4

Calibration in a Bayesian modelling framework

Michiel J.W. Jansen[#] and Thomas J. Hagenaars[#]

Abstract

Bayesian statistics may constitute the core of a consistent and comprehensive framework for the statistical aspects of modelling complex processes that involve many parameters whose values are derived from many sources. Bayesian statistics holds great promises for model calibration, provides the perfect starting point for uncertainty analysis and provides an excellent starting point for decision support. The purpose of this paper is to draw attention to problems and possible solutions. It is not our intention to introduce ready-for-use methods.
Keywords: Bayesian analysis; Monte Carlo; complex models; model calibration; uncertainty analysis; sensitivity analysis; decision support

Introduction

This paper discusses three related statistical aspects of complex modelling, namely Bayesian calibration, uncertainty analysis, and decision-making under uncertainty. Figure 1 sketches a framework that connects these aspects. French (2001) gives a similar framework.

The dark-grey boxes in the upper left-hand corner of Figure 1 represent the data analysis, given the model and the prior distribution of the parameters. Bayesian methods hold great promises for model calibration. Currently, the calibration of complex models is an art rather than science. No problems are encountered in the calibration of a simple model, such as a regression model with a small number of parameters, estimated from one single and simple data set. If the model is complex, however, like most parameter-rich ecological or economic models, and if the modeller draws information from many diverse data sets and other sources, calibration often becomes obscure. Although modellers find ways to produce values for parameters, the methods used are often intuitive and non-reproducible, while the accuracy of the estimates is unspecified. Bayesian methods may bring conceptual clarity in the calibration of complex models, especially because they enable combination of heterogeneous information about parameter values, like several types of data sets. In addition, expert judgement may be used as prior information.

Nevertheless, the use of Bayesian methods for the calibration of complex models is far from straightforward in practice. The most notorious problem is the obligation to use prior information even if there is no such information, since priors flawlessly expressing total ignorance do not exist. Apart from that, the analyst may encounter quite a few computational difficulties, of which the computer time required to run a

[#] Biometris, Wageningen University and Research Centre, P.O. Box 100, 6700AC Wageningen, The Netherlands. E-mail: michiel.jansen@wur.nl, thomas.hagenaars@wur.nl

complex model thousands of times may be the most serious one. There is also a shortage of suitable software, which necessitates the detailed implementation of many details of – fundamentally quite transparent – numerical recipes. After the above difficulties have been surmounted, the problem of convergence monitoring still remains.

