JOURNAL OF THE ROYAL SOCIETY Interface

Towards a simplification of models using regression trees

Y. Eynaud, D. Nerini, M. Baklouti and J.-C. Poggiale

J. R. Soc. Interface 2013 **10**, 20120613, published 21 November 2012

JOURNAL OF THE ROYAL SOCIETY Interface

rsif.royalsocietypublishing.org

Research

Cite this article: Eynaud Y, Nerini D, Baklouti M, Poggiale J-C. 2012 Towards a simplification of models using regression trees. J R Soc Interface 10: 20120613.
http://dx.doi.org/10.1098/rsif.2012.0613

Received: 31 July 2012
Accepted: 31 October 2012

Subject Areas:
environmental science, computational biology

Keywords:
model complexity reduction, regression trees, parameter estimation, cost function, criteria of model performance

Author for correspondence:
Y. Eynaud
e-mail: yoan.eynaud@univ-amu.fr

Towards a simplification of models using regression trees

Y. Eynaud, D. Nerini, M. Baklouti and J.-C. Poggiale

Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, Mediterranean Institute of Oceanography (MIO), UM 110, 13288 Marseille, Cedex 09, France

Over-parametrization in modelling is a well-known issue that makes it hard to identify which part of a model is responsible for a given behaviour. In line with that ascertainment, this work presents the outline of an empirical method to simplify models by decreasing the number of parameters. By using regression trees to classify outputs according to related input parameters, the method provides the modeller with an objective tool to reduce the range of the used parameters and, under certain conditions, to establish relations between them. Thereby, the complexity of the model is reduced on the basis of mathematical arguments. As an example, a dynamic energy budget-based model of a mesopelagic bacterial ecosystem is simplified using the presented method. The main benefits of such a method are thus highlighted: (i) more robust parameter estimations; (ii) less complex formulations; and (iii) fewer modelling assumptions. To conclude, the difficulties encountered are discussed, and several solutions are proposed to deal with them.

1. Introduction

In order to increase their ability to represent various situations, mathematical models in environmental sciences nowadays integrate an increasing number of variables and parameters. Because they describe physical, chemical and biological processes at different time and space scales, such models are at the interface of different disciplines. Consequently, they often result from a consensus among the various specialists involved in their development [1]. Such a consensual nature generally explains this increasing number of parameters and state variables, because each implied process is relevant from the specialist point of view.

Furthermore, because of the lack of general laws in environmental sciences, the mathematical formulation of a given process is generally an open problem. Empirical formulations are obtained by fitting data with an *a priori* chosen mathematical formula. Phenomenological models aim to reproduce a given phenomenon with the simplest mathematical expressions. Mechanistic formulations are supposed to integrate the mechanisms underlying the modelled processes. Even if each approach has its own set of interests, which depends on the questions that the model aims to address, mechanistic formulations provide a more robust approach in *conceptual modelling* (as defined in [2]), because they are built on assumptions based on a theorization of the system.

Nevertheless, the recurrent use of complex mechanistic formulations has led some modellers to be lured by *the unreasonable effectiveness of mathematics* [3], namely in situations where accuracy of the model results is rather due to its mathematical flexibility than to a good representation of the system. Indeed, if it appears clear that mechanistic formulations have given more sense to parameters, they are still constrained by assumptions that explain why several formulations can be relevant to represent the same process. Moreover, two formulations of the same process can give the model very different behaviours [4–6] and thus the choice is not always obvious. Finally, the relation between the mechanistic approach and the efficiency of models to reproduce observations may not be straightforward [7].

Royal Society Publishing
Informing the science of the future

A difficulty associated with the large number of parameters in models concerns over-parametrization [8,9]. Among other problems, it makes it difficult to identify which modelled processes are really responsible for the observed results. Indeed, when mathematical tractability is lost, the modeller begins to use a model with an unknown dynamical behaviour, a kind of mathematical black-box. There is no guarantee that the results are due to a good representation of the system: a sufficiently complex model could easily give expected outputs because of the flexibility of the model induced by the large number of parameters [3]. Finally, among other phenomena, the parameter redundancy (as defined in [10]) may lead to the difficulty to estimate parameters adequately, and being able to identify useless parameters is definitely a hard task [11].

As a consequence, it appears that when it is possible, the simplification of models may provide an interesting issue. Simplification methods generally aim to reduce complexity. However, according to the nature and the structure of the models, existing simplification methods [12–18] cannot always be directly used. Also, only some of those methodologies [13,15] allow a better understanding of the model properties through the simplification process. In order to simplify, additional assumptions may be useful. In some cases, these assumptions can directly be linked to the part of the model that can be simplified. However, in many cases, there is no direct link between assumptions made on the basis of data, observations or more generally stylized facts [19] and the way to simplify models. It would thus be interesting to develop methods that specifically simplify a model according to data, or more globally, to the purposes of the modeller, and that also enlighten some of the model inner properties. Following this objective, the need of a tool capable of identifying how the outputs of a model are linked to its input parameters appears. One goal is to identify which parameters actually drive the model to reach some specific outputs. The bifurcation theory is dedicated to this type of problem, but is nowadays impractical with a high number of parameters. Thus, in this work, we aim at simplifying non-mathematically tractable models by using a statistical approach capable of outlining a pattern of outputs relative to inputs.

Such an empirical approach, extremely dependent on the numerical realizations of the model, requires the use of an objective statistical tool. Different methods along this line have been proposed. In Raick *et al.* [13], for instance, the authors conclude that principal component analysis may be used to simplify models but with constraining conditions when nonlinear relationships occur. The approximation Bayesian computation has been used by Toni *et al.* [20,21] to develop a model selection method. By sampling the parameter space, they outline the output distribution of each compared model and then, using a Bayes criterion, define which of the compared models is the more appropriate one. However, a set of models has to be defined *a priori*, and the best model is not produced by the method itself. Thus, this work is of great help when comparing different models with the same purpose, but it does not produce a simpler and more appropriate model by itself.

In the present paper, we propose a method that also uses an outline of the outputs distribution by sampling the parameter space. Nevertheless, we now propose to classify the coupled distribution of the parameters and variables using regression trees [22]. Although the primary objective of that statistical method is to predict a response variable using a set of explanatory variables, regression trees will be used here to highlight the impact of parameters on the output distribution. Using that statistical tool in another context has already been done by Pappenberger *et al.* [23] to introduce an interesting parameter sensitivity analysis approach. Thus, we use here only the classification capabilities provided by the regression trees. In our case, the set of state variables of the model forms the response variable, and the explanatory variables are formed by the set of parameters. Note that here, we use a multivariate extension of trees because of the multivariate structure of the response variable.

A classification obtained using regression trees is in the form of a binary tree containing several final classes. Each of those final classes associates a specific part of the state variable space to a specific part of the parameter one. It is likely that at least one of those final classes of outputs reaches the modeller's expectations. Thereby, by using this association, we should be able to establish mathematical relations between parameters within the parameter space associated with this given subset of outputs and so define a new model, with less parameters, for this subset. We underline the fact that the classification process will not require any data, because the sampling of the output space will be done using randomly generated parameter sets.

The paper is organized as follows. In the next section, we present the general approach, on the basis of the regression trees method. We first describe how far we use this statistical method in our simplification approach. In the following section, we apply the method on an example to illustrate the way it functions and the results that can be obtained. Then we discuss the different problems that may occur when building the tree used for the simplification results, the interest of the simplification and the limits of the method. Finally, we conclude and provide some perspectives to extend the method to more general situations.

2. General description of the method

2.1. On the use of regression trees

Let M be a model that has m parameters and n state variables. We assume for the sake of simplicity that the model M admits an equilibrium vector $X = (x_1, \ldots, x_i, \ldots, x_n) \in \mathbb{R}^n$. We define M_∞ the map that associates the equilibrium X (the output) to the parameter vector $\Pi = (\pi_1, \ldots, \pi_i, \ldots, \pi_m) \in \mathbb{R}^m$ (the input) for a given initial condition $X_0 \in \mathbb{R}^n$. Thus, we have

$$\begin{aligned} M_\infty : \Lambda \times \{X_0\} &\to \Omega, \\ (\Pi, X_0) &\mapsto X = M_\infty(\Pi, X_0). \end{aligned}$$

The whole parameter domain $\Lambda = \Lambda_1 \times \cdots \times \Lambda_m$ forms an orthotope belonging to $\mathbb{R}^m$. Each parameter π_i takes realistic values in $\Lambda_i = [b_i^-, b_i^+]$, ranging from b_i^- to b_i^+. Those boundaries can be chosen from values found in the literature or directly from experimental data. We name $\Omega \subset \mathbb{R}^n$ the equilibrium state variable domain. Because of nonlinearities in the structure of M_∞, the set Ω is not necessarily an orthotope (figure 1). Moreover, it is likely that a part of the Ω domain is unrealistic from the studied system point of view. For example, when using an ecological model, admissible values of parameters can lead to negative biomasses, which

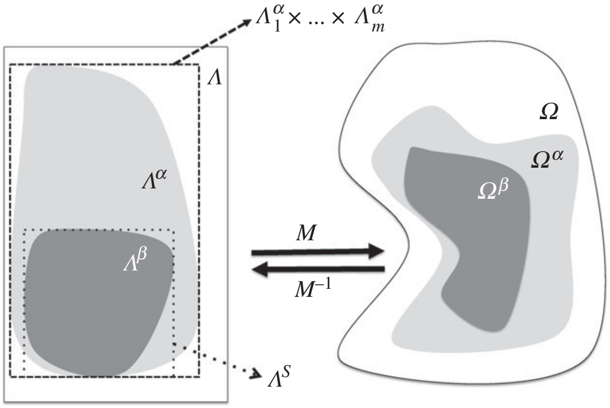

Figure 1. A qualitative representation of a model M with $n = m = 2$. We see that: $\Lambda^\beta \subset \Lambda^\alpha \subset \Lambda$ and that $\Omega^\beta \subset \Omega^\alpha \subset \Omega$.

