272

Evaluation of an evaluation list for model complexity

G.A.K. van Voorn
D.J.J. Walvoort

werkdocumenten

wot
Wettelijke Onderzoekstaken Natuur & Milieu

WAGENINGEN UR
For quality of life

Evaluation of an evaluation list for model complexity

WOt Working Document **272** presents the findings of a research project commissioned by the Netherlands Environmental Assessment Agency (PBL) and funded by the Dutch Ministry of Economic Affairs, Agriculture and Innovation (EL&I). This document contributes to the body of knowledge which will be incorporated in more policy-oriented publications such as the National Nature Outlook and Environmental Balance reports, and thematic assessments.

Evaluation of an evaluation list for model complexity

G.A.K. van Voorn

D.J.J. Walvoort

Werkdocument 272

Wettelijke Onderzoekstaken Natuur & Milieu

Wageningen, december 2011

Abstract
Voorn, G.A.K. van & D.J.J. Walvoort, 2011. *Evaluation of an evaluation list for model complexity; ondertitel.* Wageningen, Wettelijke Onderzoekstaken Natuur & Milieu, WOt-werkdocument 272. 86 blz. 4 Figs.; 3 Tabs.; 56 Refs.; 2 Annexes.

This study ('WOt-werkdocument') builds on the project 'Evaluation model complexity', in which a list has been developed to assess the 'equilibrium' of a model or database. This list compares the complexity of a model or database with the availability and quality of data and the application area. A model or database is said to be in equilibrium if the uncertainty in the predictions by the model or database is appropriately small for the intended application, while the data availability supports this complexity. In this study the prototype of the list is reviewed and tested by applying it to test cases. The review has been performed by modelling experts from within and outside Wageningen University & Research centre (Wageningen UR). The test cases have been selected form the scientific literature in order to evaluate the various elements of the list. The results are used to update the list to a new version.

Key Words: Model application, model complexity, model quality, model uncertainty, optimal model

Auteurs:
George A.K. van Voorn (PRI Biometris)
Dennis J.J. Walvoort (Alterra)

©2011 **PRI Biometris Wageningen UR**
Postbus 100, 6700 AC Wageningen
Tel: (0317) 48 60 01; fax: (0317) 41 80 94; e-mail: info.pri@wur.nl

Alterra Wageningen UR
Postbus 47, 6700 AA Wageningen
Tel: (0317) 48 07 00; fax: (0317) 41 90 00; e-mail: info.alterra@wur.nl

The Working Documents series is published by the Statutory Research Tasks Unit for Nature & the Environment (WOT Natuur & Milieu), part of Wageningen UR. This document is available from the secretary's office, and can be downloaded from www.wotnatuurenmilieu.wur.nl.

Statutory Research Tasks Unit for Nature & the Environment, P.O. Box 47, NL-6700 AA Wageningen, The Netherlands
Phone: +31 317 48 54 71; Fax: +31 317 41 90 00; e-mail: info.wnm@wur.nl;
Internet: www.wotnatuurenmilieu.wur.nl

Preface

This publication contains the results of the testing of a list for the evaluation of model or database complexity compared to the availability and quality of data and the application area of the model or database. This list was developed earlier in the project "Evaluatie modelcomplexiteit", of which the results are described in a different publication (in Dutch).

Here we would like to thank the people who have taken part in the expert evaluation of the list, or have contributed otherwise. Thanks to Patrick Bogaart, Peter Janssen, Harm Houweling, Martin Knotters, and to the various people from Team Integraal Waterbeheer (Team Integral Water Management, Environmental Sciences Group part of Wageningen UR), PBL (Netherlands Environmental Assessment Agency), and the Systems Biology Group at PRI Biometris (part of Wageningen UR), who attended the presentations on the subject and provided useful feedback.

George van Voorn & Dennis Walvoort

Contents

Summary

This publication reports the work done in the project 'Evaluation model complexity 2010' (PRI nr. 3320010210, Alterra pr.nr. 5235784-01), in collaboration with the Netherlands Environmental Assessment Agency (PBL, in Dutch 'Planbureau voor de Leefomgeving'). PBL uses many numerical models and spatial databases developed at Wageningen University & Research centre (Wageningen UR). The complexity of these models and databases increases almost continuously, as they are used for more and more applications. Contrary to the popular assumption an increased complexity does not necessarily lead to better models and databases. Instead, increased complexity also results in increased uncertainty about included (and excluded) factors and increased calculation times.

Here it is assumed that an optimal complexity exists, named 'equilibrium', where the complexity is sufficient for making adequately accurate predictions in view of a certain application, but also sufficiently restricted to minimize all kinds of practical limitations (available computer power, model insight and uncertainties, data availability). Furthermore, this complexity is supported by data of sufficient quantity and quality. Ideally, you would want your model or database to approach this optimum or 'equilibrium' as close as possible. In a previous publication (WOt-Werkdocument 226, Bogaart *et al.*, 2011; in Dutch) a tool has been developed aimed at evaluating models and databases in view of this 'equilibrium'. This tool, the 'evaluation list model complexity' EMC version 0.1, is used to ascertain if the complexity of a model or database is adequate or not with regard to the application. In this publication the list is tested in two ways: by model expert review, and by applying it to a number of somewhat arbitrarily selected test cases from the scientific literature. The results from the reviewing and testing are used to update the list to a new version.

1 Introduction

The Netherlands Environmental Assessment Agency (the Dutch 'Planbureau voor de Leefomgeving, PBL) uses many numerical models and databases for answering questions from the government, and for investigating the effects of policy implementations. Although these models and databases may differ significantly from each other, they are all used for making quantitative predictions. In contrast with 'toy models', which are mainly used as aid in the formation of hypotheses on the qualitative effects of mechanisms, the models and databases used by PBL have clear applications. To satisfy these applications the models and databases tend to be rather complex, meaning they include many process details and contain many components. For databases a larger complexity usually also entails more input (from models, expert knowledge, or output of other databases), leading to an increased number of possible combinations on how this input is fed into the database.

1.1 Are our models and databases too complex?

As a model or database increases in complexity it becomes more detailed and several factors and issues seem to start playing a role that in turn tend to lead to a further increase of this complexity. The end result may easily be a model or database that is huge, and of which nobody except for the original programmer has any clue about its functions, and hence its reliability. Some of these factors and issues are:

- The 'show off' factor: A manager will likely be more impressed by a complex model than by a simple model (Chwif *et al*., 2000).
- The 'include all' syndrome: Specifically inexperienced modellers may suffer from some sort of panic, the fear of forgetting essential details (Chwif *et al*., 2000).
- The thinking of models as 'knowledge banks', where more knowledge automatically means more accuracy in the output.
- Much expert knowledge about the system is necessary to design a proper model (Chwif *et al*., 2000). Often, however, there seems to be a lack of understanding about the modelled system.
- For models a larger complexity usually entails an increased number of variables, parameters, boundary conditions, initial conditions, non-linear interactions, and so forth. Complex numerical models and large spatial databases with many grid cells require much more data for support and calibration. In practice, this data support is almost impossible to achieve.
- The 'Concorde' effect: Investing in a model or database tends to increase conservative thinking, even beyond a point where it would be a better choice to discard (parts of) the model and start anew (Van Nes & Scheffer, 2005).
- The 'model on the shelf'-phenomenon: In a project-driven working environment it is tempting to use a model or database that is readily available ("from the shelf and push the button"), although a more detailed assessment may reveal that it is not very well suited for the job. The common approach is then to add a module to cope with the specific application within the project.
- A lack of clarity on the intended application area can easily lead to the approach of taking the 'application bounds' too wide, just to be sure that the future application will be served (Chwif *et al*., 2000). This issue may be partly countered by ignoring model properties that are irrelevant for the specific applications (Hjalmarsson, 2009).
- The 'because we can' factor: As the computer power increases continuously, model running time seems to be a disappearing limitation (Chwif *et al.*, 2000). In practice,

the complexity of most models and databases increases many times faster, so they often still have a significant calculation time.

- Related to the previous factor is the notion, that it is difficult to test, calibrate, and validate a model or database, and perform sensitivity and uncertainty analyses, with a significant run time, as these activities become too time-consuming.

All these issues eventually result in a loss of confidence in the model, as roughly depicted in Figure 1.1. Seen from the application point of view, this is unwanted. The uncertainty in models and databases would probably be no problem when they were designed for academic purposes. Rather the opposite: they would lead to discussion on the relative importance of the different model or database components, and what they represent in the real system. However, most applications require insight in the uncertainties of the predictions. Especially the long calculation times however mean that it is not appealing to make new series of calculations after every model or database adaptation, which would be necessary to guarantee the quality of the model or database and its predictions.

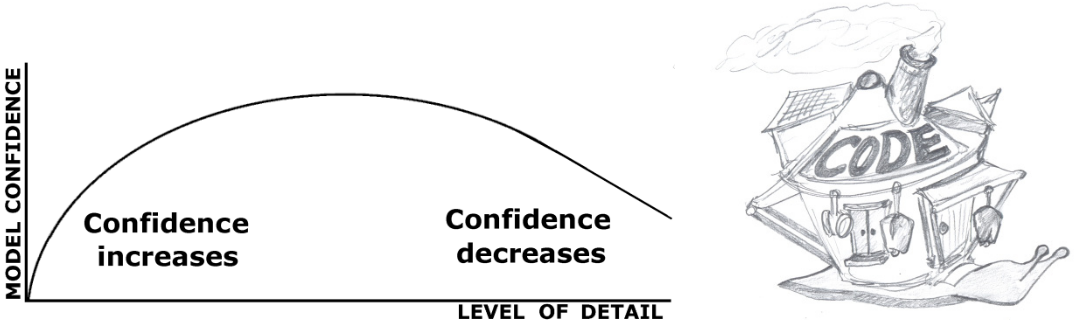

Figure 1.1. *Conceptual graphic representation of model confidence as function of the level of detail of modelling, based on Chwif et al. (2000). A simple model or database usually does not include sufficient detail to explain the modelled system. As the model includes more detail, the confidence in the model increases. Model predictions are likely to be more accurate. At some point, however, the confidence starts to decrease again. Several factors, especially having a large number of model components and huge run times (cartoon-like represented by the snail with the over-sized and cluttered house on the right), decrease the confidence in the reliability of the model or database.*

1.2 Classical model selection

This conflict between on the one hand more complexity to support quantitative predictions, and on the other hand uncertainty in model quality and predictions, is an issue already considered in the academic literature. The popular reductionist view is that a model should not be more complex than supported by the data, a principle well known as ‘parsimony’, also popularly known as ‘Occam’s razor’, which says that in general any explanation should be no more complex than is needed to explain a phenomenon. In general the concept of balancing model complexity and the required data support is considered when ‘model selection’ is performed. The standard example of model selection is ‘data fitting’ (or ‘curve fitting’): different models are compared to a sample of data points, while considering some default assumptions (like the assumption that all points are independent and identically distributed random samples, i.d.d.). The model with the best performance in terms of an objective function (a commonly used one is the ‘root mean squared error’, RMSE) is selected; this is called the ‘goodness of fit’.

The approach taken with the goodness of fit is, however, too simple. In this approach the model with the best fit, no matter how complex, is the preferred choice. There are several problems with this. Measurements are always subject to measurement and system noise. Therefore, one can put only limited confidence in the samples. Statistically speaking, any estimator of the mean and the variance generated using a sample of limited size is biased (see Annex 1). The performance of an estimator is commonly measured by the MSE (mean square error), for which it can be shown that

$$MSE = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

This suggests that there is an trade-off between the variance and the bias, and that any model will have to be balanced between the two. For example, consider Figure 1.2 in which a set of measurements is plotted (red dots), to which we want to fit a polynomial function of some order. At one end of the spectrum is a curve fit of the lowest order (a straight line), which gives a very inflexible curve. It will 'fit' almost any data set, but with a large bias, and therefore a large portion of the difference between any data and predictions by the model cannot be explained by the model. At the other end of the spectrum is a curve fit of very high order, which gives a greatly flexible curve which may fit onto the data exactly. However, in this case the model is 'over-fitted': if we would have a new set of measurements the curve would most likely not fit any more at all, as contrary to the straight line, which almost always explains at least some part of the data. A model of high order has a large variance, and hence this model is not generally applicable: It explains one set of data exactly but cannot describe other, related data sets. For one case the difference between the data and predictions by the model is minimal, but for other cases this difference is actually 'blown up' as compared to the straight line.

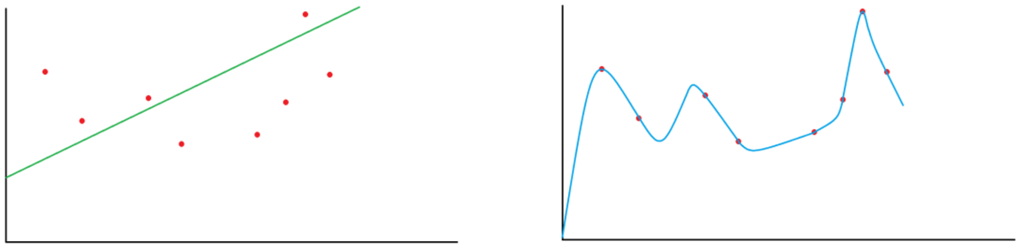

Figure 1.2. *Demonstration of the principle of trade-off between bias of an estimator and its variance. Assume we want to fit the measurements (red dots). Left a fit is made using a straight line (lowest order), right a polynomial of very high order is used to fit the data exactly. The straight line is very rigorous, and will likely fit almost any set of measurements, but at the cost of explaining little of the data. The high order curve fits this data exactly, but will show a large difference between data and model prediction for other, related data sets. Both are extremes that do not present the best option for the trade-off; this option is somewhere 'in the middle'.*

Between these two extremes there is a trade-off between the bias of the estimator and its variance. Indeed, ideally a model is developed such, that the error resulting from the bias and the variance is minimal (Ljung, 1987). This is, because the bias decreases when the model complexity increases, as the model structure becomes more flexible and fits to the data easier. On the other hand, the uncertainty at some point increases also, as there are more parameters and thus the variance increases. More complex models are more flexible and can be made to fit more sets of data. Reversing this argument, there will be no unique complex model that fits a given set of data. Rather, there will be many complex models that can fit the same data (Kooijman, 2000). Also, when using standard calibration objective functions like least squares it is common to find more than one minima (*i.e.*, for the same model several sets of dissimilar parameter values give a good fit), a phenomenon known as 'equifinality' (Beven, 2006). As such, the 'goodness of fit' is not a good aid for model selection, as it favours over-fitting (Myung, 2000; Pitt & Myung, 2002).

Several automatic criteria for model selection have been developed based on the trade-off between bias and variance to find better balanced models, such as the Akaike Information Criterion (AIC, Akaike, 1974). The AIC was developed primarily for statistical models, but has been applied also to first principle models. The AIC considers not only the goodness of fit, but also penalizes the use of more parameters. As such it provides a trade-off between model complexity (expressed as number of parameters), and the goodness of fit. The AIC may be viewed as a statistical formalisation of 'Occam's razor' (Hipel & McLeod, 1994). A thorough and general overview of measures and approaches for model selection and model averaging is given by Claeskens & Hjort (2008), and a brief overview is given by Bierens (2006). A review and comparison of several commonly used model selection tools in ecological modelling is provided by for instance Johnson & Omland (2004), and Ward (2008). The AIC will be revisited in some of the cases presented later on in this report (see Chapters 4 and 5).

1.3 A broader scope: the concept 'equilibrium'

The above-discussed use of statistically based, automated model selection criteria such as the AIC is widespread, but is also subject to serious limitations.

First of all, the model 'complexity' has a strict interpretation in the above-mentioned context, namely it is defined as the number of parameters (or 'degrees of freedom', where each parameter is the coefficient to a polynomial term). This 'definition' can be criticised in more than one way. Not only is it (implicitly) assumed that all parameters are independent, which is certainly not necessarily the case, but also does it ignore complexity of models (and databases) in the context of describing processes and interactions. Many people will have some idea of 'complexity' in terms of the 'size' of the model, the number of processes and interactions that are incorporated. A broader definition of 'complexity' is therefore required, and we will provide one later on, in the next section.

Secondly, it is not *a priori* obvious that the data to which the model is fitted is also the 'correct' data in view of the application. Maybe not all data are equally important, or maybe not all processes and interactions in the model are equally supported by data, while they play a significant role in the model because of the application area of the model. For instance, consider a model that makes predictions of water levels. Now, consider an application of this, for instance, the water level predictions are used to estimate the risk of flooding. For the application only the high(er) water levels are relevant, and the precision of the timing of these high levels. There is thus no need to put much effort in getting the lower water levels right. In this case, it is quite likely that the model which is suitable for the application is at odds with the model that is selected using a traditional model selection criterion. After all, traditional selection criteria focus on fitting the data, but we do not necessarily wish to fit all the data as closely as possible.

The flood example suggests that the use of a model selection criterion like the AIC should perhaps be adapted to fit application requirements. For instance, more weight could be assigned to the data on high water levels than to the data on low water levels. In this context it is also interesting to mention the so-called 'cost of complexity' (Hjalmarsson, 2009), which is a measure for the minimally required experimental effort that has to be invested to identify or calibrate a system. The measure provides the user with an aid to make a trade-off between increasing the experimental cost on the one hand, and on the other hand decreasing system identification abilities when trying to select an adequate complexity for the model. After all, to identify a system over its full range the costs are many times larger than when only part of the range of the system is considered. For many applications the latter option may be perfectly acceptable. In view of this it is of interest to mention the use of alternative objective functions for the calibration of hydrological models. Some criteria emphasize modelling the peak flow, while others stress the base flow (Gebremeskel *et al.*, 2005).

More recently more methods differing from the classical model selection criteria have been developed. We briefly mention only two here. First, there is the method referred to as '*landscaping*', demonstrated by Navarro *et al*. (2004) to evaluate the distinguishability of models (note: for the AIC this is done by calculating the so-called *Δ*-values, Drury & Candelaria, 2008), and also to evaluate the information content of data in distinguishing between models. The idea is that for a suite of models the output is compared pairwise. It is evaluated how much one model can reproduce the output produced by a second model, and vice versa. If a less complex model is capable of reproducing all the output from a more complex model, the less complex model might completely replace the more complex model. Landscaping has been developed as 'local model analysis' by data-fitting, and it suffers from three obvious problems:

- most data contain lots of sampling errors, obscuring information and trends;
- some experimental designs may not be very informative (e.g., too few variables are measured);
- even very good fits may not distinguish between different candidate models.

Landscaping aims to include the global picture, stepping away from the particular data set and instead focussing on model behaviour.

A second method is discussed by Crout *et al.* (2009) in a paper with the promising words "is my model too complex?" in the title. The method is aimed particularly at process based models, contrary to the AIC and similar information criteria. The general problem of process based models, in contrast with statistical models, is that usually two or more considered alternative models are non-nested. This makes it more difficult to make straightforward comparisons (but we will evaluate test cases later on where we do have nested alternative models). The idea of variable replacement is to systematically reduce 'complex' process based models by replacing variables with constants (i.e, keeping the variable fixed), while fitting to the same data, thus creating a suite of nested alternative models. The method is a further development of earlier methods, of which the references can be found in Crout *et al.* (2009). Two possible limitations are that not all possible alternative models are considered (only a sub-class of the nested ones), and that the choice of constants is somewhat arbitrary. Sensitivity analysis might be helpful in this matter.

While both methods indirectly consider a broader scope than the statistical automated model selection criteria, they still suffer from two issues:

- They are not particularly application-driven, while the role of the application should be more pronounced. This specifically holds for the previously mentioned applications, namely investigation of the effects of policy measures, for which models need to be able to give predictions, or databases need to provide information or monitoring abilities, within pre-determined margins of errors.
- They use a single criterion, while in our view any serious evaluation of complexity should consider multiple criteria.

Obviously, both of these objections are stated rather boldly, but we wish to emphasize here that 1/ models (and databases) in general tend to become too complex, and 2/ that any approach to make progress in terms of creating 'better balanced models' needs to consider a broader definition of 'complexity', and needs to explicitly consider the application. This also almost inevitably means that multiple criteria have to be considered, as any trade-off is between more than two measures, i.e., we have to consider more than for instance just the bias and the variance. A model that has been selected based on support from data, and considering model complexity, still does not have to be a usable or relevant model for a given application.

The same idea also applies to databases. Databases, containing for instance characteristics of a landscape, will contain errors. These errors may not be relevant when considering some levels of aggregation of data, for instance, when one wants to know the total amount of forest area. However, for an application that requires zooming in at a much smaller scale, for instance when one considers the forest area in some area X, these errors may lead to large uncertainties. Say, there is 95% accuracy in determining a right land use type. The data for the area of interest might be in the 5% that is wrongly characterized. Such an application would then require a larger effort to get the database right. On the other hand, one can also decide to lower the efforts on getting the data right in the database grid outside area X, and instead just focus all effort to area X.

The eventual goal is to gain confidence in the predictions, information, or monitoring abilities by models and databases. For that it is important to find some sort of 'equilibrium' between the complexity of the model or database, the quantity and the quality of the data, and the requirements from the application area. In a previous publication this 'equilibrium' was defined as the result of a trade-off between the complexity of a model or database, the availability of data, and the intended applications of the model (Bogaart *et al.,* 2011). A model or database in this context has to provide projections within some reasonably accepted uncertainty bounds. For that, sufficient understanding of the system is required, and coupled to that, a certain amount of complexity of the model or database. In turn, this complexity requires a certain quantity and quality of data for support. A more complex model may still give projections within the pre-set uncertainty bounds, but with the less desirable property of having increased 'costs', such as larger data requirements, a longer run time, or being a model that is more difficult to calibrate. A more complex database will give projections within the pre-set uncertainty bounds, but will also have a worse than necessary performance per query (SELECT-statement, storage procedure). Model or database complexity, linked to the understanding of the system, the quantity and quality of data, and the requirements of the application, are in 'equilibrium' when projections can be made within accepted bounds of uncertainty.

1.4 Defining complexity, and finding 'equilibrium'

To make the concept of 'equilibrium' operational we have to provide some definition of 'complexity'. The term 'complexity' can have related but different meanings depending on the context in which it is used. The most limited, classical definition in model selection considers complexity to be the number of fitting parameters. However, many (partly) deterministic models in environmental sciences include various possible states, boundary conditions, non-linear interactions, and other elements that might be considered also (Wagener *et al*., 2001, and references therein). In computational terms, complexity may point to the amounts of resources required to execute algorithms. There is ample literature considering problems of complexity classes, such as P, NP, etc. (for instance Blum, 1967). This aspect is also relevant for applications, as many applications desire a small running time. Also of some relevance is 'complexity' in the sense of programming complexity, which is a measure of how the various elements of a program interact. It is possible that a model or database is conceptually simple, yet the numerical implementation is highly complex, for instance because it was badly programmed, or programmed in an ill-suited computer language, or it requires this complexity for some reason. This latter is not unrelated to Kolmogorov complexity, expressing a program as a string which is the length of the shortest binary program that outputs that string (Blum, 1967). In information processing 'complexity' refers to the total number of properties that is transmitted by an object and detected by an observer. Furthermore, 'complexity' may refer to output complexity, i.e., the behavioral complexity of the model. Phenomena like chaos come to mind then.

None of the above definitions is especially suited for databases. For databases, complexity of the numerical implementation and complexity of the design should be considered. Any definition should include the possible states of the grid cells, which are often nominal in nature, for instance, cells are assigned labels like 'forest', 'water', 'roads', etc. In view of Kolmogorov complexity the number of cells itself is also an element of complexity. Imagine something like a chessboard where cells can only be black or white: Although the number of states is only two (equal to 0 or 1 in bit strings), the number of bits increases with the number of cells in the database. The permutations of the order in which input files are used in the creation of the database also adds to the complexity: When many input files are used, depending on how essential the input order is for the final content stored in the data base, there is a lot of room for possible conflicts and bugs in input order.

For the use with regard to 'equilibrium' a bit loose definition of 'model complexity' is suggested that is motivated from a practical point of view, and that focusses on the applications of models and databases. Complexity encompasses the number of states, grid cells, parameters, input order permutations, non-linear interactions, and boundary conditions, and the efficiency with which the code has been built, the number of interactions between modules, and the amount of resources required to run the model. This way, also non-typical 'models' like databases, expert knowledge systems, and cellular automata are taken into account. A model or database that is in equilibrium has a complexity that is well-suited for the application(s) it is used for, and is supported by sufficient data of adequate quality.

Now that we have a rather qualitative definition of the concept 'equilibrium', the question now is how to evaluate whether a model or database is close to equilibrium or not. What criterion or criteria would be needed for such an assessment? To this end a prototype evaluation list model complexity was developed in a previous publication by Bogaart *et al.* (2011). This list (translated from Dutch to English) can be found in Annex 2, and contains questions about the model formulation, the data, the goal and the application area of the model or database, model calibration, validation, uncertainty or sensitivity analysis, and more. The list contains many questions, and consists of two partly overlapping sub-lists, of which one contains very abstract questions. As such, the list does not have a very user friendly set-up, and it is unclear whether or not it is capable of fulfilling the goal the list was developed for, namely the 'equilibrium' evaluation of models and databases.

1.5 Goal of the study

The goal of the study reported here is to test how relevant and appropriate the evaluation list EMC v0.1 proposed by Bogaart *et al.* (2011) is for the above-mentioned purpose it was developed, and whether or not it is usable. Furthermore, it is investigated how the list can be combined into one single list, instead of two separate sub-lists, and how it can be made more user friendly. To this end, the list has been tested in two ways:

- It has been subjected to expert review, and
- It is applied to several test cases.

In the review of the list the following questions are relevant to determine its performance:

- What are the key issues with regard to 'equilibrium' and 'model complexity', and does the list contain these issues (is it appropriate)?
- Does the list have a logical set-up?
- Is the list useful?

When test cases are used, other questions are used to determine the performance of the list:

- Which questions on the list were answered in test cases (and which ones not)?
- Could all questions on the list be answered?
- If a question on the list could not be answered, why not? Was the question unclear, or irrelevant, or was there simply missing information in the case study? If a certain question continues to be left unanswered, this may mean that the question is irrelevant. However, it may also point to a 'blind spot' in modelling practices in general.
- If unclear, how should the question on the list be rephrased?
- Should there have been additional questions on the list? Which?