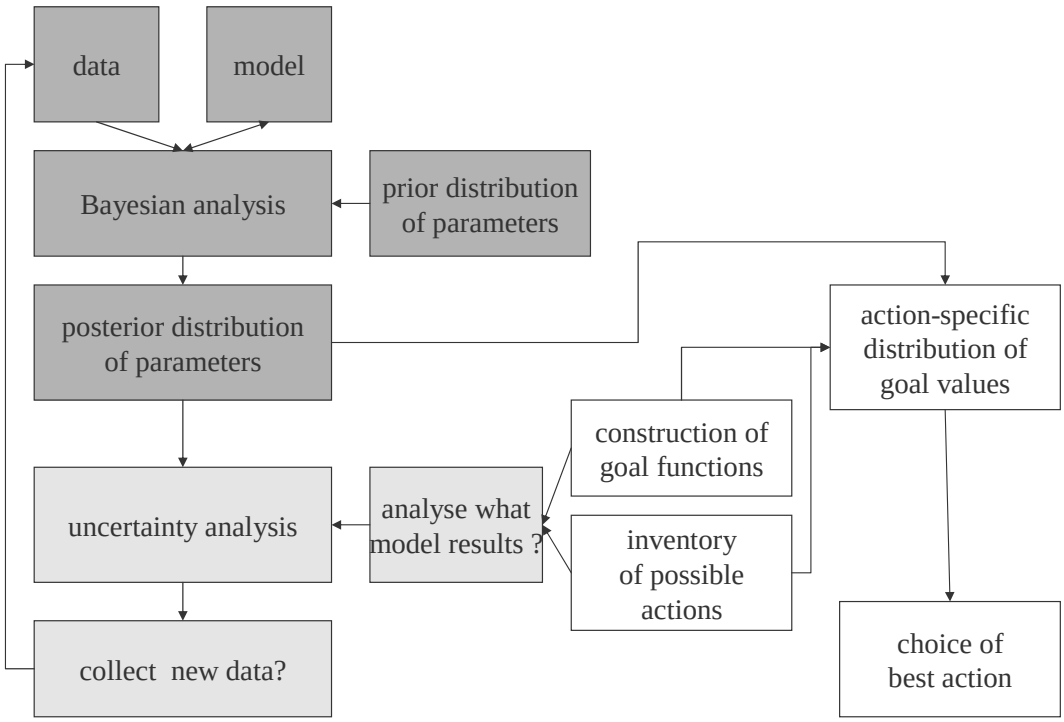


Figure 1. Framework. The dark-grey boxes in the upper-left corner represent the data analysis. The light-grey boxes in the lower-left corner show how uncertainty analysis fits in. The white boxes in the right part pertain to decision theory

The light-grey boxes in the lower left-hand corner of Figure 1 show how uncertainty analysis fits into the framework. The results of an uncertainty analysis may suggest whether the collection of new data is required for sufficiently precise model predictions, and may also suggest what kind of data are required most for sharper predictions. Uncertainty expressed as randomness, as resulting from a Bayesian data analysis, is attractive as a starting point for uncertainty analysis. The currently most common form of uncertainty analysis assumes that parameters and other uncertain quantities that specify a model and its working in a specific context, are represented as interdependent random variables (e.g. Saltelli, Chan and Scott 2000). Thus, a Bayesian analysis of the quantities that enter a model can be seamlessly linked with a subsequent uncertainty analysis. On the other hand, when results of a non-Bayesian (classical) statistical analysis of model parameters have to enter an uncertainty analysis, some forcing has to be applied in order to transform classical results into the required form.

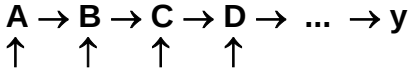
The white boxes in the right-hand part of Figure 1 pertain to decision theory. Almost invariably, decision support is based on results of a mathematical or statistical model. A Bayesian analysis of the uncertainty of model parameters and other inputs can be used quite well as starting point for decision analysis (e.g. Berger 1985).

Although there is a classical form of decision analysis under uncertainty, the Bayesian way is most widespread (e.g. Efron 1986). A decision analysis needs an inventory of possible actions that might be taken, and between which a choice has to be made. Apart from that, one has to formulate a goal function, probably compromising between several goals of several stakeholders. The decision analysis uses the posterior parameter distribution to evaluate the uncertainty of the goal values and the uncertainty of the difference in goal values between the possible actions. Moreover, the decision analysis may help identifying what model results are the most interesting subjects for an uncertainty analysis.

This paper starts with a brief sketch of uncertainty analysis, with food-chain models as example. The paper ends with Bayesian calibration of complex models as major subject. The purpose of the paper is to draw attention to problems and possible solutions. It is not the intention to introduce ready-for-use methods.

Uncertainty analysis of a food-chain model

A simple food chain may be modelled by a train of sub-models, say A, B, C, D..., the output of one sub-model serving as input of the next (Jansen 1998). The sub-models can pertain to production, transport, processing, storage and so on. The whole food-chain model yields some property of the end-product, say y, as output.



Each sub-model is specified by imperfectly known parameters and influenced by unpredictable exogenous inputs. Uncertainty entering the chain is represented by the upward arrows in the diagram, and accumulates down-stream. The purpose of an uncertainty analysis is firstly to evaluate if the accumulated random effect is still acceptably small, and secondly to pinpoint the major sources of uncertainty in order to assess the possibilities to reduce uncertainty by control of the chain or by additional research. Uncertainty analysis often takes the form of an analysis of variance, producing for instance an answer to the question: “What is the expected reduction of variance of model output y if some group of inputs would become known perfectly?” (e.g. Jansen, Rossing and Daamen 1994; Saltelli, Chan and Scott 2000).

In the most common form of uncertainty analysis – often also referred to as sensitivity analysis – uncertain quantities that specify a model and its working in a specific context, are represented as interdependent random variables. Thus, Bayesian analyses of the data used to parameterize the model produce precisely the description of input uncertainty that is needed for an uncertainty analysis. On the other hand, the results of classical data analyses require some conversion and re-interpretation before they can enter such an uncertainty analysis.

The analysis estimates the resulting distribution of one or more model outputs. Most often, the uncertainty of an output is characterized by its variance.

