to select the best sensitivity analysis for a complex model: (1) the number of
107 possible simulation runs, (2) the number and (3) type of inputs. The feasible number of runs,
108 depending on the simulation time and the number of inputs are crucial to select the relevant methods
109 (Table 1). When little is known about the model behaviour, which is often the case for complex
110 models, performing early tests to increase the knowledge of the model is needed.

111 The objective of the present paper was to accelerate and simplify a mechanistic model, by
112 implementing efficient metamodels through: (1) identification of the sensitivity analysis and
113 metamodelling methods adapted to a slow, complex model such as FLORSYS, (2) identification of the

114 important inputs, (3) simplification of the equations and reduction of the computing time. We
115 voluntarily excluded technical solutions such as parallel processing or Graphical Processing Unit. We
116 focused here on the reasoning for choosing the sensitivity and metamodeling methods and how they
117 were combined with the other steps needed to transform a metamodel into a simulation model. The
118 chosen method was fully developed in Gauchi *et al.* (2017) for a preliminary study. This global
119 sensitivity analysis method is able to deal with both dependent and independent inputs. It is based on a
120 truncated Legendre polynomial chaos expansion (PCE) whose coefficients are estimated by Partial
121 Least Squares (PLS) regression (Gauchi *et al.*, 2017) aiming to simultaneously rank inputs as a
122 function of their polynomial and total effects on outputs via the so-called PCE-PLS sensitivity indices,
123 and to provide precise and fast metamodels. Finally, the metamodels were reduced into parsimonious
124 metamodels via a LASSO-PLS regression method.

125 The methodology to accelerate and simplify FLORSYS involves several steps (Figure 1). Section 2
126 presents the target model and its submodels in order to identify the most time-consuming submodel
127 (i.e., the light-interception submodel, step 1). Section 3 first presents the metamodeling approach *per*
128 *se*, with the tests that led to the choice of the most suitable method working with a simple case study
129 (i.e. a single plant in the field, step 2), and then how we applied this method to cover all situations in
130 the model (i.e., target plants surrounded by neighbour plants, step 3). Finally, section 4 demonstrates
131 how the metamodels were integrated into FLORSYS (step 4), and section 5 uses field observations to
132 investigate which model (process-based vs metamodel) is the best in terms of simulation speed and
133 precision, depending on the model use (step 5).

134 2 Identification of the model constraints

135 2.1 Presentation of FLORSYS

136 FLORSYS (Gardarin *et al.*, 2012; Munier-Jolain *et al.*, 2013; Colbach *et al.*, 2014; Colbach *et al.*, 2021)
137 is a mechanistic (i.e. process-based) model which simulates multispecies weed dynamics depending on
138 the cropping system and pedoclimate in a “virtual field”. Its purpose is to experiment numerous
139 cropping systems to design sustainable weed management strategies that reconcile reduced herbicide
140 use, crop production and biodiversity. FLORSYS simulates the annual life-cycle of crop and weed
141 plants at a daily time step and is a combination of submodels such as plant emergence, plant growth or
142 radiation interception. FLORSYS inputs are daily weather, soil characteristics, initial weed seed bank,
143 and cropping system practices (crop succession and detailed list of cultural operations). Outputs
144 include crop yield, daily weed seed bank, plant densities and biomass (for more information see
145 supplementary materiel online S1). As FLORSYS consists of a collection of submodels, the
146 simplification should not concern the whole model, but individual submodels should be simplified
147 individually to keep modularity and access to specific submodels outputs.

148 2.2 Identification of the most time-consuming submodel in FLORSYS

149 (step 1)

150 The computing time for each FLORSYS submodel was registered for a simulation with diverse crops
151 and cultural practices over 13 years corresponding to a cropping system trial (Colbach et al., 2016a).
152 Code profiling of the C++ source code of FLORSYS showed that the 3D radiation interception
153 submodel was by far the most time-consuming submodel. This submodel predicts the
154 photosynthetically active radiation (PAR) intercepted by each plant of the crop:weed canopy volume
155 discretised into voxels (3D pixels). It used 57%, 64% and 99% of total simulation time with a voxel
156 edge size of 7, 4 and 1 cm, respectively. The second most time-consuming submodel was the
157 germination/emergence submodel which used 20%, 7% and 0.04% of the computation time for the
158 three voxel edge sizes. Consequently, we will focus here on simplifying and accelerating the radiation
159 interception submodel.

160 2.3 A short presentation of the 3D radiation interception submodel

161 The 3D radiation interception submodel (Munier-Jolain *et al.*, 2013) simulates a 3D sample of the
162 virtual field where the space is discretised into voxels (*i.e.* 3D pixels). Crop and weed plants are placed
163 onto this field, with plant position and morphology resulting from other FLORSYS submodels. Crop
164 plants can be sown in rows or broadcast (*i.e.* random position in the field); weeds can be positioned
165 randomly or in species-specific patches. The radiation interception submodel calculates the amount of
166 photosynthetically active radiation (PAR) that arrives on top of the crop:weed canopy and that trickles
167 down to the voxels in the underlying layers, depending on plant leaf areas, species radiation extinction
168 coefficients and solar angle (which depends on latitude and season). In total 14 input variables can be
169 modified in the submodel for five different outputs.

170 2.3.1 3D radiation interception inputs

171 Plants are represented as cylinders delimited by their height and width (Figure 2, Table 2.A). The leaf
172 area (LA) of the plant is distributed across the successive voxel layers of the cylinder, with 50% of the
173 cumulative leaf area below relative median leaf height (RH50) of the plant and its distribution
174 governed by the shape parameter, *b*. The species radiation extinction coefficient (*k*) and the plant leaf
175 area inside each voxel determine how much incident radiation of the voxel is absorbed by the plant's
176 leaves. The radiation absorbed by each plant (PAR_a) is the sum of the radiation absorbed by its leaves
177 in the different voxels. Other inputs describe the location: (1) the field sample, *i.e.* dimensions in the
178 north-south and in the east-west directions, as well as the grain of the discretization, *i.e.*, the voxel
179 edge size, and (2) the position of the solar angle, *i.e.* latitude of the simulated field and the Julian day.

2.3.2 3D radiation interception outputs

Outputs of this submodel (Table 2.B) are used for different purposes in FLORSYS: the photosynthetically active radiation absorbed by a plant (PARaP) drives biomass accumulation in the growth submodel, the daily shading intensity perceived by the plant (SID) drives etiolation in the morphology submodel. The relative intercepted photosynthetically active radiation (rPARi) is considered at three scales, the top of a given plant (rPARi_{top}), over the whole plant (rPARi_{plant}), and the soil surface below all plants (rPARi_{base}). These are used as proxies for herbicide penetration and interception in the canopy in the herbicide treatment submodel.

As single plants are not shaded by neighbouring plants, their relative PAR intercepted on the plant's top (rPARi_{top}) is always 1. Thus, for the single plant case, only four outputs were studied, *i.e.* PARaP, SID, rPARi_{plant}, rPARi_{base}. For the "plant in a canopy" step, all five outputs are studied (Table 2.B). The metamodels also predict PARa per cm² (PARaC), *i.e.* relative absorption efficiency for a given plant volume.

3 Simplification and acceleration of the 3D radiation interception submodel

To find the best metamodelling and sensitivity analysis method for the radiation-interception submodel, we started with the simplest possible situation for this submodel which consists of a single plant in a field, without shade due to surrounding plants (step 2 in Figure 1). Once identified, this method was then applied to more realistic but more complex occurrences relevant for the submodel, *i.e.* target plants surrounded by neighbour plants (step 3 in Figure 1).

3.1 Simplified case study with single target plants (step 2)

This section aims (1) to test the effect of the range of variation in inputs, (2) to test the effect of correlations between inputs, (3) to analyse the sensitivity indices for unshaded (single) plants, and (4) to evaluate the metamodels predicting the radiation interception variables for unshaded plants.

3.1.1 Testing the sensitivity to the range of the inputs (step 2.i)

Two input range sizes were tested following a Plackett & Burman experimental design (Plackett and Burman, 1946) with 12 combinations of the two extreme ranges for the 11 inputs with Latin Hypercube Sampling (LHS) designs of 29200 rows (supplementary material online S2 section 1): (1) a small range corresponding to France, focusing on spring and summer, and the plant morphologies most common in fields and (2) a large range for all possible plant morphologies growing, all year and all around the world (except polar regions). For each configuration of ranges, the FLORSYS radiation interception submodel was run with a single target plant located at the center of the field sample, and Sobol sensitivity indices (Saltelli, 2002) were estimated for the analysed outputs. Sobol indices are the most widely used sensitivity indices and are robust enough for complex models (Gauchi et al., 2017).

214 The Sobol decomposition leads to decompose the variance of a on-linear model output into
215 percentages of variance which can be attributed to the different inputs, discriminating the variance due
216 to the main effect of a given input from the one resulting from interactions with other inputs. The
217 Sobol indices, based on these variance percentages can be directly interpreted as measures of
218 sensitivity measured across the whole input space. The main advantages is that Sobol indices can deal
219 with nonlinear responses and interactions in non-additive systems. This design gave a set of 12
220 sensitivity indices for each of the 11 inputs. A linear regression of these sensitivity indices was fitted,
221 where the regression coefficients indicate the importance of the effect of the range.

222 Absolute values and ranking of sensitivity indices of the various inputs changed for all outputs when a
223 small input range was used instead of large one (supplementary material online S2 section 1). Not all
224 inputs were, though, concerned, depending on the analysed output, *e.g.* the range of the voxel was
225 important for the relative intercepted PAR (PAR_i) but not for the shading index (SID). Consequently,
226 for the subsequent steps, the large input ranges were used to cover all the possible input situations and
227 notably for novel combinations of species traits, *e.g.* resulting from new crop varieties or invasive
228 weed species.

229 3.1.2 Sensitivity indices estimation via Sobol-Saltelli method and via 230 Polynomial Chaos Expansion

231 The objective of this section was to compare the Sobol sensitivity indices that we estimated in section
232 3.1.1 with a method that both estimates sensitivity indices and fits a metamodel. The Polynomial
233 Chaos Expansion (PCE) method uses the same principle as Sobol sensitivity indices *via* Ordinary
234 Least Square Regression (Sudret, 2008), here shortened to PCE-OLS. For each input, the sensitivity
235 indices estimated by the polynomial chaos expansion are (1) the polynomial effect that accounts for
236 the effect of the input only (*i.e.* the main effect of the input) and (2) the total effect (*i.e.* quantifying all
237 the interactions of this input with other inputs). These indices are respectively comparable to the first-
238 order indices and the total-effect indices of Sobol indices. The large-range experimental design *via*
239 Latin Hypercube Sampling, LHS (McKay et al., 2000), created in the previous section was used to
240 estimate both indices. PCE-OLS indices were similar to Sobol indices computed on the same dataset
241 (supplementary material online S2 section 2). The largest difference was of 0.13 for the total effect of
242 the voxel on the radiation intercepted by the target plant (rPAR_{i,plant}). The ranking of the inputs was the
243 same with both methods. We thus preferred PCE in the following steps since it both estimates
244 sensitivity indices and fits a metamodel, which is needed to simplify the radiation interception
245 submodel.

246 3.1.3 Sensitivity indices with correlated inputs (step 2.ii)

247 The method for estimating PCE-OLS indices assumes that inputs are independent and uncorrelated.
248 However, some inputs of the radiation interception submodel are correlated, *e.g.* plant height and the

249 total leaf area are strongly linked (*e.g.* *Galium aparine* L. (Klem et al., 2014)). We thus tested the
250 effect of including correlations among inputs on the estimation of the sensitivity indices and decided
251 whether the method needed to be adapted. This part was fully presented in Gauchi *et al.* (2017) and
252 further details can be found in supplementary material online S2. In summary, the space filling LHS
253 design of section 3.1.1 was modified to include correlations among inputs following the Iman and
254 Conover method (Iman and Conover, 1982). These correlations (supplementary material online S2
255 section 3) were estimated on simulated plants occurring in 10 diverse cropping systems (Colbach et
256 al., 2016b). A number of 10000 runs were selected out of a total of 29200 runs to avoid an excessive
257 weight of outputs too close to the limit of the ranges. We ensured that the experimental design
258 remained orthogonal and that enough runs were kept to estimate the sensitivity indices. Adding
259 correlations to the space filling design of the inputs changed absolute values of sensitivity indices
260 PCE-OLS for all outputs and gave deviant values (< 0 or > 1).