The combined results of the expert review of the list and the application of the list in test cases should reveal whether or not the performance of the list is adequate. Also, the results are used in this study to develop and present an upgraded version of the list, namely EMC v1.0.

1.6 Document set-up

This document is set up as follows. In Chapter 2, the results of the expert review are discussed. In Chapters 3, 4, and 5 we discuss several test cases from the literature which have been subjected to the list. In Chapter 6 the combined results are discussed in order to update the list. The updated list is presented in Chapter 7, and the direction of future work is discussed in Chapter 8. Two appendices (one on the background of bias in estimators, and one containing the translated version of the old evaluation list EMC v0.1) complete this study.

2 Expert review

2.1 Philosophy behind the list

The original prototype evaluation list model complexity version 0.1 is given, translated into English, in Annex 2. The list follows the modelling cycle, asking questions about all phases in the cycle. The aspect 'scale' is treated separately, as this aspect is very general and very important. The list consists of two sub-lists. One sub-list contains questions that are related to the questions from the status A evaluation list. This latter list is used for the quality control of models and databases at the Wageningen University & Research centre (Wageningen UR). It primarily contains questions about the documentation. However, it also indirectly exposes a lot of information about model concept, assumptions, data, validation, verification, etc. that is useful for the evaluation in terms of equilibrium. Therefore, these questions have been incorporated in the prototype of the list for model complexity. The other sub-list is composed of questions that were deemed more relevant for the evaluation of equilibrium, but it is sometimes also less obvious how they should be answered. For instance, a question about the documentation of how the model was calibrated is easily answered by writing down the procedure. However, it does not tell you whether it was a 'good' procedure for calibration or not, which is the real question that should be answered to evaluate the equilibrium.

The goal behind the development of the list EMC v0.1 is to come up with an aid for modellers and users of models and databases in obtaining an sufficiently balanced model or database for the intended application. For that, the list should in the first place consist of questions that are *relevant* for the evaluation of equilibrium. Relevance implies that if the questions cannot be answered then equilibrium cannot be determined, i.e., the question addresses an aspect that bears relevance for equilibrium. This also means, that non-relevant questions should probably be removed, as they only distract the list user from the real point (making the list more complex than required, which would be contrary to its own philosophy, as a matter of speaking). It may be obvious to the reader that the relevance of a question can be indirect, as in the above-mentioned example on calibration. Secondly, it should be considered whether or not all relevant questions are present, that is, when there is a missing aspect overlooked in the current list, the list should be supplemented on this point. Thirdly, an aspect that is in practice equally important as relevance, is utility. It should be rather obvious to the user what is intended by the questions, and how they should be answered satisfactorily. If this is not the case the list cannot serve its purpose properly.

2.2 Expert review set-up

To test the relevance and the utility of the prototype evaluation list in Annex 2, the list was distributed together with an early version of the publication by Bogaart *et al.* (2011) among several experts in the field of modelling, with varying backgrounds like theoretical ecology, systems engineering and applied mathematics. Beforehand the reviewers they were asked to evaluate the list, and in their evaluation pay specific attention to a few points:

- Is the general set-up of the list logical, and is it in line with current thoughts on the modelling cycle?
- Is the list useful in its current form for the intended goal?
- On what points could the list be improved (addition or deletion of questions, clarity of questions, etc.)?

Feedback was provided in a small series of interviews with the reviewers after they had read the material, and they were interviewed on their thoughts on the subject of model complexity. The main results from the interviews have been summarized below.

2.3 Results and discussion of expert review

In general, the reviewers have judged the overall approach of the list positively, and they indicated that the strategy is logical and in line with current theory on modelling cycles used in several scientific fields, like for instance Systems Biology (Kitano, 2002; Chou & Voit, 2009) and hydrological modelling (Wagener *et al.*, 2001). The utility of the questions has been judged with mixed responses, and several critical remarks have been made. Below we mention some raised remarks and points of criticism in more detail. Furthermore, we indicate after each remark what we will do or have done with it.

1/ The approach seems to be logical. Modelling cycles are in essence the same, which makes the approach of this list a 'natural' one.

This suggests that the list should keep a format that follows the general modelling cycle.

2/ The term 'model complexity' implies that a model or database is necessarily complex, but models do not have to be very complex to successfully describe the data, for instance, power laws, linear reservoirs, and empirical models constructed following the philosophy of parsimony, such as time-series models (e.g., Box and Jenkins, 1970; Hipel and McLeod, 1994).

The point is taken, but it is difficult to come up with a different term to replace 'model complexity'. Furthermore, we expect that most models and databases are too complex for their intended applications instead of too simple. In the updated list EMC v1.0 we ask people for their judgement on the complexity of the model at the end of the evaluation, but maybe in an updated version also one of the first questions should be: "At the start of this evaluation, do you think that your model is too complex?"

3/ There are many different definitions of 'complexity' (> 30). This should be included in the list somehow.

The subject of defining 'model complexity' has been included more explicitly in the previous chapter of this study. It has not been included explicitly in the updated version EMC v1.0 of the list (see the last chapter). It is a valid point though, and an update of version 1.0 should maybe include an explicit question: "How do you define 'model complexity'?"

4/ Complexity is not necessarily a feature of the model or database itself. Model complexity can also be a feature of the behaviour of the model (for instance, chaotic behaviour). The complexity of the control is also an aspect.

This is a difficult point. We agree that even 'simple' models may generate 'difficult' output (for instance, dynamical models capable of 'tipping points' or 'chaos'). The point is considered in the previous chapter of this study (and see the previous question), but at the moment we are unsure if and how to incorporate this aspect in the list.

5/ The best way to deal with equilibrium is probably to start simple, then increase the model complexity gradually. The other way around, which would be model reduction, is generally not assumed to be a process with a high probability of finding a model in equilibrium, although this is not an unanimously supported point of view.

The approach of "forward selection" is from simple to complex. The reverse direction (from complex to simple) is sometimes used – it is named "backward elimination" (as is used in Crout *et al*., 2009, discussed in the previous chapter). A combination of the two is a third approach, an currently also "all-subset regression" is used, in which all candidate models are considered. In practice, however, most models are follow-up versions of previous models or model versions, which contain additional modules with more details, processes, data, etc. and have been suited to be used for more applications. This suggests that in most cases there will be a gradual increase in complexity, and that the history of the model should be considered in the evaluation. For this reason we ask not only for the name, but also version number and date of publication of the model. For a fuller equilibrium assessment it may even be required to consider a whole series of evaluations, one for each applied version of the current model and of its predecessors, and then evaluate this whole historical trace.

6/ System analysis is comparable to software engineering. In that sense a conceptual model or database design should already be built in a modular way.

This point can be considered when making an evaluation of a model. The list at the moment is intended as an evaluation tool, and we do not deem it prudent to suggest how people should build their models based on the list.

7/ Extending on the modular design: it is furthermore suggested that maybe equilibrium is possible in each separate step of the modelling cycle (system analysis, conceptual design, etc.).

In each question on a modelling step in the updated list EMC v1.0 the current state of affairs around that model step is challenged against the application of the model and the model requirements for that application as deemed fit by the evaluating people. For instance, a system analysis is asked, then it is asked which aspects of the system analysis bear a great relevance for the application and which ones not, and then it is asked whether these aspects have been included or not. As such, the updated list is intended to evaluate 'equilibrium' for each step of the modelling cycle, as far as it is possible.

8/ Equally important as its relevance is the utility of a question, that is, a question is only as good as it is posed. An ill-posed question cannot reveal the information that you desire to obtain. A question like "Is X true?" is very relevant in a modelling case, but can probably never be answered, not so much because it is difficult to obtain the answer, but rather because the question is very ill-defined. After all, what is 'the truth'? A question like "Has been model been calibrated?" has a higher utility.

This aspect has been largely ignored in the early version (v0.1) of the list. After all, the initial work was aimed at gaining insight in the subject of 'model complexity'. In the new version (v1.0) it has been attempted to formulate questions that are clear, useful, and relevant.

9/ It has been suggested to talk to game designers, and investigate how they perceive 'complexity'.

The suggestion has been duly noted, and the background search has been extended to include the software modelling cycle and definitions of 'model complexity' oriented more towards software.

10/ The use of a wiki-system has been mentioned. In such an environment it is likely that the best product will eventually appear and gain popularity. Obvious downside with such a system is costs, and there needs to be a supporting structure for sufficient interactions between the modellers and other stakeholders in the wiki-system.

In practice we find that many model developers compare their models with models from outside the Netherlands on similar subjects. Also, several models have been published in peer-reviewed literature. This is already in line with the 'wiki-system'. On the other hand, it has been suggested that time and financial pressures form a force in the opposite direction, which favours the 'quick and dirty'. Another point is the subjectivity of what is the 'best' product: people might judge different models to be the 'best'.

11/ Do not consider only model components, but also in- and output, and possible states of the system.

These aspects are considered more explicitly in the updated version EMC v1.0, and in the employed practical definition of 'model complexity'.

12/ Although it is a common assumption, parameters are never fully mechanistically based, nor ever fully independent. They can have a more physical basis because of confidence (e.g., the gravity constant). This is a note of awareness especially when it comes to conceptual models, but also when model complexity is expressed as the number of parameters, which is a common approach.

The remark refers to a phrase in WOt-werkdocument 226 by Bogaart *et al.* (2011) on the distinction between experimentally determined parameters and 'true' parameters like the gravity constant. Of course it is correct that at some level all parameters are determined experimentally or by fitting to data, even the gravity constant and such. The note on the caveats of defining 'model complexity' in terms of the number of parameters is certainly supported by us, and is actually one of the motivations why automatic model selection criteria such as the AIC are problematic in our view.

13/ What are 'free' parameters? How can you gain access to these parameters?

The issue is relevant when it comes to calibration. In the updated version EMC v1.0 we explicitly ask for references that clearly explain the calibration, and for a discussion about the demands of the application and the possibilities by the available data with regard to the calibration. For instance, Fisher Information (Fisher, 1922) tells you how well you can estimate parameter values from data.

14/ A decrease in variance of the bias in criteria for model selection (like Akaike's Information Criterion) does not have to be (nearly) linear. Also, there is not necessarily a clear optimal model order. In a similar fashion, there might not be a clear equilibrium for all models and databases.

The first remark specifically points to an early version of an accompanying WOt-paper, but the general note is one to consider. We explicitly assume that there is such a thing as 'equilibrium', or at least that a model or database may be 'better' or 'worse' in view of the application. Also, we believe that in practice the application of the list will lead to improved models and databases in view of the application.

15/ Models can be used as aids for data acquisition, instead of the other way round (data being used for building and calibrating models). Systems & Control theory is focussed on experimental design, in which the question is important which types of experiments would provide the most insight in the identification of parameters, assuming that the model is correct, or at least applicable.

This point may be considered when evaluating the model or database. However, we are currently unsure if and how to incorporate this explicitly in the list.

16/ A sociological aspect is the following: giving the model or database a name often leads to an increased conservative attitude. An unnamed model or database is more likely to be discarded and replaced by a new concept than a named one.

It is difficult to ask about such things directly in a checklist. However, by asking explicitly about the motivation behind the choices and assumptions made by modellers this aspect may be exposed. It is most likely an aspect that will pop up when considering the history of a model, and again, this is not something that is included in the current version of the list. The updated version EMC v1.0 explicitly asks for version number and date. Perhaps a future updated version should consider the history of a model more explicitly, as discussed earlier at remark 4.

17/ The application plays a role in almost every aspect of the modelling cycle. It is therefore not obvious to use a model or database for a different application. Of course, be mindful what is understood by a 'different' application. For most new applications a new model design is far superior to model adaptation.

The previous version of the list EMC v0.1 is not focused very much on the application(s). This point has explicitly been included in the updated version EMC v1.0. In many questions of the new version the application(s) of the model are explicitly considered.

18/ A point of dissent among the reviewers: One point of view is, that a model does not have to be 'good', it has to be sufficient for the (intended) application. A different view is that a 'good' model, based on the right first principles, should perform well in many different applications. Actually, both views seem to be supported by real examples.

We can find merit in the different views. It is something that will probably depend on the model or database under investigation. It is difficult to explicitly put this as a question on the list, but it will be a point of discussion in any evaluation.

19/ Be mindful of the effects of errors caused by the implementation of numerical methods. The errors caused by numerical methods can be larger than those caused by data (Clark & Kavetski, 2010). Also, there likely exists an optimal type of numerical method for any implementation. For instance, a rigorous numerical method like forward Euler is quick, but introduces a large error when the dynamics of the system change quickly (in a way this is analogous to bias error). Very sophisticated methods give much more reliable predictions, but become very slow. Depending on the application, this is perhaps not what you desire.

The issue seems important enough to explicitly put in a question (question 7C in EMC v1.0, see the last chapter).

20/ It was suggested the list might be applicable for model trains. The distinction between model and model train is in some way non-existent. It does not really matter whether a stream of data goes from statement to statement, subroutine to subroutine, or model to model (or database).

This point will be considered in the follow-up of the project. See Section 8.3 (Prospects and suggestions) of the last chapter.

2.4 General discussion

Overall, the expert review has led to useful suggestions for improving the list and the ‘philosophy’ behind it. Several remarks can be used directly for the formulation of questions or change in set-up of the list, while others indicate less directly that the list needs improving. Most notably there should be (more) questions in the list on comparing several aspects of the modelling (cycle) with the application(s) of the model or database.

Some remarks have been acknowledged, but have not (yet) led to changes in the list. Most notably are the questions that seem to indicate that the history of the model or database should be included. Specifically remarks 4 (gradual increasing of model complexity) and 16 (the ‘Concorde’ effect) suggest that it is necessary to have knowledge about the previous versions of the model. However, this aspect has not been included in either the prototype EMC v0.1 or the updated version EMC v1.0.

3 Test cases on yeast and beer-brewing

In this chapter, we discuss modelling and expert knowledge for the application of yeast in beer-making. For the testing of the ECM v0.1, we have to determine how the list performed. For that, an inventory should be made which questions were answered easily, with difficulty, or not at all. An explanation or some extra information should be given with all these answers. Furthermore, it should be considered if there were any issues missing. These should be indicated also. In the test cases below, we apply the prototype list model complexity as model evaluation aid. We loosely follow the order of questions on the list. In the last sub-section in each test case, we give an overview of which questions were answered how in each case study. For the explicit questions, see Annex 2.

3.1 General background

The several test cases in this chapter are concerned with different models and literature on the application of beer-making. Most models have different goals and application areas, but the main focus is on the production of ethanol in beer by yeast. This subject obviously has a great commercial interest, and thus there is a large amount of literature available with information and data about the biology of yeast and the application (> 1 million publications in Web of Sciences when combining the terms "*S. cerevisiae*" and "beer"). Several different models have been published in journals concerning yeast growth, with different levels of detail.

Beer is brewed in different steps in a batch process. One of those steps is the fermenting stage, in which many of the substances are produced that give the beer its colour, its flavour, and its scent. The most important substance is ethanol, obviously. Ethanol, CO_2 and other substances are produced by budding yeast (*Saccharomyces* sp.), of which the baker's yeast (*S. cerevisiae*) is the best known. This species is one of the most studied model species ('model' in the sense of biological example, not a mathematical model), and is used for example in the production processes of beer, wine, bread, and carbon dioxide injection mechanisms in aquaria. *S. cerevisiae* is an ascomycota, and can reproduce via ascospores. However, the species multiplies mostly via binary fission, via 'budding', a process in which an identical daughter cell separates from the mother cell.

In the production of beer yeast ferments sugars to ethanol, CO_2, and other substances. This fermenting stage takes about one to three weeks at temperatures around 20-25 °C (by the way, in the production of *lager* ale the temperature is kept low on purpose, as other budding yeast species become active then and produce different substances, which give the *lager* its specific qualities). Colonies of yeast grow to full size within two to three days. During this time, the yeast passes through a series of different physiological states. There is a lag phase, a phase of growth, accompanied by high ethanol production rates, and an end phase, in which ethanol production decreases again. The ethanol production stops before all sugar is consumed. Typically, some 2-10 g/L residual sugar remains in the beer, depending on the 'gravity' (sugar content) and the physiological condition of the yeast at the start of the fermentation (Guimarães & Londesborough, 2008). To maintain a high conversion from sugar to ethanol certain spore elements are added. Baker's yeast forms biofilms, and often the same colony of yeast is used for several subsequent fermentations. When not used for fermenting, yeast is 'cold-stored' at 3-4 °C.

During beer production there are regular malfunctions, in which temperature and/or oxygen inflow decrease. Temperature differences are about 10-20 °C, while differences in oxygen pressure are far greater (for a review on yeast responses to industrial brewery handling stresses, see Gibson *et al.*, 2007). Yeast can metabolize sugars without oxygen, but this results in the production of different (quantities of) substances, which is undesired from the commercial point of view. The temperature differences may be more problematic for the yeast itself, as some yeast species like *S. cerevisiae* become inactive at too low temperatures ('cold shock'). For industrial applications it is relevant to device models that can be used for optimization and control strategies. During fermentation, temperature and oxygen supply are constantly monitored 'on-line' and, if needed, adjusted. We consider some of this below.

Before starting with the cases, we discuss some basic models that are used frequently in ecology to describe the population dynamics of species. Conceptual and numerical models can vary from small deterministic models to large agent-based models (Shachak & Boeken, 2010, and references therein), but the Monod model (Monod, 1949) is the most commonly used unstructured model for yeast growth. The model considers a growth limiting substrate. Baker's yeast is a so-called 'supply' organism, i.e., it will always grow and propagate as fast as allowed by the most limiting resource, a notion named 'Liebig's law of the minimum'. The Monod model is given as

$$\frac{dX(t)}{dt} = \mu\, X(t)$$

$$\frac{dS(t)}{dt} = -\frac{1}{y}\,\mu\, X(t)$$

Here $X(t)$ is the population biomass size (varying in time), and

$$\mu = \mu_{max}\,\frac{S(t)}{K_s + S(t)}$$

Here μ is the specific growth rate, μ_{max} is the maximum specific growth rate, $S(t)$ the substrate concentration, K_s the substrate saturation constant, and y is the yield (which is assumed to be constant). The model is an alternative formulation of the well-known *Michaelis-Menten* kinetics. By introducing a constant $C = X(t) + y\,S(t)$, where $C = X(0) + y\,S(0)$, mass balance is introduced, which allows the reformulation of the above equations into one equation

$$\frac{dX(t)}{dt} = \frac{\mu_{max}\,\big(C - X(t)\big)X(t)}{y\,K_s + (C - X(t))}$$

The Monod model lacks consideration of either maintenance costs for cells (Pirt, 1965) or reserves, the storage of nutrients to have some form of internal control by the organism over the growth rate (Droop, 1983), and the model has been criticized for this (Kooijman, 2009). Furthermore, the unstructured nature of the equation excludes spatial variations, and any specification of the biochemical networks behind the growth and production of ethanol, carbon dioxide, etc., and the cell cycle (among others, Kurz *et al.*, 2002).

Probably the first model describing the population dynamics of a species is the well-known Verhulst-Pearl equation. The logistic model is a limit case of the more general Monod model. It is given as

$$\frac{dX(t)}{dt} = r\,X(t)\left(1 - \frac{X(t)}{K}\right)$$

This equation is a differential equation (Verhulst, 1838), and assumes the population size $X(t)$ to be an autonomous function of time. The parameter K is a parameter described as 'carrying capacity'. The carrying capacity gives an upper limit to the population size. The next parameter r is the 'intrinsic growth rate'. Together with an initial condition $X(0)$ this model has a total of three parameters. Growth first intensifies, but as the number of interactions between the different individuals increases with increasing population size, the growth rate decreases. After some time mortality rate and recruitment rate are of equal size but opposite, and the population size has reached a 'steady state' value $X^* = K$.

3.2 Control strategy model

3.2.1 Point-by-point case evaluation

The first model we consider is by Kurz *et al.* (2002). They provide much criticism on 'simpler' models for yeast propagation for use in brewing applications. Alternatively, they propose a comprehensive kinetic linear stoichiometric model that includes several metabolic reactions.

Goal/Application area

The model represents a basis for a control strategy, with the aim of providing optimal inoculum at the starting time of subsequent beer fermentations. The performance of yeast in terms of fermenting time and beer quality in the fermenting stage has been found to be influenced significantly by the state of the yeast before inoculation. This is not a strange notion, as the propagation rate by yeast will be determined by the energy/nutrient storage (much like a seed of a plant or the yolk of an egg). Variations up to 2-3 days can occur in the brewing production plan, which must be taken into account. Therefore, active control over the yeast propagation is desirable to achieve optimal organism 'quality'. The first goal is to develop a metabolic modelling approach for brewing yeast propagation including limitation and inhibition effects, especially the Crabtree-effect, and validate this model with experimental and literature data. The second goal is to evaluate the potential of the model simulations for process control strategies, using case scenarios.

System analysis, assumptions, model concept

Unstructured modelling (like the Monod model) is not deemed sufficient by Kurz *et al*. (2002) to adequately describe environmental disturbances. Structured models include certain levels of detail on cell age, and biochemical pathways, and form the basis for the model. However, the validity of the models is often limited to certain states in which the metabolism is set, like oxidative, or fermentative. Lag times, arising from the adaptation of yeast to the growth medium(e.g., for enzyme production) are also not included. Temperature and maintenance energy should be considered explicitly for the application in beer brewing, as temperature and oxygen supply changes frequently occur. Structured models that include maintenance energy are limited to oxygen-supplied situations. Combined nutrient limitations and ethanol inhibition effects are usually not considered. Ethanol is toxic for yeast in higher concentrations. Nitrogen is essential for protein and biomass synthesis.

Kurz *et al.* (2002) take into account limitations and inhibition effects for sugar, nitrogen, ethanol, and oxygen. Specifically the 'Crabtree-effect' is considered, in which the yeast produces ethanol in the presence of high external glucose concentrations, instead of via the 'normal' tricarboxylic acid cycle. Increasing concentrations of glucose lead to ATP formation via this route, reducing oxygen consumption. Glucose is fully oxidised to carbon dioxide, and 12 NADH/H$^+$ are formed. ATP are formed via oxidative phosphorylation. The biochemical reactions are described by the linear equations as given NADH/H^+ and ATP are neither accumulated in the cell nor excreted. Maintenance (Pirt, 1965) is modelled explicitly by considering the ATP, and non-growth maintenance requirements are assumed to be m_{ATP}=0.007-0.018 mol ATP/C-mol biomass and hour, as given by the literature (see references in Kurz *et al.*, 2002). Yield coefficients are assumed to be constant.

Mathematical and numerical model

The model is composed of a total of 10 components (glucose, biomass, nitrogen, dissolved oxygen, carbon dioxide, water, ethanol, glycerol, NADH/H^+, and ATP), and 8 reactions (see Tables I and II in Kurz *et al.*, 2002). The model is a linear state space formulation

$$r = A\,v$$

Here r is the 10-dimensional vector for the turnover of the single substances, A is a 10-by-8 stoichiometric matrix, and v is the 8-dimensional vector for reaction rates. The full model has 10 equations and 16 unknown rates, i.e., there are 10 variables and 16 parameters.

Kinetics for substrate and oxygen uptake are expressed in Monod terms. The uptake is modified by glucose (substrate), nitrogen, temperature, and ethanol. Ethanol inhibition is modelled by a term for non-competitive inhibition. Nitrogen limitation is modelled by an additional Monod term. Temperature dependency is included by introduction of a coefficient, that is multiplied by the specific uptake rate. As Monod kinetics are only suitable for describing the exponential and stationary states of the propagation, the lag times for adjustment to new media is described by a sigmoid function. The switch from purely oxidative to partly fermenting behaviour is modelled via a switch function. Oxygen uptake rate is also limited by ethanol and nitrogen availability, and the kinetics for the specific oxygen uptake rate is modelled as another switch function. Energy demand is assumed to be linear.

The model is implemented in AQUASIM 2.0 (see Kurz *et al.*, 2002). As this is a specific numerical tool for modelling reactions, the verification will likely have occurred in terms of checking if all reactions were written down correctly, but this is not explicitly mentioned. Metabolic turnover rates are calculated as the sum of the turnover in the single reactions. Stoichiometric coefficients are computed from the balance of single elements. Several coefficients are calculated using the known yield coefficients for purely aerobic or anaerobic growth on glucose, and purely aerobic growth on ethanol. Stoichiometric and kinetic parameter values are given in Table III (Kurz *et al.*, 2002), including references.

Data requirements

Most experiments in the beer-making literature use full media for growth to avoid nutrient limitations. For the model this is not a valid approach, as the growth medium is beer wort. This medium contains high concentrations of sugar (100 g/L), which leads to the Crabtree-effect. Temperature and dissolved oxygen are examined in the data. Data from the literature used for setting most stoichiometric and kinetic parameters are listed in Table III (Kurz *et al.*, 2002). Data for calibration and validation are discussed in more detail below.

Calibration, analysis, and validation

From the paper, it is not entirely clear which of the sources and data were used to calibrate the model, and which were used to validate it. Also, it is not reported what calibration process was used. The metabolic model was compared to literature data from baker's yeast batch propagations (references in Kurz *et al*., 2002). A sensitivity analysis indicated that yield coefficients, maximum specific oxygen uptake rate, temperature coefficient and lag-time were influential parameters. Only the latter three parameters were included in the parameter estimation; yields and parameters with low sensitivity were assumed to be fixed.

Simulation results are compared to experimental data for biomass, substrate and ethanol concentrations. In Figure 1 (Kurz *et al.*, 2002) simulation results are compared to glucose and biomass concentration measurements at different temperature settings. Temperatures are 10, 15, 20 and 25 °C for constant dissolved oxygen (0.5 ppm), values are given in Table IV (Kurz *et al*, 2002). For each temperature measurements are at 5-10 hours intervals, and Table IV suggests there are 4 replicates per temperature. The replicates per temperature setting have been clustered to mean values with standard deviation, and the simulation runs have been compared to these mean values. The measurement error on yeast count was assumed to be 5%. For the validation it was said "that the simulation represents the reference values accurately" (Kurz *et al.*, 2002), but no quantitative results are given (e.g., like an R^2).