Model calibration

Calibration of a model is often viewed in the narrow sense of adapting some of the parameters in order to get better resemblance between observations and major end-predictions in a specific situation. Still too often, calibration is performed in a totally irreproducible way, sometimes even by hand according to some eyeball criterion,

producing results of unknown quality. Even if such a calibration is performed in a reproducible way, the choice of the adjusting parameters and of the resemblance criterion remains problematic by lack of guiding principles.

From the very beginning of complex modelling until the present day, some authors speak of *degeneration* or *corruption* of mechanistic models by such a form of calibration.

De Wit (1970, p. 17-18) for instance, distinguishes an explanatory level and an explainable level in modelling. At the explanatory level, more or less basic and known laws are at work. The knowledge at this level is summarized in a model whose behaviour is used to predict phenomena at the explainable level. According to De Wit, parameters have to be estimated at the explanatory level. But "it will often be found that the results obtained with experimenting with the model and the actual system do not agree. In that case the model may be adjusted such that a better agreement is obtained. Since there are many parameters and many equations involved this is not difficult. However, it is a disastrous way of working because the model degenerates from an explanatory model into a demonstrative model [like a planetarium] which cannot be used anymore for extrapolation, and the technique reduces into the most cumbersome and subjective technique of curve fitting that can be imagined."

More recently, Beck and Chen (2000, p. 402) similarly state: "After a model structure has been created in a maximally *physically meaningful* manner, calibration may *corrupt* the original *relatively pure articulation of theory*".

An extreme example of spoiling the mechanistic nature of a model is provided by a theoretically positive parameter that takes a negative value after calibration. A subtler instance would be a parameter assuming a value that is implausible in view of experimental results or expert knowledge.

The essence of the misgivings in the above quotations seems to be that calibration in the narrow sense may corrupt a model by ignoring information. The obvious solution would be to perform calibration *in the broad sense of combining all relevant information about the parameters*. Characteristically, complex models draw their information from quite diverse sources: observations at several spatial and temporal scales, experiments with sub-models, information from literature and experts, observations of end-predictions of the model etc. A good calibration should combine the available information in a reproducible and scientifically sound way into estimates of parameters, accompanied by an indication of their accuracy.

In the context of broad-sense calibration, the term 'model' should be viewed in a broad way: firstly, it should account for measurement errors and other possible forms of randomness; moreover, it should comprise all kinds of special models for special situations, for instance, measurement models or models for experiments where only a sub-set of the parameters plays a role. By these extensions the model need not be seen as a black box producing end-results only. After the formulation of the extended model, the choice of a calibration criterion poses no problem, since no such criterion is needed any more; instead the whole Bayesian procedure follows from the extended model.

Before and during the combination of the information from the various sources, one should remain aware of the possibility that the bits of information are somehow conflicting. Parameter values that were in force during an experiment with a sub-model might differ from the values in the current situation. Experts may have drawn their experience from situations quite different from the current one, while over-confidence of experts has been reported more than once. The model for all the data used during the calibration should be considered very critically, and checked

whenever possible. If the model for the calibration data is wrong, Monte Carlo algorithms for Bayesian analyses may suffer from convergence problems. The awareness of the possibility of conflicting evidence is analogous to the care required in meta-analysis, the statistical analysis of results of individual studies for integrating their findings. Indeed, a large part of meta-analysis theory is directed towards the detection and resolution of contradictions (e.g. Hedges and Olkin 1985).

In principle, a broad-sense calibration might be performed with classical statistical methods, but such an analysis may be hampered by technical problems. It is often very difficult to combine heterogeneous information with classical statistical methods. Moreover, in a classical analysis all parameters should be identifiable from the data: the prediction of the data should be different for different values of the parameter vector. A classical statistical analysis requires several more regularity conditions (e.g. Cox and Hinkley 1974, Chapter 9): the analysis requires continuity of the model's response to parameter changes, and continuity of the derivative of the response.

Bayesian methods seem to be much more promising for calibration in the broad sense. First of all, these methods can easily integrate diverse information. Heterogeneous data sets can be analysed consecutively, the posterior of the previous analysis taking the role of prior in the next one. Expert knowledge can be used as prior information for the first analysis.

Similarly, Bayesian methods can elegantly manage 'missing data'. The term often refers to planned measurements that were not executed. Some calculations cannot be performed efficiently, or cannot be performed at all, when such data are missing. In a Bayesian analysis one can efficiently cope with missing values by treating their values as unknown parameters. This stratagem may be applied to supplement the data with any other data – planned or not – that may enable or facilitate analysis.