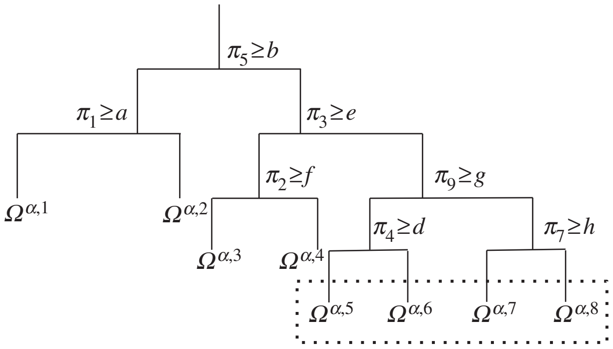

Figure 2. A seven split tree, with eight final classes. Note that $\bigcup_{i=1}^{8} \Omega^{\alpha,i} = \Omega^\alpha$, i.e. the sample of size r taken to build the tree that fulfils the α-condition. Here, the β-condition is fulfilled by $\bigcup_{i=5}^{8} \Omega^{\alpha,i} = \Omega^\beta$ (dotted rectangle).

is a nonsense. More precisely, there exists a condition that we will call the α-condition that delimits a part $\Omega^\alpha \subset \Omega$ that is coherent with basic properties of the system. Some stylized facts, as described in [19], can be used to define this α-condition. Consequently, each equilibrium $X \in \Omega^\alpha$ has been obtained from $\Pi \in \Lambda^\alpha$, the α-conditioned part of the parameter domain defined as

$$\Lambda^\alpha \times \{X_0\} = M_\infty^{-1}(\Omega^\alpha).$$

Note that $\Lambda^\alpha \subset \Lambda$. This set is not necessarily an orthotope, but $\Lambda_1^\alpha \times \cdots \times \Lambda_m^\alpha$ is the smallest orthotope in which Λ^α is contained, as shown in figure 1.

Suppose now that only a part of the equilibrium $X \in \Omega^\alpha$ fits the purposes of the modeller. Consider for instance a situation where the model outputs X have to match some experimental data within some given tolerance. That specific part of the outputs can be classified as *behaviour*, a concept coined by [24] or as an *admissible region* as defined in [25]. In our case, there exists a condition that we will call the β-condition, which delimits a part $\Omega^\beta \subset \Omega^\alpha$ that matches the purposes of the modeller (figure 1).

Similar to the α-condition, each equilibrium $X \in \Omega^\beta$ has been obtained from $\Pi \in \Lambda^\beta$, the β-conditioned part of the parameter domain. In the following, we wish to use the properties of the parameter vectors belonging to Λ^β to create a model with less parameters, and thus simpler.

2.2. Construction of the regression tree

In order to study the relation between Ω^β and Λ^β, we start by randomly sampling the space of parameters Λ and use the map M_∞ to associate an output X to an input Π, for a given X_0. As discussed before, some values of X are not realistic. We keep only the α-conditioned part of the random sample to form the sample $\Phi = \{(\Pi_k; X_k), k = 1, \ldots, r\}$ of size r, formed with the parameter vectors $\Pi_k = (\pi_1^k, \ldots, \pi_j^k, \ldots, \pi_m^k)$, $\pi_j^k \in \Lambda_j^\alpha$ and its associated output vectors $X_k \in \Omega^\alpha$.

We now define how regression trees are built and why they are of interest for simplification purposes. Regression trees are statistical models concerned with the prediction of a response variable using a set of explanatory variables. As said in the introduction, we focus only on the classification capabilities provided by the tree structured model. Here the response variable is $X \in \Omega^\alpha$. The explanatory variables are constituted with the vector of parameters $\Pi \in \Lambda^\alpha$.

Starting with sample Φ, a multivariate regression tree is constructed by recursive partitioning of the parameter space Λ, which determines subsets (classes) of Ω for which the values of X are the most homogeneous. The homogeneity of each class is measured using a deviance criterion. The formulation of the deviance has to be chosen relative to the nature of the outputs.

A multivariate regression tree is constructed by iteratively splitting the classes in order to maximize the decrease of the deviance. See appendix A for an explanation of the splitting procedure used here. The tree is grown until a quantitative termination criterion is reached. At the end of the construction, we obtain a binary tree containing q terminal classes, and a nested sequence of splitting conditions appears on the predictive variables.

Our goal here is to use this conditional classification given by the tree to enlighten links between the state variable and the parameter space.

The structure of the obtained tree may not properly represent the distribution of Ω^α. Indeed, two phenomena can appear: first, if another tree is obtained by changing the size of the sample, r, the tree is not robust. Second, if a different tree is obtained with another sample of size r, the tree is unstable. The stability of a given tree is directly linked to the nature of the analysed outputs distribution, whereas robustness is associated with the representativeness of the sample itself, and thus the value of r. Breiman *et al.* [22] advise the use of a sample of size $r = 10^n$, with n the dimension of the state variable space, to obtain a robust tree.

2.3. Analysis of the regression tree

A tree is built using a sample (discrete form) of a domain (continuous form). Hereafter, we will directly refer to the domains Ω and Λ, even though we are dealing with samples.

Let us take the example of a tree obtained from a model with $n = 6$ variables and $m = 10$ parameters, as shown in figure 2, where eight final classes have been obtained. In this example, we assume that the β-condition is fulfilled by the classes $\Omega^\beta = \bigcup_{i=5}^{8} \Omega^{\alpha,i}$. The way to define the β-situation will be discussed within the practical approach. Here, the system has 10 parameters, and the β-conditioned part of the outputs is reached after three splits of the parameter space, more precisely, after having applied restrictions on three parameter ranges $(\Lambda_5^\alpha, \Lambda_3^\alpha, \Lambda_9^\alpha)$. Those restrictions

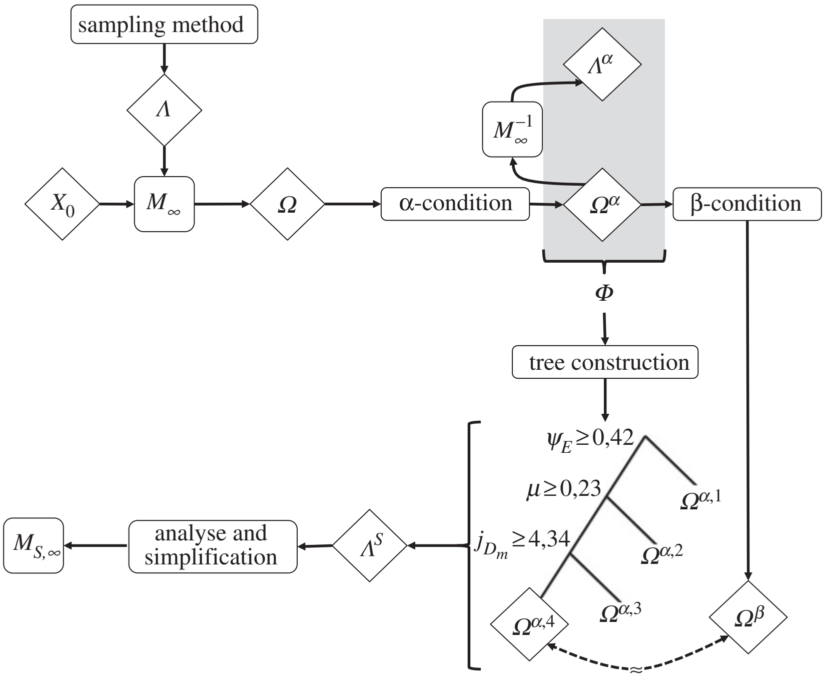

Figure 3. General scheme of the simplification method. (i) The output distribution is sampled using a large number of parameter sets. (ii) Two parts of that domain are defined, the α- and the β-conditioned one. Whereas the first is defined using stylized facts, the second is based on the observations or/and data that the model aims to reproduce. (iii) The α-conditioned domain is classified through a multivariate regression tree until the β-conditioned part of it is efficiently isolated. (iv) The β-conditioned part of the parameter domain Λ^S is then analysed to enlighten the possible links existing between parameters. (v) When it is possible, those links are used to create new parameters which decreases the model complexity: a simpler model with less parameters, M_S, is created.

define a new parameter domain, denoted Λ^S, which is the smallest m-orthotope containing Λ^β (figure 1):

$$\Lambda^S = \prod_{j=1}^{m} \Lambda_j^\beta, \quad \text{with } \Lambda^\beta \subset \Lambda^S \subset \Lambda.$$

Under those restrictions, values for the seven other parameters can therefore be freely chosen within their respective α-conditioned ranges.

Thus, on the basis of the results given by the tree, we may simplify the model. Indeed by conveniently choosing the value for the parameters kept free by the tree regression, i.e. within Λ^S, we can end up with a simpler model. By fixing relations between parameters implied in the same type of processes, we will decrease the number of parameters. Two cases can be distinguished: (i) the general case in which we define one parameter as a function of another, $\pi_{j_1} = f(\pi_{j_2}), j_* \in [1, m]$; (ii) the particular case in which we define that two parameters are replaced by a new one, $\pi_{j_s} = \pi_{j_1} = \pi_{j_2 \neq j_1}, j_* \in [1, m]$. Again, this freedom of choice highly depends on the respect of the restricted range for the parameter enlightened by the tree. For example, owing to the parameter range restriction provided by the tree, we have $\Lambda_5^\alpha = [0, h]$ and $\Lambda_5^\beta = [b, h]$.

After simplification, we get a model M_S and using the same type of notations as before, we have

$$M_{S,\infty} : \Lambda^{S_1} \times \{X_0\} \rightarrow \Omega^{\beta_S}, \; X_0 \in \mathbb{R}^n,$$

where

$$\Lambda^{S_1} = \prod_{j=1}^{m_s} \Lambda_j^\beta, \quad \text{with } \Lambda^\beta \subset \Lambda^S \subset \Lambda$$

with Λ^{S_1} the parameter domain of M_S, X_0 the initial vector, m_s the number of parameters and Ω^{β_S} the state variable output.