For the second goal (process management) the dependency of the propagation on the commonly manipulated variables temperature and dissolved oxygen concentration has to be described. The relationships are transferred into the process model by formulating the parameters as functions of the manipulated variables. Except for one source (reference in paper), Kurz *et al.* (2002) could not find reliable data for the description of temperature-dependent industrial baker's yeast growth. The one source used fixed generation times, resulting in wide variations of doubling times, even for one temperature setting.

Figure 2 (Kurz *et al.*, 2002) gives half logarithmic plots of the maximum oxygen uptake rate, specific growth rate, and temperature coefficient versus the inverse temperature (K^{-1}) of all their experiments (21 data points counted). An Arrhenius approach with activation energy E_a = 84.801 kJ mol^{-1} was used for the description of the temperature dependency. It is furthermore suggested that the activation energy does not depend on the growth medium or specific yeast strain. A similar approach is reported for estimating the temperature coefficient and maximum specific oxygen uptake rate. Experiments with different dissolved oxygen concentrations, varying from 0.1 to 0.8 ppm at 15 °C, exerted a minor influence on the biomass growth.

For the validation of the use of the model for control strategies, case scenarios were considered. Figure 3 (Kurz *et al.*, 2002) provides two scenarios in which variations in the temperature occur (oxygen fixed at 0.5 ppm), and one in which a significant drop in oxygen supply occurs (temperature at 20 °C). The simulation results for concentrations of biomass and substrate were compared with the available data for these three scenarios. Data coverage for the scenarios was ca. 10 measurements during 20 hours. A visual inspection suggests that model and data correspond very well. The reported statistics for the first scenario is a mean deviation of 8% of the simulation points from the data, with a standard deviation of 3.9%, and for the third scenario 9.3% mean with 6.4% deviation.

The first and second case scenarios involved intended changes in temperature/oxygen supply. Indeed, a lag in growth was clearly monitored for the scenario where the temperature was dropped for about 7.5 hours.

3.2.2 Discussion of the case

The two reported goals of the model were

- To develop a metabolic modelling approach considering limitation and inhibition effects of industrial wort, especially the Crabtree-effect.
- To demonstrate the potential of the model for process control strategies, specifically to obtain the inoculum time with the best organism quality.

Even if these goals were achieved reasonably, this does not mean that the model is properly balanced. It is still possible that the model is overly complex with regard to what is required for the application.

For the application, it is relevant to consider what the probable ranges of oxygen supply and temperature failures are. The considered temperature range <10-25 °C> seems to be appropriate for the application, as it is intended to model responses to temperature failures, which in practice are always temperature decreases. For the oxygen supply range of <0.1-0.8 ppm> this is less certain. Although the model performed well for the case scenarios, below 0.1 ppm the specific growth rate decreases significantly (70% at 0.5 ppm, Kurz *et al*., 2002). The temperature was fixed in the oxygen variation experiments to 15 °C. This is however somewhat strange when considering that *S. cerevisiae* often experiences cold shock with downshifts to temperatures below 20 °C (Gibson *et al.*, 2007). Only one oxygen setting was used with the experiments in which the temperature was varied, while for the experiments with variation in oxygen the temperature was fixed to an average value. This gives no real support with regard to simultaneous fluctuations, e.g., what happens if the temperature is 25 °C with oxygen at 0.1 ppm? Given the intended application, it seems that the validation is incomplete.

Kurz *et al*. (2002) point out, that the limiting effects of missing trace elements (like zinc) are not considered. It is generally accepted that growth ceases in wort fermentations because of nutrient limitations, but specifically in the case where oxygen is depleted (Guimarães & Londesborough, 2008, and their references). Many yeast strains perform poorly in 'very-high-gravity' worts, like decreased growth, slow or bad fermentation, and low viability (Huuskonen *et al*., 2010). However, in the case scenarios the match between model predictions and measurement was certainly not bad. One can argue that for the application it is acceptable that trace element limitations are not considered, provided the oxygen supply is sufficient. Kurz *et al*. (2002) also indicate, that the model cannot be properly used for temperatures below 10 °C, and thus not for applications like yeast storage (typically occurring at 3-4 °C).

The model does not discriminate between different types of substrate, and hence not between different types of uptake mechanisms. This is intended in a future expansion. Also proliferation state, key enzyme processes, and extensions concerning flocculation in non-stirred propagator systems will be considered. These extensions all appear to be premature. Instead, it can be argued that perhaps the switching mechanisms should be replaced by non-switching descriptions. These introduce additional parameters, but avoid erratic and non-smooth behaviour.

The data support for the model seems to be somewhat limited. Ethanol concentrations are not measured, which seems a fundamental omission given why yeast is used for brewing in the first place. The plots in Figure 2 are lines, but the data show some scattering, especially considering it is on a log scale, while the support and coverage seem to be rather low. The time support of the measurements with temperature variation varies somewhat, and based on Figure 1 is 5/9, 5/15, 7/30, and 9/50 measurements/hours. It was reported that single measurement values were not plausible, hence the simulation runs were compared to the mean values (Kurz *et al.*, 2002). It is not clear if the model could be fitted to separate time series when parameter values were allowed to differ between different time series. At any rate, extending the model will likely make the data shortage worse. Instead, it seems more appropriate to increase the data support by gathering extra time series on other combinations of oxygen supply and temperature within the given ranges.

Overall, the current model by Kurz *et al.* (2002) does not seem to be near 'equilibrium'. The dissolved oxygen concentration should not decrease too much for the model to lose its validity of use. The inclusion of the Crabtree-effect makes sense. Normally ethanol is only produced as by-product of anaerobic conditions (i.e., with low oxygen concentrations), but the Crabtree-effect is the name for the production of ethanol under aerobic conditions, that occurs if large amounts of sugar supersaturate the normal metabolic pathways. Brewing wort typically contains large amounts of sugar. Strangely enough, ethanol measurements are not included in the paper for calibration of the model.

3.2.3 Advice for improvement

To balance the model more with the application of process control, some points for improvement could be considered. The model is rather complex, while many of the variables play only an auxiliary role. On the other hand all equations are linear, with some switches included. This is a strange approach, as the goal was to describe the effects of limitation and inhibition, which are sometimes strongly non-linear. The model probably does not have to contain all the variables or explicit equations for the in-between metabolic pathway steps. The model formulation may be improved by a significant reduction in the number of equations, while including some non-linear effects or relationships.

Data support could be increased in several ways. Explicit ethanol measurements would be a significant improvement, as ethanol is the most important substance of interest produced in the beer-brewing process. The lower boundary of the extent of the oxygen flux should be lowered, to include data on what happens when severe failures in oxygen supply occur.

3.2.4 List fill-out

It is a peer-reviewed paper, and although this might provide a proper format for meta-information (1), there is none. The model purpose (2) is clearly described in the introduction. The conceptual model (3), assumptions and simplifications (4), and mathematical model (5) are discussed in the introduction and throughout various sections of the paper. This is actually often the case in papers and reports, and this is not a real problem. The numerical model (6) is not given, except for the remarks that the equations were implemented in AQUASIM 2.0. There is no proper reference to the package, and therefore, a verification (7) is lacking, there is no description of the program (8), and it is also not described what the required knowledge of the user is (9), although this last question is obvious as the paper requires much expert knowledge about yeast propagation and beer-making to read. The missing references also mean there is no evaluation possible of the limitations of the program (10).

The origins of data from external sources and own experiments (11) are clearly given. It is not always entirely clear from the text what data are used for calibration, validation, etc. (12), but in general the information is there. It is assumed all data manipulations (13), like log conversions, have been given. There is discussion on the quality of several sources of data (14). The paper clearly discusses which parts of the model need to be parameterized (15) after the initial assignment of parameter values from literature sources, and it is clear from the paper for which parts of the extent of the model more data is required. The output of the model (16) are substrate and biomass concentrations. There is a very limited and mostly only qualitative analysis of the uncertainty on the output (17).

The application area (18) is clearly given in the introduction of the paper. The calibration (19) and validation (20) are discussed, albeit there is some confusion which parts of the text are about calibration, and which parts are about validation. A combined sensitivity analysis (21) and calibration was performed for a few (but not all) parameters without initial values from literature sources. Identification problems (22) were addressed, by providing discussion of other literature sources (see 14). A real uncertainty analysis (23) is missing.

A system analysis (24) is given in the introduction. The issue of scale (25) is mostly limited to coverage of time and parameter space, as the model is non-spatial. It is not discussed in the paper, but the extent and coverage can be approximated from the figures and text. Some of the sub-questions on the conceptual model (26) are difficult to answer without expert knowledge on yeast biology and the fermentation process. The same goes for alternatives of the formal model (27). The issue of structured versus unstructured modelling (age structured, biochemical detail) is missing explicitly in the sub-questions. Question 28 (verification) is more or less identical to question 6, and could not be answered. There was no schematization (29), as the model is an aggregate model. A limited sensitivity analysis (30) was performed (see also 21), but no uncertainty analysis. The model was calibrated (31) using literature and experimental data (see also 11-15), although not all sub-questions can be answered from the paper. The model has been validated (32) in some way via case scenarios, although the authors do not consider this as such.

3.3 Trehalose accumulation model

3.3.1 Point-by-point case evaluation

Aranda *et al.* (2004) pointed out, that studies have shown that yeast cells can be protected to environmental stresses like heat and cold shock, osmotic shock, and starvation by trehalose, a non-reducing disaccharide. This compound also plays an important role as reserve carbohydrate, adding to the viability of yeast. Findings suggest that up to 13% of the yeast dry weight can be trehalose.

Model goal
The goal is to come up with a model that can quantitatively describe the formation and accumulation of trehalose in yeast under environmental stresses. It is investigated how trehalose is formed and accumulated in yeast cells in fed-batch fermentation under carbon or nitrogen starvation conditions. The model may also be useful to obtain estimates of the trehalose content in cells during yeast production (Aranda *et al.*, 2004).

System analysis
Intracellular trehalose accumulation occurs under two conditions: slow growth, and environmental stress, like nutrient depletion, heat, and osmotic shock (see references in Aranda *et al.*, 2004). Given the model goal, important metabolic pathway components concerned with carbon and nitrogen uptake should be involved, as well as pathway components that are involved in the trehalose formation. The biosynthesis of trehalose occurs entirely in the cytosol, and the major reactions are given by Aranda *et al.* (2004). Proper cellular pathway modelling methods should be employed. During the experiments, it is assumed that the medium has been enriched sufficiently to avoid limitation effects other than because of carbon or nitrogen shortage.

Conceptual model
Figure 1 and section 2.1 in Aranda *et al.* (2004) give the conceptual model. The yeast cell is divided into an enzymatic compartment, a cellular compartment, and a trehalose compartment. The enzymatic compartment is further subdivided into a neutral trehalase fraction and a trehalose phosphate synthase fraction. There is an influx of substrate, and an outflux of carbon dioxide and water. The bioreactor substrate concentration s is not further divided into nitrogen or carbon. The expansion of cell volume should be included for calculations of the trehalose synthesis and hydrolysis rates. The authors indicate that the model is evidently 'incomplete', as there is no model that is capable of

a complete description of the cell. However, the key components should be included, which is exactly what is desired in view of equilibrium.

Mathematical model and assumptions

The mathematical model is clearly described in sections 2.2-2.5 of Aranda *et al.* (2004), while their Table 1 gives an overview of a mathematical equations. The different sections discuss the different compartments of the system. In total, the model is made up of 17 equations, 20 parameters, and 6 initial conditions.

The biomass x and substrate s concentrations, and the working volume of the reactor v are described using ordinary differential equations. Hence, there is no spatial model component, and no schematization. Starvation conditions are mimicked by keeping the substrate concentration low, assuming $s = 0$ to rewrite the equations. Again, a constant yield is assumed. Maintenance costs are considered zero, based on reports that suggest indiscriminately low maintenance (see references in Aranda *et al.*, 2004). *Trehalose compartment*: Dilution by growth is included in a differential equation describing the trehalose fraction in cells. The rate of biosynthesis of trehalose in that equation is simplified to two reactions, including glucose 6-P and UDPG. It is assumed that both reactants are in excess, which means that the trehalose synthesis rate depends solely on the TPS active fraction. The trehalose hydrolysis rate in the equation is assumed to be governed by pseudo-first-order kinetics. *Enzymatic compartment*: Enzyme fractions of TPS and TH change in time because of gene expression, and enzyme breakdown. Two differential equations describe the fractions of neutral trehalase and trehalose phosphate synthase. An operon repression model is involved in the TPS induced production, of which the reaction is assumed to be in steady state. Hydrolysis is assumed to follow first-order kinetics. *Cellular compartment*: Includes all cellular components other than TPS and TH. Variation in cAMP-concentration is considered to be important, and is represented in an ODE. Effects of cAMP on TPS and TH are modelled via equilibrium reactions also.

Data

There is a clear description of how experiments are performed, and of how the different measurements have been done. Measured variables were trehalose, biomass, and glucose concentration. Carbon and nitrogen starvation experiments were performed in triples. Carbon starvation was induced by stopping the substrate feed, while nitrogen starvation was induced by switching the alkalinity solution for pH control. Measurement data are compared to literature findings, which generates some confidence. Data coverage is about 10 points on 16 hours, about two-thirds before the starvation phase, and one-thirds within this phase.

Numerical model and calibration

The numerical model is an implementation of the equations into MATLAB. There is no explicit reference to a verification. Nine model parameters have been estimated by calculating the minimized residuals between model predictions and measurement points. A fourth order Runge-Kutta integration method (see also Annex 2) was used, while the minimum of the criterion was searched using the well-known Nelder-Mead simplex method (Nelder & Mead, 1965), which is implemented in a standard toolbox in MATLAB (Press *et al*., 2002). Visually, the fits of biomass, working volume, and trehalose fraction in the cells (Figures 2 and 3, Aranda *et al.*, 2004) are not overwhelmingly convincing. Confidence intervals for the estimated parameters have been calculated using a Monte Carlo approach. For this, the model was used to produce *in silico* data by generating numerical output supplemented by white noise. Figures 4 and 5 (Aranda *et al*., 2004) give the comparisons between internal trehalose measurements and model results, where the experimental averages practically all lie within the confidence bounds (60% of a Student's *t*-distribution).

Validation

Model predictions were compared to data by others (references in Aranda *et al.*, 2004), generated under other conditions. About half the estimated parameters of the data by others fall within the 60% confidence limits. The authors use this as some sort of validation, but this is not really an application-oriented validation.

Sensitivity analysis

A SA, limited to 6 components and 9 parameters, was carried out. Table 5 (Aranda *et al.*, 2004) give numerical sensitivities, but there is no real critical evaluation of the results. It is found that several reaction coefficients are very influential. Several sensitivities are found to be equal, probably signifying an identification problem.

3.3.2 Discussion of the case

Similar to the model by Kurz *et al*. (2002), the model by Aranda *et al.* (2004) is made up of ordinary differential equations. Hence , it is implicitly assumed that reactants mix infinitely quick within the cell, and that the number of molecules of the different reactants are sufficiently high to validate the use of ODEs, which exclude stochastic effects. In reality, from Single Molecule Imaging (SMI, Sako, 2006, and references therein) it is now known that also small numbers of molecules can lead to measurable reactions in all kinds of cellular cascade pathways. Furthermore, it has been challenged fundamentally that reaction rates would be infinitely high. Instead, it is much more natural to consider lower reaction rates with tighter binding coefficients (Kooijman, 2009). Therefore, the results of the SA should be viewed critically, as the results depend on the assumption that the model framework is reasonable. However, as this is a very fundamental issue, and the use of ODEs for these kinds of modelling problems is perhaps the most used approach, we discontinue the discussion of this issue for now.

For the goal of understanding the formation and accumulation the demands are much lower than for the goal of estimating trehalose quantities. To be able to use the model for the latter goal, not only a proper mechanistic basis should be available on the formation and accumulation of trehalose, but also sufficiently reliable data for calibration and validation. Carbon starvation is induced by stopping the substrate feed – in the model the carbon content is given by the variable 'substrate' (s). It is not clearly explained how nitrogen starvation is supposed to work by switching the pH control. In the paper it is assumed to work via a reduction in the concentration of intracellular cAMP, based on expert knowledge. There is also no data on nitrogen concentrations available.

Although the model is of only limited complexity when compared to many Systems Biology models, the presented network may still be overly complex for the application. As data and a clear mechanism on nitrogen starvation are missing anyway, the suggestion is to try and significantly reduce the number of model equations and only include carbon starvation. This will obviously limit the application area of the model by making demands on the used method of pH control, but it is uncertain if this is a significant problem. The goal that the "model could be a useful way to get a narrow estimate of trehalose content in cells during yeast production" (Aranda *et al.*, 2004) seems overenthusiastic, as the confidence bounds are not very convincing. Compared to the previous case (Kurz *et al.*, 2002) it should be said, that in this case the substance of focus (trehalose) was explicitly included in the data.

3.3.3 Advice for improvement

Like the previous model by Kurz *et al.* (2002), the model by Aranda *et al.* (2004) seems to contain too many equations. Many variables and equations are included that describe only auxiliary steps. The main in- and outputs of the model are glucose, cell growth, and trehalose. As such, it seems that the data support is more or less in line with what is needed. On the other hand, the experimental conditions do not fully cover the goal of the model. It was intended to model both carbon and nitrogen starvation, but it is not clear how the latter is included in the model now.

The model can probably be reduced significantly by replacing many of the equations by some non-linear equations describing the main variables glucose, cell size, and trehalose, that include the starvation. A more thorough sensitivity analysis, in which the potential effects of including various alternative non-linear relationships are evaluated, could be a first step in determining how this reduction/replacement should be achieved. Furthermore, a choice should be made between two options: either the nitrogen starvation is dropped, and then the application area of the model is constricted to carbon starvation conditions only, or the application area of the model includes the nitrogen starvation, and then the mechanism for this process should be included more explicitly. In the latter case the model is likely to be 'bigger' than in the first case, albeit still certainly not as complex as it is currently.

3.3.4 List fill-out

Again, it is a peer-reviewed paper, and there is no meta-information (1), although there are several tables that combine much of the essential information. The model goal (2) is given near the end of the introduction. Model concept (3) is given. Assumptions and simplifications (4) are given throughout the description of the mathematical model (5). There is a short, implicit description of the numerical model (6) in the section on calibration (31). There is no description of any verification or test results (7), how the program works (8), or what the performance and limitations (10) are. User required knowledge (9) is implicitly present through the journal the paper is published in.

The original experiments and data (11) are clearly described, as are the data produced by others. The experimental data (12) is used for parameter estimation, the quoted data is used for some parameter values and for some validation. Data manipulations (13) are given very briefly, and it is not very clear what uncertainties might results from these (14). Glucose, trehalose, and biomass are measured, so there is not a real need for additional data (15). The program output (16) is clear from the figures, in which it is compared to the data. Confidence intervals in later figures give some measure of uncertainty about the output (17).

The possible application(s) (18) are given at the end of the introduction. The calibration (19) and validation (20) are described in section 4.2 of Aranda *et al*. (2004). A partial sensitivity analysis (21) is performed and reported. This gives limited insight in the identifiability (22) of several parameters. Question 23 overlaps with 17: confidence intervals are given that give some insight into uncertainty.

There is a very limited system analysis (24) in the introduction. Coverage of the data (25) is extracted from the figures, and further scaling issues are uncertain. For the conceptual model (26) see 3 and 4. The formal model (27) corresponds to 5. The numerical model (28) corresponds to 6. There is no schematization (29), so this question does not apply here. A limited SA/UA (30) is given, as reported for questions 17, 21 and 23. Questions 31 and 32 overlap with 19 and 20, respectively.

3.4 Biofilm formation

3.4.1 Point-by-point case evaluation

The following test case focuses on the formation of biofilms in continuous beer fermentation, instead of fed-batch fermentation. The case is extra interesting as several nested models are considered, making it an ideal test case for our study. Biofilms are aggregates of bacteria, yeast, algae, or such microorganisms in which cells adhere to each other and/or to a surface (also called ‘support’). The aggregates are usually formed by excreted polymeric substances, that form a matrix that immobilizes the cells. Biofilms are often viewed in a negative context, for instance, bacteria in biofilms are much more difficult to kill with antibiotics. In the beer-brewing context it can be positively viewed. Yeast can form biofilms, making it much more resistant against wash-out from the fermentation reactor. This resistance to wash-out obviously depends on the quality of the support. Furthermore, a biofilm is much more capable of self-regeneration, replacing dead cells by new, alive ones.

Biofilm formation has been shown to possibly lead to all kinds of changes in the cell metabolism of yeast, among those product synthesis rates like ethanol production. Biofilm thickness can be influenced by feed rate, cell physiology, diffusion limitations, and hydrodynamic reactor conditions. Because of the engineering (dis)advantages, it is relevant to study the conditions that lead to biofilm formation, and how to possibly control biofilm output like synthesized products.

In this test case a model study by Brányik *et al.* (2004) on biofilm formation on spent grain particles is evaluated. Spent grain particles is a biocatalyst that has been successfully applied in continuous beer fermentation. In the study observations are considered that the immobilized biomass shows decreasing metabolic activity, and that yeast has a finite replicative life span. Brányik *et al.* (2004) hypothesize a viable fraction of the total immobilized biomass, so not all yeast biomass becomes inactive when adhering to a surface.

Goal

The goal set by Brányik *et al.* (2004) is to develop a simple kinetic model, based on mass-balanced equations, that can predict the immobilization of yeast on spent grain particles. They hypothesize there should be an active fraction of the total immobilized brewing yeast.

System analysis

The induction phase starts at time zero, and ends when the detection limit for the immobilized biomass concentration is reached. It is followed by the exponential accumulation phase, after which the biomass accumulation starts to progress linearly. After that, the growth starts to decline until some steady state is reached. Two possible steady states exist, namely one caused by substrate limitation, in which the immobilized fraction reaches a plateau, and one caused by the maximum mechanical strength of the biofilm under the specific hydrodynamic conditions.

Brewing yeast immobilized on spent grain particles does not have a constant metabolic state (see Figure 5 in Brányik *et al*., 2004). Initially, the specific glucose consumption rate of the immobilized and the free yeast cells are comparable, but later it decreases, until it seems to reach some sort of new steady state value. Changes in the physiological state of the immobilized cells is probably due to aging. The free cell population is assumingly homogeneously mixed, also in terms of age. This is a common assumption for chemostat-like environments like bioreactors, and depending on the time scale not necessarily bad. On the other hand, immobilized cells show a significant decrease in the first derivative of NAD(P)H fluorescence signal, which is highly correlated with the viable cell count. The decrease starts after ca. 115 h at fixed dilution rate, which corresponds to ca. 30 cell cycles.

Data

Experiments were performed by Brányik *et al.* (2004) and are clearly explained in their paper. Batch experiments were performed for the comparison of growth rates of free cells and immobilized cells. The immobilized cells were liberated from spent grain taken from a continuous brewing process. Although practically inevitable, the extraction methods no doubt influenced the results, and thus the estimations of the growth rates of immobilized cells. Continuous culture experiments were done at a variety of dilution rates (D = 0.06-0.27 h^{-1}, temperature was fixed at 25 °C) to obtain measurement points of glucose concentration $[g\,L^{-1}]$, free biomass $[g\,L^{-1}]$, and immobilized biomass accumulation $[g_{IB}\,g_C^{-1}]$ (C = carrier). Inoculation was 100 mL of pre-cultured yeast suspension in medium in a continuous reactor of 440 mL working volume. The inoculation can also be identified as a source of significant uncertainty. The results are shown in Figure 2 in Brányik *et al.* (2004).

To test the biofilm detachment the dilution rate was increased from 0.27 to 1.15 h^{-1} in experiments. At the start the immobilized biomass load X_{im} varied between 0.02-0.03 $g_{IB}\,g_C^{-1}$. The immobilized yeast biomass was determined by measuring the weight of biofilm material before and after treatment to remove the cells. Cell removal was verified by microscope. Uncertainty in this data due to loss of carrier material itself was minimized by corrections from data obtained from blank experiments with clean carrier. Metabolic activity was measured by means of a NAD(P)H fluorescence spectrophotometer.

The data are graphically represented in Figures 2-7 of Brányik *et al.* (2004). In Figure 2 the model simulations are fit to measurements of immobilized biomass accumulation, free biomass, and glucose concentration at D = 0.24 h^{-1}. Figure 3 gives the fit of model simulations to measurements of immobilized biomass accumulation on spent grain particles at different dilution rates. Figure 4 shows the immobilized biomass deposition rate as function of the dilution rate, and the free cell concentration during the induction phase. Figure 5 depicts the fit of model simulations to measurements of the immobilized biomass load accumulation, specific immobilized biomass detachment rate, and specific glucose consumption rate of the immobilized yeast during the detachment experiment. Figure 6 gives the changes in NAD(P)H fluorescence during starts and stops of the MM flow. At MM flow stops, the intracellular NAD(P)H increases, as it cannot be oxidized through the respiration change. Figure 7, finally, gives the fluorescence signal derivative.

The extent and support of the different components in Figures 2 and 5 seem to be reasonable, about 10 measurements for each components over a time range of less than 200 hours. The coverage in Figure 3 is not very good, with only 3 or 4 measurements over a range of more than 200 hours. The NAD(P)H fluorescence is measured continuously. Perhaps some more dilution rates could have been considered (only 5 rates between 0.115 and 0.27 h^{-1}, and 1.15 for the detachment).

Assumptions & simplifications

The basic ODE assumptions apply. For instance, there is a homogeneously mixed free cell population in terms of age, cell physiology, and space. The biofilm is assumed to be at steady state at the end of the detachment experiment. The NAD(P)H fluorescence signal is fully correlated with the viable cell count. The maximum specific growth rate for free and immobilized cells is assumed to be equal. The cell deposition rate is neglected due to the low free cell measurements at high dilution rates.

Some assumptions are model dependent. For instance, in the simplest model there are no maintenance requirements, the biofilm has an infinite stability, there are no changes in growth rate and metabolic activity of the immobilized cells, etc. The more complex model has simplifications with regard to diffusion: the kinetic model does not incorporate limitations of diffusion inside the biofilm, as the penetration depths of oxygen and glucose are significantly higher than the maximum measured thickness of the biofilm. In the induction phase only cell deposition is taken into account.