The problem of selection of calibration parameters might be dealt with by calibrating all parameters rather than a few selected ones. Insensitive parameters can be handled: the posterior distribution of a totally insensitive parameter will be the same as the prior distribution. Moreover, with most Monte Carlo methods for Bayesian analysis, the computational burden hardly grows by the addition of insensitive parameters. If the calibration situation was insensitive to some parameters, the same insensitivity might occur in a new prediction situation, so that one can still make sufficiently sharp predictions; but in a new situation that is sensitive to a badly-known parameter, predictions will appear to be vague. In both cases the method of prediction under the given uncertainty is the same.

Fast emulator errors

Kennedy and O'Hagan (2001) describe a general Bayesian framework for the calibration of complex models. They provide the most complete treatment to date of computer-code uncertainties: errors due to replacement of the model by a fast emulator. As pointed out above, Bayesian calibration often requires thousands of model runs, and thus it may become necessary to replace the model by a quick and simple stand-in. A Bayesian approach is chosen because the authors seek to evaluate post-calibration uncertainty taking into account all kinds of causes of differences between model prediction and measured realization.

Almost-Bayesian procedures

Several authors have proposed procedures to improve on the above-sketched state of affairs in present-day model calibration. Their appraisal of the current situation is similar to ours. Their solution has a Bayesian flavour but is less formal. Beven and

Binly (1992) propose a procedure called Generalized Likelihood Uncertainty Estimation, to calibrate a hydrological model for a river catchment. The term likelihood is used in a very general loose sense and not in any formal statistical sense. Klepper (1989) claims to have calculated post-calibration parameter uncertainty with the use of the so-called Price algorithm, but no proof of the claim is given.

Examples of model calibration

In this section we briefly discuss a few examples of Bayesian calibration. Admittedly, most examples are somewhat simpler than the complex cases discussed in the previous sections. This fall in complexity reflects the fact that Bayesian calibration is still in its infancy.

Crop growth

The first example is an exercise in the calibration of a crop growth model, SUCROS, tailored for maize. The model simulates maize growth during one growing season, with daily temperature and solar radiation as input. A prior distribution for 21 of the parameters had been obtained from a meta-analysis of worldwide literature results (Metselaar 1999). Possibly, this prior distribution may be viewed as a description of the variation of the parameters of different maize cultivars over the world. This prior information was supplemented with a fictitious harvest measurement of 15 ton/ha dry-weight with a standard error of 1 ton/ha, with a specific cultivar in The Netherlands in 1985.

The Bayesian analysis is based on sampling from the posterior distribution of the parameters. A sample of 1000 independent draws was constructed with a very basic Monte Carlo method. In a for loop, independent draws from the prior parameter distribution were proposed, and accepted into the posterior sample by chance with probability proportional to the likelihood. As soon as the required posterior sample size was reached, the for loop ended.

The effect of combining this generic prior information with cultivar-specific data will be illustrated by making model predictions under slightly different weather conditions, namely those of 1986, while the measurement pertained to 1985 weather. Figure 2 (left) shows a histogram of a size-1000 sample from the prior predictive harvest distribution for 1986; only the prior information is used here. The histogram shows that this prior information is insufficient by itself: the distribution is very vague. The effect of the information about the 1985 harvest on prediction for 1986 is illustrated by Figure 2 (right), which shows the histogram of a size-1000 sample from the posterior predictive harvest distribution for 1986, taking into account the 1985 harvest measurement in the way described above. The histogram is much sharper, which shows that combining the two kinds of information has a positive effect. Moreover, the mean of the posterior predictive distribution is greater than the prior mean, which might express that the cultivar in question is adapted to the local climate.

Since the data consist of merely one observation, whereas 21 parameters have to be estimated, these parameters cannot be estimated with a classical analysis, which requires that the prediction of the data should be different for different values of the parameter vector. The Bayesian approach enables us to compensate for this lack of information in the data by taking prior information into account. Even if the data would have been sufficiently numerous for a classical analysis, a Bayesian analysis has an advantage that is worth mentioning particularly in the context of calibrating