Defining Λ^{S_1} as a Cartesian product is a requirement because we need to associate with each parameter a continuous range of values to maintain the exportability of the model. The relevance of such an approximation is dependent on the size of $\Lambda^\beta \cap \Lambda^S$; this remark will be discussed further. Similarly, the validity of the chosen simplification depends on the size of $\Omega^{\beta_S} \cap \Omega^\beta$. Roughly, the method follows the steps summarized in figure 3.

3. An example: the mesopelagic layer in marine systems

It is not rare in theoretical ecology to encounter numerical and computational issues. By giving an example of how to simplify a model using our method, we will try to show the commonly encountered problems and how to avoid, or at least minimize, them.

3.1. A model to simplify

As an example, the method has been applied to simplify a model of the mesopelagic ecosystem. This ecosystem plays a strong role in the global carbon cycle because it is where a large part of biomineralization made by bacteria occurs [26]. However, a lot of questions concerning the carbon cycle in this ecosystem are still unanswered [27]. Here, we aim at understanding the bacterial contribution to the carbon cycle through dissolved (DOC) and particulate (POC) organic carbon uptakes. We have chosen to represent such a system following the dynamic energy budget (DEB) theory [28,29]. DEB theory describes the rates at which an organism acquires resources from the environment and subsequently uses the nutrients and energy for production and maintenance. Thus, this theory gives us a framework that, in addition to being general and easily adaptable, fits our interest for bacterial metabolism. Indeed, it has given good results in representing physiological behaviour of prokariotic

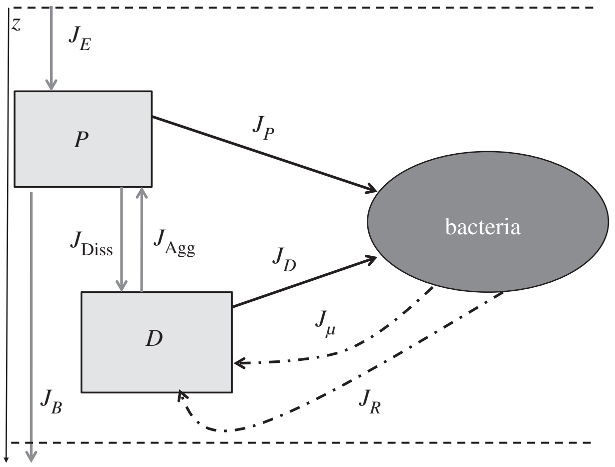

Figure 4. Schematic of M. This four-dimensional model represents a three-compartment trophic web. Two substrates, P and D, are both assimilated by bacteria. That latter compartment is divided in two types of biomass: the structure, M_V, and the reserve, $M_E = m_E M_V$. That biomass type distinction is made following the DEB theory [28,29]. Flux and compartment definitions are given in table 1.

organisms [5,30–32]. We here work at the scale of a bacterial cell, where only three trophic compartments can be identified: the particulate matter itself, the dissolved matter and the bacteria compartment. The model is not spatialized; thus, the particulate matter enters and exits the system through two different fluxes. At such a scale, data are very hard to acquire and are usually extrapolated from larger scales. Nevertheless, qualifying which process is primordial at the very beginning of the bacterial carbon remineralization in a deep marine ecosystem is mandatory. Thus, the model used here is very conceptual and our result analysis will be qualitative. Thereby, we focus on bacteria following a DEB model, called M, representing the dynamics of a three-compartment trophic web as represented in figure 4. The dynamics of the carbon mass within the system is given by the following system of ODEs:

$$\left.\begin{aligned} \frac{\mathrm{d}P}{\mathrm{d}t} &= \psi_E M_\varphi - \psi_M P + \alpha_{P,D} D - \alpha_{D,P} P - \frac{j_{Pm} P}{P + K_P} M_V, \\ \frac{\mathrm{d}D}{\mathrm{d}t} &= \mu M_B + (1 - y_A)\frac{j_{Pm} P}{P + K_P} M_V - \frac{j_{Dm} D}{D + K_D} M_V \\ &\quad - \alpha_{P,D} D + \alpha_{D,P} P, \\ \frac{\mathrm{d}M_V}{\mathrm{d}t} &= \frac{k m_E - j_M}{m_E + \frac{1}{y_{V,E}}} M_V - \mu M_V \\ \text{and} \quad \frac{\mathrm{d}m_E}{\mathrm{d}t} &= \frac{j_{Dm} D}{D + K_D} + y_A \frac{j_{Pm} P}{P + K_P} - k m_E - r_B m_E - \mu m_E. \end{aligned}\right\} \tag{3.1}$$

Parameter ranges and definitions are given in tables 1 and 2. Other compartments could have been added to the model if we had chosen to work on a larger scale. However, because our purpose is the understanding of the remineralization of carbon (i.e. the use of P and D) by bacteria, such a simple model appears to be a good first step towards that objective.

3.2. Building the tree

In this example, it is considered that bacteria, dissolved and particular organic carbon coexist, which means that state variables representing them have to keep strictly positive values: it is the α-condition of the system. We are not looking at the dynamics of the system but rather at its equilibrium

Table 1. Compartment and flux definitions for M, M_S, M_{S^2}.

symbol	interpretation	unit
M_V	bacteria structure biomass	mmolC m^{-3}
m_E	bacteria reserve density	mmolC mmolC^{-1} m^{-3}
D	dissolved organic carbon density	mmolC m^{-3}
P	particular organic carbon density	mmolC m^{-3}
J_E	flux of P entering the system	mmolC m^{-3} d^{-1}
J_B	flux of P leaving the system	mmolC m^{-3} d^{-1}
J_{Diss}	flux of D from P dissolution	mmolC m^{-3} d^{-1}
J_{Agg}	flux of P from D aggregation	mmolC m^{-3}d^{-1}
J_D	flux of D assimilated by the bacteria compartment	mmolC m^{-3} d^{-1}
J_P	flux of P captured by the bacteria compartment	mmolC m^{-3} d^{-1}
J_μ	flux of D from the bacteria compartment mortality	mmolC m^{-3} d^{-1}
J_R	flux of D from P catabolization	mmolC m^{-3} d^{-1}

states for a given X_0, denoted:

$$X = (P_\infty, D_\infty, M_{V,\infty}, m_{E,\infty}).$$

Since the state space has four dimensions, we should need a minimum of $r = 10^4$ equilibrium points fulfilling the α-condition: $\forall i \in [1, r], X_{i,\infty} \in]0.01; 10[^4$.

That condition comes from field observations [33] that somehow characterize the mesopelagic ecosystem. To obtain robust results, and to study the stability of the tree, we first take a value of r much larger than recommended by Breiman *et al.* [22], which creates a finer grid in this four-dimensional space. Also, in order to enhance the heterogeneous distribution nature of the sample and exhibit the different classes more easily, we have chosen to use log-transformed outputs, $\log(\Omega^\alpha + 1)$, to build the tree. We have selected such a transformation function for its inner properties and monotonic nature. Any other transformation that respects the original distribution structure but leads the algorithm to be more efficient could be used. We then launch the binary regression tree algorithm on the modified sample. We have here used the R-part package of the software R64 [35,36]. The setting options are: (i) no final classes containing less than $r/30$ points; (ii) no split for class containing less than $r/10$ points; and (iii) no split if the deviance of the new buckets decreases less than 1 per cent.

Let us remember that those criteria are deeply important, because splitting parameters determine the shape of the tree. We obtain the tree presented in figure 5.

One of the obtained classes, $\Omega^{\alpha,4}$, presents an interesting distribution of the state variables: high and almost equal values for P and D, low values of M_V and high values for m_E. It is a meaningful pattern because it drives us to think of an ecosystem with a high level of resource but also a high level of *maintenance*, from a DEB point of view.

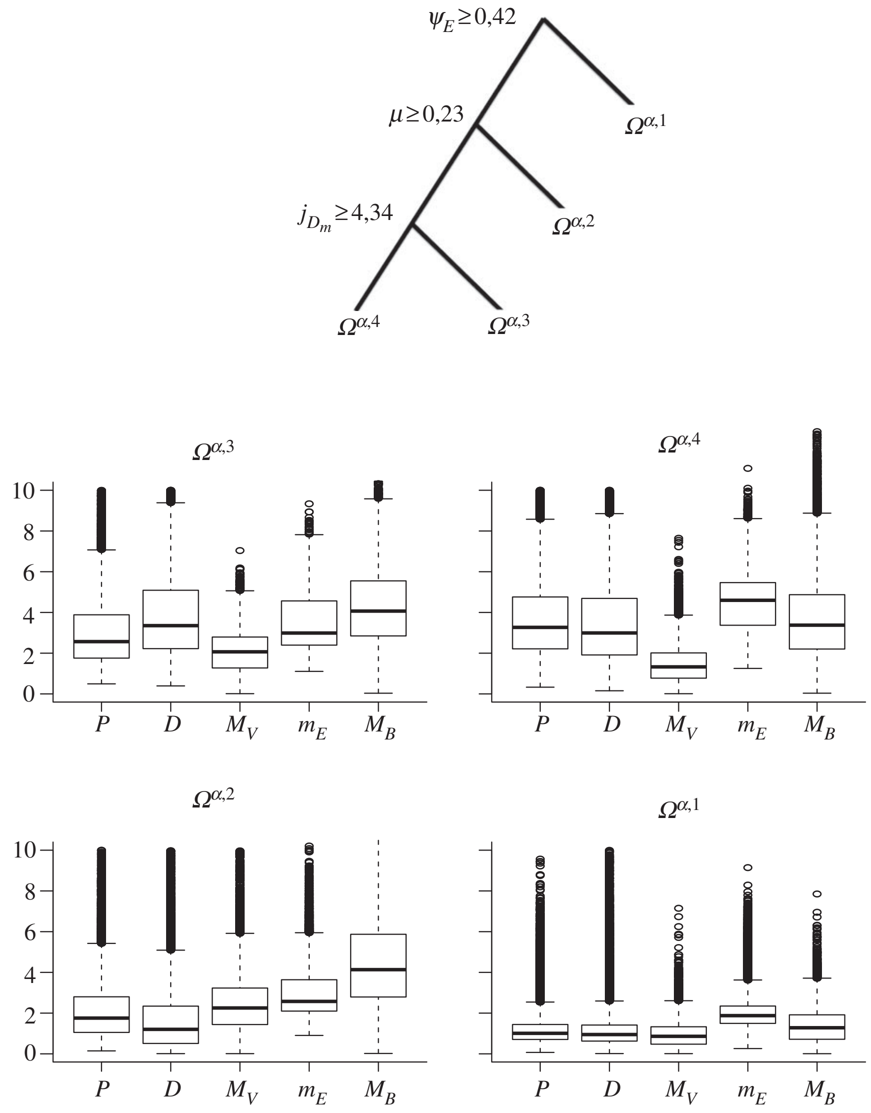

Figure 5. We obtain a tree with four final classes. For each class, the distribution of outputs is given using with boxplots. We have included the distribution of $M_B = M_V * (m_E + 1)$, the total biomass, because it is interesting biological information commonly measured. Indeed, we see that M_B has approximately the same distribution in $\Omega^{\alpha,4}$ and $\Omega^{\alpha,3}$, but that of M_V and m_E greatly differs from one class to the other. Such a pattern of the distribution underlines the heterogeneity of the outputs given by the model M.