An empirical term is used to describe cell deposition, as it is difficult to separate the contribution of growth and detachment from the deposition rate. The different processes that are responsible for adhesion of cells into a biofilm are simplified to one, non-explicit description: the cell deposition coefficient. The decrease of cell deposition due to a reduction in carrier surface area is simplified to a term $(1 - X_{im}/X_{im}^{max})$. An actively growing fraction is assumed to exist in the total immobilized biomass. It does not contain a separate term for cell activation. The viable immobilized biomass fraction is considered constant during biomass accumulation.

The specific steady-state detachment rate is valid under constant hydrodynamic conditions and reactor design. The experiments are kept to fulfil these conditions. It is assumed there is no selection advantage for free or immobilized subpopulations whatsoever.

Mathematical models

A list of model parameters, a short description, values and units is given in Table 1 in Brányik *et al.* (2004). The model equations are given in Eq. (1-6) on p. 1737 in Brányik *et al*. (2004). Four models are considered (for a graphical overview see Figure 3.1).

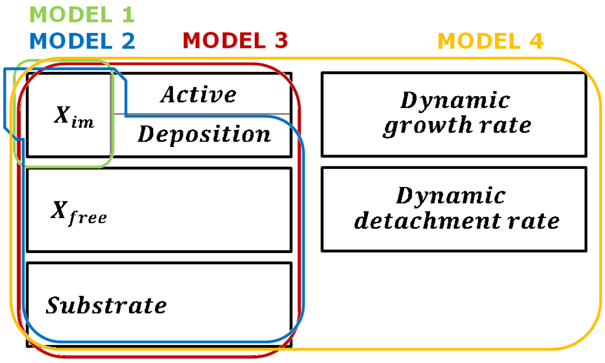

Figure 3.1. *Graphical overview of the four different biofilm models by Bráynik et al. (2004), of increasing complexity. Model 1 only contains a dynamical equation for the immobilized yeast biomass. Model 2 contains also dynamical equations for free biomass and glucose. Model 3 includes also an active fraction in the immobilized yeast population. Model 4 includes dynamical equations for the maximum specific growth rate and the specific steady-state detachment rate.*

The first, simplest model is covered by Eq. (1) in Brányik *et al.* (2004), which considers the overall mass balance of the immobilized population. It is given as

$$\frac{dX_{im}}{dt} = R_{dep} - R_{grw} - R_{det}$$

Here R_{dep} is the cell deposition rate, R_{grw} is the immobilized biomass growth rate, which can be replaced by the function $\mu_{im} X_{im}$, and R_{det} is the immobilized biomass detachment rate, which is replaced by the function $k_{det} X_{im}$. Several phenomena are ignored in this formulation, such as changes in the growth rate and metabolic activity (aging) of the immobilized cells, maintenance costs (Pirt, 1965; Kooijman, 2009; see section 3.1), and that the stability of the biofilm is only finite.

The set of Eqs. (2-6) gives a more complex model, including variations in the free biomass and glucose, and varying maximum specific growth rate based on the Monod model (see section 3.1). The second and third model are comprised of only Eqn. (2-4). Under excessive glucose concentration (for instance, in the detachment experiment at high dilution rate), the dynamic parameters in Eqn. (5-6) are replaced by fixed values (found in Table 1 in Brányik *et al.*, 2004). In the second model only the cell deposition term in Eq. (2) is considered for the biomass accumulation process, simplifying the model even further. This second model is valid during the induction phase. The third model then consists of the full Eqn. (2-4). The fourth and most complex model is the full set of Eqn. (2-6), so including the dynamic specific growth rate.

Parameters

The values of the parameters are given Table 1 on p. 1738 in Brányik *et al*. (2004). The reasoning behind the choices are given in the discussion at p. 1737 of that paper. Some additional parameters for the complex model Eqn. (2-6) are given below those equations.

Calibration & validation

The complex model shows a good agreement with the experimental data in Figures 2 and 3 in Brányik *et al.* (2004) starting from X_{im} = 0.03 $g_{IB}\, g_{C}^{-1}$. Slightly higher experimental values in the free biomass are explained by the authors as disturbed material balance due to sampling. Model simulations of the biomass accumulation match very well with the measurements (Figure 5 in Brányik *et al.*, 2004) for some initial condition. The model predictions of the specific glucose consumption rate of immobilized biomass is systematically lower than the measured data points. The authors explain this as substrate overutilization (with an added reference).

3.4.2 Discussion of the case

The different models are compared, and their use is bounded by application ranges. In general, the impression in view of 'equilibrium' is very good. Most model predictions match very well (at least visually) to the measurements. Most assumptions appear to be reasonable, especially from a practical point of view. The model framework is well supported by experimental data, as all considered state variables have been measured explicitly (free and immobilized cells, glucose) or implicitly (dynamic parameters).

The explanation given by the authors for the discrepancy between predictions and measurements of free biomass (disturbed material balance due to sampling) seems reasonable. On the other hand, the explanation given for the systematically under-predicted specific glucose consumption rate of immobilized biomass, namely substrate overutilization, seems rather peculiar, as it is unclear how such a phenomenon should work. More important, however, is whether this discrepancy is relevant for the application of the model. A critical visual inspection of Figure 2 in Brányik *et al.* (2004) reveals that the model systematically underestimates the glucose concentrations during the continuous experiments. But, apart from that, the model fits for immobilized and free yeast biomass seem to be very good. The goal set by Brányik *et al.* (2004) is the development of a simple kinetic model for predicting the immobilization of yeast on spent grain particles. The immobilized yeast biomass predictions correspond well to the measurements (see Figure 2 in Brányik *et al.*, 2004), just not the predictions related to glucose consumption. This will introduce uncertainties in the predictions, but probably they are acceptable for some application range. Also, it is reasonable to assume that a sensitivity analysis and consecutive re-calibration may produce better fits and increase confidence in the model(s).

3.4.3 Advice for improvement

The model by Brányik *et al.* (2004) actually seems to be quite well balanced in view of the application (estimating the viable cell load in biofilm). A modest model selection procedure was performed, and the model is adapted (changed) to fit ranges of experimental conditions (like under excessive glucose concentrations). The number of equations and variables is limited, and all main variables are measured. Some re-calibration may improve the fits to the data. Assuming that verification and validation are properly performed and give convincing results, this test case might very well be near 'equilibrium'.

3.4.4 List fill-out

Meta-information (1) as such is missing, again as it is a paper. The model goal (2) is given near the end of the introduction. The conceptual model (3) and the assumptions and simplifications (4) can be collected from several parts of the text. The mathematical models 1-4 (5) are clearly given. There is no information whatsoever on the numerical model (6), verification (7), program description (8), or program limitations (10). Again, user expert knowledge (9) is implicitly assumed via the journal.

The origin and use of data is clearly described (11, 12), as well as how it is manipulated (13). There is no clear testing of the data (14), e.g., by comparing it to other sources. The data supply seems adequate (15). The output (16) is probably what is plotted in the Figures when compared to data. For several sources it is unclear what uncertainties (17) there might be. Model application (18) is not very well described, but becomes clear from the introduction. There is no real description of the calibration (19), while validation (20) is comprised of comparing of model simulations to data points.

No SA (21) or UA (23) is performed, and there are only very limited remarks on identification of the parameters (22). System analysis (24) and conceptual model (26) are spread out across the text (compare to 3 and 4). The coverage of data (25) becomes clear from the Figures. The formal model (27) is clear (see 5), while the numerical model is skipped (28, see 6-8, 10). There is no schematization (29). 30-32 have been covered.

3.5 Lessons learned

The evaluation list EMC v0.1 has been applied to a number of yeast growth models in the field of beer-brewing. The different test cases consider different goals and applications in this field, but all have significant overlap in theory, which decreases the time to gain 'expert knowledge' on the subject, and makes it easier to compare the test cases. For each test case it has been attempted to fill out all the questions of the list, to evaluate if a model would be near 'equilibrium' or not, and to consider what could be improved in either the test case or the list. Obviously we did not have *a priori* evaluations of 'equilibrium' available. Nevertheless, in each of the test cases the application of the list led to some estimation of whether or not the evaluated model is far off from 'equilibrium': in two cases the model complexity was found to be significantly unbalanced with regard to the intended application, in one case the evaluation gave the impression that the model was reasonably balanced compared with data support and the intended application.

3.5.1 Model equilibrium

The model by Kurz *et al.* (2002, control strategy for optimal inoculum at the starting time of subsequent beer fermentations) was found to be significantly off from any 'equilibrium'. The complexity of the model is significant in terms of the number of (linear) equations, switches, and parameters, and it does not seem to be well balanced with regard to the application. In view of the goal (limitation and inhibition effects) the support from the data should be increased, so that this at least includes ethanol measurements. In view of the application (process control including failures in oxygen supply) additional measurements should be made for the system at lower oxygen supply than currently covered in the data.

The model by Aranda *et al*. (2004) contains even more equations than the model by Kurz *et al*. (2002), while some of them are also non-linear. The ratio of complexity to data support seemed to match that of the model by Kurz *et al*. (2002). The results of some limited SA and UA reduced the confidence in the predictive capabilities of the model. The important carbon-starvation mechanism is clear, but the important nitrogen-starvation mechanism is not. As such, the overly complex model does not even cover the full intended application area, and is very much off balance.

The test case by Brányik *et al.* (2004, predicting biofilm dynamics), in which four models of increasing complexity were studied, seems to be properly balanced. In the simplest model only immobilized yeast growth and a mass balance is explicitly considered, while many properties are ignored. The more complex models incorporate dynamics for glucose and free biomass, while the most complex model allows for variations in the maximum specific growth rate and specific steady-state detachment rate. The comparison between model simulations and measurements appear to be very good, and also relevant for the intended application. The second model was selected as best model for some ranges (excessive glucose concentrations), while the third and fourth model were most applicable for other ranges (induction phase and the remainder, respectively). An advice for this test case is to perform a re-calibration and SA/UA to increase model confidence.

In all three cases it seems that a rather clear judgement could be made about the model balance with the aid of the evaluation list EMC v0.1. However, these conclusions have a qualitative nature, and are prone to being challenged by different evaluators, modellers, stakeholders, or otherwise. A future task will be to try and develop a way of making more quantitative means.

3.5.2 List evaluation

Several conclusions can already be drawn on the properties of many questions and the overall performance of the EMC v0.1.

Papers do not provide meta-information, which renders question 1 of EMC v0.1 useless in these cases. The question is also not directly relevant for determining ‘equilibrium’. The reason why it was inserted by Bogaart *et al.* (2011) was to get an overview of the available documentation in every test case. For the application (model balance assessment) it is for the most part a redundant question. Nevertheless, it helps to gain a quick overview of the model and prevent confusion about which version of the model was used for which application(s).

In these cases, the questions 6 through 10 failed consistently. The numerical model description, verification and testing, program description, user knowledge, and program performance and limitations are often ignored in papers and other types of publications, yet they are (highly) relevant for ‘equilibrium’. A numerical model or database that is not described, tested, and verified is philosophically and often also practically speaking rather doubtful. In part the ignorance seems to be caused by common confusion over what verification means as compared to validation, and in part by a common failure to rigorously test a numerical model or database for all sorts of reasons, or by simple refusal to report these tests. The latter may be rooted in the observation, that papers and other academic publications should have something interesting to tell, while testing and verification are ‘not considered interesting’ in itself, unless new testing and verification techniques are considered explicitly. Nevertheless, a verification and testing is necessary to ascertain if the model is a proper translation of the conceptual model. Required user knowledge (question 9) is implicitly present through the journals the models have been published in. This issue should be automatically solved when the list is applied to models and databases that have fulfilled status A requirements. Question 8 (program description) is not directly important. However, in an adapted form it should remain in the list (perhaps implicitly) as it gives background information of the functioning of the program and what the model user should know to handle it well.

Question 14 on the testing of data quality fails often too. Data are scarce, as it is often difficult or costly to perform experiments. Only the paper by Kurz *et al*. (2002) frequently discusses comparisons between different types of literature-cited and/or experimental data. Nevertheless, it seems that it is important to obtain insight in the quality and utility of data, especially in view of the application. The question should also be reformulated to include this aspect of application, and probably it can be merged with question 22 for that matter.

Question 17 does not perform well either. In general, an uncertainty analysis on the output is not present. However, most applications in policy evaluation require some form of uncertainty bounds. If some uncertainty analysis has been performed (*e.g*., the model by Aranda *et al.*, 2004) the results are not encouraging. The question could be merged with question 23, and maybe 30. It proved difficult to obtain an answer to question 19 in these test cases. This is surprising, as the calibration of a model is one of the essential steps for model applications. A description of the calibration is important, as it can reveal reasons why a good validation fails. Also, it allows users to re-calibrate the model for better performance in applications.

Question 25 on scale often yields unsatisfactory answers. Data extent and support is mostly reported only implicitly through Figures, and although often the experiments are described rather meticulously, there is rarely any insight if there are scaling issues or not. Scaling issues arise as most attributes like time and space are infinite, but are measured on discrete scales with limited extent: imagine the 'measuring' of a continuous normal distribution using limited samples divided into histogram classes, and the histogram has an upper and lower boundary. For spatial discretization into grid cells (for example, of river catchment areas) it is commonly assumed that the values of variables are the average of the measurements at the corners of the grid cells (although there are more sophisticated approaches, like 'kriging'). More information on scaling issues can be found in Bierkens *et al.* (2000; but see also Bogaart *et al.*, 2011, in Dutch). We argue that scaling is still a very important issue in modelling and questions of the list should focus on it.

There is a great deal of overlap between several questions, especially those focused at calibration, validation, sensitivity and uncertainty analysis, and on model descriptions. These questions could be merged, but furthermore, they need to be more focused on the application. This brings us to the most important point not covered well by the list at the moment: the application itself. The application area is not considered explicitly before question 18. Clearly, questions on the application should be moved to the early part of the list. There is also yet no clear distinction between the intended and the actual application area, i.e., what could the model be used for, and for what is it currently used? Model goal, intended and actual application areas should be compared, and they should show sufficient overlap. Next to that, it is relevant to note that if a model or database is used for different applications, that the demands by each application may be different. This will make it necessary to consider the equilibrium again for each application. The test cases already indicate, that it is a very real possibility that a model may be well balanced for one application, but not for another. The list will have to be adapted to include these changes.

4 Spatial test case model

The examples in beer-making did not explicitly consider a spatial element. Therefore, in the following example ecological range expansion models are discussed. The test case is a suite of models by Drury and Candelaria (2008), who considered the range expansion rates of the sea otter (*Enhydra lutris*) in the water near California during the 20th century. Furthermore, they used the AICc for the selection between different candidate models, which is interesting for comparison with the ECM v0.1.

4.1 Background

Models describing the invasion rates of organisms invading new habitats explicitly consider spatial dimensions. These models can be divided into two groups, which are discussed in short below.

4.1.1 Reaction-diffusion equation models

Reaction-diffusion (RD) equation models date back to the first half of the 20th century (Fisher, 1937; Skellam, 1951; Drury & Candelaria, 2008). The equations provide an intuitive description of interacting individuals, and furthermore, they may be solved analytically when interaction terms are simple enough. RD systems describe the changes of concentrations of substances in space and time under the influence of local reactions in which the substances are formed, and diffusion which causes the distribution of the substances over a surface.

RD equations are given in the form

$$\frac{\partial N(x,t)}{\partial t} = f\big(N(x,t)\big) + D\nabla^2 N(x,t)$$

The first right hand side term is the reaction term. Here, $N(x,t)$ is the population density, with indices for space and time, $f(N(x,t))$ is a population growth function, and t is time. The second right hand side term of the equation is the diffusion term. Parameter D is the diffusion constant, and ∇^2 is the Laplacian, giving the second order spatial derivative. Observe, that a regular ordinary differential equation is a collapsed form of this type of equation, with the index x removed (the space is then perceived as homogeneously mixed) and the diffusion term gone.

Some (biological, mathematical) assumptions and simplifications behind RD models are the same as behind ODEs, like continuous reproduction of the population, and large numbers. In practice, RD models must be numerically approximated using grid cells. Within these cells there is still homogeneous mixing. Also, the Laplacian assumes that dispersal distances are normally distributed.

RD models are known to produce spatial patterns, that have been used to explain several biologically observed phenomena. An established name in this context is that of Alan Turing (Turing, 1952). Nevertheless, Drury & Candelaria (2008) point out there are two general fallacies with RD models. The first is that many invasions start slowly, but then proceed at an increased rate. This is not covered by the RD models, that predict a single, constant spread rate. The second is that many empirical distributions are *leptokurtic*, i.e., they have a lower probability than a normally distributed variable of values near the mean, and thicker tails, so a higher probability than a normally distributed variable of extreme values. These issues can be addressed informally by making the diffusion time and space dependent.

4.1.2 Integrodifference equation models

Integrodifference equations (IDE, not to be confused with integro-differential equations) have evolved with the recent advances in desktop computers. The basic type of equation is given as

$$N_{t+1}(x) = \int_{\Omega} k\ (x,y) f\big(N_t(y)\big) dy$$

In this type of equation the population size at the next time period at location x is related to the population growth at each spatial location y in the current time period, and the probability of arriving at x from y. The domain Ω is typically taken from minus to plus infinity. Again, $f\big(N_t(x)\big)$ describes the local population growth at location x. In most applications, for any $y \in \Omega$ there is a probability density function (pdf) $k(x,y)$, that describes the probability of moving from point y to point x. Often, $k(x,y)$ is referred to as the *dispersal kernel*. Currently, integrodifference equations are most commonly used to describe the spread of *univoltine* populations, species that have one brood or generation per year. This in contrast to RD models, which are most commonly used for species with continuous procreation.

In one spatial dimensional, the dispersal kernel often depends on the distance between the source and the destination, simplifying the above equation to

$$N_{t+1}(x) = \int_{-\infty}^{+\infty} k\ (|x-y|) f\big(N_t(y)\big) dy$$

This formulation corresponds to the one given by Drury & Candelaria (2008; and see Kot & Schaffer, 1986). Here, the dispersal kernel can be any appropriate pdf, and the probability of moving from y to x depends only on the distance $|x-y|$. The population function $f\big(N_t(y)\big)$ is analogous to the reaction term in the RD equations. The dispersal kernel can make very different predictions about the dispersal rate, which is an obvious advantage compared to the RD models.

4.2 Test case

4.2.1 Point-by-point case evaluation

The sea otter (*Enhydra lutris*) is a marine mammal species that used to be distributed form the northern part of Japan, via the Bering Strait, and then from Alaska downward to California. The species has become endangered as a result of the great fur hunt between 1741 and 1911, the threat of oil spills, and the competition of fisheries. Currently, the Californian sea otter is limited to a region around San Francisco and Santa Barbara County, and recently a new threat has arisen in the form of a toxicant, which is produced by cyanobacteria. In this study case, data of the sea otter are compared to several different RD and IDE models. To our knowledge, Drury & Candelaria (2008) are the first to compare RD and IDE models to analyse the same data. With these comparisons, the authors use the AIC.

Data & system analysis

In 1914, a remnant population of otters was found at Point Sur, CA (see references in Drury & Candelaria, 2008). As a result of their protection this otter population expanded both north- and southward along the coast. Range expansion can be assumed to be one-dimensional, as otters generally stick to a narrow band of coastline and to shallow water levels. Range expansion seems to occur gradually or by ‘jumps’ of up to 127 km for adults and 187 km for sub-adults (see references in Drury & Candelaria, 2008). Otters get roughly one pup each year. The reproduction of the otter is more or less continuous, but some peak behaviour has been reported. There is variation in reported population growth rates and survival data. Time scale is therefore partiallly related to population growth. Drury & Candelaria (2008) mention an average rate of population increase of 5-6%.

The used data consist of sea otter population numbers from the period of 1914 through 1986. The sea otter population around California has increased significantly since the end of the hunt in 1911. However, there is a sharp transition in the data in the years 1972-73. The rates of spread before and after this event differ significantly. Therefore, one aspect in the modelling that is considered is allowing the diffusion rates to vary with time, or separating the model into two separate equations for the two distinct periods in time. Habitat differences occur along the coastline. The northward spread of the otter seems to be significantly slower than the southward spread. Therefore, this element is also considered in the modelling by allowing space variation in the diffusion.

The intrinsic population growth rate was estimated to be r = 0.056 yr^{-1}, obtained by best-fitting a straight line through the log population size roughly for the period 1940-1960. This value is compared to several literature values (see references in Drury & Candelaria, 2008). The carrying capacity K = 7.54 km^{-1} was obtained by averaging the population density during times for which population density was more or less constant, i.e., the population growth and range increase were such that the density remained constant.

Scale seems to be a problem. The data form a time series with an extent from 1914 through 1986. The support is a year, i.e., there are yearly aggregated population number measurements. The coverage is very poor at some points. In fact, the period from 1914 through 1938 has no coverage at all! Furthermore, in the second world war counting otters did not have a high priority, obviously. Only from 1950 onwards there seems to be a regular yearly update. However, for all data points there is a clear division into a northern and southern subpopulation.

Units

The output (range expansion speeds) is given in terms of km/yr and diffusion constants in terms of km^2/yr.

Goal & application

The goal in the paper is to demonstrate the need for quantitative model selection criteria to differentiate between the potential of different spatial models to describe spread data. We convert this here to a goal of finding an appropriately complex model for describing sea otter densities and expansion along the Californian coast, with some predictive power. The application would then be to be able to predict the sea otter numbers and locations for the near future, given unchanged circumstances.

Conceptual model, assumptions & simplifications

See the data & system analysis above. Basic assumptions and simplifications apply for IDE and RD models.

Mathematical model(s)

The basic population growth model adopted by Drury & Candelaria (2008) is the well-known logistic model (the Pearl-Verhulst equation). A continuous formulation was used for the RD models the, and for the IDE models the integral form mentioned in sub-section 3.4.4. Several variants are discussed, based on the evaluation in data & system analysis. The simplest model, let us name it model 1rd and 1ide, is spatially and temporally *homogeneous*. The RD-formulation (Eqn. (1) & (13), Drury & Candelaria, 2008) and IDE-formulation (Eqn. (14-15), Drury & Candelaria, 2008) are, respectively

$$\frac{\partial N(x,t)}{\partial t} = r\,N(x,t)\left(1 - \frac{N(x,t)}{K}\right) + D\nabla^2 N(x,t)$$
$$N_{t+1}(x) = \int_{-\infty}^{+\infty} k\,(|x-y|)\,\frac{N_0 e^{rt}}{(1+(N_0/K)e^{rt})}\,dy$$

Here N_0 is the initial population size, r is the intrinsic population growth rate, and K is the 'carrying capacity'.

The second model, model 2rd (Eqn. (16-17), Drury & Candelaria, 2008), assumes different spread rates along the northern or southern coastline, so *space*. In this model, the diffusion rate varies according to

$$\frac{\partial N(x,t)}{\partial t} = r\,N(x,t)\left(1 - \frac{N(x,t)}{K}\right) + D(x)\nabla^2 N(x,t)$$
$$D(x) = D_n, x > \tilde{x}\,; D(x) = D_s, x \leq \tilde{x}$$

Here x_n represents northward diffusion, x_s represents southward diffusion, and $\tilde{x}$ is the location of the initial population.

The southward bias can also be accounted for by including an advection term in the RD equation (model 2ard, Eq. (18), see references in Drury & Candelaria, 2008)

$$\frac{\partial N(x,t)}{\partial t} = r\,N(x,t)\left(1 - \frac{N(x,t)}{K}\right) + D\nabla^2 N(x,t) + a\frac{\partial N(x,t)}{\partial x}$$

Here the diffusion rate is the same in north- and southward directions. Instead, there are habitat differences that possible hinder northward expansion.

Model 2ide (Eq. (3), (14), (15) & (19), Drury & Candelaria, 2008, where Eq, (19) is analogous to Eq. (17) in the RD case) is given as

$$N_{t+1}(x) = \int_{-\infty}^{+\infty} k\,(z)\,\frac{N_0 e^{rt}}{(1+(N_0/K)e^{rt})}\,dy$$
$$k(z) = k_s = \exp\left(-\alpha_s z^{\beta_s}\right), x < \tilde{x};\; k(z) = k_n = \exp\left(-\alpha_n z^{\beta_n}\right), x \geq \tilde{x}$$

Observe, that this division into a 'northern' and 'southern' kernel comes with twice as much parameters as for the RD model formulation.

In the third case a division in *time* is assumed, more particularly a division between pre-1973 (e = early) and post-1973 (l = late). Model 3rd (Eqn. (20-21), Drury & Candelaria, 2008), is given as

$$\frac{\partial N(x,t)}{\partial t} = r\,N(x,t)\left(1 - \frac{N(x,t)}{K}\right) + D(t)\nabla^2 N(x,t)$$
$$D(t) = D_e, t < \tilde{t}\,; D(t) = D_l, t \geq \tilde{t}$$

In the IDE-formulation, model 3ide (Eq. (22), Drury & Candelaria, 2008), the temporal division is given by

$$k(z) = k_e = \exp\left(-\alpha_e z^{\beta_e}\right), t < \tilde{t};\ k(z) = k_l = \exp\left(-\alpha_l z^{\beta_l}\right), t \geq \tilde{t}$$

Models have also been formulated to include both a partitioning in *time and space*. Model 4rd (Eqn. (23-24), Drury & Candelaria, 2008) is a combination of the previous formulations for time and space separately

$$\frac{\partial N(x,t)}{\partial t} = r\,N(x,t)\left(1 - \frac{N(x,t)}{K}\right) + D(x,t)\nabla^2 N(x,t)$$
$$D(x,t) = D_{n,e}, x > \tilde{x}, t < \tilde{t}$$
$$D(x,t) = D_{n,l}, x > \tilde{x}, t \geq \tilde{t}$$
$$D(x,t) = D_{s,e}, x \leq \tilde{x}, t < \tilde{t}$$
$$D(x,t) = D_{s,l}, x \leq \tilde{x}, t \geq \tilde{t}$$

For the IDE formulation (Eq. (25), Drury & Candelaria, 2008) of model 4ide there are four different kernels

$$k(z,t) = k_{n,e} = \exp\left(-\alpha_{n,e} z^{\beta_{n,e}}\right), x > \tilde{x}, t < \tilde{t}$$
$$k(z,t) = k_{n,l} = \exp\left(-\alpha_{n,l} z^{\beta_{n,l}}\right), x > \tilde{x}, t \geq \tilde{t}$$
$$k(z,t) = k_{s,e} = \exp\left(-\alpha_{s,e} z^{\beta_{s,e}}\right), x \leq \tilde{x}, t < \tilde{t}$$
$$k(z,t) = k_{s,l} = \exp\left(-\alpha_{s,l} z^{\beta_{s,l}}\right), x \leq \tilde{x}, t \geq \tilde{t}$$

Models 5rd and 5ide are as the combined models, but then with a temporal partitioning into the periods 1939-72 and 1973-86. Models 6rd and 6ide are as the combined models, but now for only the data of 1973-86. A summary of the different models is given in Table 4.1.