261 Consequently, it was essential to find a method better adapted to correlated inputs. Gauchi *et al.*
262 (2017) proposed to calculate the sensitivity indices (*i.e.* polynomial effect and total effect) by
263 estimating the coefficients of Polynomial Chaos Expansion using a Partial Least Squares method,
264 namely a Partial Least Squares Regression (PCE-PLS, see (Wold et al., 2001)). Here, the resulting
265 PCE metamodels were though too complex to speed up FLORSYS computations. We thus built more
266 parsimonious and faster metamodels, using a method developed by Gauchi et al (2017) who tested it
267 on a single FLORSYS output. These parsimonious faster metamodels were built with a LASSO
268 regression (Tibshirani, 1996) to select monomials via GLMSELECT (SAS). With the selected
269 monomials we performed a new PLS regression for the final parsimonious metamodel (hence, "fast"
270 metamodel). This combination of methods was hence referred to as LASSO-PLS. The resulting single
271 plant PCE-PLS metamodels (full and fast) were evaluated via a PLS specific criterion, the Q^2_{cum}
272 (Tenenhaus, 1998; Lazraq et al., 2003) for fitting and prediction qualities. We used the same principle
273 and stopping rule as in Gauchi *et al.* (2017) giving a $Q^2_{cum}(h^*)$ referred to as Q^2_{cum} in this paper. This
274 cross-validated fitting prediction criterion is bounded between 0 and 1; the closer to 1 it is, the better
275 the metamodel is in terms of prediction and fitting. The prediction error was evaluated via the relative
276 mean squared error in prediction RRMSEP (supplementary material online S5 section 1). This method
277 was used for the sensitivity analysis and metamodeling of the single-plant case (sections 3.1.5) and
278 then for the more complex case with target plants surrounded by neighbouring plants (section 3.2).

279 3.1.4 Identifying the key inputs that drive radiation interception of 280 single plants (step 2.iii)

281 The sensitivity analysis based on PCE-PLS showed that that the most important inputs for the
282 photosynthetically active radiation absorbed by the plant (PARaP, which drives plant growth) were
283 voxel size and plant width (Figure 3). The third most important inputs were the target-plant
284 characteristics driving potential leaf area absorption ability, *i.e.* total plant leaf area and species

285 extinction coefficient. Total leaf area and plant volume (determined by its width and height) affected
286 PARaP more than leaf distribution (RH50 and b). The environmental variables (latitude and day) as
287 well as field size had small but non-negligible impacts. All inputs strongly interacted, with interactions
288 making up between 46 % (voxel edge size) and almost 100% of the total effect (all others except plant
289 width). Consequently, the sign of the main regression coefficient of an input was useless to assess how
290 an output varied with an input. Graphs of outputs vs. inputs confirmed that interactions made it usually
291 impossible to identify general tendencies, except that PARaP tended to decrease with increasing plant
292 height and width, indicating a self-shading effect (supplementary material online S2, Figure 2 in
293 section 5).

294 The same general tendencies as for PARaP were observed for the other outputs, *i.e.* all inputs matter,
295 voxel edge size and target plant variables mattered more than physical variables and field size (though
296 voxel size could be less important for some outputs such as the shading index, SID); plant volume
297 (though the most relevant variable could be height rather than volume) and leaf area mattered more
298 than plant shape and leaf distribution (supplementary material online S2 section 4).

299 This analysis also showed that large voxel edge sizes (*i.e.*, 10 cm and above) frequently led to weird
300 outputs values, such as an abnormal concentration of 0.5 values for the $rPARi_{plant}$ (supplementary
301 material online S2 section 5). This is due to the computational effect of the voxel-based algorithms,
302 notably when voxels are so large that they include the whole plant (further details in supplementary
303 material online S2 section 5). For small voxels, this only occurs during the 1-2 days after plant
304 emergence, at a time when shading and light absorption have no influence on later plant growth
305 because young plants do not respond to shading and their biomass accumulation only depends on
306 temperature in FLORSYS.

307 3.1.5 Metamodels for a single plant (step 2.iv)

308 The metamodels for a single target plant in the field included all inputs as the sensitivity analysis
309 indicated that all were influential, albeit to varying degrees. The full metamodels included 4367
310 monomials resulting in a good (*i.e.* close to 1) Q^2_{cum} (0.93 - 0.98) (Table 3.A, lines 1, 3, 5, 7) and a
311 low prediction error ($RRMSEP = 0.15 - 0.25 MJ \cdot MJ^{-1}$) (supplementary material online S5 section 3).
312 LASSO-PLS selection produced simpler and faster metamodels, with only 25 to 27 monomials,
313 resulting in a quite good Q^2_{cum} (0.70 - 0.90) but a slightly worse prediction error ($RRMSEP = 0.35 -$
314 0.55 , Table 3.A, lines 2, 4, 6, 8). Regardless of the metamodeling approach (fast or full), radiation
315 interception at the base of the target plant ($rPARi_{base}$, a proxy for the total herbicide penetration into
316 the canopy) is the least well predicted output. This was also the only output that was not calculated at
317 the scale of the plant but at the field scale.

318 3.1.6 Summary for single target plants

319 We tested different sensitivity-analysis methods that increased our knowledge of the 3D radiation
320 interception submodel. This resulted in a more appropriate method that accounted for the correlated
321 inputs. We then proposed a handy solution for more parsimonious and faster metamodels. Part of the
322 methods were developed in a previous study on a single output (Gauchi et al., 2017) and were
323 completed here before being applied to a larger set of outputs.

324 3.2 Case for a target plant inside a canopy (step 3)

325 Fields (or even field portions) rarely only comprise a single plant. Step 4 thus focused on radiation
326 interception of target plants surrounded by neighbouring plants. The method developed in the previous
327 step to analyse and metamodel radiation interception from target-plant, environmental and precision
328 inputs was adapted to (1) include contrasting canopies representing the diversity in crop:weed
329 canopies in arable fields in the simulation plan while (2) limiting the amount of additional inputs
330 needed to describe the canopy surrounding the target plant.

331 3.2.1 Simulation plan

332 A canopy is a complex set of plants of different species, sizes, widths, positions, etc. To set up diverse
333 plant canopies in our virtual field, we needed to vary many variables: plant density (crop density, weed
334 density, amount of bare field area), the position of weeds (random or in patches, number of patches in
335 the field), the position of crop plants (row vs broadcast sown, inter-row width), canopy structure (e.g.,
336 presence and diameter of canopy gaps surrounding target plants), the heterogeneity of plant
337 morphology (mean and variation coefficient of target plant characteristics), weed populations being
338 more heterogeneous than crop population (*i.e.* presenting a larger range of variation) (see
339 supplementary material online S3). These preliminary inputs were used in a LHS design of 20440
340 rows. Correlations were added in the same way as for the single-plant study, with the Iman and
341 Conover method (Iman and Conover, 1982). The diverse canopies were built by placing the plants on
342 a virtual field and attributing morphologies, and then radiation interception and absorption were
343 simulated with the FLORSYS radiation interception submodel. Cases with outlying values were
344 removed as well as output values too close to the range limits (*i.e.* 0, 1 or 100 depending on the
345 output) to avoid side effects due to computation errors; 2536 canopies remained after the sorting. The
346 PCE-PLS method was used to metamodel and perform the sensitivity analysis.

347 3.2.2 Describing the canopy

348 Many detailed variables are needed to create contrasting canopies in FLORSYS, but only a limited
349 number of inputs are allowed keeping the metamodel simple. The detailed canopy variables were thus
350 aggregated into five mean canopy inputs (Table 2), to account for the canopy effect in the metamodel.
351 The nearer the neighbours are to the target, the more their characteristics contribute to the variables
352 describing the average canopy characteristics, here the example of the canopy height (cm):

353 Eq. 1
$$mean\ height = \frac{\sum_{i=1}^n (\frac{1}{d_{i+1}} height_i)}{\sum_{i=1}^n (\frac{1}{d_{i+1}})}$$

354 where d_i is the distance (m) of the target plant to the closest neighbour plant i (+1 to account for a zero
 355 distance when the neighbour is located in the same voxel as the target), $height_i$ is the height (cm) of
 356 neighbour i and n the number of neighbour plants in the field sample. For the equations of the other
 357 canopy variables, see supplementary material online S3 section 4.

358 In addition to these aggregated canopy inputs, we added: (1) the plant density and the maximum
 359 distance between the target plant and neighbour plants, (2) target plant variables (as in the single plant
 360 case) and (3) two environmental variables (latitude and day), resulting in 15 metamodel inputs (Table
 361 2). To reduce the number of inputs, field dimensions (Xmax and Ymax) whose effect was shown to be
 362 slight in the single-plant sensitivity analysis of the single plant (section 3.1.4) were both fixed at 8 m
 363 which allowed having large plants in the virtual field sample. The voxel size was shown to be
 364 important for most outputs (section 3.1.4), but to simplify and accelerate the simulation plan, we kept
 365 it constant. Additional simulations (supplementary material online S2 section 6) showed that a voxel
 366 edge size of 4 cm was the best compromise between the precision of the radiation interception
 367 submodel output and the computation time.

368 3.2.3 Sensitivity indices (step 3.i)

369 The sensitivity analysis of radiation interception outputs to inputs depicting target plant, physical
 370 environment and neighbour plants showed that input effects were almost entirely due to interactions
 371 among inputs (Figure 4). Globally, target-plant inputs had the most and neighbour-plant inputs the
 372 least impact. Inputs of a given type had similar effects, except for the relative PAR intercepted by the
 373 target plant ($rPAR_{i_{plant}}$) whose height effect was several times the effect of any other inputs. As for the
 374 single-plant scenario (section 3.1.4), the interactions among inputs were generally too complex to
 375 identify general tendencies, whether from the signs of the polynomial effects or from graphs
 376 (supplementary material online S2 section 5). And again, outputs were sensitive to all inputs *via*
 377 interactions with other inputs and none of latter could be set at a default value in the following
 378 metamodels.

379 3.2.4 Metamodels (step 3.ii)

380 The metamodels for target plants surrounded by neighbour plants included all inputs, *i.e.* for
 381 describing the target plant, the physical environment and the biological environment due to the
 382 neighbour plants. The polynomial degree of these metamodels was smaller than for the single-plant
 383 ones, except for the relative intercepted PAR $rPAR_i$, (Table 3.B, lines 14-15); the $Q2cum$ was always
 384 lower and the prediction error higher (Table 3.B vs A). Further increasing the polynomial degree did
 385 not improve the $Q2cum$ or reduce the prediction error (results not shown). The need for a higher
 386 polynomial degree for the $rPAR_i$ points to more and more complex interactions among inputs. The fast

387 metamodels usually needed a higher polynomial degree and more complex monomials (Table 3.B)
388 than full metamodels to optimize the Q2cum. The latter though remained low (0.27-0.56) and
389 prediction error was much larger than for single plants (0.85 and 0.65). Radiation interception and
390 absorption by a plant surrounded by neighbour plants is thus much harder to simplify *via* a small
391 metamodel than for single plants.

392 3.2.5 Summary for plants in the canopy

393 The metamodels for radiation interception and absorption by a plant surrounded by neighbour plants
394 were simple enough to be implemented into FLORSYS but with a poorer prediction quality than for a
395 single plant. The canopy creates a complex interaction with the radiation that cannot be easily
396 simplified at a scale as large as the plant. Strong interactions between all inputs prevented us from
397 setting the least important inputs to constants.

398 4 Combining the metamodels into a FLORSYS submodel 399 (step 4)

400 As we had developed metamodels for two situations, *i.e.* single plant and plant in a canopy, it was
401 necessary to establish rules to determine when to use which metamodel in a simulation using the
402 whole FLORSYS including the metamodels (hereafter called FLORSYS-ML). This section presents how
403 the metamodels were combined and what else was needed to cover all likely canopy scenarios with
404 FLORSYS-ML.

405 4.1 Principle

406 Even when there is more than one plant in a field, some of these plants can be considered as single if
407 they do not interfere with each other's radiation interception, which depends on plant sizes, solar angle
408 and distance between plants. Consequently, each day, for each target plant (crop or weed), rules are
409 needed to determine whether a target plant can be considered as single or as surrounded by neighbour
410 plants (supplementary material online S4 section 2).

411 When building the metamodels, a large number of runs were eliminated because outputs were too
412 close to the limits of the range or because their combination was biologically impossible and resulted
413 in deviant values (section 3.2.1). This also reduced the ranges accepted by the metamodels for several
414 key inputs such as target leaf area, making it impossible to predict radiation interception for newly
415 emerged seedlings (*i.e.* with almost nil height, width and leaf area), voluminous single plants (having
416 reached the maximum height and width possible for the species) or mature plants with dried leaves
417 (with a near zero leaf area). But such plant morphologies are frequent in any cropping system. To
418 remedy this, further metamodels were built for the particular case of small seedlings, and for the
419 remaining outlying situations, equations were added to predict radiation interception and absorption
420 from ecophysiological knowledge, or from likely constants (section 4.3). Figure 6 summarizes how the

421 different rules, equations and metamodels were aggregated. Finally, the calculation loops over
422 neighbour plants needed to calculate the aggregated canopy variables are often time-consuming. As a
423 consequence, alternative methods to compute aggregated neighbour were tested (section 4.4).