Table 2. Parameter ranges and definitions for M taken and calculated from [26,32–34].

parameter	range	interpretation	unit
k	0.5–0-9	turnover rate of reserves	day^{-1}
j_M	0–1	specific maintenance rate	mmolC mmolC^{-1} d^{-1}
$y_{V,E}$	0.01–1	yield of structure from reserves	mmolC mmolC^{-1}
μ	0.05–1	mortality rate	day^{-1}
M_φ	0–10	biomass of the epipelagic compartment	mmolC m^{-3}
ψ_E	0–1	sinking rate of P within the epipelagic layer	d^{-1}
ψ_M	0–1	sinking rate of P within the mesopelagic layer	d^{-1}
$\alpha_{P,D}$	0–1.2	coagulation rate of D	d^{-1}
$\alpha_{D,P}$	0–1.2	dissolution rate of P	d^{-1}
y_A	0–1	yield of reserves from P	mmolC mmolC^{-1}
K_P	2–3.5	half saturation constant for P uptake	mmolC m^{-3}
K_D	2–3.5	half saturation constant for D uptake	mmolC m^{-3}
j_{Dm}	2–7	maximum specific D uptake rate	mmolC mmolC^{-1} d^{-1}
j_{Pm}	2–7	maximum specific P uptake rate	mmolC mmolC^{-1} d^{-1}

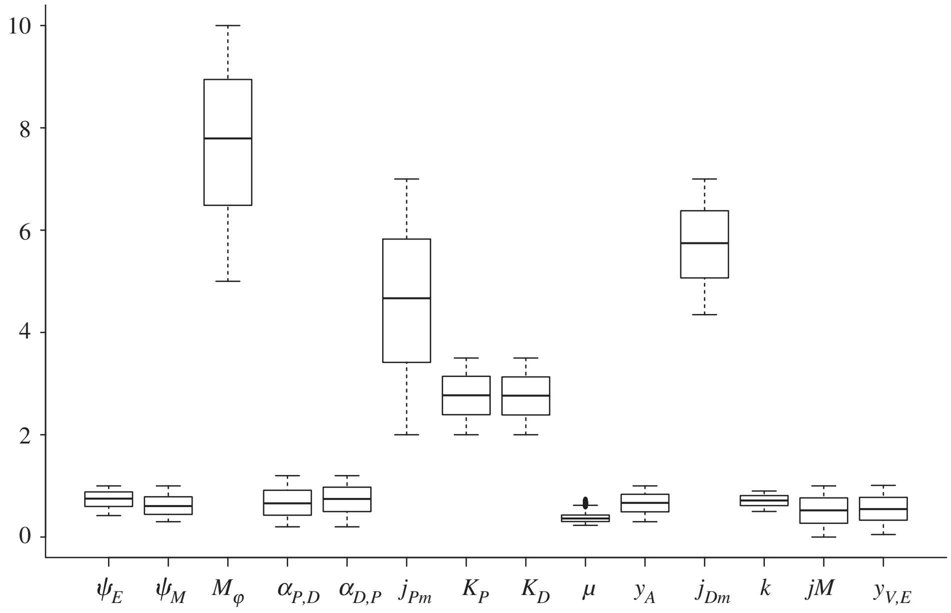

Figure 6. A boxplot representation of Λ^β. Each parameter only reaches a specific range of values under the β-condition. It appears that three couples of parameters reach very close ranges of values: ψ_M and ψ_E, $\alpha_{P,D}$ and $\alpha_{D,P}$, K_P and K_D.

Indeed, by analytically solving the equation $\mathrm{d}Mv/\mathrm{d}t = 0$, we see that $m_{E,\infty}$ can be expressed using other parameters such that $m_{E,\infty} = (j_M y_{V,E} - \mu)/y_{V,E}(k - \mu)$. Thus, at the equilibrium state, $m_{E,\infty}$ (the reserve density) increases with j_M (the specific maintenance rate). Because $\Omega^{\alpha,4}$ exhibits a high $m_{E,\infty}$ and considering the range of the other parameters within the equation, especially μ, j_M has to be high. Such a result draws the interest because it appears coherent when compared with some data acquired at a much larger scale [37]. It leads us to define $\Omega^{\alpha,4}$ as the class fulfilling the β-condition, and rewrite $\Omega^{\alpha,4} = \Omega^\beta$. Although supported by ecological arguments, the choice is here mainly methodological and has to be seen as a β-condition defined over a qualitative stylized fact. The definition of the β-condition is one of the major challenges of the method and must be done accurately.

Table 3. Parameter ranges and definitions for M_S.

parameter	range	unit
k	0.5–0.9	day^{-1}
j_M	0–1	mmolC mmolC^{-1} d^{-1}
$y_{V,E}$	0.01–1	mmolC mmolC^{-1}
μ_b	0.23–1	day^{-1}
M_φ	0–10	mmolC m^{-3}
ψ_s	0.42–1	day^{-1}
α_s	0–1.2	day^{-1}
y_A	0–1	mmolC mmolC^{-1}
K_s	2–3.5	mmolC m^{-3}
j_{Dm}	4.34–7	mmolC mmolC^{-1} d^{-1}
j_{Pm}	2–7	mmolC mmolC^{-1} d^{-1}

3.3. A particular case: the equalization of parameters

After having selected the class that fulfils the β-condition, Ω^β, we analyse its associated parameter range, $\Lambda_1^\beta \times \cdots \times \Lambda_{14}^\beta$, as represented in figure 6.

Because for any set of parameters chosen in Λ^β the β-condition will be respected, a good way to simplify the model is to equalize the parameters with really close distributions. To keep a mechanistic approach, only parameters with the same unit and representing the same kind of phenomenon are able to be equalized two by two. Here, three couples of parameters describe same ranges of values: (i) ψ_M and ψ_E; (ii) $\alpha_{P,D}$ and $\alpha_{D,P}$; (iii) K_P and K_D. We now define three parameters instead of six by saying: (i) $\psi_s = \psi_M = \psi_E$; (ii) $\alpha_s = \alpha_{P,D} = \alpha_{D,P}$; (iii) $K_s = K_P = K_D$. The model M is then rewritten in a simpler form as M_S:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \psi_s(M_\varphi - P) + \alpha_s(D - P) - M_V\frac{j_{Pm}P}{P + K_s},$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \mu M_B + (1 - y_A)\frac{j_{Pm}P}{P + K_s}M_V - \frac{j_{Dm}D}{D + K_s}M_V + \alpha_s(P - D),$$

$$\frac{\mathrm{d}M_V}{\mathrm{d}t} = M_V\frac{km_E - j_M}{m_E + \frac{1}{y_{V,E}}} - \mu M_V$$

and $$\frac{\mathrm{d}m_E}{\mathrm{d}t} = \frac{j_{Dm}D}{D + K_s} + y_A\frac{j_{Pm}P}{P + K_s} - km_E - r_B m_E - \mu m_E,$$

and we define the parameter range for the new model, Λ^{S_1}, as given in table 3. This new model respects the same mechanistic approach as the original one, but has only 11 instead of 14 parameters.

3.4. The general case: $\pi_{j_1}{}^\beta = f(\pi_{j_2}{}^\beta)$

We have seen that parameters involved in similar processes in different compartments can be directly equalized to simplify the model. However, for other couples of such types of parameters, such as j_{Dm} and j_{Pm}, a direct equalization is impossible (figure 6). Nevertheless, by looking closer at the range of these parameters, a correlation may sometimes appear, leading to another kind of simplification.

If we want to keep decreasing the number of parameters, the only way to simplify the model using such a property of the range is to define a relation where no new parameter

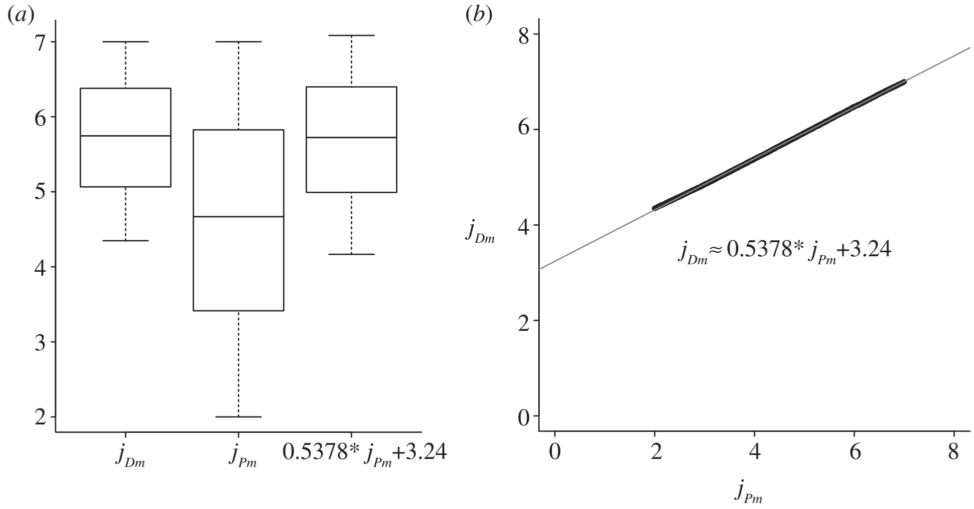

Figure 7. (*a*) A boxplot representation of the distribution of j_{Dm}, j_{Pm} and the simplification regression. (*b*) A quantile–quantile plot of j_{Dm} and j_{Pm} distribution. The grey line represents the regression which has led to the simplification equation. It appears clear that a linear model links the distribution of j_{Dm} and j_{Pm}.