Table 4.1. *An overview of the different models by Drury & Candelaria (2008), including the number of parameters to fit and the number of data points.*

Model	North/South	1914-72/73-86	1939-72/73-86	1973-86	k	n
1rd/1ide	No	No	No	No	1/2	44
2rd/2ide	Yes	No	No	No	2/4	44
3rd/3ide	No	Yes	No	No	2/4	42
4rd/4ide	Yes	Yes	No	No	4/8	42
5rd/5ide	Yes	No	Yes	No	4/8	40
6rd/6ide	Yes	No	No	Yes	2/4	22
2ard*						

*This model was not considered further.

Numerical implementation

The mathematical models are implemented as numerical models in Matlab. Forward-Euler methods with Neumann boundary conditions are used for the integration. The integrals in the integrodifference equations are convolved by using the fast Fourier transform (FFT) in Matlab.

Calibration & validation

The Nelder-Mead simplex optimization algorithm "fminsearch" is used to minimize the log-likelihoods (Nelder & Mead, 1965; Press *et al.*, 2002). Initial model solutions were discarded, and models where fitted to the data after the solutions approached their long-term (asymptotic; steady state) behaviour. This in itself is also an important assumption, which may be challenged. Spatial discretization was Δx = 0.01, temporal discretization was Δt = 0.0001.

Flexibility for the IDE models was increased by multiplying the integer-valued observed range extent by the constant 1.815, a factor that is introduced by Drury & Candelaria to allow for fractional predicted range increases, and hence smoother model predictions. Constant speeds were calculated for the RD models by iterating them until a constant speed was obtained. Diffusion constants were calculated by rearranging

$$c = 2\sqrt{r\,D}$$

Where c equals distance divided by time, to obtain

$$D = \frac{1}{r}\left(\frac{c}{2}\right)^2$$

The initial distribution is ã, the 11 km remnant population at Point Sur. Wave speeds and diffusion constants were calculated using the entire data set 1914-1986, leaving no data for validation.

Other aspects

The models are not validated via separate data or the output of comparable models. There is no sensitivity or uncertainty analysis.

4.2.2 Discussion of the case

In the case by Drury & Candelaria (2008) there is no validation or real application of the model(s). However, apart from the explicit spatial dimensionality this test case is of interest because different models are tested against the same data set, which consists of a time series of yearly sea otter population number estimates. We assume that the goal of the model is to describe the data as well as possible for applications of projections on population numbers and spread rates. Drury & Candelaria (2008) found that the IDE models described the data significantly better than the RD models, with the exception of the transition between 1972-1973. This also goes for the IDE models that have separate epochs and north/south directions. Despite the significant increase in parameters, the AIC-value of the more complex model is the best score.

The lack of difference in likelihoods between RD and IDE models for the late epoch 1973-1986 led Drury & Candelaria to separate the data into more epochs (models 5 and 6). Based on the results, they conclude that for the late epoch 1973-1986 a RD model actually suffices. The comparable likelihoods lead to a preference of the RD model over the IDE model based on parsimony.

Drury & Candelaria (2008) remark that a further separation of the data set into epochs is not fruitful, as the amount of data per epoch then becomes too small. But it seems that in general additional data would be required to make predictions on otter population numbers and spread with some level of confidence. Habitat quality may depend on different factors, like food availability, resource competition from other species, and habitat pollution (including of anthropogenic origin). For example, (quantitative) estimates of available food sources may be useful data. Perhaps also data could be used from other species, that are somehow related in the sense that they behave similarly in terms of spreading and live under comparable conditions.

One point of concern is the transition in the data in the years 1972-1973. None of the models is capable of capturing this transition. It is unclear what causes this switching point, but it seems to be a ‘real’ thing, as there is a clear difference in the spread rates of the period before 1972 and the period after 1973. This may suggest that none of the models is actually ‘good’ enough. Observe, that this does not mean that a model that is capable of describing this data set properly is also necessarily more complex than the considered models. Drury & Candelaria (2008) furthermore point out that populations cannot spread indefinitely, which renders all the proposed models useless at some point. However, they suggest that for limited forward predictions the spread rate can be assumed to be constant, as the spread rates in the period 1938-72 and 1973-86 appear to be more or less constant. Although the transition point 1972-73 generates uncertainty, model 4 (separate north/south, pre- and post-1973) might actually be rather reliable for making future predictions for a limited time frame. Observe, however, that the application area would be very limited to short-term predictions on the spread rate of the sea otter population in California only. Maybe model 4 is a model near 'equilibrium', but one with very little use.

4.2.3 List fill-out

Meta-information (1) is missing, although most specifics are clear from the paper. Goal (2) is defined in the abstract, and re-defined here. The conceptual model (3) and the assumptions and simplifications become clear from sections 3 and 5 of the paper by Drury & Candelaria (2008). The different mathematical models are given in section 5 of the paper. There is a short, global description of the numerical model (in Matlab) in section 3.3 of the paper. Questions 7-10 cannot be answered, although the required user knowledge (9) is again rather clear from the journal context.

Data origin (11), use (12), and manipulations (13) for parameter estimation are clear. The data do not seem to be tested (14), but the data and parameter estimations are compared to other data from comparable sources. The most important data (15), namely a time series of otter population estimates, are available. It is not clear if other data should be used. The output description (16) is otter population size estimations. Uncertainty of the output does not seem to have been estimated (17). There is no real application area (18), although it is attempted to obtain and reproduce the spread rate of otters. The calibration (19) is shortly but clearly described. There is no validation (20) to a separate data set. Sensitivity analysis (21), parameter identifiability (22), and uncertainty analysis (23) are missing.

The system analysis (24) is rather extensive. Some points on scale (25) are included (like the poor coverage in the early part of the data). Again, conceptual (26), formal (27) and numerical models (28) are covered by earlier questions. The schematization (29) is limited to separation into north/south, and early/late, but plays an important role in this case. The other questions are also covered by earlier ones.

4.3 Lessons learned

Despite the innovation of applying the AIC to spatially explicit dynamical models, the 'classical' perspective of this test case meant it was not very helpful in determining the usefulness and performance of the list. There was no real application defined, and no further validation of the 'calibrated' model(s) occurred. Nevertheless, what is learned from the test case is that a spatially explicit model formulation may result in a significant increase in the complexity of a model, foremost in the number of parameters. In this case the more complex models had the best AIC scores, but at most these models had two spatially distinct components (north and south). It is not unlikely that more complex models, i.e. with more than two spatially distinct components, would quickly score worse in terms of AIC, or 'equilibrium' for that matter. Some future test cases should elaborate on this issue by focusing on spatially heterogeneous models and/or databases.

Questions 1, 7, 8, 9, and 10 (meta-information, required user knowledge, testing and verification of the numerical model) are all poorly covered again. The questions on testing and specifically verification are still useful indirectly, as a proper verification indicates that a model or database is a reasonable translation from model or design ‘on paper’ to software. In this test case, many questions could not be properly answered.

5 Aquatic biogeochemical model selection test case

A wide variety of complex aquatic biogeochemical models exist that are used as tools for understanding and predicting the response of aquatic systems to different inputs, such as nutrient loading, toxicants, and climate change. In a recent paper, McDonald & Urban (2010) applied the AIC to a suite of one-dimensional algal dynamical models, and found an optimum between increased data-fitting capacity and increased risk of over-fitting with increased model complexity. One common problem with large biogeochemical models is that there is usually a serious lack of data support. The necessary result is almost always an over-fitting of the models. Nevertheless, there is a common census to incorporate as much 'mechanistically correct' mechanisms as is deemed of interest (see references in McDonald & Urban, 2010), which in general results in significantly increased model complexity. McDonald & Urban (2010) pointed out, that using the AIC makes sense, because it is poised at a generalization of the data. The authors understandably consider this an important feature in view of the predictive capability of models.

Here we re-evaluate the reported results by McDonald & Urban (2010). The above-mentioned lack of data support and desire to include physical mechanisms makes their study an ideal test case for our list on model complexity.

5.1 Background

Aquatic ecosystem modelling, both of marine and freshwater systems, aims at describing the interactions of aquatic biota with each other and with important abiotic factors, such as light and organic nutrients. Models often have one explicit spatial dimension (depth). Light comes from above, and is a main resource for phytoplankton, while nutrients well up from below. Usually, there is an inflow of water with nutrients, and an outflow with nutrients and organisms. During a year various groups of plankton species 'bloom' and then disappear, to be followed by a new group. Most aquatic modelling is aimed at making predictions on when these blooms occur and for which functional groups (groups of species with the same 'ecological function', e.g., diatoms).

Here we focus on lake systems. Lakes are important sinks for nitrogen (Trolle *et al.*, 2008), and nitrogen may be a limiting factor for plankton growth. Phosphorus may also be a limiting factor for plankton growth, although not necessarily. A sudden influx of phosphorus can lead to algal or cyanobacteria blooms that cover whole water bodies. Lakes are divided into five ecological quality classes based on their nitrogen and phosphorus availability (Trolle *et al.*, 2008). Various types of models and approaches have been used for lake modelling. Mooij *et al.* (2010) discussed these in a review.

5.2 Test case

5.2.1 Point-by-point case evaluation

The test case is based on the 1D hydrodynamic model DYRESM-CAEDYM (Dynamic Reservoir Simulation Model – Computational Aquatic Ecosystem Dynamics Model) package (see http://www.cwr.uwa.edu.au, and references in Trolle *et al.,* 2008, and McDonald & Urban, 2010). After registration the model is freely available from the webpage, but memberships (in three classes, varying in price) give you the right to additional support, downloads, and code changes. You are asked to provide a purpose for which you need the code, and what changes you plan to make.

The original goal of DYRESM-CAEDYM was to better understand the relative importance of internal and external phosphorus loading for management purposes (Pelgrim, 2007). Phosphorus is an important driver for blooms of cyanobacteria, that produce toxins that pollute freshwater and drinking water supplies, and algae. The benefit of the original model is that the actual contribution of internal phosphorus load to the surface of a lake could be estimated more accurately, to get greater confidence in the outcome of management efforts to reduce internal or external phosphorus loading. Unfortunately, the modeling approach requires a lot of data support.

The current model versions seem to have 'outgrown' this original goal. Furthermore, via the webpage there is a free download available of all documentation regarding the software. The model therefore seems to be ideal for a model complexity evaluation. In this test case it is applied to the plankton community in Trout Lake, which is dominated by diatoms and chrysophytes. Two consecutive annual cycles of chlorophyll a have to be described as accurate as possible by four different models, all generated by DYRESM-CAEDYM. The models are of different levels of complexity.

Meta-information

The theoretical framework for DYRESM–CAEDYM can be found in references cited by Trolle *et al*. (2008, p. 221) and Mooij *et al*. (2010, p. 638). CAEDYM is a process-based suite of water quality, biological and geochemical sub-models. It can be driven by two other models, of which one is DYRESM, a 1D Lagrangian vertical stratification model. The model contains a temperature stratification, and from this stratification the dissolved oxygen and phosphorus dynamics are developed. Mixing of layers can be simulated. The most recent version of CAEDYM (v3.3, Hipsey & Hamilton, 2008) includes model compartments for suspended solids, oxygen and organic and inorganic nutrients (C, N, P and Si), various phytoplankton functional groups, zooplankton, fish, benthic biological communities (macroalgae, macrophytes and benthic invertebrates), pathogens, geochemistry (including ions, pH, redox and metals), and sediment oxygen, nutrient and metal fluxes. They are implemented as mass-balanced differential equations.

Goal

The goal of the models in this test case is to describe two consecutive annual cycles of chlorophyll a measurements as accurately as possible.

Application area

DYRESM-CAEDYM can be used for long-term simulations. It is not very applicable when a higher spatial resolution is required to focus on more complex horizontal circulation and transport processes. Instead, ELCOM-CAEDYM should be used. CAEDYM has been widely applied to study nutrient loading effects, nutrient cycling, plankton successions, and algal/cyanobacteria blooms (Mooij *et al.*, 2010, and references therein).

Data

The data used in this study consist of the following sets:

- Daily outflow measurements from stream gauges on the outflow of the Trout Lake Basin, obtained from http://waterdata.usgs.gov (with open access);
- Hourly meteorological forcing data from the nearby Woodruff Airport, obtained from http://lter.limnology.wisc.edu (access requires password);
- 10 years of water temperature measurements (1992-2002), obtained via the same source, for calibration;
- Additional data to constrain parameter estimations, i.e., expert judgment from literature and CAEDYM default values (references in McDonald & Urban, 2010);
- Field measurements by McDonald & Urban (2010) for initial conditions;
- A time series of 785 consecutive days of chlorophyll a measurements (starting on January 1st, 1992), also obtained via http://lter.limnology.wisc.edu.

Data on plankton production are often collected as GIS-data on the production of chlorophyll a, which occurs universally in all phototrophic species, in contrast to other molecular species of chlorophyll (b occurs mostly in plants, c1 and c2 in various algae, d and f in cyanobacteria). The data are used for comparison to the model output, to calculate the RSS, which in turn are used in the calculations of the AIC-values.

Assumptions & simplifications

Besides the standard assumptions (see references), some additional assumptions and simplifications are made for this case. Significant periods of freezing (135 days of annual ice cover on average) influence the Great Northern Lakes. However, the DYRESM model is not explicitly applicable for winter ice cover. Winter meteorological forcing data were therefore modified, and the model was stabilized by a combination of applying a lower threshold of 6 °C to air temperatures and fixing the wind speed at 3 m s^{-1}. A general assumption about the data is that chlorophyll a is proportional to the existing biomass and may hence be used for estimations of the total biomass.

Mathematical & numerical models

CAEDYM requires 12 state variables. Four different models, of increasing complexity, were generated from this model by the elimination of state variables via decoupling, and were specified in a nested structure. Model 1, the least complex model, considers only the effect of light on the growth of phytoplankton, while all losses are due to respiration (maintenance) and settling

$$\frac{\delta\, chl\ a}{\delta t} = \left\{ \mu_{max} \left[\frac{I}{I_S} \exp\left(1 - \frac{I}{I_S}\right) \right] - k_r \pm \frac{v_s}{\Delta z} \right\} chl\ a$$

Here μ_{max} is the maximum phytoplankton growth rate, I_S the light saturation according to Beer's Law, k_r the respiration rate coefficient, v_s the phytoplankton settling velocity, and Δz indicates the depth. All parameters and their values are given in Annex A2 of McDonald & Urban (2010). Model 2 adds the effect of temperature on metabolic processes, but still is limited to one equation. Model 3 encompasses four equations, and includes nutrient (phosphorus) limitation on growth, and adds variables for particulate and dissolved organic phosphorus (abbreviated as POP and DOP, resp.), and phosphate in the water column. Model 4 contains five equations, and adds summer sediment release (P) and sediment pools of phosphorus. The four mathematical models (sets of state equations) are given in Annex A1 of McDonald & Urban (2010), while the nesting is graphically represented in Figure 5.1, which is after Figure 1 in McDonald & Urban (2010).

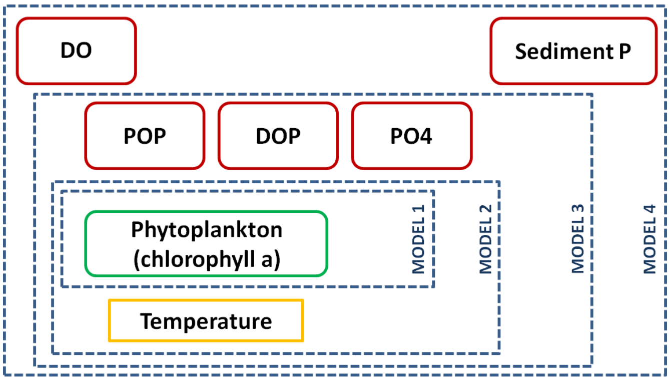

Figure 5.1. *An graphical overview of the different models by McDonald & Urban (2010), after their Figure 1. The core variable is the chlorophyll a as measure of the phytoplankton density. Additional variables are in red. Limitation by temperature is include in models 2 through 4.*

DO = dissolved oxygen
POP = particulate organic phosphate
DOP = dissolved organic phosphate

Calibration & model forcing

The 'pre-calibrated' models were forced by daily outflow measurements from stream gauges on the outflow of the Trout Lake Basin, and by hourly meteorological forcing data from the nearby Woodruff Airport. The hydrodynamic part of the models was calibrated to 10 years of water temperature measurements (1992-2002) by optimizing the value of the extinction coefficient. The used optimization function was the normalised root mean square error (RMSE) of the temperature output vs. observations. Additional biological data from literature and CAEDYM default values was used to constrain parameter estimations.

The data on chlorophyll a was used to further calibrate the 'pre-calibrated' models, from which the results would be used for the calculations of the AIC-values. For models 1 and 2 the RSS was used, while for models 3 and 4 first a different initial objective function was used (see McDonald & Urban, 2010, p. 430), only switching to the use of RSS after stabilization of the algorithm.

Other aspects

The paper by McDonald & Urban (2010) does not report any verification study, sensitivity or uncertainty analysis, or validation. This is understandable given the goal of the paper.

Results of AIC calculations

The results from the calibration of the different models and the consecutive calculations of the AIC are presented in Table 1 in McDonald & Urban (2010), and are reproduced here in Table 5.1 for easy reference. Their results show that the complexity of models 1 and 2 was insufficient to properly describe the chlorophyll a measurements. Also, inspections of the output showed that fitting led to unrealistic parameter values. Model 1 can only fit somewhat to the data by adopting a quick dispersal of the spring production throughout the water column. Model 2 can only fit to the data when the light saturation coefficient is very large. This reduces the effect of light, instead leading to a strong correlation with temperature patterns. This corresponds to what was put in by model 2 as compared to model 1, namely the effect of temperature on metabolic processes.

Table 5.1. *A reproduction of Table 1 in McDonald & Urban (2010), containing the results of their study. While the most complex model (model 4) gave the best fitting results (lowest RSS), the middle model (model 3) gave the best trade-off between fitting and model complexity.*

Model	k	RSS	AIC	$AICc$
Model 1	7	2932	677	677
Model 2	12	2003	579	580
Model 3	24	689	300	304
Model 4	39	638	308	321

Model fit improved significantly by introducing phosphorus in model 3. Spring bloom timing, location, chlorophyll a maximum, and fall bloom all agreed with the data. Modelled concentrations were similar to the measured values. The winter months still gave a bad fit. However, this is not surprising, as the annual ice cover period is not covered well by the model anyway, as discussed in the "assumptions & simplifications" above. From the point of the application this could be considered irrelevant if the model is not used for the months with ice cover.

Model fit increased a bit more when more details are added on phosphorus dynamics in model 4. Graphically, the fit of model 4 is not better than that of model 3 (McDonald & Urban, 2010). Sub-thermocline chlorophyll a concentrations in the summer of 1993 are notably better in the fit by model 4.

5.2.2 Discussion of the case

This test case explicitly uses the AIC to evaluate the complexity of the models. The use of the AIC (or a similar criteria) only makes sense when the model under evaluation is compared to similar models for the same data. Selecting a data set that is appropriate for the application area will help a lot, obviously. McDonald & Urban (2010) pointed out that the fit and the consecutive AIC-values lead to a greater reliability of model 3 only based "on the data [they] have available [...] referring to chlorophyll a data" (McDonald & Urban, 2010). Although the use of chlorophyll a data is common for lake and marine ecosystem modelling, the main goal of the original DYRESM-CAEDYM model was evaluating the management of lakes with regard to phosphorus loading. It would have been interesting to consider also some measurements of phosphorus in this test case, then re-evaluate the performance of the different models. Based on the use of chlorophyll a data, the verdict is that model 3 is the best balanced model in terms of model complexity and data-fitting. However, it is unclear if model 3 is also the best balanced model when the management application is considered. Model 1 and 2 obviously are not in equilibrium, as the phosphorus is not considered at all. Models 3 and 4 do consider phosphorus, but only indirectly as the data and the evaluation of model complexity focus on the biota.

The one-dimensional spatial model formulation with dynamic equations is a common approach, and probably a proper one considering the application. Lakes are not exactly chemostat devices, as they 'leak' water and resources to the surrounding soil (and vice versa), and there is no active mixing. Nevertheless, there is usually a certain amount of mixing caused by thermoclines to homogenize the water at fixed depths, which validates a similar mathematical description. It is doubtful if it would add much to the realism of the model predictions of chlorophyll a when the model would be three-dimensional instead of one-dimensional. The vertical stratification makes sense in order to explicitly consider the effect of light. After all, light is one of the main 'resources' for planktonic growth. Unfortunately, the 'equilibrium' evaluation of this test case remains somewhat limited, as no alternatives like zero- or three-dimensional alternative models, and/or models with explicit interactions between bottom soil, sediment, and water, are considered.

In all considered model formulations additional data is used for forcing (outflow measurements, meteorological data, water temperature), replacing explicit dynamic descriptions or boundary conditions. This seems prudent, although in some cases the scale is 'off', e.g., the meteorological data is on an hourly base, which is probably too detailed compared to the daily chlorophyll measurements. And again, the important variable phosphorus is missing in the data support. The winter forcing is very artificial, and it is unclear what this means for the model choice or use for the application. The application area of the original model does not include explicit winter ice cover modeling. Therefore, in this test case there is an obvious mismatch between intended and actual application area, but it is uncertain what the implications for equilibrium are. There is no uncertainty analysis that might shed some light on the uncertainties that are introduced via this artificial fix.

5.2.3 Advice for improvement

In view of the data support desired by any management application, the model would be better balanced by incorporating time series of phosphorus measurements. There are some uncertain issues that could be dealt with in any evaluation of this test case. There are many possible modelling options that have not been investigated, because of the limitations of the modular set-up of DYRESM-CAEDYM, and probably to limit the number of models to select from. The current status of this case is such, that it is not clear how far off the 'best' model (model 3) is from equilibrium. The best advice to give at the moment would be to include phosphorus measurements, and to consider some of the mentioned model alternatives.

5.2.4 List fill-out

Meta-information (1) can be found on the webpage http://www.cwr.uwa.edu.au, and there are many documented references. The original purpose of the model (2) is given by (Pelgrim, 2007). The conceptual (3) and mathematical model (5), as well as the assumptions and simplifications (4), can be found in McDonald & Urban (2010) and references therein, and on the webpage in the interactive guide. The numerical model (6) and program description (8) is available with supporting Matlab scripts after payment. User knowledge (9) is available, also through courses. Verification (7) and tests (10) will probably have been performed, but documentation regarding these issues cannot be found.

There is a description of the input and data (11), for what it is used (12), and what manipulations (13) have been done (as far as it can be checked). It is unclear if the data has been tested (14). From the test case it is rather obvious which data is missing (15). Output is described (16) on the webpage. Uncertainties (17) are unclear. The application area (18) is water quality management. The calibration (19) is described in the paper. There has been no real validation (20) for this specific test case. There is also no information on SA, UA, or the identifiability (21-23).

The system analysis (24) is clear from the paper. Extent, support, and coverage (25) of the data are clear from the paper or additional sources. This test case has a one-dimensional schematization (29), namely lake depth layers, which is described in the documentation on the webpage. Other questions (26-28, 30-32) have been covered already.

5.3 Lessons learned

The evaluation list EMC v0.1 misses explicit questions about for instance the use of AIC. As already indicated before, the AIC is not very explicitly aimed at considering the application area of the model, other than assessing the prediction performance of a model. Nevertheless, it may be useful in any assessment of the complexity of a model to consider the AIC as a first indication whether the complexity is too low, probably sufficient, or too high. Therefore, this point might be incorporated in the upgraded version of the list, if nothing else just to stimulate modellers to consider the complexity of the model. On the other hand, it is not clear how the AIC or other such criteria would be of relevance when trying to assess the complexity of a database. A future test case might be aimed at evaluating this. This requires first of all a useful definition of 'complexity' for databases. As already proposed in the introduction, initially one can think that properties like the number of grid cells, the number of possible cell states, and the number of permutations on the input of the database have a role to play in this, but we have not yet come up with a satisfactory and practical definition. Possibly information-theoretic ideas like the 'minimum description length' principle can be of importance when working this out (see Grunwald, 2007).

The test case shows that several points are superfluous, or at least too prominent, while others should get more attention. Again, it became apparent from the evaluation that the application should be more important and explicit.

This test case was closest to what would be a 'real' case in terms of the use of a model or database for an application. Indeed, there is a webpage with clear documentation, user support, and version numbering, indicating the model is probably well maintained. Unfortunately, there are many issues about which we are uncertain. Alternative model formulations have been suggested, but could not be tried, as not all data are freely/easily accessible. The positive side is that the list prompted all these issues, i.e., "the questions have been asked". This test case therefore perhaps best demonstrates the effects of raised awareness on the subject of model complexity stimulated by using the list.

6 Discussion of the results

For the testing of the ECM v0.1, we have to determine how the list performed. For that, an inventory is made for each test case in the preceding chapters on which questions were easily answered, answered with difficulty, or not answered at all. These results are combined in Table 6.1 below. This provides some overview of information on the validity and utility of the different questions. An explanation or some extra information will be given with all these answers if necessary. Furthermore, it is discussed if there were any issues missing. Furthermore, we reconsider the results of the expert review from Chapter 2.