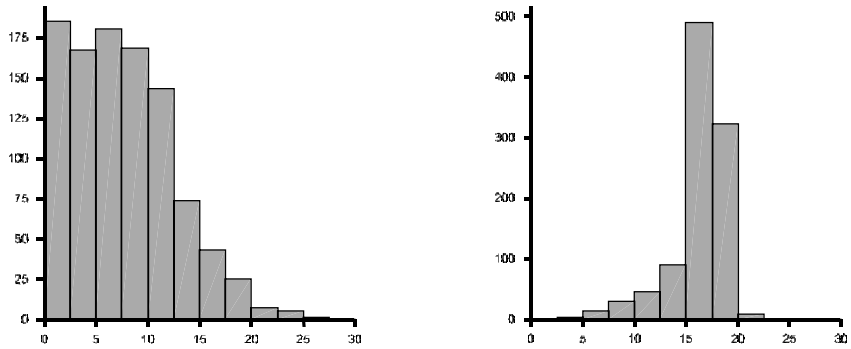


Figure 2. *Left*: Histogram of size-1000 *prior* predictive sample for 1986 maize harvest in kg/ha. *Right*: Histogram of size-1000 *posterior* predictive sample for 1986 maize harvest in kg/ha

complex models. Usually, model outcomes are not equally sensitive to changes in all 21 parameters considered over their range of variation. For a non-Bayesian calibration, one would normally restrict the calibration to a small number of sensitive parameters with a wide range. Such a restriction, which always suffers from some arbitrariness, is not necessary in most Monte-Carlo-based Bayesian analyses, since insensitive parameters hardly increase the workload.

Markovian meta-population model

O'Hara et al. (2002) and Ter Braak and Etienne (2003) analyse a meta-population model, an often-used model describing a population dispersed over several distinct patches. The model has 5 parameters, defining extinction probability, colonization probability and connectivity. The data describe for N patches over T years whether the patches were occupied that year. The patches-by-year data matrix typically contains many missing values, which is the major reason to choose a Bayesian approach. This is because in a Bayesian analysis, as mentioned above, unknown observations can be treated as parameters, which can simply be sampled in a Monte-Carlo-based calculation of the posterior distribution. Using simulated data sets O'Hara et al. (2002) show that the approach can successfully deal with missing data for this important type of ecological model. Ter Braak and Etienne (2003) improve on their approach by additionally augmenting the data with a past period preceding the first year of observation. This obviates the need to condition the likelihood on the first-year observations, and thus uncovers the information in those observations, a gain worthwhile in data sets that typically cover only a few years. The attached past period has to be chosen sufficiently long for an approximate establishment of quasi-stationarity in the meta-population model, such that the likelihood is virtually independent of the fixed initial state at the start of the attached period. This results in a well-founded and efficient estimation of parameters, with a posterior distribution expressing post-calibration uncertainty.

Correlated random walk

Another important type of ecological models is that of the correlated random walk, briefly CRW (Turchin 1998). CRW models describe animal movement within and/or between habitats, and as such they can be used to inform the construction and parameterization of the meta-population models discussed in the previous example.
Data from animal-tracking experiments are perhaps the most important type of information used to calibrate CRW models. In this type of data missing observations

tend to be abundant, so again the best-choice calibration approach seems to be the Bayesian one. In Figure 3 we present results for a very simple model variant describing movement in a single homogeneous vegetation type, in which the animal movement is described by two distributions, one for the step length and one for the turning angle between consecutive steps. The simulated data are for an exponentially distributed step length (specified by one parameter, the mean step length λ) and a turning angle following a wrapped-normal distribution centred at zero angle (again specified by a single parameter, namely the turning angle standard deviation σ). The results suggest that a Bayesian analysis can indeed be very powerful in estimating CRW model parameters from animal-tracking data with many missing observations (Hagenaars, Goedhart and Jansen in prep.). In the Markov-chain Monte Carlo algorithm underlying the results of Figure 3b, the unobserved data points are sampled by drawing CRW steps using the last-sampled CRW parameters. The standard deviation of the posterior sample is only 1.2 times larger than that of the sample in Figure 3a, showing that the effort of including the unobserved animal positions in the estimation procedure is paying off.