appears. A classical linear regression shows here that the two parameters can be related by a first-order equation:

$$j_{Dm} = f(j_{Pm}) = 0.5378 j_{Pm} + 3.24.$$

Figure 7 shows details of that regression. Note here that these numerical values result from internal properties of the model as well as from specific definitions of both the α- and β-conditions. The model M_S is then rewritten as M_{S^2}:

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \psi_s(M_\varphi - P) + \alpha_s(D - P) - M_V \frac{j_{Pm}P}{P + K_s},$$

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \mu M_B + (1 - y_A)\frac{j_{Pm}P}{P + K_s}M_v$$

$$\frac{(0.5378 j_{Pm} + 3.24)D}{D + K_s}M_V + \alpha_s(P - D),$$

$$\frac{\mathrm{d}M_V}{\mathrm{d}t} = M_V\frac{km_E - j_M}{m_E + \frac{1}{y_{V,E}}} - \mu M_V$$

and $$\frac{\mathrm{d}m_E}{\mathrm{d}t} = \frac{(0.5378 j_{Pm} + 3.24)D}{D + K_s} + y_A\frac{j_{Pm}P}{P + K_s} - km_E - r_B m_E - \mu m_E,$$

and, as previously, we define the parameter range for the new model, Λ^{S_2}, as given in table 4.

Again, this new model respects the same mechanistic approach as the original one, but has only 10 instead of 14 parameters.

If we consider that M_{S^2} is the model that best fits our expectations, then the simplification has enlightened that some assumptions are not supported by the data/observations: (i) there is no change of density for the POC through the mesopelagic layer since ψ_M and ψ_E have been replaced by a unique parameter; (ii) there is no differentiation between the coagulation rate of the DOC and the dissolution rate of the POC, indeed $\alpha_{P,D}$ and $\alpha_{D,P}$ have been replaced by a unique parameter; and (iii) K_P and K_D have been equalized as the bacteria affinity for the DOC and the POC is the same, but their assimilation process is quantitatively different because J_{Pm} has been expressed as a function of J_{Dm}. Note again that, even though some of the primary assumptions have been left, the global approach stays mechanistic, and the structure of the model is not affected.

Table 4. Parameter ranges and definitions for M_{S^2}.

parameter	range	unit
k	0.5–0.9	day^{-1}
j_M	0–1	mmolC mmolC^{-1} d^{-1}
$y_{V,E}$	0.01–1	mmolC mmolC^{-1}
μ_b	0.23–1	day^{-1}
M_φ	0–10	mmolC m^{-3}
ψ_s	0.42–1	day^{-1}
α_s	0–1.2	day^{-1}
y_A	0–1	mmolC mmolC^{-1}
K_s	2–3.5	mmolC m^{-3}
j_{Pm}	2–7	mmolC mmolC^{-1} d^{-1}

4. Discussion

Starting from a given stationary model, it has been shown in the previous sections how to build a tree that allows one to sort the model outputs with criteria defined by critical values of some parameters. The statement of the β-condition restricts the model output range as well as the parameter range. Under those conditions, it has been shown that simplified models could be derived. Unfortunately, building the tree is not straightforward, which is why inner difficulties of the method will be discussed in the following section. In the example used in this paper, two simplified models have been obtained and their ability to act as an alternative to the complete model will be assessed in what follows. Beyond the specific results concerning the reduction of the number of parameters, we discuss how this reduction can be used in usual model applications.

4.1. Determining the β-condition

A critical point of the method lies in the choice of the β-condition. The way the condition is expressed is essential because it determines the required number of final and intermediate classes, and thus the way the tree is built. The more restrictive the β-condition is, i.e. $\mathcal{V}_{\Omega^\alpha} \gg \mathcal{V}_{\Omega^\beta}$ with $\mathcal{V}_i$ the volume of the domain i, the greater the restriction of parameters is likely to be, leading the modeller to build a large and complex tree, highly dependent of the sampling. Thus, a good β-condition has to be reached (i.e. isolated by the tree regression) after a few splits. Furthermore, even if that latter condition is fulfilled, the isolated class Ω^{β} has to occupy a non-negligible part of Ω^{α} otherwise some primary model assumptions should be questioned: (i) the parameter domain is too large, Λ^{α} has to be reduced and (ii) the model structure itself is not adapted to the purpose of the modeller since too many restrictions on the parameters are necessary to fit the β-condition, even with a reduced Λ^{α}.

The β-condition can be seen as a measure of the constraints supplied by the data. It is important to underline that data play a key role when defining the β-condition, especially when that latter is quantifiable. Thus, it is the nature of data, albeit weighted by the expectations of the modeller, which determines the β-condition. It is important to see that, by determining the β-condition, the modeller actually determines the complexity of the final model. If the model is expected to be highly predictive, the β-condition will be highly restrictive. Whereas, for a model with a more explicative interest, the β-condition will probably be less restrictive, allowing the model to make quantitative but not qualitative mistakes. Such a difference explains why, with this last kind of model, simplifications will be more easily applicable, thanks to wider parameter ranges. In our example, the isolation of Ω_{β} has been straightforward; unfortunately it is not the general case.

4.2. Isolating the β-conditioned part of Ω

The isolation of the β-conditioned part of Ω in one of the final classes of the tree has to be seen as the event that will stop the construction of the tree. First, it is important to quantitatively define when Ω_{β} can be considered as isolated. In most cases, it is up to the modeller. Indeed, because we use a sample of the output space, any empirical criterion that satisfies the modeller's needs can be chosen. For example, by saying that if 90 per cent of a given class fulfil the β-condition, then Ω_{β} has been successfully isolated.

Secondly, it is possible that where Ω_{β} has not been entirely isolated in one class, the next split will separate Ω_{β} in two sub-classes, making impossible its meaningful isolation. Such a phenomenon could occur when the structure of the distribution of Ω_{β} is not closely related to the Ω one. By using a transformation function on Ω, which enhances the heterogeneity of the distribution, that effect could be partially minimized. However, the use of such a function has to be taken into account when defining the splitting criteria based on the deviance decrease. Indeed, it will be greatly affected by a quantitative change of the output distribution. That is why, in addition to the previous cited methods, changing the measure of the deviance itself could bring a solution. Eventually, it could be possible that, despite many efforts, Ω_{β} would not be properly isolated in one class. When confronted with such a situation, being sure of having well-sampled Ω is essential to ensure a good representation of its distribution.

4.3. On the choice of the size of the sample, r

Four classes are obtained on the sample of size much larger than 10^4. We do not know if such a size for the sample is large enough to guarantee the stability of the selected tree. As written before, a sample of the order of 10^4 should be large enough. However, the robustness of the obtained tree also depends on the number of final classes. For example, a tree with the highest number of final classes, r, is both unstable and non-robust because intrinsically dependent on the sample. Thus, the minimum sample size advised by Breiman *et al.* [22] can only be seen as an order of magnitude: larger samples will be required for a high number of final classes. We highlight the fact that the number of final classes depends on the used algorithm and its tuning. Indeed, the choice of the building parameters is a key step to build a robust and stable tree that will properly fit the modeller's needs.

This order of magnitude depends on the required number of final classes. We here choose to work with a seven class tree and look more closely at the influence of the size of the sample on the structure of the obtained tree. The results are shown in figure 8. Such results comfort us in choosing a larger sample than previously recommended, of at least $r = 10^{n+1}$, to build the regression tree. Eventually, the minimum size of the sample is directly determined by: (i) the nature of the state variable distribution since heterogeneous distributions, even poorly sampled, are more easily classified; (ii) the requested number of final classes, because high numbers of final classes are deeply linked to the sample itself, it appears reasonable to select a tree where the number of points per class is, at least, equal to $r/30$, which limits its complexity; and (iii) the number of parameters. A large number of parameters increases the number of possible splits, resulting in an increase of the sample size r to ensure the stability of the tree. We want to emphasize that the size of the sample is crucial in constructing the tree. Indeed, using a non-stable tree highly compromises the efficiency of the presented method.

4.4. Density function of the model output

We first discuss the comparison of the distributions of the different model outputs. The presented results have been obtained as follows: we first sample Λ, Λ_S and Λ_{S^2} using a quasi-random method to optimize computation time. Then, for the three models, we determine the asymptotic behaviour for each sampled parameter set by solving the steady-state equation using a Newton–Raphson method. Results are presented in figure 9. We can see that the results for the simplified models are satisfying because the distributions for $(P_{\infty}, D_{\infty}, M_{V,\infty}, m_{E,\infty})$ are conserved. Moreover, if we consider the pattern described by the β-condition (high and almost equal value for P and D, low value of M_V and high value for m_E), we can measure the good performance of the simplified models for which $m_{E,\infty}/M_{V,\infty}$ is even higher than for the original model. Because a simplified model will only be acceptable for a purpose that fits the defined β-condition, it turns out that from a general model fulfilling the α-condition, several simplified models can be obtained using several β-conditions. To summarize, the simplified models obtained with the method can be seen as local approximations of the global model, for a given β-condition.