424 4.2 Rules for deciding whether to use the single plant or plant in a 425 canopy metamodel

426 4.2.1 Method

427 The PAR intercepted at the top of a target plant ($rPAR_{i_{top}}$) is relevant to identify whether radiation
428 interception by the target plant is impacted by neighbour plants, because this output is always 1 for
429 single targets and decreases in the presence of shading neighbours. To establish decision rules to
430 discriminate these two situations, a regression tree was built from the data sets of sections 3.1 and 3.2,
431 using the inputs listed in Table 2. As the metamodels in the previous sections showed that it was
432 difficult to take account of all effects and interactions with these inputs, some were transformed and
433 others added in the present analysis. The environmental variables were transformed to emphasize the
434 effect related to solar angle: latitude was transformed into degrees to the equator (*i.e.* absolute latitude)
435 and Julian days into days from the summer solstice to the winter solstice (*i.e.* between solstice days).
436 The distance from target plant to its closest neighbour was also used as input (with distances
437 calculated between plant centres), and all other inputs were weighted by the inverse of this distance to
438 take into account that closer neighbours shade more than farther neighbours. Finally, the target height
439 relative to the canopy height (overtaking percentage) was integrated *via* the ratio of the difference
440 between the two heights (eq. 6 supplementary material online S3 section 5).

441 The CART method (Breiman et al., 1984) was used to build a classification tree to determine the
442 decision rules. This method successively splits the data set into two subsets along a threshold value of
443 an input (*e.g.* distance to the closest neighbour) in order to maximize the difference between subsets in
444 terms of output. Branches are combinations of input values that lead to output predictions contained in
445 leaf nodes. CART also ranks the input according to their importance to explain the output.

446 The output analysed in the trees was not directly the $rPAR_{i_{top}}$ but a binary variable indicating whether
447 the target plant was considered single or inside canopy, depending on whether its $rPAR_{i_{top}}$ was
448 respectively \geq or $<$ a threshold value. In addition to the theoretical value of 1, ten other thresholds
449 were tested, ranging from 0.90 to 0.99 (incremented by 0.01), in order to increase the number of single
450 plant cases compared to canopy cases and thus the robustness of the tree. Among the 11 trees, the one
451 corresponding to the 0.98 threshold was chosen. This threshold is close to 1 (*i.e.* the theoretical value
452 of $rPAR_{i_{top}}$ in single plants) and it identified the most situations when to use the single-plant
453 metamodel. The latter allows accelerating calculations because the single-plant metamodels were
454 simpler and did not need to calculate the aggregated canopy variables.

4.2.2 Decision tree to determine where a target is shaded by neighbours

The rules determining whether a target plant can be considered as single are shown in Figure 5. For example, if the nearest neighbour is further than 1.6 m, and the target plant is taller than the neighbouring canopy, the target can be considered as single. Surrogates of the tree (*i.e.* variables correlated to the variable in the tree that could also explain the segmentation, but to a lesser degree) and ranking of variables in their order of importance (supplementary material online S3 section 6) showed that nearest neighbour distance (either alone or in interaction with other variables) are predominant to determine whether the target plant is single or within a canopy.

Based on expertise, we added a further logical rule: if there are no neighbours whose height exceeds the distance separating the outer rims of the neighbour and target plants, the target is considered as single. In that case, even if the sun is low on the horizon, the closest neighbour is too far to shade the target (supplementary material online S4). The combination of the decision tree and this additional rule constitute step A in Figure 6.

4.3 Adding equations at the limits of the input ranges

The input ranges of the metamodels missed small seedlings for which good prediction is essential as their initial growth determines which plants outgrow the others. Consequently, we ran a further simulation plan to build a third metamodel focusing on small seedlings (Step C, Figure 6), using the method developed in section 3 (supplementary material online S4 section 4). This additional metamodel was still inadequate for fresh seedlings whose leaf area was lower than the metamodel's accepted input range. In that case, as there is neither shading nor self-shading, the PAR_a absorbed by the plant is the product of the incident PAR_a, the plant leaf area times its extinction coefficient, based on Beer's law (Monsi and Saeki, 1953, 2005) (step B in Figure 6). This works fine for single plants that are unshaded by neighbours. To include either small plants surrounded by neighbours or any plants by small neighbours outside the canopy metamodel range, a linear combination of predictions for single plants (either small or large) and plants in canopy was used, step G in Figure 6. This was particularly true for canopy leaf area whose lower range limit was extremely high (Table 2). Single-plant predictions and target-in-canopy predictions were weighted by respectively 1 and the canopy leaf area, and divided by the same of these weights (supplementary material online S4 section 4).

The metamodels do not include voluminous or mature leaf-less plants either. As these have finished their growth, outputs were simply fixed either to a minimum or maximum value, or linked with a simple regression if one input was out-of-range (step Figure 6). The values were based on graphs of outputs *vs.* inputs from the complete data set including the outliers that were ousted during metamodel construction (supplementary material online S4 section 4). If several inputs were out of range, the output was estimated based on the analysis of the most influential input, with the strongest polynomial

490 effect in the sensitivity analysis (supplementary material online S4 section 4). For example, if a target
491 plant surrounded by neighbours is taller than 254.8 cm, then its relative intercepted PAR is 0.00649
492 MJ·MJ⁻¹.

493 4.4 Different methods to aggregate neighbour plants into canopy 494 variables

495 We proposed three different methods to calculate the aggregated neighbour variables of each target
496 plants: (1) all neighbours close to the target are used for the computation ("local" neighbours), (2) all
497 plants in the field are averaged and the same aggregated variables were used for all target plants
498 ("average" neighbours), (3) a mix between the previous two methods, using average canopy variables
499 when the plant density exceeds 500 plants.m⁻², and local neighbours otherwise. The effect of the
500 aggregation method on prediction error and simulation speed of the whole FLORSYS-ML was
501 evaluated in section 5.

502 5 Evaluation of the simplified FLORSYS-ML with field 503 observations (step 5)

504 5.1 Objective

505 Sections 3.1.5 and 3.2.4 evaluated the prediction quality of the individual metamodels. Here, the
506 objective was to evaluate how good and fast the predictions produced by FLORSYS-ML compared to
507 the process-based FLORSYS, by comparing simulations to field observations following the methods
508 developed in a previous paper (Colbach et al., 2016b). Different voxel edge sizes and the three
509 methods for aggregating neighbour plants were tested.

510 5.2 Material and methods

511 5.2.1 Field observations and features common to all simulations

512 Observations were taken from the INRAE long-term field experiment at Dijon-Epoisses (Burgundy)
513 (Chikowo et al., 2009) where weed and crop variables (plant and seed densities, plant biomass, yield)
514 were monitored from 1999 to 2012. Details can be found in (Colbach et al., 2016b). This trial included
515 ten fields with diverse crop rotations, ranging from intensive herbicide-based to herbicide-free systems
516 and varying degrees of tillage and mechanical weeding. Weed flora was assessed, with species
517 identification, plant density, above-ground biomass and seed bank measurements. Crop yield was also
518 estimated.

519 5.2.2 Simulation plan

520 The simulation combined (1) the FLORSYS version (metamodelled or process-based), with (2) the
521 voxel edge size (1, 4 or 7 cm) which determined the precision of plant location (all FLORSYS versions)
522 and plant morphology (processed-based version). The FLORSYS-ML version moreover tested

523 (3) different methods for aggregating neighbour plants (local, average, or mixed, see section 4.4), and
524 the process-based version tested (4) field sample areas (1m × 1 m, 3 m × 3 m, and 6 m × 3 m) with a
525 7-cm voxel. Unless otherwise indicated, field area was 6 m × 3 m. In total, nine scenarios were run
526 with the FLORSYS-ML version and six with the process-based one. Each of the ten field histories was
527 simulated over 13 years, using the weather measured at the local weather station (INRAE Climatik
528 platform) and starting with the weed seed bank observed at the onset of the field experiment. Each
529 scenario was repeated ten times, to account for stochastic effects. Outputs were produced for all the
530 days where observations were carried out in the fields. Simulations were run with a computer with two
531 2GHz processors and 16 Gb RAM and their simulation time was recorded and averaged over
532 repetitions for the different methods.

533 5.2.3 Evaluation criteria

534 Simulations with the modelled FLORSYS-ML and process-based FLORSYS were compared to field
535 observations. Prediction error was assessed with the relative root square mean squared error of
536 prediction (RRMSEP) corrected for variability in observations (due to measurement errors and intra-
537 field variability) and simulations (due to stochasticity) (Colbach et al., 2016b). This error was
538 calculated relative to the range of variation of the observations (details can be found in supplementary
539 material online S5 section 1). Outputs were analysed at two temporal scales, either corresponding to
540 the individual observation dates (daily scale), or values averaged over the simulation (multiannual
541 scale).

542 5.3 Results

543 5.3.1 Mean simulation time

544 The simulation time of the process based FLORSYS for all cropping systems tested, decreased with
545 voxel edge size. When voxel edge size increased from 1 to 4 cm, simulation time was divided by
546 approximately 20 (Figure 7.A). Increasing voxel size further from 4 to 7 cm decreased simulation time
547 by an additional 43%. Increasing voxel size from 7 to 10 cm did not decrease simulation time any
548 further. The slowest scenario took 259 times more time than the fastest. The fastest scenario with the 7
549 cm voxel edge size and 1-m² field sample took 4 minutes for a repetition of the 13 year long cropping
550 system, compared to more than 18 hours for the slowest, with the 1-cm voxel and the 18-m² area.
551 Conversely, simulation time increased with field sample area (supplementary material online S5
552 section 2). Increasing area from 1 to 9 m² multiplied the simulation time by approximately 8; doubling
553 the field sample area to 18 m² only increased the simulation time by a further 10%. The field size
554 multiplies the simulation time by 1.15 for every m² of a 13-year simulation

555 The simulation time of FLORSYS-ML remained stable for all voxel sizes, but it depended on the
556 method for calculating neighbouring canopy variables (Figure 7.A). The FLORSYS-ML with average
557 neighbours was fastest and the one combining local and average neighbours was nearly as fast.

558 Always using local neighbours made simulations considerably slower than with the process-based
559 model, and simulation time even increased with voxel edge size. Indeed, in FLORSYS-ML, the voxel
560 determines plant location, and the larger the voxel is, the more plants are in each voxel. So, when
561 FLORSYS-ML searches through the voxels surrounding the target plant to compute the canopy inputs,
562 it must compute more plants, which takes longer. FLORSYS-ML was considerably faster than the
563 process-based model with small voxel edge sizes, *i.e.* 28 times faster for FLORSYS-ML with average
564 neighbours.

565 5.3.2 Prediction error

566 In process-based simulations, the prediction error tended to increase slightly with increasing voxel size
567 (Figure 7.B). The same trend was observed for prediction error in FLORSYS-ML simulations with
568 larger voxels, suggesting a sensitivity to plant position, which is less precise if the voxel is large.
569 Simulations with a 1-m² field sample produced slightly better results than 18-m² and larger areas (*e.g.*
570 for the multiannual weed density for all species summed, the RRMSEP for 1 × 1, 3 × 3 and 6 × 3 m²
571 field samples was respectively 63, 113 and 116 MJ·MJ⁻¹, details in supplementary material online S5
572 section 3), probably because it increased interspecific competition between weed species by increasing
573 the probability of overlapping species patches. However, small fields potentially miss rare species, and
574 overestimate interspecific competition in case of high weed densities.

575 Generally, the error was larger for metamodel-based vs. process-based simulations, particularly for
576 weed plant biomass (Table 4), and it varied more among repetitions (supplementary material online S5
577 section 3). Error was often smaller than the variability in observations, pointing to a negligible
578 prediction error, and making it impossible to calculate the relative variation in error for metamodelled
579 vs process-based simulations (Table 4). Conversely, FLORSYS-ML was better than the process-based
580 FLORSYS to predict multiannual weed plant densities.

581 Usually, FLORSYS-ML using either local or average neighbours respectively had the smallest and
582 largest errors, whereas errors were intermediate when using both average and local neighbours (Figure
583 7.B, Table 4). Regardless of the evaluation criteria, there was no model version (process or
584 metamodel-based, approach for calculating canopy variables in metamodels) or precision level (voxel
585 size, field sample area) that optimized the precision of all model outputs.

586 6 Discussion

587 6.1 Simplifying a complex process-based model

588 In this article, we presented a method to accelerate and simplify a complex process-based model. The
589 paper is of interest for non-statisticians that want to metamodel complex models and are often baffled
590 by statistical methods and how to apply them in their real-life complicated situation. Another
591 particularity of our work was that we did not use the metamodel as such but integrated it into a larger

592 model and evaluated the latter with independent field data, two steps that have, to the best of our
593 knowledge, been rarely carried out in the past.