6.1 Evaluation of the expert review

As already mentioned in Chapter 2 the expert review has led to several useful suggestions with regard to improving the list, and with regard on what issues should be considered in any evaluation. As it was considered logical to follow the modelling cycle, at the highest level the set-up of the list has remained the same. However, several remarks led us to reconsider the whole set-up of the questions, specifically where it comes to the role of the application in 'equilibrium'. Several remarks are directly usable for the formulation or reformulation of questions. The specifics will become clear below, where we will consider which questions should be kept, which are superfluous or should be changed, and which points are still missing. Then, in the next chapter we will provide the new version of the list.

Some remarks by the reviewers have been acknowledged, but have not (yet) led to changes in the list. Most notably are the questions that seem to indicate that the history of the model or database should be included. Specifically remarks 4 (gradual increasing of model complexity) and 16 (the 'Concorde' effect) suggest that it is necessary to have knowledge about the previous versions of the model. However, this aspect has not been included in either the prototype EMC v0.1 or the updated version EMC v1.0.

6.2 Evaluation of the test case results

The results of all the test cases are summarized in the colour-coded Table 6.1 on the next page. The first column gives the number of the question, corresponding to the number of the question in the EMC v0.1. The second column gives the clarity of the question. Most questions were quite clear in view of the test cases. For each test case there are two columns. The first column indicates whether or not the question was difficult to answer. The second column indicates qualitatively the relative contribution of the answer to the eventual determination whether the model was well balanced or not as per our judgement. The lowest row summarizes the final verdict in each test case. For more details see the respective test case. Three colours are used for classification: blue represents that the question was clear, easy to answer, or/and had a significant contribution to the final evaluation of 'equilibrium', while red means the opposite, and yellow is obviously in between the two.

Table 6.1. *Colour-coded overview of how questions were answered in the different test cases. Three elements are discriminated: Clarity of the question (Q), the possibility to answer the question in the test case (P), and the qualitative contribution of the (answer to the) question to the understanding of 'equilibrium' (U). The results have been separated into three classes: Blue: high score (clear, relevant); Yellow: middle score (not entirely clear, not very important); Red: bad score (completely unclear, no information at all, unimportant); NA: not applicable. If a question scores red or orange in Q or P, it usually automatically becomes irrelevant to continue the evaluation. The question can then not lead to understanding of 'equilibrium' anymore. A cell with "> no." refers to that question, and implicitly indicates that the question may be superfluous. The last row summarizes the judgement of the cases based on the evaluations, where blue indicates that the model is probably properly balanced, while red means the model is far from 'equilibrium'. X: The information was available in principle, but not used. XX: With the co-notation that both model and application are very limited. XXX: Model 3 was the best model, but no phosphorus data and no alternative models were considered.*

Case		**3.2 Control**		**3.3 Trehalose**		**3.4 Biofilm**		**4 Sea otter**		**5 Lake**	
Question	Q	P	U	P	U	P	U	P	U	P	U
1	🟦	🟥		🟥		🟥		🟥		🟦	🟥
2	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
3	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
4	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
5	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
6	🟦	🟨		🟨		🟥		🟦	🟦	🟦	🟥 X
7	🟦	🟥		🟥		🟥		🟥		🟥	
8	🟦	🟥		🟥		🟥		🟥		🟦	🟥 X
9	🟦	🟨		🟨		🟨		🟨		🟦	🟦
10	🟦	🟥		🟥		🟥		🟥		🟥	
11	🟦	🟦	🟨	🟦	🟨	🟦	🟨	🟦	🟨	🟦	🟨
12	🟦	🟨	🟨	🟦	🟨	🟦	🟨	🟦	🟨	🟦	🟨
13	🟦	🟦	🟨	🟨		🟦	🟨	🟦	🟨	🟦	🟨
14	🟦	🟦	🟨	🟥		🟥		🟨		🟥	
15	🟦	🟦	🟨	🟦	🟨	🟦	🟨	🟦	🟦	🟦	🟨
16	🟦	🟦	🟨	🟨	🟨	🟨	🟨	🟦	🟦	🟦	🟦
17	🟨	🟨	🟨	🟨	🟨	🟥		🟥		🟥	
18	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟨	🟦	🟦	🟦
19	🟦	🟨	🟦	🟦	🟦	🟥		🟦	🟦	🟦	🟦
20	🟦	🟨	🟦	🟨	🟦	🟦	🟦	🟥		🟥	
21	🟦	🟨	🟦	🟦	🟦	🟥		🟥		🟥	
22	🟦	🟦	🟦	🟨	🟦	🟥		🟥		🟥	
23	🟦	🟥		> 17		🟥		🟥		🟥	
24	🟨	>3,4		>3,4		>3,4		>3,4		>3,4	
25	🟨	🟨	🟦	🟨	🟦	🟨	🟦	🟦	🟦	🟦	🟦
26	🟨	🟨	🟨	>3,4		>3,4		>3,4		>3,4	
27	🟨	🟨	🟨	>5		>5		>5		>5	
28	🟨	> 6		>6		>6		>6		>6	
29	🟨	NA		NA		NA		🟦	🟦	🟦	🟦
30	🟨	>17,21,23		>17,21,23		>17,21,23		>17,21,23		>17,21,23	
31	🟨	>19		>19		>19		>19		>19	
32	🟨	>20		>20		>20		>20		>20	
Verdict		🟥		🟥		🟦		🟦 XX		🟨 XXX	

Looking at Table 6.1, there are some surprising results. For instance, consider the biofilm test case. Despite the large number of questions that were difficult or impossible to answer due to lack of information, the final judgment of the model(s) is very positive. In part this may be because of the very solid description of the data, which is more rich than the data supplied by the other two test cases on beer-brewing. Few or none of the model parameters were 'calibrated' via other sources with possible different meanings or different application validity, which gives some form of confidence in the validity of the parameter values. For another part this may be contributed to the 'cleanness' of the models. All important dynamical model outputs are measured in time. And for an important part this may be because the test case performed a model selection experiment, starting at a very minimal model which only describes biofilm yeast, then increasing the model complexity in small and motivated steps. Ironically, it seems that despite the lack of 'modelling background' of the test case (e.g., no description of the calibration) the results do not have to be bad. Nevertheless, we have to be critical about this. Suppose the results would not have been so good. Assuming that the list is applicable, in that case it would seem a good starting point for improvement of the model to "fill in the gaps".

6.3 Performance of questions and missing issues

Several questions were found to be superfluous, but many proved useful in some way. However, not all of those questions contributed equally and/or directly to obtaining an 'equilibrium' evaluation.

Question 1 on the meta-information is a typical question for documentation evaluation (it was copied from the status A evaluation list for quality assurance of documentation). For the most part the question gives information that is not or only very indirectly relevant for determining equilibrium, except for the version numbering of the model or database. In practice different model and database versions are used for different applications, as often a new version is produced to cope with the new application. It should be asked explicitly which version is used for which application(s). The rest of the question is superfluous.

Questions 2 through 6 provide important information, albeit indirectly. The application plays an important role in the steps from system analysis through formal model, but now it is only considered implicitly during these steps. The expert review already suggested that the application(s) should be considered more explicitly, for instance, as there may exist 'equilibrium' during the whole modelling cycle. This will require to challenge the model or database to the application at every stage. In practice the application determines the output, e.g., one is interested in values of some variable per square km per hour. The system analysis, schematization, scale, conceptual and formal model provide the background to the means how to obtain these values, namely the numerical model or database, and what data and input is required. The questions should be kept, but it makes more sense to merge all these questions into more application-oriented questions on in- and output.

The questions on verification (7) and testing (10) are relevant for the application, to determine if the numerical model or database is a proper translation of the formal model. If this is the case, only then the model or database can really be validated to evaluate whether or not it is an acceptable representation of the modelled system (no matter what the application is). But they are also only implicit.

Question 9 on the user required knowledge is hardly available, but very important. This knowledge indicates what a model or database user should know, not only to deal with the numerical model, but also to use it properly for any application.

Although all questions 11-17 are relevant, they do not directly address the point of application and balanced model complexity. Rather, they provide background information to the evaluation, and even more so than the questions on system analysis, etc. Nevertheless, they need to remain in the list.

Calibration and validation are important issues in view of any application (although for databases calibration is in most cases meaningless). A model is always parameterized to be used for some application. A proper validation should help to determine if a model or database is an acceptable representation of a modelled system, or at least an acceptable representation in view of the application. Sensitivity analyses often play a role in the calibration of a model, and help to gain insight in what are important variables, parameters, etc. in a model. An uncertainty analysis is useful to make explicit what inputs, data sources, etc. introduce a lot of uncertainty about the model output. The application will determine what level of uncertainty is acceptable, and an UA may reveal on which points the data and model or database can be improved. It is not really clear if and how a model or database is calibrated or validated in terms of the application(s). A model may be validated for some part which is not the most relevant part for the application(s) for which it is currently used. This aspect should be included more prominently. The questions on data manipulation (13), data quality (14), and model and parameter identification (22) are related and could be merged in some fashion. There is also a great deal of overlap between questions 21, 23 and 30 (SA and UA), and between questions 11-15 and 19-20/31-32 (data description, and calibration and validation). They should be merged and rephrased. The (questions about) calibration and validation should be focussed on the application, and it that sense questions 19 and 20 are not bad.

There is significant overlap between questions 24, 26, and 27, with 3 through 5 (on the conceptual and formal model formulation). Furthermore, question 29 (schematization) and also 25 (scale) overlap with 3-5 if applicable. The later questions ask for more specific details, and also require to consider alternative formulations.

There are also several missing issues. In view of ‘equilibrium’ of model complexity it makes sense to consider explicitly the matter of redundancy, and the efficiency and utility of the model or database. Are all assumptions and simplifications that were made necessary? Why (not)? What effect did they have on the modelling framework (e.g., some assumption led to a boundary condition for X)? What is the effect of discarding them (e.g., X becomes unbounded)? Is that problematic in view of the application? Efficiency (of the numerical code) is valued in view of the application. In theory a conceptual model or database design can be implemented in different ways. Possibly one implementation is much more complex in terms of, say, running time, or utility then another. These elements should be included in the updated version of the list.

There is also no distinction between intended and actual application area. The list currently does not distinguish between the different applications of a model or database. A model or database may very well be balanced for one application, while very much off balance for a different one. Also, the goal should match with the application(s). What is the degree of overlap of the goal with application 1, application 2, etc.? How is this quantified? If it is not ca. 100%, how much of a problem is this? Is it not better to limit the scope of the model, and device separate models for some of the applications?

The same is applicable for the data and input, and for comparing the current output of the model or database with what is actually desired in view of the application. What is the degree of overlap of data requirements for application 1 (and others) with current data supply? If not a 100%, is this problematic, and why (not)? What could be done to improve this overlap, if necessary? What is the degree of overlap of output requirement for application 1 (and the others) with current model output? If not a 100%, are there missing or superfluous outputs? If superfluous, are there uncertainties introduced in the relevant output? If missing, how much should be added to the model to produce these outputs (i.e., adapt the model to the application)? It is probably best to split this evaluation for each distinct application, as applications will likely differ in their output requirements. Other questions that may be considered: Has any analysis been done already to assess the data quality or complexity trade-off (like AIC)? What kind of data could contribute positively to the model in some way? The answer to this question depends for a great deal on the (intended) application(s). Such an evaluation would require expert knowledge on the subject, but the mere presence of the question may invoke a contribution to 'equilibrium'.

6.4 List adaptations

Based on the findings, the set-up of the list had to be reconsidered. The questions of the second part of the list are actually very much composite questions, and they will have to be separated into sub-groups of questions. Some questions do not involve the application at all, while others do. There is much partial overlap between questions on certain aspects in which the same aspect is covered by two or three questions, for instance, there are more than one question focussed on calibration or validation.

In the updated version of the list the questions are divided into subsections (that we will henceforth refer to as 'questions'), based on the steps of the modelling cycle. Each 'question' is then divided into 'sub-questions', which roughly follow the same pattern (excluding the introductory question):

- First the model step (say, the system analysis) is discussed,
- Then an inventory is made of the requirements by the application(s), and if applicable the possibilities by the data,
- And then the current state is compared to this inventory, and conclusions are drawn with regard to any (mis)matches.

This set-up of questions that are divided into sub-questions with specific focus on data and application requirements has specifically been designed to generate a 'conflicting atmosphere', in order to magnify weak points. For instance, should there be a mismatch between the intended application(s) of the model and the goal it was initially designed for, then the set-up of the questions should reveal this mismatch.

The questions have also been reformulated to make practical gain. The questions ask specifically for citations in order to avoid confusion. It has also been attempted to reformulate the questions as clear as possible to be useful by modellers in general. Short examples are provided where they were deemed necessary to aid the users of the list.

In Chapter 7 we present the updated version of the list, EMC v1.0.

7 Updated list EMC v1.0

In this chapter we provide the update version of the evaluation list, ECM v1.0. The list is divided into 12 questions, with further subdivisions into sub-questions. Here we provide the English version of the EMC v1.0; the Dutch version is published in WOt-paper 11 by Van Voorn *et al.* (2011) that accompanies this publication.

1 Basic data

Provide the following data: Model/database, name, version, revision number, date of publication. By model or database we mean the source code plus input data.

2 Goal & applications

A What was the goal of the model or database?
B What was the intended application area, and how do we confirm this?
C What are the current applications? Does the model or database play a role in any model chains? If yes, which ones and what role?
D Does the intended application area overlap all current applications?
E If relevant, how much do the current application areas overlap each other?

3 System analysis

A Provide an analysis of the system that is being modelled. This may be a reference to some publication. Aspects of a system analysis are: What are the important attributes and processes in the system? What feedbacks exist? How have the boundaries on the system been determined? How are relevant attributes and processes outside the system boundaries handled, e.g., are they ignored, replaced by constants, by forcing data, etc.? About which processes, attributes, feedbacks, etc. If the knowledge is uncertainty, how is this uncertainty determined?
B Per application, determine which system analytical aspects are directly relevant for the application? And which ones are less or not important? Motivate this relevance.
C Have all aspects that are important for the application been included in the model or database? And which ones are missing?
D Have also less or not important aspects been included?
E Give an initial judgement of 'equilibrium' based on the answers to questions 3C and 3D.

4 Conceptual model or database design

A Provide the conceptual model or the conceptual design for the database (like a flow chart). This may be a reference. The conceptual model determines the relationships between system components in a relatively informal way, for instance, an ecological food web. Aspects that may be included are: What assumptions and approximations have been used? What are the spatial dimensions? What spatial-temporal aggregation levels are used? What aspects are important for a database? What rules should be used for (dis)aggregation of data? For models: What model type is used?
B Consider per application at which (spatial-temporal) aggregation levels answers are required, given the application. What spatial dimensions are required? What model type or design of the database is required?
C Do the spatial dimensions, model type, aggregation levels, etc. of the model or database match those desired by the application?

5 Data

A What data are required, given the application, and explain why? Be specific with regard to resolution, accuracy, scale, etc. and avoid general and unsupported remarks like “File X is needed”.
B What data are available, and how are they used (e.g. as forcing data, or parameter values)?
C What is eventually the input of the model or database?
D Does the input match the data requirements of the application? Compare the answers to questions 5A, 5B, and 5C. Pay attention to units and the three elements of scale (coverage, extent, support; see Bierkens *et al*., 2000). Does the data have the right dimensions and units? If not, is there are (dis)aggregation method available? Scaling issues arise as most attributes like time and space are infinite, but are measured on discrete scales with limited extent: imagine the ‘measuring’ of a continuous normal distribution using limited samples divided into histogram classes, and the histogram has an upper and lower boundary. For spatial discretization into grid cells (for example, of river catchment areas) it is commonly assumed that the values of variables are the average of the measurements at the corners of the grid cells (although there are more sophisticated approaches, like ‘kriging’). More information on scaling issues can be found in Bierkens *et al*. (2000; but see also Bogaart *et al.*, 2011, in Dutch).

6 Formal model *(not for databases)*

A Provide the formal model and motivate, or give specific references (page numbers) to published reports, papers, books, etc. The formal model is the conversion from concept to e.g. mathematical equations, or a set of rules. The choice of formal model depends on the intended application and data, and is partly determined by the required uncertainty margins.
B What data are required for this formal model?
C What data are available for this model?
D Have the data that are required and available also been used?
Use the answers of questions 5A through D to address questions 6B through D.

7 Numerical model or database

A Provide the numerical model or database (the implementation of the formal model or database design in software) and motivate, or give specific references (page numbers) to published reports, papers, books, etc.
B Describe the verification, or give specific references (page numbers) to published reports, papers, books, etc. The verification should not be confused with validation. In a verification study it is evaluated if the numerical code is a proper conversion of the formal model or database design. It does **not** evaluate whether or not the code is a proper conversion of the system for which it is applied.
C Describe and motivate the choice of numerical (calculation) methods and the used discretization, or give specific references (page numbers) to published reports, papers, books, etc. Usually for numerical methods there is a trade-off between calculation time and accuracy.

8 Schematization

The schematization is e.g. the division of the soil into separate layers, or the division of a geographical region into different smaller regions like land use patches. The level of detail of the schematization depends on the application.
A If applicable describe the schematization and motivate, or give specific references (page numbers) to published reports, papers, books, etc.
B The level of detail of the schematization should also be sufficiently supported by the data. Is the schematization sufficiently supported? Motivate the answer.

9 Sensitivities, uncertainties & numerical coding

This question is focussed on assessing which parameters, forcing or inputs, parts of the numerical coding (for instance, the used integration scheme, or some module), etc. are important, and which ones may be superfluous, or introduce a large quantity of uncertainty about the output.

A If available describe performed sensitivity and/or uncertainty analyses, or give specific references (page numbers) to published reports, papers, books, etc., and analyse and discuss the results. Some factors may not be important for the application. These may be skipped from the calibration, or be used for model reduction. Other factors may be essential.

B ***(not for databases)*** What numerical integration method has been used? Discuss its performance. The errors and hence uncertainties caused by numerical methods can be larger than those caused by data (Clark & Kavetski, 2010).

C What is the role of each module of the code, for instance a crop growth module of a hydrological model, or steps in a flow chart? Discuss these roles in view of the application. Are there any superfluous modules? Are there missing modules? Motivate the answers.

D Judge the numerical model or database complexity in terms of calculation time and efficiency. What is the running time? How much (diagnostic or final) output is there? What is the size of the data stream? Could the code have been more efficient or could the running time be reduced, e.g. by selection, aggregation, implementing different methods, different modular design, etc.?

10 Calibration *(in principle not for databases)*

Calibration is the process in which parameters, initial conditions, etc. are given values. Calibration is often preceded by sensitivity analysis.

A If available describe the performed calibration, or give specific references (page numbers) to published reports, papers, books, etc., and analyse and discuss the results. Indicate what objective functions (local, global, deterministic, stochastic) were selected, if and how the confidence of the calibrated parameters is quantified, with which data, what (cross) validation has taken place, etc. with available references to the literature. Be mindful of over-fitting, multiple local optima, identifiability problems.

B What demands does the application have with regard to the accuracy with which parameter values, etc. are being determined.

C What possibilities do the data provide with regard to the accuracy with which parameter values, etc. are being determined. Consider the resolution, accuracy, amount of data, aggregation level(s), etc., and be mindful of over-fitting.

11 Validation

There exist several definitions of validation, but it is defined here as evaluating if a model or database is a proper representation of the modelled system. This is ideally done on the basis of independent data, i.e., data that have not been used in a previous stage of the modelling, and moreover which are representative for the intended application(s).

A If available describe the performed validation studies, or give specific references (page numbers) to published reports, papers, books, etc.

B Evaluate the value of the different validation studies. How relevant are the different studies for the different applications? Are all applications covered by the validation studies? For which applications should additional validation studies be performed?

12 Conclusion

This last question is aimed at combining the answers to the different questions and the opinions of the different evaluators involved. It may indirectly lead to advice on how to improve the model or database, if applicable.

A Judge the amount of general confidence in the model or database, also based on what has been done with regard to testing, verification, calibration, validation, sensitivity and uncertainty analysis, and motivate this.

B Are things missing? Are there many points for which the model or database was judged to be ‘too simple’?

C Are there superfluous components? Are there many points for which the model or database may be simplified?

D Is there sufficient support from the data for the complexity of the model or database? How is this motivated?

E What specific suggestions for model or database improvement and data provision emanate from the foregoing analysis?

8 General discussion and prospects

In this final chapter we discuss the prospects with regard to the further use and development of the list, and the project on ‘evaluation of model complexity’ in general.

8.1 List update discussion

The test cases and expert review of the prototype of the list EMC v0.1 have revealed that the general approach is sound, but that the set-up of the prototype suffers from various shortcomings. The approach by going through the various steps of modelling (going from analysing the system to validating the final model) makes sense. The thought behind it is when you consider ‘equilibrium’ in each separate modelling step then the end result will have a much better chance of being at ‘equilibrium’ also. Various questions on the prototype of the list are valid and important, but the overall goal of being able to assess the model complexity is not fulfilled by the prototype. Most importantly, the role of the application is underappreciated. This is understandable, as the original analysis of the subject by Bogaart *et al.* (2011) was heavily rooted in the part of the academic literature that focusses on the philosophical aspects of model complexity. However, the list is intended to evaluate models that have clear applications for predictions and policy evaluations. Therefore, in the updated evaluation list EMC v1.0 the role of the application is now very explicitly considered.

The set-up of the list has been redesigned to ‘generate conflict’ by dividing different questions into sub-questions, *i.e.*, generally it is asked in one sub-question what *should* be used, and then in the next sub-question what *has* been used. This type of questioning is more likely to expose weaknesses of models and databases. It has also been attempted to adopt a more ‘open’ structure to include also databases and other types of models, like cellular automata. Various questions on different steps of the modelling cycle have been reformulated to be as practical as possible, for instance, by asking explicitly for references.

8.2 Remaining issues

Several issues remain, which will have to be tackled in the continuation of the project:

1/ It remains to be seen how user friendly the list EMC v1.0 really is. For now it is advised to use the list in a small setting including at least one of the original developers of the list and at least one original model developer of the model that is being evaluated. This will work as a dual learning environment, and will avoid confusion about what is meant by each question. The authors of the list can provide background information to motivate why a question is part of the list. Furthermore, keeping a setting from becoming too large ensures that the effects of ‘peer pressure’ are avoided. It is more likely that relevant issues and also conflicting points of view will emerge from the analysis when the different involved model developers, users, etc. are interviewed individually, instead of when they are together.

It is probably useful to expose either the list EMC v1.0 or the upgraded version to expert review again. This review should consist not only of a review of the contents (the technical side), but also a review of the presentation. This latter review might be performed by people from social sciences. Review issues for a presentation review may concern the optimal number of questions, and the way they are ideally posed. The number of questions is a constraint, as too few questions will likely not generate enough in-depth information for a proper evaluation, but too many questions become a burden on the concentration of the evaluators. It is important to get the right technical information from the list, but modellers should not become discouraged by too large a number of (difficult) questions.

2/ The list has not been subjected to any test cases regarding databases, meta-models, or model chains (combinations of models and databases containing coupling between these different components). In the follow up of this research these gaps will be addressed by selecting suitable test cases. Two test cases will be MetaSWAP (Van Walsum & Groenendijk, 2008), a hydrological model that is part of the NHI (the Netherlands Hydrological Instrument, "Nederlands Hydrologisch Instrumentarium", see www.nhi.nu), and the Nature Planner (in Dutch the "Natuurplanner", Van der Hoek & Bakkenes, 2007), which is a model chain used for evaluation of the policy regarding the sustainability of Dutch nature. This latter case involves a set of models and databases with a clear application in policy evaluation, which makes it an ideal test case for testing the list with respect its application to model chains. A suitable meta-model and suitable database case still needs to be selected at this point. Observe, that these test cases will have a dual function: both the performance of the EMC list v1.0 and the model, database, or model chain will be evaluated. It is probable that some questions on the list will have to be adapted or additional questions will have to be formulated to deal better with meta-models, model chains and data bases.

With the planned evaluation of the Nature Planner, one important point that needs to be considered for the evaluation of model trains is the strategy that needs to be followed. In practice, the step from one individual model or database to a model chain is not trivial. There seem to be several strategies available. The first one is to first perform an analysis of the whole model chain, and then a more detailed analysis of the individual models and databases involved in the chain. The second one is to first perform an analysis of the individual models and databases involved in the chain, and then an analysis of the whole model chain. In theory, all individual models and databases might be well-balanced in view of equilibrium while the model chain is not. For instance, when the individual models and databases are well-balanced but have different application areas, then it is not very likely that the model chain is well-balanced. The other extreme, in which not all individual models and databases are well-balanced while the model chain is, seems less likely. The most application-oriented approach for the Nature Planner is to start with the main application of the model chain and see what type of output and decision criteria are desired. Then, determine how the output and decision rules come about, and what demands this puts on the model complexity.

3/ Although we discuss several viewpoints on defining 'model complexity' and trade-off measures for model complexity, the list does not explicitly contain any definition or quantitative trade-off measure with regard to complexity or 'equilibrium', not to mention a way of quantifying the overlap between goal and application areas (as is asked for in question 2 of the list EMC v1.0). It is therefore difficult to establish what quantitative conclusions may be drawn with regard to how far off a model or database is from 'equilibrium'. The conclusions from the test cases in Chapter 3 were very qualitative in nature, and may be contested by others. It may be necessary to add a way of quantifying the answers to the different questions. However, this approach is not very desirable, as there will be a lot of detailed information in the diversity of opinions which is lost when simply grading the questions from, say, 1 to 10 and averaging the different grades. Also, a possible extension is to add the question "How do you (the modeller) define model complexity?". Although this will not directly solve the issue on what is defined as model complexity in the list, or on how it should be defined, it may reveal more insight into the subject, both for the interviewing and the interviewed people.

8.3 Prospects and suggestions

Apart from dealing with the above-mentioned major shortcomings, there are several other points to mention, and possible research directions in which this project can be taken:

1/ It is the intention that the evaluation list EMC eventually becomes a part of the quality control system of the WOt N&M ('Wettelijke Onderzoekstaken Natuur & Milieu'). Currently the WOt N&M runs a project to ensure a minimum quality level of the numerical models and databases developed at the Wageningen University & Research Centre (see http://www.wotnatuurenmilieu.wur.nl/UK/Model and Data Quality/). This quality evaluation mainly focusses on documentation, maintenance and management (so-called status A). In the near future it is intended that the quality control will also incorporate reviewing the contents of a model or database (so-called status AA). The EMC-list will provide (a part of) the tools to do this.