When several simplified models are derived with the method depicted in this paper, the simplest is not necessarily the most appropriate. Among the list of simplified models obtained with the method, let us call M_{S*} the best simplified

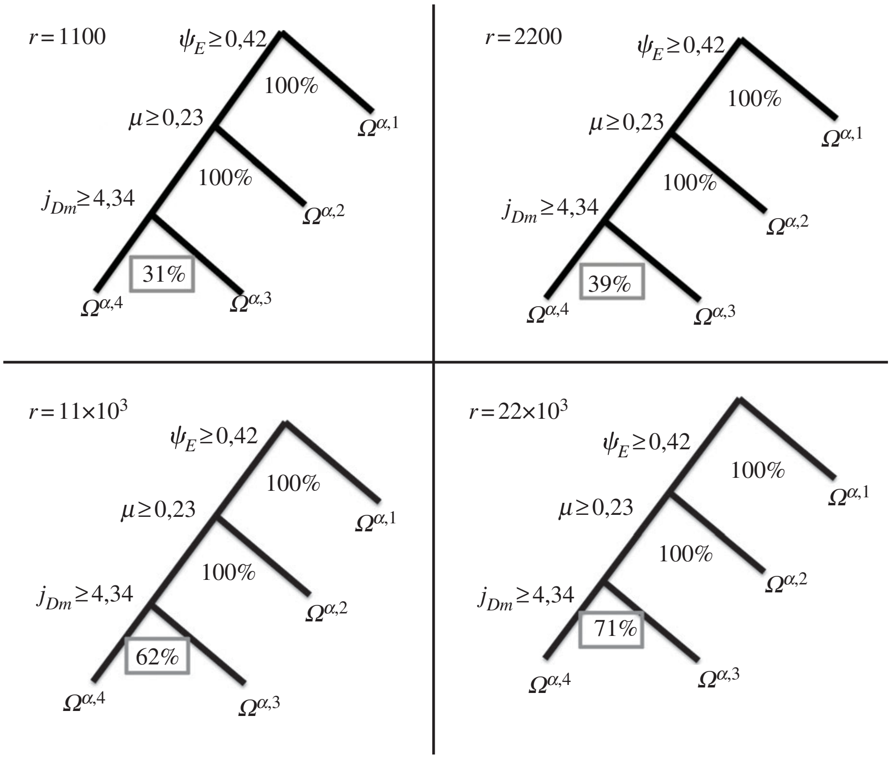

Figure 8. For four different values of r (1100, 2200, 11×10^3, 22×10^3), 300 samples of size r of the original sample have been used to build 300 trees for each value of r. The percentage of appearance of each class has been calculated. The third class is still unstable (i.e. appears in less than 95% of the trees) when $r = 22 \times 10^3$. Stability occurs with a sample of a minimum size of $r = 10 \times 10^4$ (not shown here).

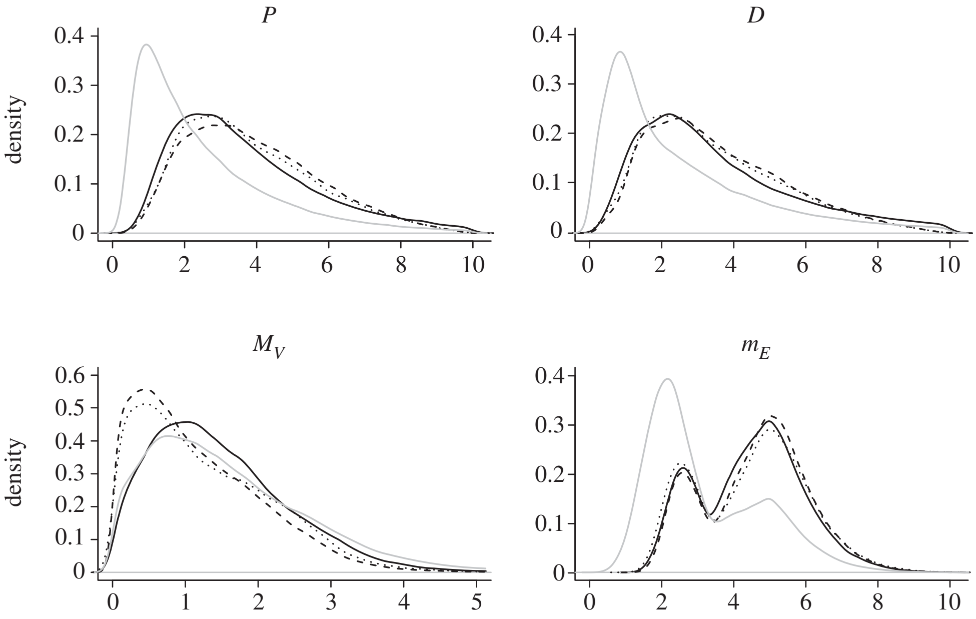

Figure 9. Density estimates of the three model outputs: Ω_S (dashed line), Ω_{S^2} (dotted line) and Ω^{β} (solid black line). For information, Ω^{α} is also given (solid grey line).

model according to a given criterion. Several approaches are possible using different criteria (Akaike information criterion, Schwarz criterion, etc.), the choice depending on the nature of the outputs of the model and on the β-condition (see [38] for a description of those criteria). For instance, if the method presented in [20,21] does not allow one to build simplified models, it is likely to be of great help when searching M_{S*} among several possibilities. Such quantitative selection methods are meaningful with a quantitative definition of the β-condition. Nevertheless, for qualitative β-conditions, the choice of M_{S*} should be also balanced by non-computational arguments.

4.5. Parameter estimation

The present method has another benefit in the estimation of model parameter values from measured output data. This operation is made easier with a simplified rather than a complex model. Usually, the values of the searched parameters are provided by the minimization of a cost function

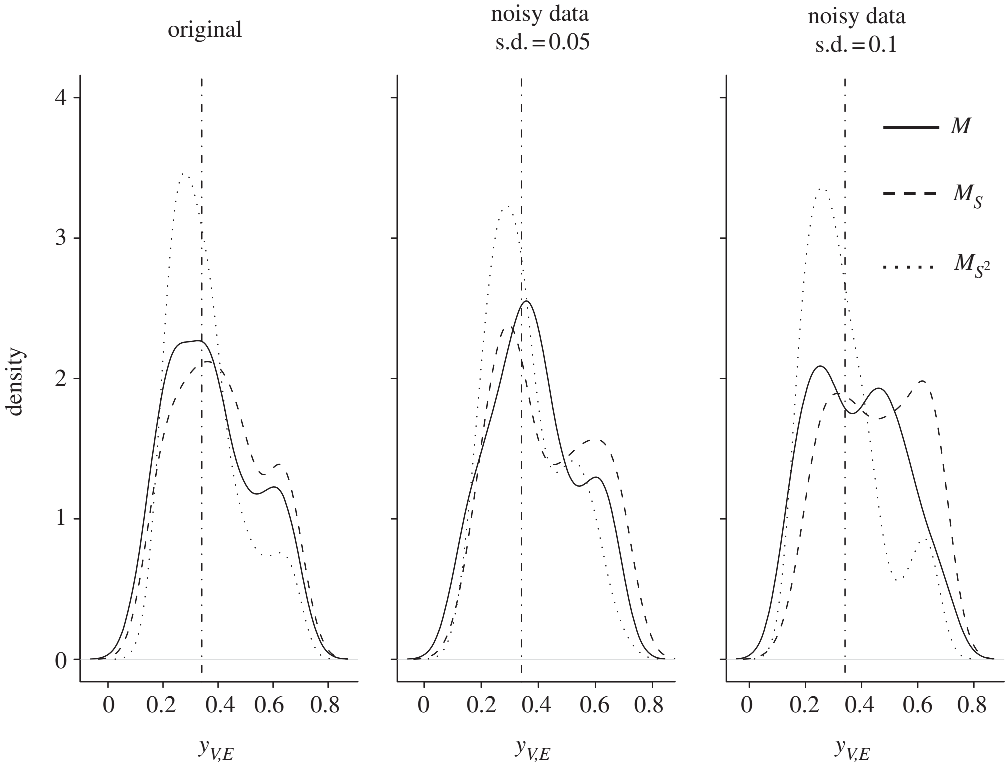

Figure 10. 100 parameter estimations have been carried out using a hybrid method of optimization (L-BFGS-B + SAAN [40,41]) for a time series of 80 values reaching the asymptotic state $X_\infty(P_\infty = 1.81, D_\infty = 1.34, M_{V,\infty} = 1.69, m_{E,\infty} = 4.93)$. The density of the optimized parameters is here represented for $y_{V,E}$, a hardly measurable parameter. Also, the original dataset has been noised following a normal law with a standard deviation of successively 0.05 and 0.1. For each of those noised datasets, parameter estimations have also been made. The original value of the parameter that has given the studied dataset is denoted by the vertical dashed line. It appears that the estimation is more robust for M_{S^2} (dotted line) as we observed higher density value for the dominant modality (3.5 at $\sim y_{V,E} = 0.27$) whereas the *original* value ($y_{V,E} = 0.34$) is recovered using M (solid line) and M_S (dashed line). Adding noise to the dataset randomly leads to different, more or less robust, estimations for M and M_S while M_{S^2} ones stay equally robust around the same value.

representing the error between the model outputs and measured data. It appears that the higher the number of parameters is, the more complex the cost function will be because its degrees of freedom are increased. When the cost function is highly complex, its minimization is made difficult and leads one sometimes to choose a local rather than the global minimum of the cost function, thus providing a set of parameters that is not optimal. In decreasing the number of parameters, the method directly decreases the number of degrees of freedom of the cost function, and its complexity, which is highly interesting from a computational point of view. As an example, let us consider a situation where 10 values per parameter are chosen; so a total of 10^m simulations with m parameters. Let us assume that one simulation takes reasonably 10^{-6} seconds; it will take more than 3 years to estimate the cost function of an $m = 14$ parameters model, whereas 30 h would be sufficient with $m = 11$ parameters.

In order to perform a parameter estimation in our example, artificial data have been generated using the model M (see equations (3.1)) that have been integrated until reaching the asymptotic state. For each state variable, the last 80 values have been considered as a set of data. That time series has been used to build a cost function based on the sum of the squares of the errors. Parameter estimation has been done using a hybrid method of optimization mixing both a stochastic (simulated annealing) and a deterministic (Newton's method) search. The parameters of the three models (the original model and the simplified ones) have been estimated to compare the ability of all models to retrieve the original parameter set. Results are presented in figures 10 and 11. The yield parameter, $y_{V,E}$, has been chosen as an example since its estimation has practical interests: it is a non-measurable parameter that is usually estimated from other parameters (for a method of parameter estimation in DEB theory, see [29,39]). Figure 10 shows that only one of the simplified models, M_S, is able to recover the original value of that parameter. However, for a modeller who does not know *a priori* the value of a parameter, the robustness of the estimation is what matters. The use of the second simplified model, M_{S^2}, leads to a more robust estimation since we observe higher density values for the dominant modality, whereas the original value is recovered using M and M_S. In this case, M_{S^2} represents the best alternative to M. Eventually, let us underline the complex nature of the cost function, which has trapped the estimation method into local minima, very close to 0.