594 From a more technical point of view, the originality of the approach lies in (1) the choice of
595 metamodelling only the most time-consuming part of the model (*i.e.* the 3D light-interception
596 submodel), (2) the choice of an innovative metamodelling method that handles correlated inputs and
597 selects monomials, (3) the integration of the metamodelled submodel into the complex model and (4)
598 the description of the nearby canopy with a limited number of inputs. This work did not compare
599 different metamodelling methods (Villa-Vialaneix et al., 2012) but provided practical guidelines for
600 choosing and tuning metamodelling methods with respect to the complex model constraints (*e.g.*
601 correlated inputs). It extended what was done in the previous paper (Gauchi et al., 2017) by showing
602 the whole approach to simplify a complex model. Usually, metamodelling via polynomial chaos
603 expansion allows reducing the number of inputs in the model by setting the inputs to average values
604 (Luo et al., 2013; Rothenberg and Wang, 2016). Here, however, no input could be omitted because all
605 either influenced radiation interception outputs directly or in interaction with other inputs.

606 Usually, the whole model is metamodelled, avoiding the need to integrate the metamodel into a larger
607 model (Cohen and Prinn, 2011; Luo et al., 2013). Here, we metamodelled a single time-consuming
608 submodel in order to accelerate the simulations of the whole FLORSYS model, and we thus had to
609 integrate the metamodels, together with complementary equations, into FLORSYS. In SIRIUS (Brooks
610 et al., 2001), only a few equations were metamodelled. No implementation of the metamodel was
611 needed, as the metamodel was as good as the whole SIRIUS to predict the yield, which was the study's
612 goal. The constraints of this approach were manageable for the 3D radiation interception of FLORSYS
613 even though the number of inputs needed to be decreased with the help of aggregated canopy
614 variables. But these constraints probably make it impossible to apply this metamodelling method to
615 bigger models like the whole FLORSYS with its many more inputs and correlations.

616 6.2 Experimental design for analysing a complex model

617 The numerical space filling design, Latin Hypercube Sampling (LHS), is usually appropriate to
618 explore the whole space of possible input combinations. For our biological example, we also used the
619 Iman and Conover method (Iman and Conover, 1982) to apply a correlation matrix to the LHS, to
620 increase the biologically realistic plant variable combination. It worked less well for the dynamic
621 FLORSYS model, especially at the outer bounds of input ranges that were not sampled enough, despite
622 having tested the best minimum row number in the LHS design. This was particularly problematic at
623 the onset of the plants' life-cycle (*i.e.* for small plants) as imprecise early predictions would amplify
624 the next days' prediction errors, thus setting off the plants' growth and development in entirely the
625 wrong direction. We improved the metamodelling by using separate experimental sampling designs,
626 combining with a simplified process-based approach, discriminating three types of plants that differ in

627 terms of light interception, based on their age and size: (1) a standard LHS and metamodel covering
628 mostly large and older plants, (2) a second LHS and metamodel specifically targeting younger smaller
629 plants, (3) a simplified ecophysiological equation for newly emerged tiny plants. The latter approach
630 was acceptable here as these plants do not self-shade and are rarely shaded by neighbours. No such
631 effort was made at the other extreme of plant-size range, i.e. large adult plants with dried leaves.
632 Indeed, these do not photosynthesize anymore and misestimating their absorbed light would have little
633 impact on their future. Colleagues aiming to similarly metamodel complex could use a similar
634 approach, i.e. keep a process-based (albeit simplified) approach for the most sensitive
635 stages/components (here, the newly emerged tiny plants) and combine separate sampling plants and
636 metamodels to ensure that the more sensitive stages/components are adequately covered (here, the
637 small plants).

638 The inability of the metamodels to correctly predict small plants is explained by three combined
639 reasons: (1) we chose a broad input range to cover all possible plant morphologies in the field, which
640 reduced the probability of drawing many low input variables, (2) as the simulated plants were the
641 combination of several inputs, the probability of drawing a small plant combining low values of all
642 inputs (*e.g.* low height, width and leaf area) was even lower, particularly as the space filling design
643 was balanced, (3) the equilibrated design also drew plants combining high values for some inputs with
644 low values for others, resulting in biological impossible morphologies (*e.g.* tiny plants with an
645 enormous leaf area) and non-logical output values. These plants had to be removed from the data set,
646 decreasing even more the occurrence of extreme input values used in the metamodels. For models with
647 a high number of inputs it is thus better to sample stepwise rather than have a unique sampling design.
648 Surprisingly, adding correlation to inputs did not help to ensure many small and plausible plants.

649 6.3 Which method for which application?

650 To metamodel and perform a sensitivity analysis, many methods exist, which have been assessed in
651 comparative studies. We thus decided to detail here the entire path when choosing and applying a
652 method to transform a complex slow model into a faster metamodel. Polynomial chaos metamodeling
653 accepts only a small number of inputs, hence the aggregation of neighbour plant variables into a small
654 number of synthetic canopy variables. Unfortunately, it is the aggregation step, particularly the loop
655 computing the plants close to the target plant, which cancelled out the simulation time saved thanks to
656 the metamodels.

657 Another way to speed up simulations would be to use the initial process-based interception submodel
658 and to decrease the precision of the canopy structure by increasing the voxel edge size, which governs
659 the precision of plant locations and volumes as well as leaf distribution along plant height. This
660 approach led to less precision loss than expected. Indeed, FLORSYS does not explicitly represent plant
661 architecture in detail, with each organ (*e.g.*, leaf, stem) simulated. If that had been the case, enlarging

662 voxels would indeed decrease prediction quality considerably. Actually, FLORSYS distributes leaf area
663 in voxel layers, without considering leaf size, position or inclination. Very small voxels do thus not
664 necessarily place leaf area in the correct voxel layer and downsizing voxels cannot be simply
665 approximated here to differential equations that govern light transmission. We already discussed the
666 fact that model quality does not decrease below a certain voxel size and is not best for the smallest
667 voxel (i.e. the highest precision) in Munier-Jolain et al (2013). Moreover, there are several stochastic
668 functions in FLORSYS, particularly for plant location in the canopy. This explains why differences
669 between scenarios differing solely in terms of voxel size are not solely due to differences in voxel size.

670 We thus identified two ways to save simulation time depending on the simulation goal: either by
671 decreasing the precision of the plant and canopy description (process-based light interception
672 submodel with a large voxel edge size) or that of the light interception (metamodelled submodel with a
673 small voxel edge size). Choosing rapidity over precision can be appropriate, for example when
674 needing quick simulations for workshops with farmers to co-design cropping systems (Bergez et al.,
675 2010). The choice of the approach then depends on the target output (Table 5). When plant location is
676 essential (e.g., when testing site-specific weed management, small sowing interrows, row-only
677 nitrogen fertilization) (Berge et al., 2013), then a voxel size of 1 cm is needed, and the metamodelled
678 FLORSYS-ML would allow faster and thus more simulations than the process-based FLORSYS. When
679 both moderate simulation time and prediction quality are needed, the process based FLORSYS with a
680 voxel size of 4 cm would be best. For cropping system tests, a quantitative precision is less essential as
681 long as the management recommendations are correct (Renton, 2011).

682 6.4 What other solutions to speed up a complex model?

683 Instead of only simplifying the process-based approach for tiny newly-emerged plants, we also
684 thought about using a simplifying process-based light interception model. This approach was used in
685 early crop-weed competition models (Graf et al., 1990; SOYWEED: Wilkerson et al., 1990;
686 ALMANAC: Kiniry et al., 1992; e.g. INTERCOM: Kropff and Spitters, 1992; Kropff et al., 1992) as
687 well as in intercrop models (Gaudio et al., 2019). However, these models only work for homogeneous
688 2-species canopies and cannot grasp the complexity of heterogeneous multispecies crop-weed
689 canopies (in terms of location, emergence timing and morphology), as shown by comparisons of
690 simulations with such models to independent field observations (Debaeke et al., 1997; Deen et al.,
691 2003). Consequently, the recent trend in crop-weed competition modelling goes towards more
692 complexity rather than less (Renton, 2013). A recent review of multispecies canopy models even
693 concluded that the FLORSYS approach was a good compromise between simplicity and accounting for
694 canopy heterogeneity (Gaudio et al., 2019).

695 There are also technical solutions for speeding up simulations, for instance parallelising the execution
696 of the source code or using graphical processing units. Unfortunately, these solutions make it difficult

697 to maintain a unique source code for any type of computer or server. A "portable" solution is to run
698 multiple FLORSYS clones simultaneously on a single computer or server, either manually or
699 automatically via scripts. The speed gain then depends on the number of logical processors of the
700 computer. Another avenue is similar to the large-voxel solution, i.e. reduce the size of the simulated
701 field sample. This solution was already assessed in a previous paper (Colbach et al., 2016b) where we
702 determined the minimum acceptable size. Again, the more complex (e.g., many species, large
703 interrows), the larger the field sample needs to be.

704 So, there are several avenues for speeding up a complex model (Table 6). The best choice depends on
705 the objective and situation of use, and several solutions can be combined for an even better result.

706 6.5 Towards a larger simplification of FLORSYS

707 The simplification of the radiation interception was easier for a single plant in a bare field, than for a
708 plant located inside a canopy. Indeed, (1) the interaction with the canopy is harder to metamodel, and
709 (2) the aggregated inputs simplify the canopy too much. Simplifying a complex model with many
710 inputs is a principal issue when metamodeling. The complexity of the relationship between inputs is
711 also an issue; for the 3D radiation interception, even small variations in outputs need to be accurately
712 predicted, because small errors amplify over time as a result of the daily retro-acting interactions of
713 light interception and growth. Metamodels based on polynomials are efficient to model all the single
714 variations of the function (Hussain et al., 2002), hence were adapted for the submodel. However, for a
715 general trend, metamodeling based on polynomials cannot provide such a smooth answer. The present
716 study suggests that the polynomial chaos expansion metamodeling, even when performed step by step
717 and improved with expert knowledge, would be inadequate to metamodel the whole FLORSYS model,
718 with its many and diverse inputs. To build a metamodel and estimate sensitivity indices, this method
719 was the most suitable as there is no method that can handle many inputs, metamodeling and
720 estimation of sensitivity indices at the same time.

721 Consequently, for a global emulation of FLORSYS, in order to synthesize and make available to
722 farmers the knowledge comprised in FLORSYS to help with decision making (Wilkerson et al., 2002),
723 other methods need to be considered. In that case, non-parametric methods can be helpful. Villa-
724 Vialaneix *et al.* (2012) showed that metamodeling methods based on machine learning have good
725 results for medium and large data sets. This is particularly true for Random Forests (Breiman, 2001)
726 which provide the best trade-off between speed and accuracy. Moreover, non-parametric methods can
727 tolerate heterogeneous data sets. This is crucial as FLORSYS with its numerous inputs precludes
728 building a suitable experimental design as the one needed for the present approach.

729 The global emulation of FLORSYS will be a necessary step to make the model accessible for farmers
730 and crop advisors, particularly for a use in participatory workshops (Colas et al., 2020). Indeed, none

731 of the avenues proposed in Table 6 will be fast enough or compatible with computers used in such as
732 situation.

733 7 Conclusion

734 The present study demonstrated that the frequent practice of developing statistical methods on rather
735 simple case studies makes them difficult and sometimes impossible to apply to more complex real-life
736 situations. Latin Hypercube Sampling (LHS) was used for numerical space filling design, followed by
737 Partial Least Squares regression, combined with a polynomial expansion chaos model and selection of
738 the most influential monomials, to produce simple metamodels. The individual methods used here had
739 trouble handling all the constraints and the domain of validity needed: they (1) eliminated many data
740 close to the limits of the domain of validity of the metamodel (e.g. tiny plants tiny with near-zero leaf
741 areas immediately after emergence) from the simulated data set based on LHS in section 3.1.3,
742 (2) insufficiently accounted for correlations among inputs (e.g. 2-m-tall and 1-cm-narrow plants do not
743 exist) despite using an adapted LHS sampling plan. But both these extreme cases and correlations are
744 frequent in real life and essential for correctly predicting the agroecosystem. Notably, (3) the
745 complexity of radiation transmission and interception inside crop-weed canopies, particularly due to
746 shading by neighbour plants, made it difficult to directly predict radiation absorption at the plant scale.
747 (4) This made it necessary to add functions here, which slowed down simulations again considerably
748 and made us lose most of the speed gain due to the metamodel.

749 So, to simplify a complex process-based weed dynamics model such as FLORSYS, it is essential to
750 combine different methods of sensitivity analysis and model simplification to cover the whole range of
751 relevant stages/morphologies and take account of the complex interactions between plant objects and
752 the many feedbacks during their life cycle. We used a hybrid approach, using a process-based (albeit
753 simplified) approach for the most sensitive plant stage (newly emerged tiny plants) and separate
754 sampling plans and metamodels to ensure that the more sensitive stages/components were adequately
755 covered (small plants). By evaluating the various approaches with independent field observations, we
756 assessed the trade-off between prediction accuracy and simulation speed to identify which modelling
757 approach was best, depending on the objective of the model use.