2/ For the project continuation there are several possible considerations for extending the list. The first is to add functionality with regard to possible directions for model improvement, after the list has revealed issues. The list may provide guidelines and useful references for improvement of an evaluated model or database, this also with regard to the future development of Status AA. For example, if the evaluated model is being used for too many application(s) it might be advisable to split the development lines of the model, in effect creating different models for different applications. If the data support is insufficient, how can this be solved? Can other data be used instead as a proxy? What extra uncertainties are incorporated then? Is that still acceptable in view of the application area(s)? Can model components be reduced, replaced, or removed? For instance, it may be possible to replace a numerical method by a faster one. At the moment the list can be used for giving advice, but only in an unofficial capacity.

A second possible extension might be to include the already mentioned aspect of history of the model or database. How did the model develop? What decisions were made, and why? This in view of the reviewer's remarks in Chapter 2, specifically remarks 5 (a gradual increase of model complexity) and 16 (the 'Concorde' effect), which suggest that it is necessary to have knowledge about the previous versions of the model. A possible scenario is that a model at some point was better balanced than its current version, for example, because it has been adapted to fit more applications. However, this aspect has not been included in either the prototype EMC v0.1 or the updated version EMC v1.0.

3/ If the list is to be expanded in the above-intended way(s) it is wise to bear in mind all the factors mentioned in Chapter 1 with regard to increasing complexity: it should be avoided to come up with a list that in itself is an overly complex 'snail'. Also in the case of the list the application plays an important role. Perhaps a separation into 'list modules' is possible, or several distinguishable lists should be developed, one for each application (e.g., model evaluation, model improvement, etc.). It might be an interesting test case to evaluate the list 'to itself'. One might argue if all questions are equally important: perhaps a list of only a few questions will in practice already provide a useful answer with regard to 'equilibrium' in a large percentage of the cases.

Literature

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716-723.

Aranda, J. S., E. Salgado & P. Taillandier (2004). Thehalose accumulation in *Saccharomyces cerevisiae* cells: experimental data and structured modeling. *Biochemical Engineering Journal* 17, 129-140.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology* 320, 18-36.

Bierens, H. J. (2006). *Information criteria and model selection*. Pennsylvania State University.

Bierkens, M. F .P., P.A. Finke & P. de Willigen (2000). *Upscaling and Downscaling Methods for Environmental Research*. Kluwer Academic Publishers, Dordrecht.

Blum, M. (1965). On the size of machines. *Information and Control* 11, 257-265.

Bogaart, P. W., G. A. K. van Voorn & W. Akkermans (2011). *Evenwichtsanalyse modelcomplexiteit – een verkennende studie*. WOt-werkdocument 226. WOT Natuur & Milieu, Wageningen UR, Wageningen.

Box, G. E. P. & G. M. Jenkins (1970). Time Series Analysis – Forecasting and Control. San Francisco: Holden Day.

Brányik, T., A. A. Vicente, G. Kuncová, O. Podrazký, P. Dostálek & J. A. Teixeira (2004). Growth model and metabolic activity of brewing yeast biofilm on the surface of spent grains: a biocatalyst for continuous beer fermentation. *Biotechnol. Prog.* 20, 1733-1740.

Chou, I-C & E. O. Voit (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences* 219, 57-83.

Chwif, L., M. R. P. Barretto & R. J. Paul (2000). On simulation model complexity. *Proceedings of the 2000 winter simulation conference*, 449-455, J. A. Joines, R. R. Barton, K. Kang & P. A. Fishwick, eds.

Claeskens, G. & N. L. Hjort (2008). *Model selection and model averaging.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Clark, M. P. & D. Kavetski (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research* 46, W10510, doi:10.1029/2009WR008894.

Crout, N. M. J., D. Tarsitano, A. T. Wood (2009). Is my model too complex? Evaluating model formulation using model reduction. *Environmental Modelling & Software* 24, 1-7.

Droop, M. R. (1983). 25 years of algal growth kinetics. Bot. Mar. 26, 99-112.

Drury, K. L. S. & J. F. Candelaria (2008). Using model identification to analyze spatially explicit data with habitat, and temporal, variability. *Ecol. Model.* 214, 305-315.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A* 222, 309-368.

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Ann. Eugenics* 7, 255-369.

Gebremeskel, S., Y. B. Liu, F. de Smedt, L. Hoffmann, L. Pfister (2005). Assessing the hydrological effects of Landuse changes using distributed hydrological modelling and GIS. *Intl. J. River Basin Management* 3, 261-271.

Gibson, B. R., S. J. Lawrence, J. P. R. Leclaire, C. D. Powell & K. A. Smart (2007). Yeast responses to stresses associated with industrial brewery handling. *FEMS Microbiol. Rev.* 31, 535-569.

Grunwald, P. (2007). *The Minimum Description Length Principle*. MIT Press, June 2007. MIT Press. Link: *The Minimum Description Length Principle*

Guimarães, P. M. R. & J. Londesborough (2008). The adenylate energy charge and specific fermentation rate of brewer's yeasts fermenting high- and very-high-gravity worts. *Yeast* 25, 47-58.

Hipel, K. W. & A. I. McLeod (1994). Time Series Modelling of Water Resources and Environmental Systems. Freely available at
http://www.stats.uwo.ca/faculty/aim/1994Book/

Hipsey M. R. & D. P. Hamilton (2008). *Computational aquatic ecosystem dynamic model: CAEDYM v3 science manual*. Centre for Water Research Report, University of Western Australia, Nedlands.

Hjalmarsson, H. (2009). System identification of complex and structured systems. *European journal of control* 15, 275-310.

Huuskonen, A., T. Markkula, V. Vidgren, L. Lima, L. Mulder, W. Geurts, M. Walsh & J. Londesborough (2010). Selection from industrial lager yeast strains of variants with improved fermentation performance in very-high-gravity worts. *Applied and Environmental Microbiology* 76, 1563-1573.

Johnson, J. B. & K. S. Omland (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution* 19, 101-108.

Jones, H. L., A. Margaritis & R. J. Stewart (2007). The combined effects of oxygen supply strategy, inoculum size and temperature profile on very-high-gravity beer fermentation by *Saccharomyces cerevisiae*. *J. Inst. Brew.* 113(2), 168-184.

Kitano, H. (2002). Systems Biology: A brief overview. *Science* 295, 1662-1664.

Kooijman, S. A. L. M. (2000). Dynamic Energy and Mass Budgets in Biological Systems, Cambridge University Press, Cambridge, 2nd ed.

Kooijman, S. A. L. M. (2009). *Dynamic Energy Budget Theory for Metabolic Organisation*. Cambridge University Press, 3rd ed.

Kot, M. & W. M. Schaffer (1986). Discrete-Time Growth Dispersal Models. *Mathematical Biosciences* 80, 109-136.

Kurz, T., J. Mieleitner, T. Becker & A. Delgado (2002). A model based simulation of brewing yeast propagation. *J. Inst. Brew.* 108(2), 248-255.

Ljung, L. (1987). *System identification: theory for the user*. Prentice Hall PTR, New Jersey.

McDonald, C. P. & Urban, N. R. (2010). Using a model selection criterion to identify appropriate complexity in aquatic biogeochemical models. *Ecol. Model.* 221, 428-432.

Mooij, W. M., D. Trolle, E. Jeppesen, G. Arhonditsis, P. V. Belolipetsky, D. B. R. Chitamwebwa, A. G. Degermendzhy, D. L. DeAngelis, L. N. De Senerpont Domis, A. S. Downing, J. A. Elliott, C. R. Fragoso Jr., U. Gaedke, S. N. Genova, R. D. Gulati, L. Håkanson, D. P. Hamilton, M. R. Hipsey, J. 't Hoen, S. Hülsmann, F. H. Los, V. Makler-Pick, T. Petzoldt, I. G. Prokopkin, K. Rinke, S. A. Schep, K. Tominaga, A. A. Van Dam, E. H. Van Nes, S. A. Wells, J. H. Janse (2010). Challenges and opportunities for integrating lake ecosystem modelling approaches. *Aquat. Ecol*. 44, 633-667.

Monod, J. (1949). The growth of bacterial cultures. *Ann. Rev. Microbiol.* 3, 371-394.

Myung, I.J. (2000). The importance of complexity in model selection. *J. Math. Psychol*. 44, 190-204.

Navarro, D. J., M. A. Pitt & I. J. Myung (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology* 49, 47-84.

Nelder, J. A. & R. Mead (1965). A simplex-method for function minimization. *Computer Journal* 7(4), 308-313.

Pelgrim, K. (2007). *One Dimentional Lake Modeling Using DYRESM-CAEDYM*. New Stormwater Treatment Technology Information, March 2007, obtained via http://www.rwmwd.org/

Pirt, S. J. (1965). The maintenance energy of bacteria in growing cultures. *Proc. Royal Soc. London (B)* 163, 224-231.

Pitt, M. A. & I. J. Myung (2002). When a good fit can be bad. *Trends in cognitive sciences* 6(10), 421-425.

Press, W., S. Teukolsky, W. Vetterling & B. Flannery (2002). *Numerical Recipes in C++: The art of Scientific Computing*. 2nd ed., Cambridge University Press, Cambridge.

Sako, Y. (2006). Imaging single molecules in living cells for systems biology. *Nature: Molecular Systems Biology*, doi:10.1038/msb4100100.

Shachak, M. & B. R. Boeken (2010). Patterns of biotic community organization and reorganization: A conceptual framework and a case study. *Ecological complexity* 7, 433-445.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika* 38, 196-218.

Trolle, D., T. B. Jørgensen & E. Jeppesen (2008). Predicting the effects of reduced external nitrogen loading on the nitrogen dynamics and ecological state of deep Lake Ravn, Denmark, using the DYRESM–CAEDYM model. *Limnologica* 38, 220-232.

Turing, A. M. (1952). The chemical basis of morphogenesis. *Phil. Transact. Royal Soc. B* 237, 37-72.

Van der Hoek, D. C. J. & M. Bakkenes (2007). Natuurplanner 3.0 – Beschrijving en handleiding (in Dutch; English abstract). MNP rapport 500067002/2007.

Van Nes, E. H. & M. Scheffer (2005). A strategy to improve the contribution of complex simulation models to ecological theory. *Ecol. Model.* 185, 153-164.

Van Voorn, G. A. K., D. J. J. Walvoort, M. Knotters, P. W. Bogaart, H. Houweling, P. Janssen (2011). Een beoordelingslijst voor de complexiteit van modellen en bestanden (in Dutch). WOt-paper 11. WOt Wettelijke Onderzoekstaken Natuur & Milieu, Wageningen UR, Wageningen.

Van Walsum, P. E. V. & P. Groenendijk (2008). Quasi steady-state simulation of the unsaturated zone in groundwater modeling of lowland regions. *Vadose Zone Journal* 7(2), 769-781.

Verhulst, P. F. (1838). Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathématique et physique* 10, 113-121.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta & S. Sorooshian (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences* 5, 13-26.

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecol. Model.* 211, 1-10.

Annex 1 Bias and variance error

The bias of an estimator is given as

$$Bias\,[\hat{\theta}] = E[\hat{\theta}] - \theta = E[\hat{\theta}] - \theta$$

In here θ is the 'true' parameter set, $\hat{\theta}$ an approximation of this true set, and E the expectancy. When a set of samples is taken from for instance a normal distribution, the true population parameters mean μ and variance σ^2 are approximated by the average of measurements

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

respectively.

It can be shown that S^2 is a biased estimator by looking at the expectancy

$$E[S^2] = E\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} \left((X_i - \mu) - (\bar{X} - \mu)\right)^2\right] =$$
$$E\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - 2\,(\bar{X} - \mu)\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu) + (\bar{X} - \mu)^2\right] =$$
$$E\left[\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^2 - (\bar{X} - \mu)^2\right] = \sigma^2 - E[(\bar{X} - \mu)^2] < \sigma^2$$

(as the term $E[(\bar{X} - \mu)^2] > 0$). This is also why variances have to be normalised.

Annex 2 Prototype evaluation list model complexity

The original prototype of the evaluation list given in the publication by Bogaart *et al.* (2011) is in Dutch, and is divided into two sub-lists. The first sub-list contains questions ported from the existing evaluation list for Status A, which are easily answered, but perhaps superfluous or not sufficient. The second sub-list contains questions which are likely very applicable, but perhaps difficult or impossible to answer properly. The list, indicated as ECM version 0.1, is translated here in English, and expanded to explicitly include databases. Next to the evaluation of relevance and utility of questions, it is also important to merge these two sub-lists into one list.

Prototype list part 1

1. *Is there a document with meta-information of the model or database, and providing a short overview of tests, calibration, validation and sensitivity analysis?* Meta-information at least contains acronym (short name), version number, distribution date, short description of workings, goal, application area, scale (time, space), required in- and output, programming language, user interface, and platform. If applicable, provide the meta-information of the meta-model and point out the differences between model and meta-model. A short overview should be provided that indicates that the model or database has been tested, calibrated, validated, and a sensitivity analysis has been performed.

2. *What is the purpose/goal of the model/database?* Explain why the model or database has been developed. A model could be more or less complex than required for its goal. The goal should match with the intended and practical application area of the model or database.

3. *Is there a proper description of the conceptual model?* The conceptual model is the coarse model, without description of the exact framework. No mathematical equations are asked for, only the important concepts and couplings. Be mindful of aggregation, scale, and feedbacks. Graphical representations are very helpful. For the evaluation of equilibrium it should be clear what the concept of a model or database is.

4. *Is there an overview of assumptions and simplifications?* Assumptions and simplifications are being made for all measurement data, data maps, and models. These should be explicit and clear. Several experts consider the assumptions to be the core of modelling. Assumptions and simplifications will influence the application area of the model or database.

5. *Is there a description of the mathematical model or database?* The framework in which the conceptual model is translated is usually some type of mathematics (e.g., differential equations, a cellular automata) or a design language for databases (e.g., ArcGIS). This framework should be given, including a table that explains mathematical variables, parameters and constants, with a short description of each of them and the correct dimensions and units. This is an essential point in the evaluation of equilibrium, as the question will be, if things could be simpler than the given framework, or should be more complex.

6. *Is there a description of the numerical model or database core?* Practically all models are too complex to solve algebraically. Therefore, they are implemented as computer programs. The grid cells of most databases can, in theory, be filled manually. However, the amount of work and the probability for errors to occur are too large. Calculations occur in the 'core' of the model or database. This core should be properly described. Again, a table explaining variables and parameters (including computer-specific variables and parameters) helps a lot. The implementation of a model or database may be overly complex, or not sufficient. Furthermore, this question is a set-up for the verification of the model.

7. *Is there a verification, and a description of the test results?* A numerical model or database should be properly tested and verified. Verification means, that the numerical model or database is indeed a proper ('correct') implementation of the mathematical model or database design. (Note: this is not the same thing as validation, which means, that the numerical model or database should give results that agree with 'real life', i.e., the observations of the system.) The hardware should be appropriate. Many good test results give confidence in the translation from mathematical to numerical model, or design to database.

8. *Is there a description of how the program works?* Next to the calculation core, there will be program modules on data handling, output, diagnostics, etc., which have to be described properly. This point is relevant to equilibrium mostly because very often (des)aggregation methods of data and output are used, hence, it touches the most important aspect of modelling: scale.

9. *Is there a description of the required knowledge of the user?* Provide a description of the required non-trivial knowledge of an user, both scientific and computer-technical. For instance, a program that requires a lot of knowledge but does not 'do' much could be perceived as too complex. For the evaluation of equilibrium the knowledge of the user should agree with the application of the model or database.

10. *Is there an evaluation of the performance and limitations of the program?* The complete program, including the calculation core and in- and output handling, has a certain performance in terms of running time, required capacity, etc., and should be appropriately tested for both outcome as functionality. Running time is an essential feature for sensitivity analysis and the calibration of the model. Bugs and workarounds should be reported. Bugs could be resided in program parts that are non-essential.

11. *Is there a description of the origin of data and input?* Provide a short overview of data sources for parameter values, initial conditions, forcing, expert judgement, etc., and include the proper references. Possible sources are scientific papers, books, experiments, expert judgement, other models, or otherwise. The quality of data can be derived in part from its source. The uncertainty of model output may be improved by an improved quality of the data. For the evaluation of equilibrium it is furthermore necessary to determine if data is perhaps superfluous, or if essential data or input is missing.

12. *For what is the data and input used?* Describe the purpose for the described data and input. Which data was used for calibration, model design, validation, etc., and which input is required for which modules. Data used for validation should fit the application area of the model or database.

13. *Is there a description of the manipulations on data?* Data may have to be manipulated to be suitable for different applications. For instance, raw data can be

transformed into percentages, logarithmic values, distributions with derived parameters like mean and variance, or otherwise. Possibly errors and uncertainties have been introduced through these manipulations. Some manipulations may not be relevant for the application of the model or database.

14. *Is there an overview of tests on quality of data?* Describe how data has been tested on reliability, uncertainties, information content for parameter and model identification, and other things. If the data are output from other models or databases, refer to their sources. Data can be unsuitable for parameter identification, unreliable, or contain large errors or be biased. Some applications require a larger level of reliability than others.

15. *Is it clear which parts of the model or database still require data?* Give an overview of the parts for which data is still missing. Often one or more parameters cannot be given values because of lack of proper data. The uncertainties caused by these 'missing parts' can be large or small. For instance, an uncertainty analysis might have revealed that the uncertainty is small, and that these data are not required to meet the standards for a given application.

16. *Is there a description of the output?* Clearly describe the output produced by the program. Again, a table with variables, parameters, dimensions, units, and short descriptions is helpful. For the application it should be clear what output is produced, and how it is used. Even a verified model or database can still produce output that is not relevant for the application.

17. *Is the uncertainty about the output clear?* Describe which tests have been performed to determine the confidence about the output. It should be clear to an user if the output is usable for other models or databases, and what the level of uncertainty is in view of applications?

18. *Is there a description of the application area(s) of the model or database?* Describe for which application(s) the model or database is used or may be used. Also, explicitly describe for which related application areas it cannot be used. The application area of the model or database should agree with the goal it was designed for. Many models and databases provide input for other models and databases in a model train. If a model or database is used for more than one application, this is very relevant for the concept of equilibrium. The model or database may not be equally suitable for each of the applications. Quite possibly the model or database equilibrium varies per application.

19. *Is there an description of the calibration of the model?* This question may not be relevant for databases. Describe for which applications the model has been calibrated, e.g., a hydrological model may be calibrated for an application concerning one river while being used for another. The calibration should be clearly described, for instance, many hydrological models are calibrated by using PEST. The calibration method for one model or even one application does not have to be suitable for another model or application. An easily calibrated model may be generally valid and well-built, while a model that has to be calibrated again for each new application may be mechanistically ill-posed, although this is not a general truth. This point obviously is relevant for equilibrium.

20. *Is there a description of the validation of the model or database?* Give an overview of the applications for which the model of database has been validated, and describe how. The data used for validation should preferably not be the data used for calibration. A positively verified program is not necessarily a good and 'correct'

projection of the 'real world'. Also, the program output is not necessarily relevant or usable for the application(s). It may be possible that output has not been validated, but that this does not matter for the application. This result may indicate a superfluous component in terms of equilibrium.

21. *Is there an overview of the performed sensitivity analyses?* There should be an overview and description of the sensitivity analyses that have been performed on (parts of) the model or database. In a sensitivity analysis the parameter values, threshold values, boundary condition values, initial conditions, etc. are varied, and the sensitivity of the model or database for changes in these values is determined. Sensitivity analyses are useful to evaluate the relative role of these factors, and have previously been used to analyse model complexity. These results are important to consider the importance of knowing certain parameters values, etc., and for the identification possibilities of parameters and such. Parameters, etc., may for instance be badly identifiable for an application, while they are not very important. The effort to identify them properly does not have to be invested in that case.

22. *Is it clear which parameters and variables are (not/badly) identifiable?* Describe which parameters and variables in the model or database can be identified, and which not. The possibilities for identification might depend on the data, the model structure, the application, or something else. This is relevant for equilibrium. Unimportant factors that are unidentifiable are probably superfluous. On the other hand, significant effort may need to be invested to properly identify important factors.

23. *Is there an overview of performed uncertainty analyses?* Describe the uncertainty analyses that have been performed for (parts of) the model or database. An uncertainty analysis determines which factors (parameters, threshold values, etc.) generate most uncertainty about the output. Large quantities of uncertainty may be caused by unimportant factors, or by important ones. In the latter case the model or database may lose its value for the application. The reduction of uncertainty about model and database predictions is the ultimate goal behind the concept of equilibrium.

Prototype list part 2

24. **System analysis.** Which attributes of the reality and which processes have direct relevance for the eventual application(s)? Which feedbacks are present? In what way are these feedbacks relevant for the application? How have the boundaries of the system been defined? How are relevant attributes and processes outside the system (boundary conditions, constants, etc.) handled? What data are available for this?

25. **Scale.** What are the extent, the support and the coverage of the used data sets? And of the application(s)? And of the model/database? Which transformations are being used? Which heterogeneities exist? Is there a scaling problem, and if yes, where is it?

26. **Conceptual model.** At which aggregation level is the output desired? What are the dominant fluxes in the system? What are the spatial interactions? What data are available? Does the data have the proper dimensions (mind the units)? If not, is there a method for disaggregation available? What conceptualization of reality is used, given the application(s)? How useful is this conceptualization? What are the

implications for the level of scale, in time and space? What are the implications for the data supply?

27. **Formal model.** Which types of equations are available? Or for a database, what possible designs exist? Is there an a priori preference, given the application(s)? What need for which data is there? Can we expect a problem with the identification of certain factors? Which level of uncertainty is acceptable? What is the level of heterogeneity of the domain?

28. **Numerical model.** Has the formal model or database design been properly translated into a numerical model or database? Was the code verified? Can the required and available input and data be used in the operational process? What are the run times of the model or database? Are they within the available operational capacity? Is there capacity for an adequate calibration, validation, sensitivity and/or uncertainty analysis?

29. **Schematization.** Which schematization is desired considering the application? Can this schematization be achieved, given the run time of the model or database? Is it possible and/or desirable to combine (spatial) units into 'plots'? Is there sufficient data available to perform this schematization? Is the data of the right scale? If not, is there a scaling technique available? Can it be applied?

30. **Sensitivity/uncertainty analysis (SA/UA).** Has a SA/UA been performed? If yes, where is the description of the SA/UA and its results?

31. **Calibration.** *This part probably does not apply to databases.* How have the values of the parameters of the model been determined? Directly, manually, a numerical algorithm, expert knowledge? What were the used sources of information? What are the uncertainties in these sources? Is there a scaling problem? If yes, how has it been solved? Is there incommensurability between source and model parameters (i.e., do the source and the model assume the same definitions for the same terminology)? Which measures or measures of agreement are being used during the calibration? Which ones would be desirable given the application(s)? Which parts of the dynamical extent of the model are relevant for the application? Are all model parameters independent? Is the information content of the data sufficient to identify the different parameters? Do there exist options for multi-criteria calibration, and are these desired? Are there (desired) options for 'soft' data? For each variable: In what way is the agreement between calibration data and model output determined? Which outlier tests or residual plots have been made? Is there information about the reliability of calibrated parameter values, and how is this information obtained?

32. **Validation and cross-validation.** Has the whole model or database or parts of it been validated? Which parts have not been validated? Has the model or database been validated in the context of the intended application(s)? Which assumptions have been made for the validation? Has 'goodness-of-fit' been used in the case of cross-validation? For which parameters? Are there any meta-parameters? If yes, has a double cross-validation been used?