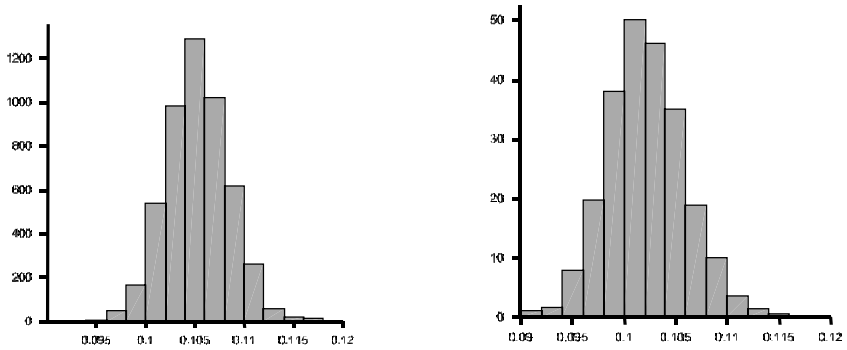


Figure 3. *Left*: Histogram of size-5000 posterior sample for CRW parameter σ. Calculated using a simulated data set of 5 correlated random walks of 100 steps each. The data set was generated with σ=0.1, and the sample of turning angles has an exact standard deviation σ=0.1051. *Right:* Histogram of size-5000 posterior sample for CRW parameter σ. Calculated using the same simulated data set as on the left side, of which 50% (randomly drawn) of data points (animal positions) are now considered unobserved.

Discussion and conclusions

The current state of affairs with respect to the calibration of complex models is far from ideal. We have argued that Bayesian methods provide a promising avenue along which this situation may be improved. The examples of the paper illustrate some of the advantages of the Bayesian approach, but they also show that still much research and development are needed in order to harvest these advantages in the calibration of truly complex models.

References

Beck, M.B. and Chen, J., 2000. Assuring the quality of models designed for predictive tasks. *In:* Saltelli, A., Chan, K. and Scott, E. M. eds. *Sensitivity analysis*. Wiley, Chichester. Wiley series in probability and statistics.

Berger, J.O., 1985. *Statistical decision theory and Bayesian analysis*. 2nd edn. Springer, New York. Springer series in statistics.

Beven, K. and Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes*, 6 (3), 279-298.

Cox, D.R. and Hinkley, D.V., 1974. *Theoretical statistics*. Chapman and Hall, London.

De Wit, C.T., 1970. Introduction: dynamic concepts in biology. *In: Prediction and measurement of photosynthetic productivity: proceedings of the IBP/PP technical meeting, Trebon, 14-21 September 1969*. Pudoc, Wageningen, 17-23.

Efron, B., 1986. Why isn't everyone a Bayesian? (with discussion). *The American Statistician*, 40 (1), 1-11.

French, S., 2001. Modelling, making inferences and making decisions: the roles of sensitivity analysis. *In:* Prado, P. and Bolado, R. eds. *Proceedings of SAMO 2001: third international symposium on sensitivity analysis of model output, Madrid, June 18-20, 2001*. CIEMAT, 45-48.

Hagenaars, T.J., Goedhart, P.W. and Jansen, M.J.W., in prep. Bayesian calibration of correlated random walk models.

Hedges, L.V. and Olkin, I., 1985. *Statistical methods for meta-analysis*. Academic Press, Orlando.

Jansen, M.J.W., 1998. Uncertainty analysis of food-chain models. *In:* Tijskens, L. M. M. and Hertog, M. L. A. T. M. eds. *Proceedings of the international symposium on applications of modelling as an innovative technology in the agri-food-chain, Model-It, Wageningen, Netherlands, 29 November-2 December, 1998. Acta Horticulturae ISHS*, Vol. 476. 33-40.

Jansen, M.J.W., Rossing, W.A.H. and Daamen, R.A., 1994. Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. *In:* Grasman, J. and Van Straten, G. eds. *Predictability and nonlinear modelling in natural sciences and economics*. Kluwer, Dordrecht, 334-343.

Kennedy, M.C. and O'Hagan, A., 2001. Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society. Series B*, 63, 425-464.

Klepper, O., 1989. *A model of carbon flows in relation to macrobenthic food supply in the Oosterschelde estuary (S.W. Netherlands)*. PhD thesis, Landbouwuniversiteit Wageningen.

Metselaar, K., 1999. *Auditing predictive models: a case study in crop growth*. PhD thesis, Landbouwuniversiteit Wageningen.

O'Hara, R.B., Arjas, E., Toivonen, H., et al., 2002. Bayesian analysis of metapopulation data. *Ecology*, 83 (9), 2408-2415.

Saltelli, A., Chan, K. and Scott, E.M. (eds.), 2000. *Sensitivity analysis*. Wiley, Chichester. Wiley series in probability and statistics.

Ter Braak, C.J.F. and Etienne, R.S., 2003. Improved Bayesian analysis of metapopulation data with an application to a tree frog metapopulation. *Ecology*, 84, 231-241.

Turchin, P., 1998. *Quantitative analysis of movement: measuring and modeling population redistribution in animals and plants*. Sinauer, Sunderland.