758 8 Acknowledgements

759 The present work was funded by INRAE (EA and MIA divisions), the French project CoSAC (ANR-
760 14-CE18-0007) and the Burgundy Region. The cropping-system trials were conducted at the INRAE
761 experimental station of Dijon-Époisses, and the weed observations were carried out by Dominique
762 Meunier and colleagues (INRAE Agroécologie Dijon). The paper was greatly improved thanks to the
763 comments of an anonymous reviewer on an earlier version of the manuscript.

764 9 References

765

- 766 Bagavathiannan, M., Beckie, H., Chantre, G., González-Andújar, J., León, R., Neve, P., Poggio, S.,
767 Schutte, B., Somerville, G., Werle, R., Van Acker, R., 2020. Simulation models on the ecology and
768 management of arable weeds: Structure, quantitative insights, and applications. *Agronomy* 10, 1611.
- 769 Berge, T.W., Goldberg, S., Kaspersen, K., Netland, J., 2013. Towards machine vision based site-
770 specific weed management in cereals. *Computers and Electronics in Agriculture* 81, 79-86.
- 771 Bergez, J.E., Colbach, N., Crespo, O., Garcia, F., Jeuffroy, M.H., Justes, E., Loyce, C., Munier-Jolain,
772 N., Sadok, W., 2010. Designing crop management systems by simulation. *Eur. J. Agron.* 32, 3-9.
- 773 Bizouard, G., 2012. *Métamodélisation : état de l'art et application.*, UFR Sciences et techniques.
774 Université de Bourgogne, Dijon, France, p. 60.
- 775 Breiman, L., 2001. Random Forests. *Machine Learning* 45, 5-32.
- 776 Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A., 1984. *Classification and Regression Trees.*
777 CRC Press, New York.
- 778 Brooks, R.J., Semenov, M.A., Jamieson, P.D., 2001. Simplifying Sirius: sensitivity analysis and
779 development of a meta-model for wheat yield prediction. *Eur. J. Agron.* 14, 43-60.
- 780 Chikowo, R., Faloya, V., Petit, S., Munier-Jolain, N., 2009. Integrated Weed Management systems
781 allow reduced reliance on herbicides and long term weed control. *Agriculture, Ecosystems and*
782 *Environment* 132, 237-242.
- 783 Cohen, J.B., Prinn, R.G., 2011. Development of a fast, urban chemistry metamodel for inclusion in
784 global models. *Atmos. Chem. Phys.* 11, 7629-7656.
- 785 Colas, F., Cordeau, S., Granger, S., Jeuffroy, M.-H., Pointurier, O., Queyrel, W., Rodriguez, A.,
786 Villerd, J., Colbach, N., 2020. Co-development of a decision support system for integrated weed
787 management: contribution from future users. *Eur. J. Agron.* 114, 126010.
- 788 Colbach, N., 2010. Modelling cropping system effects on crop pest dynamics: how to compromise
789 between process analysis and decision aid. *Plant Sci.* 179, 1-13.
- 790 Colbach, N., Bertrand, M., Busset, H., Colas, F., Dugue, F., Farcy, P., Fried, G., Granger, S., Meunier,
791 D., Munier-Jolain, N., Noilhan, C., Strbik, F., Gardarin, A., 2016a. Uncertainty analysis and
792 evaluation of a complex, multi-specific weed dynamics model with diverse and incomplete data sets.
793 *Environmental Modelling and Software* 86, 184-203.
- 794 Colbach, N., Bertrand, M., Busset, H., Colas, F., Dugué, F., Farcy, P., Fried, G., Granger, S., Meunier,
795 D., Munier-Jolain, N.M., Noilhan, C., Strbik, F., Gardarin, A., 2016b. Uncertainty analysis and
796 evaluation of a complex, multi-specific weed dynamics model with diverse and incomplete data sets.
797 *Environmental Modelling & Software* 86, 184-203.
- 798 Colbach, N., Colas, F., Cordeau, S., Maillot, T., Queyrel, W., Villerd, J., Moreau, D., 2021. The
799 FLORSYS crop-weed canopy model, a tool to investigate and promote agroecological weed
800 management. *Field Crops Research* 261, 108006.
- 801 Colbach, N., Colas, F., Pointurier, O., Queyrel, W., Villerd, J., 2017. A methodology for multi-
802 objective cropping system design based on simulations. Application to weed management. *European*
803 *Journal of Agronomy* 87, 59-73.
- 804 Colbach, N., Collard, A., Guyot, S.H.M., Mézière, D., Munier-Jolain, N.M., 2014. Assessing
805 innovative sowing patterns for integrated weed management with a 3D crop:weed competition model.
806 *Eur. J. Agron.* 53, 74-89.
- 807 Colbach, N., Debaeke, P., 1998. Integrating crop management and crop rotation effects into models of
808 weed population dynamics: a review. *Weed Sci.* 46, 717-728.
- 809 Cox, G.M., Gibbons, J.M., Wood, A.T.A., Craigon, J., Ramsden, S.J., Crout, N.M.J., 2006. Towards
810 the systematic simplification of mechanistic models. *Ecological Modelling* 198, 240-246.
- 811 Debaeke, P., Caussanel, J.P., Kiniry, J.R., Kafiz, B., Mondragon, G., 1997. Modelling crop:weed
812 interactions in wheat with ALMANAC. *Weed Research.* 37, 325-341.
- 813 Deen, W., Cousens, R., Warringa, J., Bastiaans, L., Carberry, P., Rebel, K., Riha, S., Murphy, C.,
814 Benjamin, L.R., Cloughley, C., Cussans, J., Forcella, F., Hunt, T., Jamieson, P., Lindquist, J., Wang,

815 E., 2003. An evaluation of four crop: weed competition models using a common data set. *Weed Res.*
816 43, 116-129.

817 Faivre, R., Iooss, B., Mahévas, S., Makowski, D., Monod, H., 2013. Analyse de sensibilité et
818 exploration de modèles. Editions Quae, Versailles, France 352 pp.

819 Freckleton, R.P., Stephens, P.A., 2009. Predictive models of weed population dynamics. *Weed Res.*
820 49, 225-232.

821 Gardarin, A., Dürr, C., Colbach, N., 2012. Modeling the dynamics and emergence of a multispecies
822 weed seed bank with species traits. *Ecol. Modelling* 240, 123-138.

823 Gauchi, J.P., Bensadoun, A., Colas, F., Colbach, N., 2017. Metamodeling and global sensitivity
824 analysis for computer models with correlated inputs: A practical approach tested with a 3D light
825 interception computer model. *Environmental Modelling & Software* 92, 40-56.

826 Gaudio, N., Escobar-Gutierrez, A.J., Casadebaig, P., Evers, J.B., Gérard, F., Louarn, G., Colbach, N.,
827 Munz, S., Launay, M., Marrou, H., Barillot, R., Hinsinger, P., Bergez, J.-E., Combes, D., Durand, J.-
828 L., Frak, E., Pagès, L., Pradal, C., Saint-Jean, S., Werf, W.V.D., Justes, E., 2019. Modeling mixed
829 annual crops: current knowledge and future research avenues. A review. *Agron. Sustain. Dev.* 39, 20.

830 Graf, B., Gutierrez, A.P., Rakotobe, O., Zahner, P., Delucchi, V., 1990. A simulation model for the
831 dynamics of rice growth and development: Part II-The competition with weeds for nitrogen and light.
832 *Agricultural Systems* 32, 367-392.

833 Harrington, P.d.B., Urbas, A., Wan, C., 2000. Evaluation of Neural Network Models with Generalized
834 Sensitivity Analysis. *Analytical Chemistry* 72, 5004-5013.

835 Holst, N., Rasmussen, I.A., Bastiaans, L., 2007a. Field weed population dynamics: a review of model
836 approaches and applications. *Weed Res.* 47, 1-14.

837 Holst, N., Rasmussen, I.A., Bastiaans, L., 2007b. Field weed population dynamics: a review of model
838 approaches and applications. *Weed Res.* 47, 1-14.

839 Hossard, L., Souchere, V., Pelzer, E., Pinochet, X., Jeuffroy, M.H., 2015. Meta-modelling of the
840 impacts of regional cropping system scenarios for phoma stem canker control. *Eur. J. Agron.* 68, 1-12.

841 Hussain, M.F., Barton, R.R., Joshi, S.B., 2002. Metamodeling: Radial basis functions, versus
842 polynomials. *European Journal of Operational Research* 138, 142-154.

843 Iman, R.L., Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among
844 input variables. *Communications in Statistics - Simulation and Computation* 11, 311-334.

845 Kiniry, J.R., Williams, J.R., Gassman, Debaeke, P., 1992. A general, process-oriented model for two
846 competing plant species. *Transactions of the ASAE* 35, 801-810.

847 Kleijnen, J.P.C., Sargent, R.G., 2000. A methodology for fitting and validating metamodels in
848 simulation I. *European Journal of Operational Research* 120, 14-29.

849 Klem, K., Rajsnerova, P., Novotna, K., Urban, O., Marek, M.V., 2014. Effect of the relative time of
850 emergence on the growth allometry of *Galium aparine* in competition with *Triticum aestivum*. *Weed*
851 *Biol. Manag.* 14, 262-270.

852 Kropff, M.J., Spitters, C.J.T., 1992. An ecophysiological model for interspecific competition, applied
853 to the influence of *Chenopodium album* L. on sugar-beet. 1. Model description and parameterization.
854 *Weed Res.* 32, 437-450.

855 Kropff, M.J., Spitters, C.J.T., Schnieders, B.J., Joenje, W., Degroot, W., 1992. An ecophysiological
856 model for interspecific competition, applied to the influence of *Chenopodium album* L. on sugar-beet.
857 2. Model evaluation. *Weed Res.* 32, 451-463.

858 Lazraq, A., Cléroux, R., Gauchi, J.-P., 2003. Selecting both latent and explanatory variables in the
859 PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems* 66, 117-126.

860 Liebman, M., Gallandt, E.R., 1997. Many Little Hammers: Ecological Management of Crop-Weed
861 Interactions, in: Jackson, L.E. (ed.), *Ecology in Agriculture*. Academic Press, pp. 291-343.

862 Luo, Z., Wang, E., Bryan, B.A., King, D., Zhao, G., Pan, X., Bende-Michl, U., 2013. Meta-modeling
863 soil organic carbon sequestration potential and its application at regional scale. *Ecological*
864 *Applications* 23, 408-420.

865 Marie, G., Simioni, G., 2014. Extending the use of ecological models without sacrificing details: a
866 generic and parsimonious meta-modelling approach. *Methods in Ecology and Evolution* 5, 934-943.

867 Marshall, E.J.P., Brown, V.K., Boatman, N.D., Lutman, P.J.W., Squire, G.R., Ward, L.K., 2003. The
868 role of weeds in supporting biological diversity within crop fields. *Weed Res.* 43, 77-89.

869 McKay, M.D., Beckman, R.J., Conover, W.J., 2000. A comparison of three methods for selecting
870 values of input variables in the analysis of output from a computer code. *Technometrics* 42, 55-61.
871 Monsi, M., Saeki, T., 1953. Über den Lichtfaktor in den Pflanzengesellschaften und seine Bedeutung
872 für die Stoffproduktion. *Japanese Journal of Botany* 14, 22-52.
873 Monsi, M., Saeki, T., 2005. On the factor light in plant communities and its importance for matter
874 production. *Annals of Botany* 95, 549-567.
875 Munier-Jolain, N.M., Guyot, S.H.M., Colbach, N., 2013. A 3D model for light interception in
876 heterogeneous crop:weed canopies. Model structure and evaluation. *Ecol. Modelling* 250, 101-110.
877 Oerke, E.-C., 2006. Crop losses to pests. *Journal of Agricultural Science* 144, 31-43
878 Petit, S., Boursault, A., Le Guilloux, M., Munier-Jolain, N., Reboud, X., 2011. Weeds in agricultural
879 landscapes. A review. *Agron. Sustain. Dev.* 31, 309-317
880 Plackett, R.L., Burman, J.P., 1946. The Design of Optimum Multifactorial Experiments. *Biometrika*
881 33, 305-325.
882 Renton, M., 2011. How much detail and accuracy is required in plant growth sub-models to address
883 questions about optimal management strategies in agricultural systems? *AoB PLANTS* 2011, plr006-
884 plr006.
885 Renton, M., 2013. Shifting focus from the population to the individual as a way forward in
886 understanding, predicting and managing the complexities of evolution of resistance to pesticides. *Pest*
887 *Management Science* 69, 171-175.
888 Rothenberg, D., Wang, C., 2016. Metamodeling of Droplet Activation for Global Climate Models.
889 *Journal of the Atmospheric Sciences* 73, 1255-1272.
890 Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. *Computer*
891 *Physics Communications* 145, 280-297.
892 Stoate, C., Baldi, A., Beja, P., Boatman, N.D., Herzon, I., van Doorn, A., de Snoo, G.R., Rakosy, L.,
893 Ramwell, C., 2009. Ecological impacts of early 21st century agricultural change in Europe - A review.
894 *Journal of Environmental Management* 91, 22-46.
895 Sudret, B., 2008. Global sensitivity analysis using polynomial chaos expansions. *Reliability*
896 *Engineering & System Safety* 93, 964-979.
897 Tenenhaus, M., 1998. *La régression PLS: théorie et pratique*. Editions Technip.
898 Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal*
899 *Statistical Society. Series B (Methodological)* 58, 267-288.
900 Villa-Vialaneix, N., Follador, M., Ratto, M., Leip, A., 2012. A comparison of eight metamodeling
901 techniques for the simulation of N₂O fluxes and N leaching from corn crops. *Environmental*
902 *Modelling & Software* 34, 51-66.
903 Waggoner, J., Henneberger, P., Kullman, G., Umbach, D., Kamel, F., Beane Freeman, L., Alavanja,
904 M.R., Sandler, D., Hoppin, J., 2013. Pesticide use and fatal injury among farmers in the Agricultural
905 Health Study. *International Archives of Occupational and Environmental Health* 86, 177-187.
906 Wilkerson, G.G., Jones, J.W., Coble, H.D., Gunsolus, J.L., 1990. SOYWEED - a simulation-model of
907 soybean and common cocklebur growth and competition. *Agronomy Journal* 82, 1003-1010.
908 Wilkerson, G.G., Wiles, L.J., Bennett, A.C., 2002. Weed management decision models: pitfalls,
909 perceptions, and possibilities of the economic threshold approach. *Weed Science Society of America*,
910 Lawrence, KS, ETATS-UNIS 14 pp.
911 Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics.
912 *Chemometrics and Intelligent Laboratory Systems* 58, 109-130.