Verschenen documenten in de reeks Werkdocumenten van de Wettelijke Onderzoekstaken Natuur & Milieu vanaf 2009

Werkdocumenten zijn verkrijgbaar bij het secretariaat van Unit Wettelijke Onderzoekstaken Natuur & Milieu, te Wageningen. T 0317 – 48 54 71; F 0317 – 41 90 00; E info.wnm@wur.nl
De werkdocumenten zijn ook te downloaden via de WOt-website www.wotnatuurenmilieu.wur.nl

2009

126 *Kamphorst, D.A.* Keuzes in het internationale biodiversiteitsbeleid; Verkenning van de beleidstheorie achter de internationale aspecten van het Beleidsprogramma Biodiversiteit (2008-2011)

127 *Dirkx, G.H.P. & F.J.P. van den Bosch.* Quick scan gebruik Catalogus groenblauwe diensten

128 *Loeb, R. & P.F.M. Verdonschot.* Complexiteit van nutriëntenlimitaties in oppervlaktewateren

129 *Kruit, J. & P.M. Veer.* Herfotografie van landschappen; Landschapsfoto's van de 'Collectie de Boer' als uitgangspunt voor het in beeld brengen van ontwikkelingen in het landschap in de periode 1976-2008

130 *Oenema, O., A. Smit & J.W.H. van der Kolk.* Indicatoren Landelijk Gebied; werkwijze en eerste resultaten

131 *Agricola, H.J.A.J. van Strien, J.A. Boone, M.A. Dolman, C.M. Goossen, S. de Vries, N.Y. van der Wulp, L.M.G. Groenemeijer, W.F. Lukey & R.J. van Til.* Achtergrond-document Nulmeting Effectindicatoren Monitor Agenda Vitaal Platteland

132 *Jaarrapportage 2008.* WOT-04-001 – Koepel

133 *Jaarrapportage 2008.* WOT-04-002 – Onderbouwend Onderzoek

134 *Jaarrapportage 2008.* WOT-04-003 – Advisering Natuur & Milieu

135 *Jaarrapportage 2008.* WOT-04-005 – M-AVP

136 *Jaarrapportage 2008.* WOT-04-006 – Natuurplanbureaufunctie

137 *Jaarrapportage 2008.* WOT-04-007 – Milieuplanbureaufunctie

138 *Jong de, J.J., J. van Os & R.A. Smidt.* Inventarisatie en beheerskosten van landschapselementen

139 *Dirkx, G.H.P., R.W. Verburg & P. van der Wielen.* Tegenkrachten Natuur. Korte verkenning van de weerstand tegen aankopen van landbouwgrond voor natuur

140 *Annual reports for 2008; Programme WOT-04*

141 *Vullings, L.A.E., C. Blok, G. Vonk, M. van Heusden, A. Huisman, J.M. van Linge, S. Keijzer, J. Oldengarm & J.D. Bulens.* Omgaan met digitale nationale beleidskaarten

142 *Vreke, J.,A.L. Gerritsen, R.P. Kranendonk, M. Pleijte, P.H. Kersten & F.J.P. van den Bosch.* Maatlat Government – Governance

143 *Gerritsen, A.L., R.P. Kranendonk, J. Vreke, F.J.P. van den Bosch & M. Pleijte.* Verdrogingsbestrijding in het tijdperk van het Investeringsbudget Landelijk Gebied. Een verslag van casusonderzoek in de provincies Drenthe, Noord-Brabant en Noord-Holland

144 *Luesink, H.H., P.W. Blokland, M.W. Hoogeveen & J.H. Wisman.* Ammoniakemissie uit de landbouw in 2006 en 2007

145 *Bakker de, H.C.M. & C.S.A. van Koppen.* Draagvlakonderzoek in de steigers. Een voorstudie naar indicatoren om maatschappelijk draagvlak voor natuur en landschap te meten

146 *Goossen, C.M.,* Monitoring recreatiegedrag van Nederlanders in landelijke gebieden. Jaar 2006/2007

147 *Hoefs, R.M.A., J. van Os & T.J.A. Gies.* Kavelruil en Landschap. Een korte verkenning naar ruimtelijke effecten van kavelruil

148 *Klok, T.L., R. Hille Ris Lambers, P. de Vries, J.E. Tamis & J.W.M. Wijsman.* Quick scan model instruments for marine biodiversity policy

149 *Spruijt, J., P. Spoorenberg & R. Schreuder.* Milieueffectiviteit en kosten van maatregelen gewasbescherming

150 *Ehlert, P.A.I. (rapporteur).* Advies Bemonstering bodem voor differentiatie van fosfaatgebruiksnormen

151 *Wulp van der, N.Y.* Storende elementen in het landschap: welke, waar en voor wie? Bijlage bij WOt-paper 1 – Krassen op het landschap

152 *Oltmer, K., K.H.M. van Bommel, J. Clement, J.J. de Jong, D.P. Rudrum & E.P.A.G. Schouwenberg.* Kosten voor habitattypen in Natura 2000-gebieden. Toepassing van de methode Kosteneffectiviteit natuurbeleid

153 *Adrichem van, M.H.C., F.G. Wortelboer & G.W.W. Wamelink **(2010).*** MOVE. Model for terrestrial Vegetation. Version 4.0

154 *Wamelink, G.W.W., R.M. Winkler & F.G. Wortelboer.* User documentation MOVE4 v 1.0

155 *Gies de, T.J.A., L.J.J. Jeurissen, I. Staritsky & A. Bleeker.* Leefomgevingsindicatoren Landelijk gebied. Inventarisatie naar stand van zaken over geurhinder, lichthinder en fijn stof

156 *Tamminga, S., A.W. Jongbloed, P. Bikker, L. Sebek, C. van Bruggen & O. Oenema.* Actualisatie excretiecijfers landbouwhuisdieren voor forfaits regeling Meststoffenwet

157 *Van der Salm, C., L. .M. Boumans, G.B.M. Heuvelink & T.C. van Leeuwen.* Protocol voor validatie van het nutriëntenemissiemodel STONE op meetgegevens uit het Landelijk Meetnet effecten Mestbeleid

158 *Bouwma, I.M.* Quickscan Natura 2000 en Programma Beheer. Een vergelijking van Programma Beheer met de soorten en habitats van Natura 2000

159 *Gerritsen, A.L., D.A. Kamphorst, T.A. Selnes, M. van Veen, F.J.P.van den Bosch, L. van den Broek, M.E.A. Broekmeyer, J.L.M. Donders, R.J. Fontein, S. van Tol, G.W.W. Wamelink & P. van der Wielen.* Dilemma's en barrières in de praktijk van het natuur- en landschapsbeleid; Achtergronddocument bij Natuurbalans 2009

160 *Fontein R.J, T.A. de Boer, B. Breman, C.M. Goossen, R.J.H.G. Henkens, J. Luttik & S. de Vries.* Relatie recreatie en natuur; Achtergronddocument bij Natuurbalans 2009

161 *Deneer, J.W. & R. Kruijne. **(2010).*** Atmosferische depositie van gewasbeschermingsmiddelen. Een verkenning van de literatuur verschenen na 2003

162 *Verburg, R.W., M.E. Sanders, G.H.P. Dirkx, B. de Knegt & J.W. Kuhlman.* Natuur, landschap en landelijk gebied. Achtergronddocument bij Natuurbalans 2009

163 *Doorn van, A.M. & M.P.C.P. Paulissen.* Natuurgericht milieubeleid voor Natura 2000-gebieden in Europees perspectief: een verkenning

164 *Smidt, R.A., J. van Os & I. Staritsky.* Samenstellen van landelijke kaarten met landschapselementen, grondeigendom en beheer. Technisch achtergronddocument bij de opgeleverde bestanden

165 *Pouwels, R., R.P.B. Foppen, M.F. Wallis de Vries, R. Jochem, M.J.S.M. Reijnen & A. van Kleunen,* Verkenning LARCH: omgaan met kwaliteit binnen ecologische netwerken

166 *Born van den, G.J., H.H. Luesink, H.A.C. Verkerk, H.J. Mulder, J.N. Bosma, M.J.C. de Bode & O. Oenema,* Protocol voor monitoring landelijke mestmarkt onder een stelsel van gebruiksnormen, versie 2009

167 *Dijk, T.A. van, J.J.M. Driessen, P.A.I. Ehlert, P.H. Hotsma, M.H.M.M. Montforts, S.F. Plessius & O. Oenema.* Protocol beoordeling stoffen Meststoffenwet- Versie 2.1

168 *Smits, M.J., M.J. Bogaardt, D. Eaton, A. Karbauskas & P. Roza.* De vermaatschappelijking van het Gemeenschappelijk Landbouwbeleid. Een inventarisatie van visies in Brussel en diverse EU-lidstaten

169 *Vreke, J. & I.E. Salverda.* Kwaliteit leefomgeving en stedelijk groen

170 *Hengsdijk, H. & J.W.A. Langeveld.* Yield trends and yield gap analysis of major crops in the World

171 *Horst, M.M.S. ter & J.G. Groenwold.* Tool to determine the coefficient of variation of DegT50 values of plant protection products in water-sediment systems for different values of the sorption coefficient

172 *Boons-Prins, E., P. Leffelaar, L. Bouman & E. Stehfest (**2010**)* Grassland simulation with the LPJmL model

173 *Smit, A., O. Oenema & J.W.H. van der Kolk.* Indicatoren Kwaliteit Landelijk Gebied

2010

174 *Boer de, S., M.J. Bogaardt, P.H. Kersten, F.H. Kistenkas, M.G.G. Neven & M. van der Zouwen.* Zoektocht naar nationale beleidsruimte in de EU-richtlijnen voor het milieu- en natuurbeleid. Een vergelijking van de implementatie van de Vogel- en Habitatrichtlijn, de Kaderrichtlijn Water en de Nitraatrichtlijn in Nederland, Engeland en Noordrijn-Westfalen

175 *Jaarrapportage 2009.* WOT-04-001 – Koepel

176 *Jaarrapportage 2009.* WOT-04-002 – Onderbouwend Onderzoek

177 *Jaarrapportage 2009.* WOT-04-003 – Advisering Natuur & Milieu

178 *Jaarrapportage 2009.* WOT-04-005 – M-AVP

179 *Jaarrapportage 2009.* WOT-04-006 – Natuurplanbureaufunctie

180 *Jaarrapportage 2009.* WOT-04-007 – Milieuplanbureaufunctie

181 *Annual reports for 2009;* Programme WOT-04

182 *Oenema, O., P. Bikker, J. van Harn, E.A.A. Smolders, L.B. Sebek, M. van den Berg, E. Stehfest & H. Westhoek.* Quickscan opbrengsten en efficiëntie in de gangbare en biologische akkerbouw, melkveehouderij, varkenshouderij en pluimveehouderij. Deelstudie van project 'Duurzame Eiwitvoorziening'

183 *Smits, M.J.W., N.B.P. Polman & J. Westerink.* Uitbreidingsmogelijkheden voor groene en blauwe diensten in Nederland; Ervaringen uit het buitenland

184 *Dirkx, G.H.P. (red.).* Quick responsefunctie 2009. Verslag van de werkzaamheden

185 *Kuhlman, J.W., J. Luijt, J. van Dijk, A.D. Schouten & M.J. Voskuilen.* Grondprijskaarten 1998-2008

186 *Slangen, L.H.G., R.A. Jongeneel, N.B.P. Polman, E. Lianouridis, H. Leneman & M.P.W. Sonneveld.* Rol en betekenis van commissies voor gebiedsgericht beleid

187 *Temme, A.J.A.M. & P.H. Verburg.* Modelling of intensive and extensive farming in CLUE

188 *Vreke, J.* Financieringsconstructies voor landschap

189 *Slangen, L.H.G.* Economische concepten voor beleidsanalyse van milieu, natuur en landschap

190 *Knotters, M., G.B.M. Heuvelink, T. Hoogland & D.J.J. Walvoort.* A disposition of interpolation techniques

191 *Hoogeveen, M.W., P.W. Blokland, H. van Kernebeek, H.H. Luesink & J.H. Wisman.* Ammoniakemissie uit de landbouw in 1990 en 2005-2008

192 *Beekman, V., A. Pronk & A. de Smet.* De consumptie van dierlijke producten. Ontwikkeling, determinanten, actoren en interventies.

193 *Polman, N.B.P., L.H.G. Slangen, A.T. de Blaeij, J. Vader & J. van Dijk.* Baten van de EHS; De locatie van recreatiebedrijven

194 *Veeneklaas, F.R. & J. Vader.* Demografie in de Natuurverkenning 2011; Bijlage bij WOt-paper 3

195 *Wascher, D.M., M. van Eupen, C.A. Mücher & I.R. Geijzendorffer,* Biodiversity of European Agricultural landscapes. Enhancing a High Nature Value Farmland Indicator

196 *Apeldoorn van, R.C., I.M. Bouwma, A.M. van Doorn, H.S.D. Naeff, R.M.A. Hoefs, B.S. Elbersen & B.J.R. van Rooij.* Natuurgebieden in Europa: bescherming en financiering

197 *Brus, D.J.,, R. Vasat, G. B. M. Heuvelink, M. Knotters, F. de Vries & D. J. J. Walvoort.* Towards a Soil Information System with quantified accuracy; A prototype for mapping continuous soil properties

198 *Groot, A.M.E.& A.L. Gerritsen, m.m.v. M.H. Borgstein, E.J. Bos & P. van der Wielen.* Verantwoording van de methodiek Achtergronddocument bij 'Kwalitatieve monitor Systeeminnovaties verduurzaming landbouw'

199 *Bos, E.J. & M.H. Borgstein.* Monitoring Gesloten voer-mest kringlopen. Achtergronddocument bij 'Kwalitatieve monitor Systeeminnovaties verduurzaming landbouw'

200 *Kennismarkt 27 april 2010;* Van onderbouwend onderzoek Wageningen UR naar producten Planbureau voor de Leefomgeving

201 *Wielen van der, P.* Monitoring Integrale duurzame stallen. Achtergronddocument bij 'Kwalitatieve monitor Systeeminnovaties verduurzaming landbouw'

202 *Groot, A.M.E.& A.L. Gerritsen.* Monitoring Functionele agrobiodiversiteit. Achtergronddocument bij 'Kwalitatieve monitor Systeeminnovaties verduurzaming landbouw'

203 *Jongeneel, R.A. & L. Ge.* Farmers' behavior and the provision of public goods: Towards an analytical framework

204 *Vries, S. de, M.H.G. Custers & J. Boers.* Storende elementen in beeld; de impact van menselijke artefacten op de landschapsbeleving nader onderzocht

205 *Vader, J. J.L.M. Donders & H.W.B. Bredenoord*. Zicht op natuur- en landschapsorganisaties; Achtergronddocument bij Natuurverkenning 2011

206 *Jongeneel, R.A., L.H.G. Slangen & N.B.P. Polman.* Groene en blauwe diensten; Een raamwerk voor de analyse van doelen, maatregelen en instrumenten

207 *Letourneau, A.P, P.H. Verburg & E. Stehfest.* Global change of land use systems; IMAGE: a new land allocation module

208 *Heer, M. de.* Het Park van de Toekomst. Achtergronddocument bij Natuurverkenning 2011

209 *Knotters, M., J. Lahr, A.M. van Oosten-Siedlecka & P.F.M. Verdonschot.* Aggregation of ecological indicators for mapping aquatic nature quality. Overview of existing methods and case studies

210 *Verdonschot, P.F.M. & A.M. van Oosten-Siedlecka.* Graadmeters Aquatische natuur. Analyse gegevenskwaliteit Limnodata

211 *Linderhof, V.G.M. & H. Leneman.* Quickscan kosteneffectiviteitsanalyse aquatische natuur

212 *Leneman, H., V.G.M. Linderhof & R. Michels.* Mogelijkheden voor het inbrengen van informatie uit de 'KRW database' in de 'KE database'

213 *Schrijver, R.A.M., A. Corporaal, W.A. Ozinga & D. Rudrum.* Kosteneffectieve natuur in landbouwgebieden; Methode om effecten van maatregelen voor de verhoging van biodiversiteit in landbouwgebieden te bepalen, een test in twee gebieden in Noordoost-Twente en West-Zeeuws-Vlaanderen

214 *Hoogland, T., R.H. Kemmers, D.G. Cirkel & J. Hunink.* Standplaatsfactoren afgeleid van hydrologische model uitkomsten; Methode-ontwikkeling en toetsing in het Drentse Aa-gebied

215 *Agricola, H.J., R.M.A. Hoefs, A.M. van Doorn, R.A. Smidt & J. van Os.* Landschappelijke effecten van ontwikkelingen in de landbouw

216 *Kramer, H., J. Oldengarm & L.F.S. Roupioz.* Nederland is groener dan kaarten laten zien; Mogelijkheden om 'groen' beter te inventariseren en monitoren met de automatische classificatie van digitale luchtfoto's

217 *Raffe, J.K. van, J.J. de Jong & G.W.W. Wamelink **(2011).*** Scenario's voor de kosten van natuurbeheer en stikstofdepositie; Kostenmodule v 1.0 voor de Natuurplanner

218 *Hazeu, G.W., Kramer, H., J. Clement & W.P. Daamen (2011).* Basiskaart Natuur 1990rev

219 *Boer, T.A. de.* Waardering en recreatief gebruik van Nationale Landschappen door haar bewoners

220 *Leneman, H., A.D. Schouten & R.W. Verburg.* Varianten van natuurbeleid: voorbereidende kostenberekeningen; Achtergronddocument bij Natuurverkenning 2011

221 *Knegt, B. de, J. Clement, P.W. Goedhart, H. Sierdsema, Chr. van Swaay & P. Wiersma.* Natuurkwaliteit van het agrarisch gebied

2011

222 *Kamphorst, D.A. & M.M.P. van Oorschot*. Kansen en barrières voor verduurzaming van houtketens

223 *Salm, C. van der & O.F. Schoumans.* Langetermijneffecten van verminderde fosfaatgiften

224 *Bikker, P., M.M. van Krimpen & G.J. Remmelink.* Stikstof-verteerbaarheid in voeders voor landbouwhuisdieren; Berekeningen voor de TAN-excretie

225 *M.E. Sanders & A.L. Gerritsen (red.).* Het biodiversiteitsbeleid in Nederland werkt. Achtergronddocument bij Balans van de Leefomgeving 2010

226 *Bogaart, P.W., G.A.K. van Voorn & L.M.W. Akkermans.* Evenwichtsanalyse modelcomplexiteit; een verkennende studie

227 *Kleunen A. van, K. Koffijberg, P. de Boer, J. Nienhuis, C.J. Camphuysen, H. Schekkerman, K.H. Oosterbeek, M.L. de Jong, B. Ens & C.J. Smit **(2010).*** Broedsucces van kustbroedvogels in de Waddenzee in 2007 en 2008

228 *Salm, C. van der, L.J.M. Boumans, D.J. Brus, B. Kempen & T.C van Leeuwen.* Validatie van het nutriëntenemissiemodel STONE met meetgegevens uit het Landelijk Meetnet effecten Mestbeleid (LMM) en de Landelijke Steekproef Kaarteenheden (LSK).

229 *Dijkema, K.S., W.E. van Duin, E.M. Dijkman, A. Nicolai, H. Jongerius, H. Keegstra, L. van Egmond, H.J. Venema & J.J. Jongsma*. Vijftig jaar monitoring en beheer van de Friese en Groninger kwelderwerken: 1960-2009

230 *Jaarrapportage 2010.* WOT-04-001 – Koepel

231 *Jaarrapportage 2010.* WOT-04-002 – Onderbouwend Onderzoek

232 *Jaarrapportage 2010.* WOT-04-003 – Advisering Natuur & Milieu

233 *Jaarrapportage 2010.* WOT-04-005 – M-AVP

234 *Jaarrapportage 2010.* WOT-04-006 – Natuurplanbureaufunctie

235 *Jaarrapportage 2010.* WOT-04-007 – Milieuplanbureaufunctie

236 *Arnouts, R.C.M. & F.H. Kistenkas*. Nederland op slot door Natura 2000: de discussie ontrafeld; Bijlage bij WOt-paper 7 – De deur klemt

237 *Harms, B. & M.M.M. Overbeek*. Bedrijven aan de slag met natuur en landschap; relaties tussen bedrijven en natuurorganisaties. Achtergronddocument bij Natuurverkenning 2011

238 *Agricola, H.J. & L.A.E. Vullings.* De stand van het platteland 2010. Monitor Agenda Vitaal Platteland; Rapportage Midterm meting Effectindicatoren

239 *Klijn, J.A.* Wisselend getij. Omgang met en beleid voor natuur en landschap in verleden en heden; een essayistische beschouwing. Achtergronddocument bij Natuurverkenning 2011

240 *Corporaal, A., T. Denters, H.F. van Dobben, S.M. Hennekens, A. Klimkowska, W.A. Ozinga, J.H.J. Schaminée & R.A.M. Schrijver*. Stenoeciteit van de Nederlandse flora. Een nieuwe parameter op grond van ecologische amplitudo's van de Nederlandse plantensoorten en toepassingsmogelijkheden

241 *Wamelink, G.W.W., R. Jochem, J. van der Greft, C. Grashof-Bokdam, R.M.A. Wegman, G.J. Franke & A.H. Prins.* Het plantendispersiemodel DIMO. Ter verbetering van de modellering in de Natuurplanner (werktitel)

242 *Klimkowska, A., M.H.C. van Adrichem, J.A.M. Jansen & G.W.W. Wamelink.* Bruikbaarheid van WNK-monitoringgegevens voor EC-rapportage voor Natura 2000-gebieden. Eerste fase

243 *Goossen, C.M., R.J. Fontein, J.L.M. Donders & R.C.M. Arnouts.* Mass Movement naar recreatieve gebieden; Overzicht van methoden om bezoekersaantallen te meten

244 *Spruijt, J., P.M. Spoorenberg, J.A.J.M. Rovers, J.J. Slabbekoorn, S.A.M. de Kool, M.E.T. Vlaswinkel, B. Heijne, J.A. Hiemstra, F. Nouwens & B.J. van der Sluis.* Milieueffecten van maatregelen gewasbescherming

245 *Walker, A.N. & G.B. Woltjer.* Forestry in the Magnet model.

246 *Hoefnagel, E.W.J., F.C. Buisman, J.A.E. van Oostenbrugge & B.I. de Vos.* Een duurzame toekomst voor de Nederlandse visserij. Toekomstscenario's 2040

247 *Buurma, J.S. & S.R.M. Janssens.* Het koor van adviseurs verdient een dirigent. Over kennisverspreiding rond phytophthora in aardappelen

248 *Verburg, R.W., A.L. Gerritsen & W. Nieuwenhuizen.* Natuur meekoppelen in ruimtelijke ontwikkeling: een analyse van sturingsstrategieën voor de Natuurverkenning. Achtergronddocument bij Natuurverkenning 2011

249 *Kooten, T. van & C. Klok.* The Mackinson-Daskalov North Sea EcoSpace model as a simulation tool for spatial planning scenarios

250 *Bruggen van, C., C.M. Groenestein, B.J. de Haan, M.W. Hoogeveen, J.F.M. Huijsmans, S.M. van der Sluis & G.L. Velthof.* Ammoniakemissie uit dierlijke mest en kunstmest 1990-2008. Berekeningen met het Nationaal Emissiemodel voor Ammoniak (NEMA)

251 *Bruggen van, C., C.M. Groenestein, B.J. de Haan, M.W. Hoogeveen, J.F.M. Huijsmans, S.M. van der Sluis & G.L. Velthof.* Ammoniakemmissie uit dierlijke mest en kunstmest in 2009. Berekeningen met het Nationaal Emissiemodel voor Ammoniak (NEMA)

252 *Randen van, Y., H.L.E. de Groot & L.A.E. Vullings.* Monitor Agenda Vitaal Platteland vastgelegd. Ontwerp en implementatie van een generieke beleidsmonitor

253 *Agricola, H.J., R. Reijnen, J.A. Boone, M.A. Dolman, C.M. Goossen, S. de Vries, J. Roos-Klein Lankhorst, L.M.G. Groenemeijer & S.L. Deijl.* Achtergronddocument Midterm meting Effectindicatoren Monitor Agenda Vitaal Platteland

254 *Buiteveld, J. S.J. Hiemstra & B. ten Brink.* Modelling global agrobiodiversity. A fuzzy cognitive mapping approach

255 *Hal van R., O.G. Bos & R.G. Jak.* Noordzee: systeemdynamiek, klimaatverandering, natuurtypen en benthos. Achtergronddocument bij Natuurverkenning 2011

256 *Teal, L.R..* The North Sea fish community: past, present and future. Background document for the 2011 National Nature Outlook

257 *Leopold, M.F., R.S.A. van Bemmelen & S.C.V. Geelhoed.* Zeevogels op de Noordzee. Achtergronddocument bij Natuurverkenning 2011

258 *Geelhoed, S.C.V. & T. van Polanen Petel.* Zeezoogdieren op de Noordzee. Achtergronddocument bij Natuurverkenning 2011

259 *Kuijs, E.K.M. & J. Steenbergen.* Zoet-zoutovergangen in Nederland; stand van zaken en kansen voor de toekomst. Achtergronddocument bij Natuurverkenning 2011

260 *Baptist, M.J.* Zachte kustverdediging in Nederland; scenario's voor 2040. Achtergronddocument bij Natuurverkenning 2011

261 *Wiersinga, W.A., R. van Hal, R.G. Jak & F.J. Quirijns.* Duurzame kottervisserij op de Noordzee. Achtergronddocument bij Natuurverkenning 2011

262 *Wal J.T. van der & W.A. Wiersinga.* Ruimtegebruik op de Noordzee en de trends tot 2040. Achtergronddocument bij Natuurverkenning 2011

263 *Wiersinga, W.A. J.T. van der Wal, R.G. Jak & M.J. Baptist*. Vier kijkrichtingen voor de mariene natuur in 2040. Achtergronddocument bij Natuurverkenning 2011

264 *Bolman, B.C. & D.G. Goldsborough.* Marine Governance. Achtergronddocument bij Natuurverkenning 2011

265 *Bannink, A.* Methane emissions from enteric fermentation in dairy cows, 1990-2008; Background document on the calculation method and uncertainty analysis for the Dutch National Inventory Report on Greenhouse Gas Emissions

266 *Wyngaert, I.J.J. van den, P.J. Kuikman, J.P. Lesschen, C.C. Verwer & H.H.J. Vreuls.* LULUCF values under the Kyoto Protocol; Background document in preparation of the National Inventory Report 2011 (reporting year 2009)

267 *Helming, J.F.M. & I.J. Terluin.* Scenarios for a cap beyond 2013; implications for EU27 agriculture and the cap budget.

268 *Woltjer, G.B.* Meat consumption, production and land use. Model implementation and scenarios.

269 *Knegt, B. de, M. van Eupen, A. van Hinsberg, R. Pouwels, M.S.J.M. Reijnen, S. de Vries, W.G.M. van der Bilt & S. van Tol.* Ecologische en recreatieve beoordeling van toekomstscenario's van natuur op het land. Achtergrond-document bij Natuurverkenning 2011.

270 *Bos, J.F.F.P., M.J.W. Smits, R.A.M Schrijver & R.W. van der Meer*. Gebiedsstudies naar effecten van vergroening van het Gemeenschappelijk Landbouwbeleid op bedrijfseconomie en inpassing van agrarisch natuurbeheer.

271 *Donders, J., J. Luttik, M. Goossen, F. Veeneklaas, J. Vreke & T. Weijschede.* Waar gaat dat heen? Recreatiemotieven, landschapskwaliteit en de oudere wandelaar. Achtergronddocument bij Natuurverkenning 2011.

272 *Voorn G.A.K. van & D.J.J. Walvoort.* Evaluation of an evaluation list for model complexity.

273 *Heide, C.M. van der & F.J. Sijtsma.* Maatschappelijke waardering van ecosysteemdiensten; een handreiking voor publieke besluitvorming. Achtergronddocument bij Natuurverkenning 2011

275 *Os, J. van; T.J.A. Gies; H.S.D. Naeff; L.J.J Jeurissen*. Emissieregistratie van landbouwbedrijven; verbeteringen met behulp van het Geografisch Informatiesysteem Agrarische Bedrijven.

276 *Walsum, P.E.V. van & A.A. Veldhuizen*. MetaSWAP_V7_2_0; Rapportage van activiteiten ten behoeve van certificering met Status A.