4.6. Comparison with other simplification and reduction methods

Since this method tackles a very important question, several publications have already proposed some approaches. It is important to underline what is original in the presented method. In Apri *et al.* [25], the authors propose a simplification method very similar in principle to the present one. More particularly, their concept of the *admissible region* is close to our definition of the β-condition. They also analyse the parameter ranges associated with that specific region to define possible simplifications. In the presented method, we aim at simplifying the model without affecting the nature of its structure; by only gathering parameters that are implied in the same process and by taking care to not change original

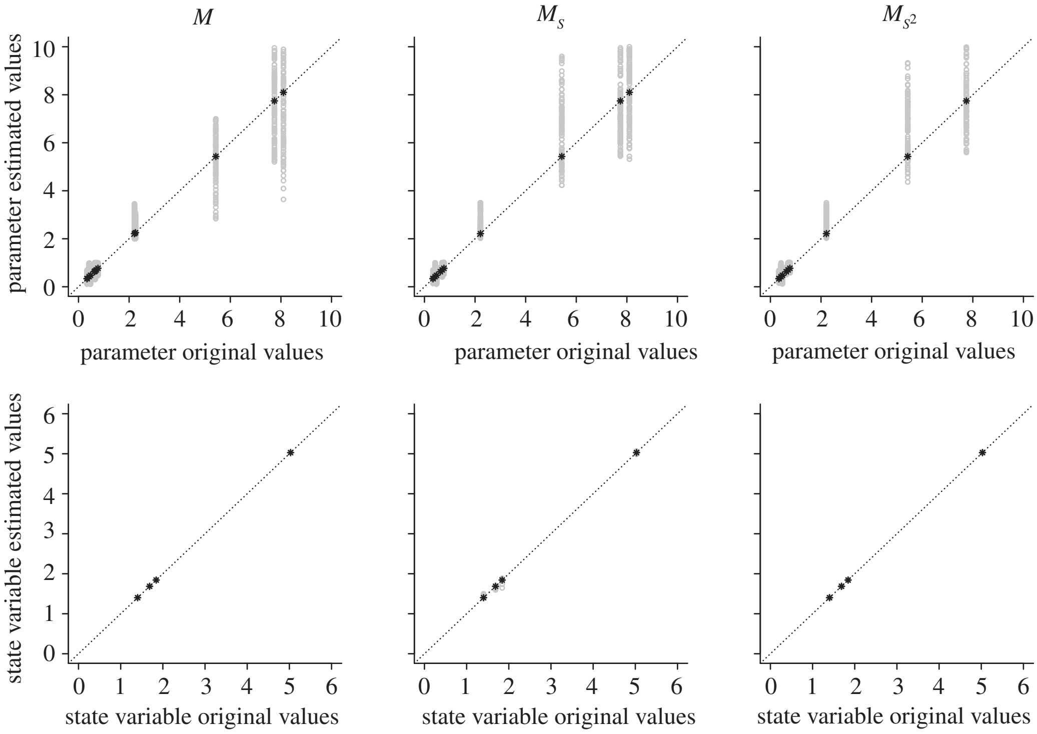

Figure 11. For each model, 100 parameter estimations have been carried out using an hybrid method of optimization (L-BFGS-B + SAAN [40,41]) for a time series of 80 values reaching the asymptotic state $X_\infty(P_\infty = 1.81, D_\infty = 1.34, M_{V,\infty} = 1.69, m_{E,\infty} = 4.93)$. The estimated values have been plotted against the original ones for both the parameters and the state variables. For all models, the quality of the fit for the state variables is equally extremely good. Parameter estimations reveal the presence of multiple local minima for the cost function.

formulations. Thus, a mechanistic model will stay mechanistic. The same methodological divergence appears when comparing our work with some other systematic simplification methods [12,14,18]. Eventually, we underline how far the classification given by a regression tree approach is useful to understand how some parameters shape the distribution of the outputs. The method presented in this paper can thus be seen as some kind of parameter sensitivity analysis of a model with the purpose of simplifying it.

Finally, one methodological point remains to be mentioned. Once the β-condition has been determined, attention must be paid to the most efficient way to tune the tree-building algorithm. Indeed, as said before, defining when the tree has efficiently isolated Ω_β is not a trivial part in the application of the method. The difficulty here lies in the measure of the deviance between a mother-class and its two daughters. Indeed, if it appears tempting to classify Ω until Ω_β is efficiently isolated, one would not have to forget that a classification without a control on the rate of the decrease of the deviance is hardly meaningful. Such a control is fundamental because without a stopping criterion based on the deviance measurement, the algorithm could end up creating classes among almost homogeneous distribution of outputs. Also, both Λ^α and Ω^α will vary depending on the formulation of the model. Both the deviance measure and the tree building parameter values may change to obtain a meaningful classification. Examples of other kinds of deviance measurement are given in [42].

5. Conclusion and perspectives

Complex models are usually intractable from a mathematical point of view. Thus, it appears difficult to isolate the part of such a complex model that is actually responsible for its dynamical behaviour. This task is not easy to reach, although it would obviously be of great help for modellers of complex systems. The method presented in this article does not fully reach this goal, but is a first step towards it. Indeed, in the case of models with asymptotically stable states, the method described here establishes an explicit relation between the model outputs and the subsets of the parameter space leading to these outputs.

The method thus has several direct benefits. It simplifies the model by reducing the number of parameters, which leads to less parameter estimations, a great advantage when thinking of estimations obtained from experiments and/or field observations. Parameter optimization is made easier and can be sometimes more robust. Such a benefit is meaningful especially for non-measurable parameters. When some parameters cannot be estimated in the simplified model chosen through a quantitative selection method (Akaike, least squares, etc.), the choice of the simplified model must be done by making the balance between a selection based on qualitative interests (e.g. loss of a non-measurable parameter, more robust parameter estimation) and quantitative criteria. Finally, it allows a quantitative evaluation of the model capacity to reach its objectives. Indeed, as said previously, the size of Ω^β has to be quite large compared with that of Ω^α; otherwise some parts of the M model have to be reconsidered. It is noticeable that the method has only been tested for steady-state models. However, by construction, an extension to more complex dynamics is possible. Indeed, the only difference for the application of the method will lie in the splitting criterion used in the tree construction algorithm. If we here used the Euclidean distance between steady-state outputs in the state variable space for evaluating the deviance criteria, comparing two sets of aperiodic curves deserves another approach, as discussed in [42]. In the case of other types of attractors (limit cycles or even more complex attractors), the use of the Haussdorf distance in the phase space could be used to evaluate the deviance.

Another issue appears when thinking about the classification tree method itself. It is currently based on an orthogonal structure of the hyperplanes used to split the parameter domain. It is not correct to approximate a highly non-convex Λ^{β} with an m-orthotope. Oblique splits, also known as linear combination splits, as described in [43] could have improved the method in such a case since their polytopial structure would allow one to maximize $\Lambda^{\beta} \cap \Lambda^{S}$. Nonetheless, using them to create a tree will lead the modeller to express some parameters as a function of others without controlling that they are still involved in the same type of processes. Thus, oblique splits would affect the structure of the model and that is why they are not of interest from our point of view.

Also, the question of the exportability of the simplified model arises. It is important to underline that the obtained simplified model, once validated, only makes sense when used to give the same kind of outputs as Ω_{β}. Again, the simplified model has to be seen as an approximation of the general one for a given β-condition. Thus, in the presented simplification example, it would be meaningless to work with the transitory phase of M_S or M_{S^2} since only the asymptotic phase of the dynamics has been used for the tree construction. It appears clear that our approach highly depends on the definition of the α- and β-conditions, which determine both the domain on which the tree will be built and the possible degree of simplification applicable to the original model. Thus, being able to quantify the effect of the definition of those two conditions on the simplification capabilities of our approach is definitely a topic for future investigations.

The authors thank two anonymous reviewers for their very constructive comments that helped improve the manuscript. The PhD scholarship for Y.E. came from the French Ministry of Research and Education. The authors thank Claude Manté, Christian Tamburini, Mehdi Boutrif and Starrlight Augustine for fruitful discussions during the work.

Appendix A

In this work, the deviance formulation for each class p of the tree is

$$\hat{R}(p) = \sum_{X_k \in p} \| X_k - \bar{X}_p \|^2,$$

where $\bar{X}_p$ is the average of the observations of X belonging to region p, with $\| \,.\, \|$ the usual norm in $\mathbb{R}^n$. Starting with the whole sample Φ, let us consider a splitting variable π_j and a threshold s on this variable. We then define the region p_1 for which $\pi_j \leq s$, $p_1 = \{X_{k=1,\ldots,r} | \pi_j \leq s\}$, and the region p_2 for which $\pi_j > s$, $p_2 = \{X_{k=1,\ldots,r} | \pi_j > s\}$, such that $p = p_1 \cup p_2$ and $p_1 \cap p_2 = \varnothing$. The within class sum of squares can be calculated in each of these parts of m-orthotopes:

$$\hat{R}(p_1) = \sum_{X_k \in p_1} \| X_k - \bar{X}_{p_1} \|^2$$

and $$\hat{R}(p_2) = \sum_{X_k \in p_2} \| X_k - \bar{X}_{p_2} \|^2 .$$

For any split s belonging to the set S of all candidate splits, p is subdivided into p_1 and p_2, and the variation in deviance is given by

$$\Delta\hat{R}(s,p) = \hat{R}(p) - [\hat{R}(p_1) + \hat{R}(p_2)].$$

The selected split s^* of p into p_1 and p_2 is the split that most decreases $\hat{R}(p_1) + \hat{R}(p_2)$ such that

$$\Delta\hat{R}(s^*,p) = \max_{s \in S} \Delta\hat{R}(s,p).$$

The decrease in $\hat{R}(p)$ when splitting a region p into p_1 and p_2 is guaranteed because the following property is verified for any p:

$$\hat{R}(p) = \hat{R}(p_1) + \hat{R}(p_2) + \frac{r_1 r_2}{r_1 + r_2} \| \bar{X}_{p_1} - \bar{X}_{p_2} \|^2$$

with r_1 and r_2 the number of observations respectively in p_1 and p_2. This property arising from the decomposition of the inertia and the Huyghens theorem is verified because the criterion $\hat{R}(p)$ is a sum of squared distances.