913

914 10 Illustrations

915

916

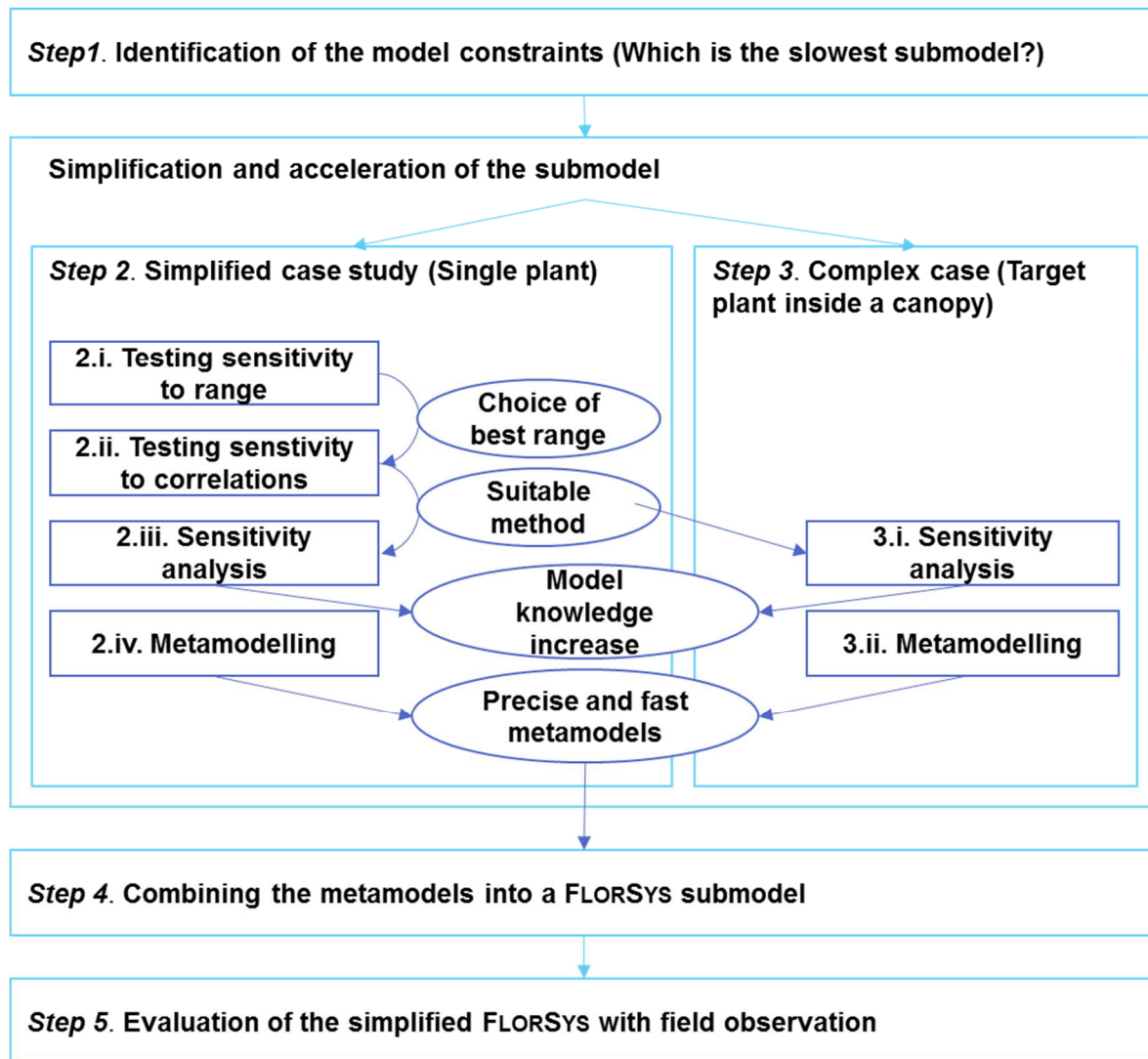
917 Table 1: Compilation of different sensitivity analysis methods for independent variables depending on complex model's proprieties. From (Tenenhaus, 1998;
 918 Harrington et al., 2000; Bizouard, 2012; Faivre et al., 2013; Gauchi et al., 2017)

919

	ANOVA	Sobol-Saltelli	FAST	PCE-OLS	PCE-PLS	CART; random forest	Neural network
Model characteristics							
Inputs number > 10	difficult to test all interactions	Yes	difficult, too heavy	Yes	Yes	Yes	Yes
Possible run number > 1000	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Accepts correlated inputs	No	No	No	No	Yes	Yes	Yes
Properties of sensitivity methods							
Estimates sensitivity indices	Yes	Yes	Yes	Yes	Yes	No	No
Evaluates inputs for their importance	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Provides a metamodel	Yes	No	No	Yes	Yes	Yes	Yes
Simulation design available from: LHS, Sobol sequence, Monte-Carlo, Hadamard, Full factorial design, Morris, OAT, numerous data from different sources	all	LHS, Sobol sequence, Monte-Carlo	Monte- Carlo	LHS, Sobol sequence, Monte-Carlo	LHS, Sobol sequence, Monte-Carlo	all	all

920

921



922

923 Figure 1: Schematic representation of the steps of the simplification and acceleration of the model
 924 FLORSYS. The Arabic numbers and roman numbers correspond to, respectively, the sections and sub-
 925 sections of the paper (Floriane Colas © 2017).

926

927

928 Table 2: Definition, range variation and unit of the inputs and outputs of the 3D radiation interception
 929 submodel.

930 A. Inputs

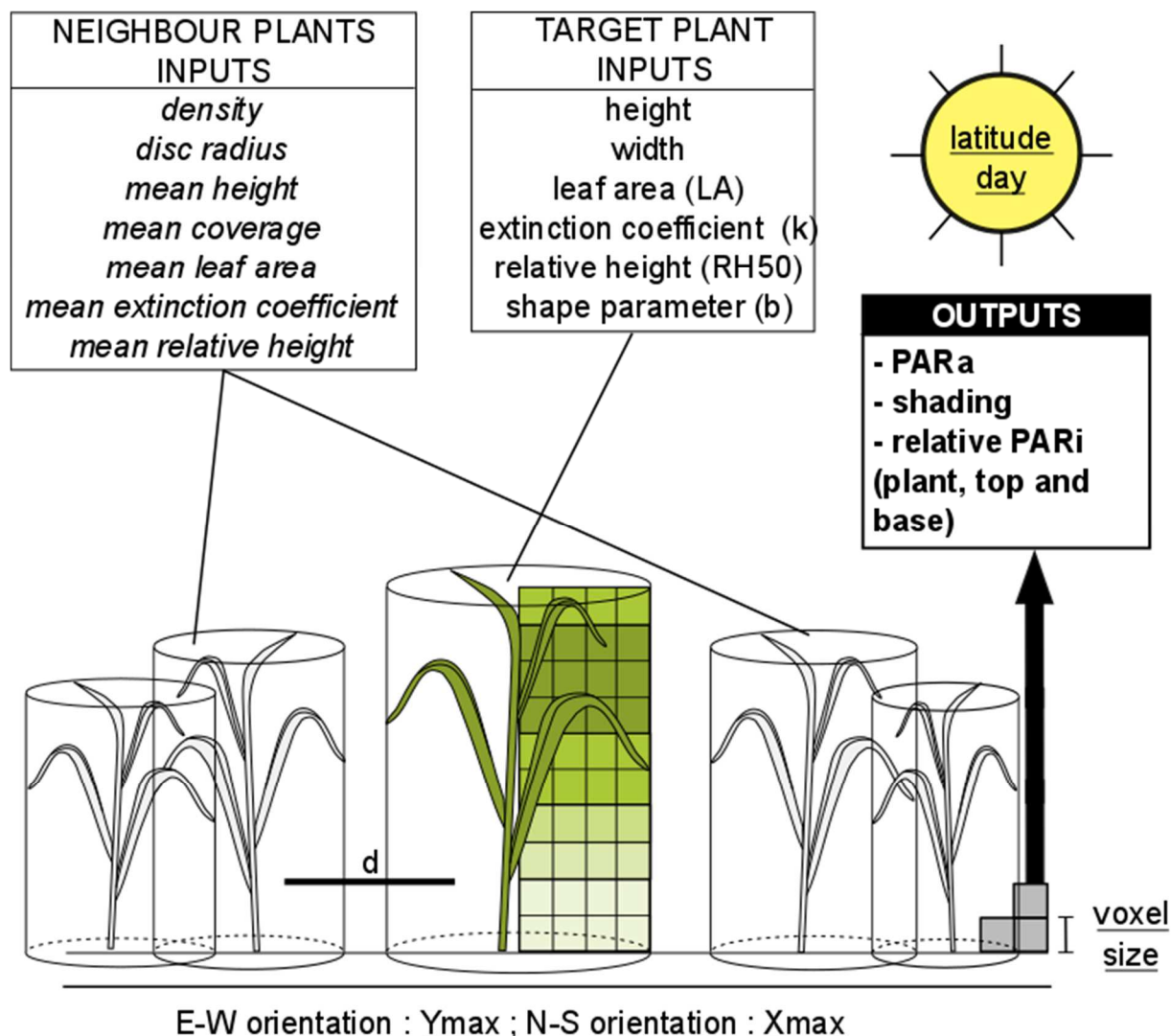
Type	Name	Short explanation	Step [§]	Range of variation	Unit
Physical environment	Latitude	Latitude of the simulated field	both	[-66; +66] single plan [0; +66] plant in a canopy	angle degree
	Day	Julian day	both	[1; 365]	no unit
Model precision	Xmax	Field sample size in the East-West direction	SP	[1; 4]	m
	Ymax	Field sample size in the North- South direction	SP	[1; 4]	m
	Voxel	Voxel edge size	SP	[1; 20]	cm
Target-plant variables	Height	Plant height	both	[1; 250]	cm
	Width	Plant width	both	[1; 200]	cm
	LA	Total plant leaf area	both	[1; 10 ⁵]	cm ²
	k	Species radiation extinction coefficient	both	[0.01; 1.1]	no unit
	RH50	Relative median leaf height below which is located half of the leaf area	both	[0.01; 1]	cm·cm ⁻¹
	b	Shape parameter for leaf distribution vs. plant height	both	[0.01; 6]	no unit
Neighbour mean plant variables	Density	Total plant density of the disc of plants (crops + weeds), including the target plant	PIC	[0.1; 3000]	plant.m ⁻²
	Distance to neighbour	Distance of the target plant to the furthest neighbour	PIC	[0.1; 3]	m
	Height	Plant height averaged over all neighbours and weighted by the inverse of distance to target plant	PIC	[0; 240]	cm
	Cover	Plant base area (superposed plants are added to the value) averaged over all neighbours and weighted by the inverse of distance to target plant	PIC	[0; 20000]	cm ²
	LA	Plant leaf area averaged over all neighbours and weighted by the inverse of distance to target plant	PIC	[0; 100000]	cm ²
	k	Species extinction coefficient averaged over all neighbours and weighted by the inverse of distance to target plant	PIC	[0; 0.7]	no unit
	RH50	Plants relative height averaged over all neighbours and weighted by the inverse of distance to target plant	PIC	[0; 115]	cm

931 [§] Input used in the "Single Plant" step (SP), the "Plant Inside a Canopy" step (PIC) or both
 932

933 B. Outputs (for target plant)

Use for FLORSYS	Name	Short explanation	Step [§]	Range of variation	Unit
Growth submodel	PARaP	Proportion of PAR ^{&} absorbed by the plant at the plant scale compared to the PAR above canopy.	both	[0; 1]	MJ cm ⁻² MJ ⁻¹ cm ² plant ⁻¹
	PARaC	Proportion of PAR absorbed by the plant for 1 cm ³ compared to the PAR above canopy	both	[0; 1]	MJ cm ⁻² MJ ⁻¹ cm ² cm ⁻³
Morphology submodel	SID	Daily Shading Intensity , <i>i.e.</i> proportion of incident radiation above canopy that does not reach the plant	both	[0; 1]	MJ MJ ⁻¹
Herbicide treatment submodel	rPARI _{planta}	Proportion of radiation intercepted by the plant relative to incident radiation above canopy	both	[0; 1]	MJ.cm ⁻² MJ ⁻¹ cm ²
	rPARI _{top}	Proportion of radiation intercepted by the top of the plant relative to incident radiation above canopy	PIC	[0; 1]	MJ.cm ⁻² MJ ⁻¹ cm ²
	rPARI _{base}	Proportion of radiation intercepted by the base of the plant relative to incident radiation above canopy	both	[0; 1]	MJ.cm ⁻² MJ ⁻¹ cm ²

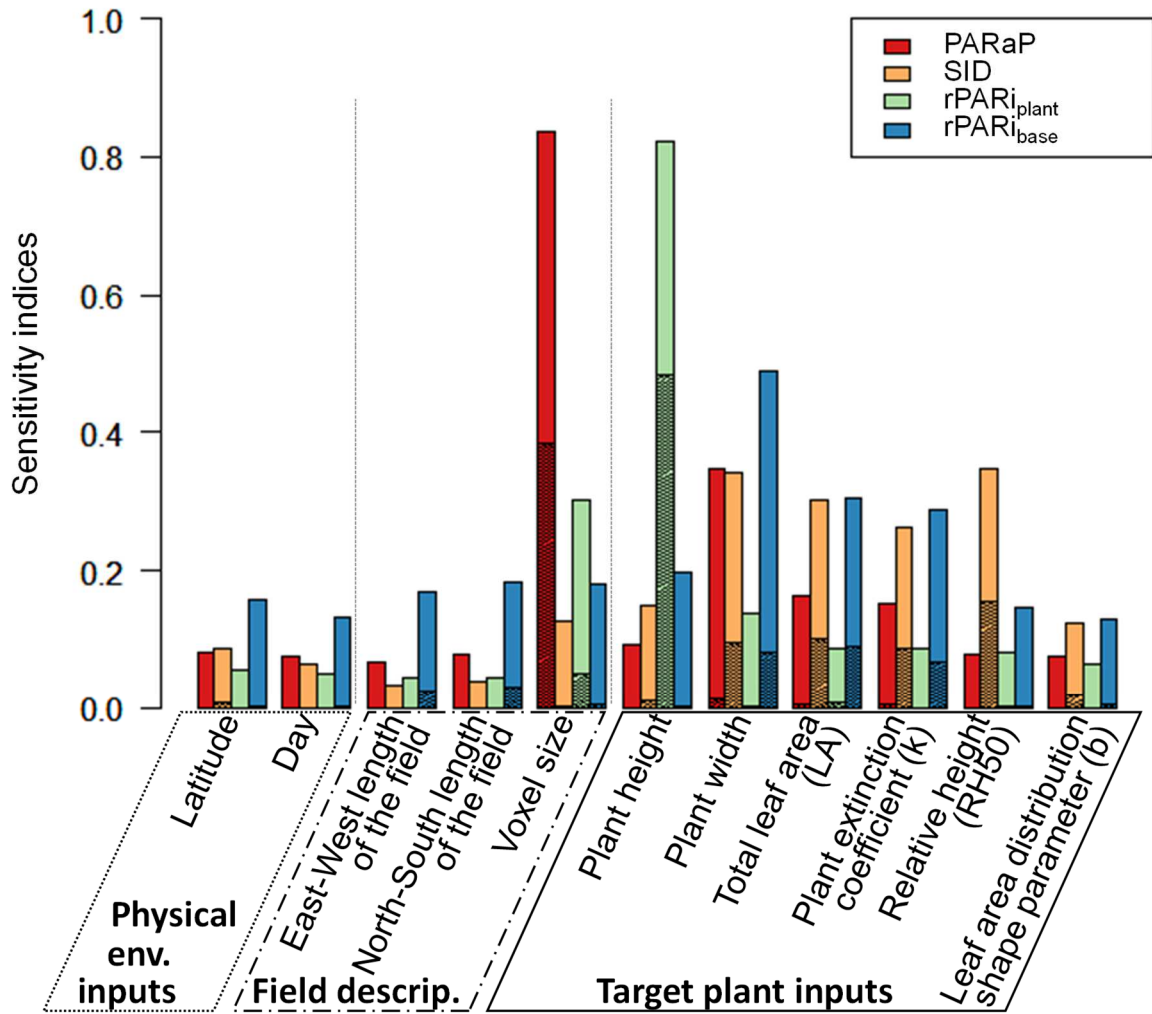
934 [&] PAR: Photosynthetically Active Radiation; [§] Output computed for the "Single Plant" step (SP), the
 935 "Plant Inside a Canopy" step (PIC) or both
 936



938

939 Figure 2: Schematic representation of the inputs and outputs of the 3D radiation interception submodel
 940 of FLORSYS (MUNIER-JOLAIN ET AL., 2013), with environmental and precision inputs (underlined),
 941 *plant in a canopy inputs* (italics), single plant common inputs (standard font) and outputs (bold). For
 942 abbreviations, see Table 2. (Floriane Colas © 2017 updated from (Gauchi et al., 2017))

943



944

945 Figure 3 : Overall view of sensitivity indices for radiation interception outputs of a target plant, in the
 946 absence of any shading neighbour plants. In hatched colours polynomial effects (*i.e.* disregarding
 947 interactions), in plain colours total effect (including interactions) of the inputs, environmental and
 948 precision inputs (underlined) and single plant input (normal font). The outputs are the Photosynthetic
 949 Active Radiation (PAR) absorbed by the target plant (PARaP), shading index (SID), relative PAR
 950 intercepted by the whole plant ($rPARi_{plant}$) or on soil surface ($rPARi_{base}$). (Floriane Colas © 2018)

951

952

953 Table 3: Synthesis of the different 3D radiation interception metamodels (fast and full) computed *via*
 954 polynomial chaos expansion (PCE) and Partial Least Squares (PLS) regression. Fast metamodels
 955 result from full metamodels via a LASSO-PLS monomials selection.

956 A. Single target plant without shading neighbouring plants

	Radiation- interception model output	Metamodel type	Polynomial degree	Monomial number	Fitting prediction Q2cum	Prediction error RMSEP [§]
[1]	PARaC	full	5	4367	0.96	0.19
[2]	PARaC	fast	5	26	0.85	0.39
[3]	SID	full	5	4367	0.98	0.15
[4]	SID	fast	5	26	0.82	0.43
[5]	rPARI _{plant}	full	5	4367	0.95	0.22
[6]	rPARI _{plant}	fast	5	27	0.90	0.32
[7]	rPARI _{base}	full	5	4367	0.93	0.25
[8]	rPARI _{base}	fast	5	25	0.70	0.55

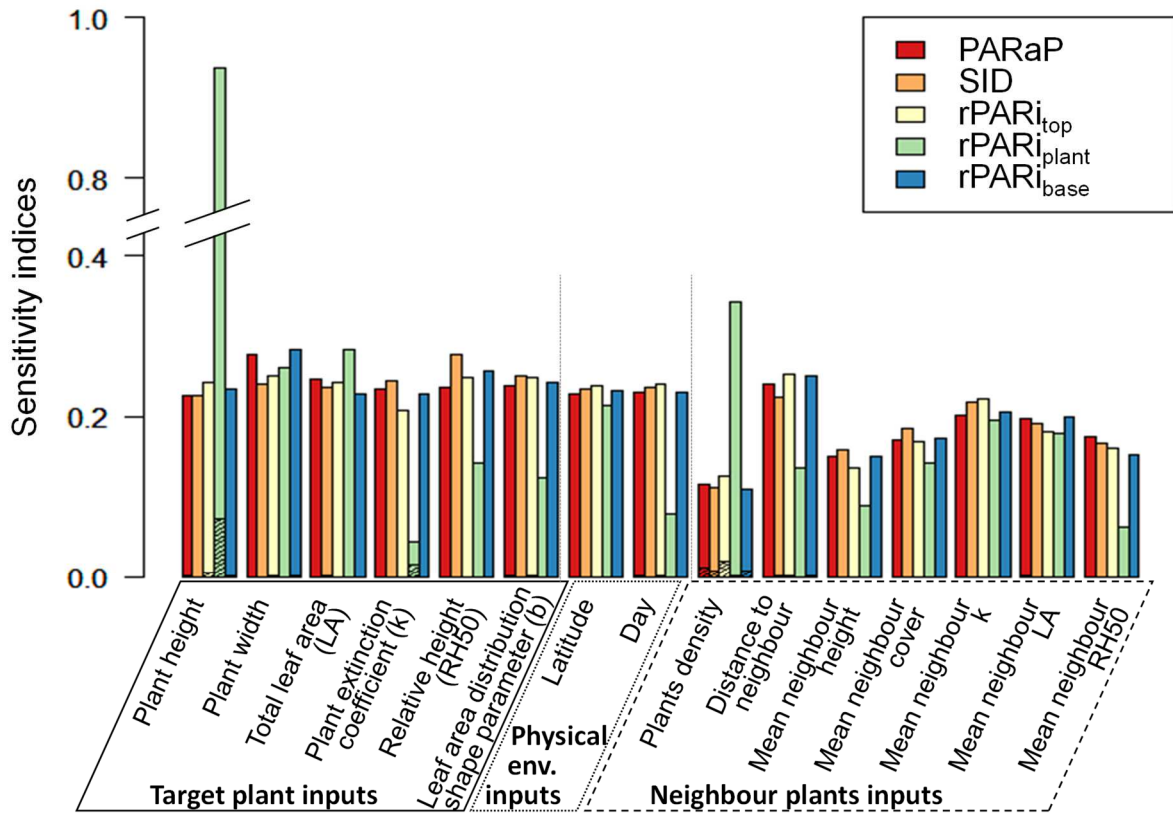
957

958 B. Target plant inside a canopy

	Radiation- interception model output	Metamodel type	Polynomial degree	Monomial number	Fitting prediction Q2cum	Prediction error RMSEP [§]
[9]	PARaC	full	4	3875	0.83	0.33
[10]	PARaC	fast	5	30	0.56	0.65
[11]	SID	full	4	3875	0.75	0.42
[12]	SID	fast	5	29	0.30	0.83
[13]	rPARI _{top}	full	4	3875	0.71	0.48
[14]	rPARI _{top}	fast	5	28	0.27	0.85
[15]	rPARI _{plant}	full	7	4000	0.82	0.36
[16]	rPARI _{plant}	fast	5	35	0.52	0.69
[17]	rPARI _{base}	full	4	3875	0.76	0.43
[18]	rPARI _{base}	fast	5	35	0.37	0.79

959 [§] root mean squared error predictor

960



961

962 Figure 4: Overall view of sensitivity indices for radiation interception outputs of a target plant
 963 surrounded by neighbour plants. Total effects (plain colours) and polynomial effects (*i.e.* disregarding
 964 interactions, hatched colours) of inputs of the FLORSYS radiation interception submodel. The outputs
 965 are the Photosynthetic Active Radiation (PAR) absorbed by the target plant (PARaP), shading index
 966 (SID), relative PAR intercepted at the summit of the target plant (rPARI_{top}), by the whole plant
 967 (rPARI_{plant}) or at the base of the target plant (rPARI_{base}). (Floriane Colas © 2017)