References

1. Demongeot J, Françoise J-P, Nerini D. 2009 From biological and clinical experiments to mathematical models. *Phil. Trans. R. Soc. A* **367**, 4657–4663. (doi:10.1098/rsta.2009.0187)
2. Petrovskii S, Petrovskaya N. 2012 Computational ecology as an emerging science. *Interface Focus* **2**, 241–254. (doi:10.1098/rsfs.2011.0083)
3. Anderson TR. 2010 Progress in marine ecosystem modelling and the unreasonable effectiveness of mathematics. *J. Mar. Syst.* **81**, 4–11. (doi:10.1016/j.jmarsys.2009.12.015)
4. Fussmann GF, Blasius B. 2005 Community response to enrichment is highly sensitive to model structure. *Biol. Lett.* **1**, 9–12. (doi:10.1098/rsbl.2004.0246)
5. Poggiale J-C, Baklouti M, Queguiner B, Kooijman SALM. 2010 How far details are important in ecosystem modelling: the case of multi-limiting nutrients in phytoplankton–zooplankton interactions. *Phil. Trans. R. Soc. B* **365**, 3495–3507. (doi:10.1098/rstb.2010.0165.)
6. Cordoleani F, Nerini D, Gauduchon M, Morozov AY, Poggiale J-C. 2011 Structural sensitivity of biological models revisited. *J. Theor. Biol.* **283**, 82–91. (doi:10.1016/j.jtbi.2011.05.021)
7. Arhonditsis GB, Brett MT. 2004 Evaluation of the current state of mechanistic aquatic biogeochemical modeling. *Mar. Ecol. Prog. Ser.* **21**, 13–26. (doi:10.3354/meps271013)
8. Reichert P, Omlin M. 1997 On the usefulness of overparameterized ecological models. *Ecol. Model.* **95**, 289–299. (doi:10.1016/S0304-3800(96)00043-9)
9. Raick C, Soetaert K, Gregoire M. 2006 Model complexity and performance: how far can we simplify? *Prog. Oceanogr.* **70**, 27–57. (doi:10.1016/j.pocean.2006.03.001)
10. Seber GAF, Wild CJ. 1989 *Nonlinear Regression*. New York, NY: Wiley.
11. Ward B, Friedrichs MAM, Anderson TR, Oschlies A. 2010 Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. *J. Mar. Syst.* **81**, 34–43. (doi:10.1016/j.jmarsys.2009.12.005)
12. Cox GM, Gibbons JM, Wood ATA, Craigon J, Ramsden SJ, Crout NMJ. 2006 Towards the systematic simplification of mechanistic models. *Ecol. Model.* **198**, 240–246. (doi:10.1016/j.ecolmodel.2006.04.016)
13. Raick C, Beckers J-M, Soetaert K, Grégoire M. 2006 Can principal component analysis be used to predict the dynamics of a strongly non-linear marine

biogeochemical model? *Ecol. Model.* **196**, 345–364. (doi:10.1016/j.ecolmodel.2006.02.014)

14. Lawrie J, Hearne J. 2007 Reducing model complexity via output sensitivity. *Ecol. Model.* **207**, 137–144. (doi:10.1016/j.ecolmodel.2007.04.013)
15. Auger P, Bravo de la Parra R, Poggiale JC, Sánchez E, Sanz L. 2008 Aggregation methods in dynamical systems and applications in population and community dynamics. *Phys. Life Rev.* **5**, 79–105. (doi:10.1016/j.plrev.2008.02.001)
16. Lawrie J, Hearne J. 2008 A method for aggregating state variables in large ecosystem models. *Math. Comp. Simul.* **79**, 368–378. (doi:10.1016/j.matcom.2008.01.001)
17. Crout N, Tarsitano D, Wood A. 2009 Is my model too complex? Evaluating model formulation using model reduction. *Environ. Model. Softw.* **24**, 1–7. (doi:10.1016/j.envsoft.2008.06.004)
18. Gibbons JM, Wood ATA, Craigon J, Ramsden SJ, Crout NMJ. 2010 Semi-automatic reduction and upscaling of large models: a farm management example. *Ecol. Model.* **221**, 590–598. (doi:10.1016/j.ecolmodel.2009.11.006)
19. Sousa T, Domingos T, Kooijman SALM. 2008 From empirical patterns to theory: a formal metabolic theory of life. *Phil. Trans. R. Soc. B* **363**, 2453–2464. (doi:10.1098/rstb.2007.2230)
20. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202 (doi:10.1098/rsif.2008.0172)
21. Toni T, Stumpf MPH. 2010 Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**, 104–110. (doi:10.1093/bioinformatics/btp619)
22. Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984 *Classification and regression trees*. New York, NY: Chapman & Hall.
23. Pappenberger F, Iorgulescu I, Beven KJ. 2005 Sensitivity analysis based on regional splits and regression trees (SARS-RT). *Environ. Model. Softw.* **21**, 976–990. (doi:10.1016/j.envsoft.2005.04.010)
24. Hornberger GM, Spear RC. 1981 An approach to the preliminary analysis of environmental systems. *J. Environ. Manag.* **12**, 7–18.
25. Apri M, de Gee M, Molenaar J. 2012 Complexity reduction preserving dynamical behavior of biochemical networks. *J. Theor. Ecol.* **304**, 16–26. (doi:10.1016/j.jtbi.2012.03.019)
26. Robinson C *et al*. 2010 Mesopelagic zone ecology and biogeochemistry: a synthesis. *Deep-Sea Res. II, Trop. Stud. Oceanogr.* **57**, 1504–1518. (doi:10.1016/j.dsr2.2010.02.018)
27. Burd A *et al*. 2010 Assessing the apparent imbalance between geochemical and biochemical indicators of meso- and bathypelagic biological activity: What the @$#! is wrong with present calculations of carbon budgets? *Deep-Sea Res. II, Trop. Stud. Oceanogr.* **57**, 1557–1571. (doi:10.1016/j.dsr2.2010.02.022)
28. Nisbet RM, Muller EB, Lika K, Kooijman SALM. 2000 From molecules to ecosystems through dynamic energy budget models. *J. Anim. Ecol.* **110**, 913–926. (doi:10.1046/j.1365-2656.2000.00448.x)
29. Kooijman SALM. 2010 *Dynamic energy budget theory for metabolic organisation*. Cambridge, UK: Cambridge University Press.
30. Eichinger M, Kooijman SALM, Sempéré R, Lefèvre D, Grégori G, Charrière B, Poggiale J-C. 2009 Consumption and release of dissolved organic carbon by marine bacteria in a pulsed-substrate environment: from experiments to modelling. *Aquat. Microb. Ecol.* **56**, 41–54. (doi:10.3354/ame01312)
31. Muller EB, Kooijman SALM, Edmunds PJ, Doyle FJ, Nisbet RM. 2009 Dynamic energy budgets in syntrophic symbiotic relationships between heterotrophic hosts and photoautotrophic symbionts. *J. Theor. Biol.* **259** 44–57. (doi:10.1016/j.jtbi.2009.03.004)
32. Eichinger M, Sempéré R, Grégori G, Charrière B, Poggiale J-C, Lefèvre D. 2010 Increased bacterial growth efficiency with environmental variability: results from DOC degradation by bacteria in pure culture experiments. *Biogeosciences* **7**, 1861–1876. (doi:10.5194/bg-7-1861-2010)
33. Martin P, Lampitt RS, Perry M-J, Sanders R, Lee C, D'Asaro E. 2011 Export and mesopelagic particle flux during a North Atlantic spring diatom bloom. *Deep-Sea Res. I, Oceanogr. Res. Pap.* **58**, 338–349. (doi:10.1016/j.dsr.2011.01.006)
34. Anderson TR, Tang KW. 2010 Carbon cycling and POC turnover in the mesopelagic zone of the ocean: insights from a simple model. *Deep-Sea Res. II, Trop. Stud. Oceanogr.* **57**, 1581–1592. (doi:10.1016/j.dsr2.2010.02.024)
35. Therneau TM, Atkinson B, Ripley B. 2011 rpart: recursive partitioning and regression trees. See http://CRAN.R-project.org/package=rpart.
36. R Development Core Team. 2011 R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. See http://www.R-project.org/.
37. Aristegui J, Gasol J, Duarte C. 2009 Microbial oceanography of the dark ocean's pelagic realm. *Limnol. Oceanogr.* **54**, 1501–1529.
38. Johnson JB, Omland KS. 2004 Model selection in ecology and evolution. *Trends Ecol. Evol.* **19**, 101–108. (doi:10.1016/j.tree.2003.10.013)
39. Lika K, Kearney MR, Freitas V, van der Veer HW, van der Meer J, Wijsman JWM, Pecquerie L, Kooijman SALM. 2011 The covariation method for estimating the parameters of the standard dynamic energy budget model I: philosophy and approach. *J. Sea Res.* **66**, 270–277. (doi:10.1016/j.seares.2011.07.010)
40. Byrd RH, Lu P, Nocedal J, Zhu C. 1995 A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16**, 1190–1208.
41. Belisle CJP. 1992 Convergence theorems for a class of simulated annealing algorithms on Rd. *J. Appl. Probab.* **29**, 885–895.
42. Nerini D, Ghattas B. 2007 Classifying densities using functional regression trees: applications in oceanology. *Comput. Stat. Data Anal.* **51**, 4894–4993. (doi:10.1016/j.csda.2006.09.028)
43. Breiman L. 1996 Stacked regressions. *Mach. Learn.* **24**, 49–64.