968

969

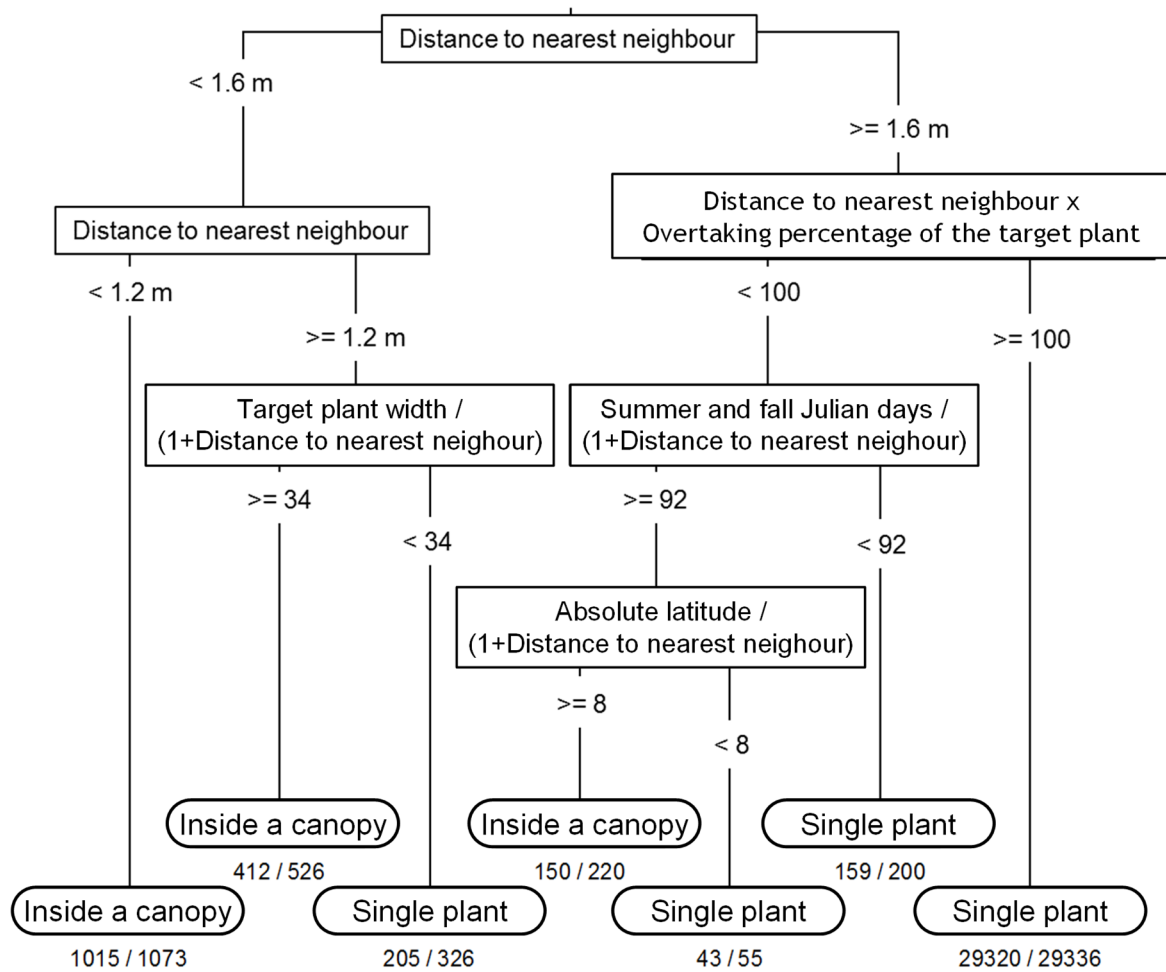
970 Table 4: Synthesis of the variation in prediction error in simulations with the metamodelled vs.
 971 process-based model. Relative root mean squared error predictor (RRMSEP) in relation to variation
 972 range of observation (max-min)/2.

Output	Species scale	Time step	Type of neighbours used for calculating canopy variables		
			Local	Mixed	Average
Weed density (plants·m ⁻²)	By species	Day	+9%	++% ^s	+10%
		Multiannual	-81%	-7%	-52%
	Sum of all species	Day	+9%	++% ^s	-85%
		Multiannual	-50%	-8%	+152%
Weed biomass (g·m ⁻²)	By species	Day	+294%	+417%	+580%
		Multiannual	++% ^s for process-based model		
	Sum of all species	Day	++% ^s	+327%	+723%
		Multiannual	+1351%	+10353%	+12391%
Seedbank (seeds·m ⁻²)	By species	Day	+164%	+163%	+84%
	Sum of all species	Day	++% ^s for process-based model		
Crop yield (T·ha ⁻¹)	By species	Day	+61%	+6%	79%

973 ^s RRMSEP of metamodelled simulation was >> 0 and RRMSEP of process-based simulation was <
 974 variability in observations, *i.e.* ~0, and no relative variation in RRMSEP could be calculated

975

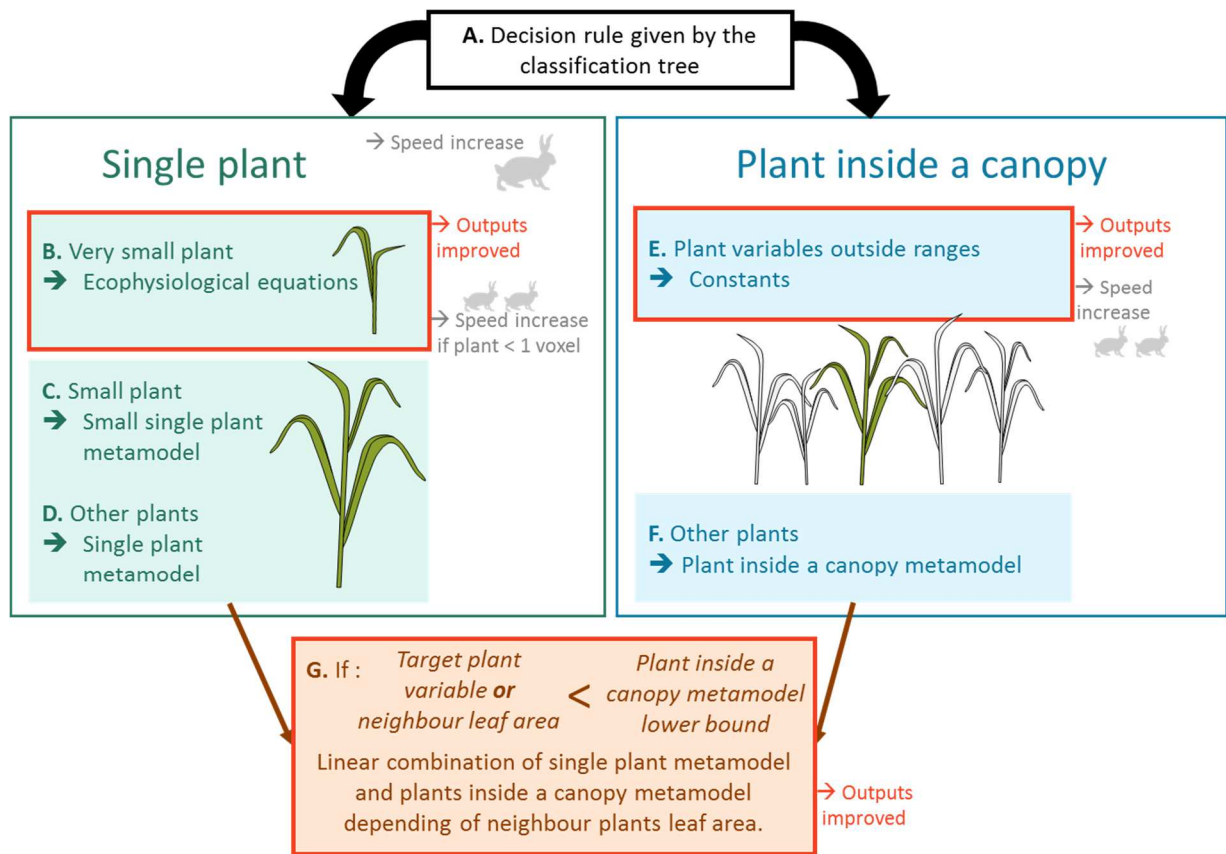
976



977

978 Figure 5: Classification tree (CART) to decide whether a target plant is single or inside a canopy. The
 979 segmentation is based on relative photosynthetically active radiation on target-plant top $rPAR_{top} >$
 980 0.98. The adjustment error (or training error) was 0.24, the cross validation error was 0.28 (standard
 981 deviation = 0.01). (Floriane Colas © 2017)

982

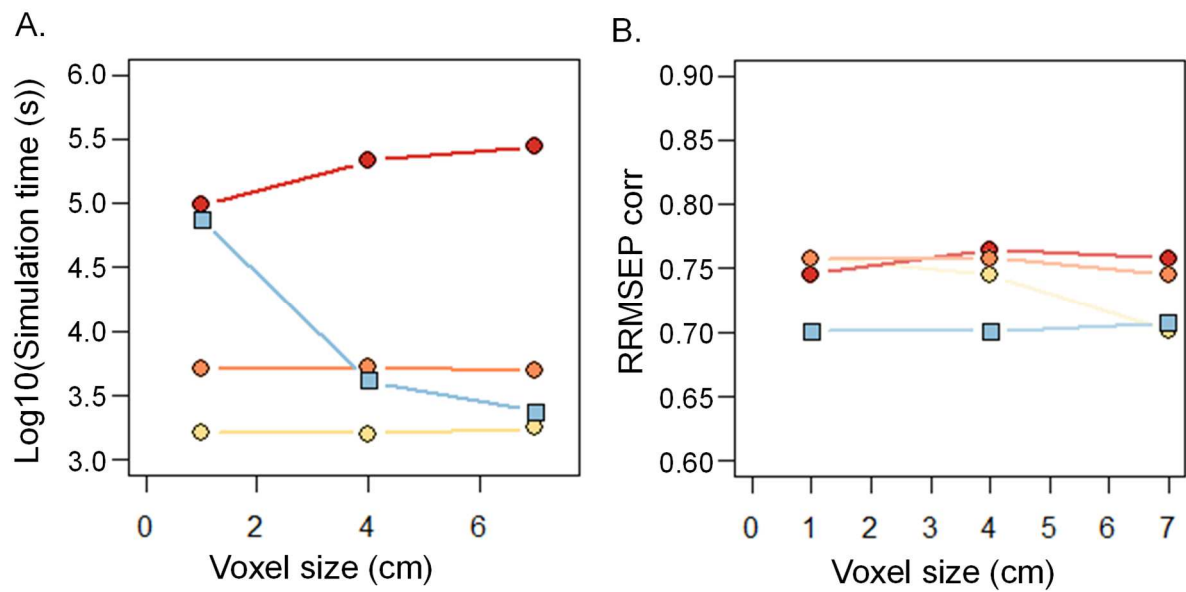


983

984 Figure 6: The different metamodels and when they are used in FLORSYS-ML depending on target plant
 985 variables, neighbour plant variables and environmental variables. (Floriane Colas © 2017)

986

987



988

989 Figure 7: Simulation time (A) and prediction error (relative root mean squared error predictor
990 RRMSEP, B) of the daily weed seedbank by species for the different FLORSYS versions (squares:
991 process-based, circles: metamodelled FLORSYS-ML), neighbour-aggregating methods (dark red: local
992 neighbours, light yellow: average, orange: mixed) and voxel edge sizes. Relative error in relation to
993 variation range of observation (max-min)/2 (Floriane Colas © 2017)

994

995 Table 5 : Synthesis table to guide the choice of the best simulation method with the smaller prediction error (relative root mean squared error predictor
 996 RRMSEP) depending on the goal and the target output.

Output	Species scale	Time step	Simulation goal	
			Farmer's workshops (fast simulations: 7 cm voxel, 6x3 m ² field)	Site-specific weed management (precise simulations : 1 cm voxel, 6x3 m ² field)
Weed density (plants·m ⁻²)	By species	Day	Process-based	Process-based #
		Multiannual	Metamodelled with average neighbours	Metamodelled with average neighbours *
	Sum	Day	Process-based	Process-based #
		Multiannual	Process-based	Process-based
Weed biomass (g·m ⁻²)	By species	Day	Process-based	Metamodelled with local + average &
		Multiannual	Metamodelled with local neighbours	Metamodelled with local + average neighbours
	Sum	Day	Process-based	Metamodelled with local + average &
		Multiannual	Metamodelled with local neighbours	Metamodelled with local + average neighbours
Seedbank (seeds·m ⁻²)	By species	Day	Process-based *	Process-based
	Sum	Day	Metamodelled with local + average &	Metamodelled with local neighbours
Crop yield (T·ha ⁻¹)	By species	Day	Process-based	Process-based

997 Other methods that are also close in the RRMSEP value: * all of the other methods; & metamodel with average neighbour; # metamodel with local neighbours

998

1000 Table 6 : Summary of avenues for speeding up complex models such as FLORSYS

Method	Advantage	Disadvantage	Best use for...	
Model modification				
1	Simpler light interception models	Simpler process-based	Cannot represent heterogeneous crop canopies	Homogeneous (single-species) canopies, uniform field management
2	Metamodel	Simpler light interception, identification of key factors for light interception	Lose connection of processes, cannot wholly grasp the complexity of plant-plant interactions	Precision agriculture, very heterogeneous multispecies canopies,
Input choice				
3	Increase voxel size, decrease field-sample area	No source modification needed	Precision loss	Canopies with little heterogeneity and few species, uniform field management
Technical solutions				
4	Parallel source processing, graphical processing units	Remain process-based	Not "portable" to all computers	Research, all cropping systems
5	Powerful computers, calculation servers	No source modification needed	Not accessible to non-researchers	Research, all cropping systems

1001

1002