

# Uncertainties in process-based modelling for environmental assessment

Marcel van Oijen

Centre for Ecology and Hydrology (CEH-Edinburgh), U.K. ([mvano@ceh.ac.uk](mailto:mvano@ceh.ac.uk))

**Abstract.** Many different process-based models of ecosystems are in use today. The majority of these models are parameter-rich, deterministic dynamic models, which require considerable input information and computation time. These characteristics, combined with the fact that the models tend to be parameterised at the point-support spatial scale, have made their use for larger regions problematic. Quantifying the uncertainties caused by incomplete knowledge of model inputs and structure, as well as uncertainty due to upscaling, is a difficult task. Various examples of model application and uncertainty quantification are presented here and the possibility to use a Bayesian approach to uncertainty quantification is discussed.

## 1 INTRODUCTION

Process-based models (PBMs) are increasingly common tools for analysing and predicting the impact of environmental change on ecosystems (Ogle & Barber 2008). This reflects the realisation that ecosystems are dynamic systems that change over time in response to the environment. Here, we focus on PBMs that simulate the dynamics of soil-vegetation systems such as forests and crops. The possible applications of such PBMs depend on their input-output relationships. A model's inputs include the environmental factors whose impacts can be assessed, whereas its outputs are the measures of ecosystem performance that we may be interested in. The environmental inputs that drive the models are atmospheric conditions, such as weather and CO<sub>2</sub>, and soil conditions, such as carbon, nitrogen and water content. The outputs from the models are time series of variables like productivity, carbon sequestration in soil and biomass, soil loss in erosion, and other variables that can be related to ecosystem services (Fig. 1).

Most PBMs for forests and crops are complex deterministic simulators that are highly nonlinear. The models simulate the cycling of carbon, water and nutrients within vegetation and across the boundaries with soil and atmosphere. In contrast to hydrological models, ecosystem models focus on vertical transport, i.e. geographically they operate at point-support. The PBMs are written as sets of differential equations which are solved numerically, making the models computationally demanding. For that reason, the application of PBMs to large, spatially heterogeneous areas has been difficult, but advances in computing capacity have made it possible to use PBMs for predicting the impact of regional and global environmental change. However, the uncertainties associated with such model applications have often not been quantified and analysed fully. Here, we shall review some recent work on PBM uncertainties and try to identify areas where more progress needs to be made.

## 2 UNCERTAINTIES IN PROCESS-BASED MODELLING

All uncertainty in the outputs of PBMs for any specific site derives from uncertainty about model structure and about model inputs. Uncertainty about the appropriate structure of ecosystem PBMs is large, as is evident from the large number of very different models that have been proposed in the literature. For example, the Register of Ecological Models (REM; <http://ecobas.org/www-server>) currently holds 78 forest models and 34 grassland models, and more can be found in the literature. The second type of uncertainty, about model inputs, is

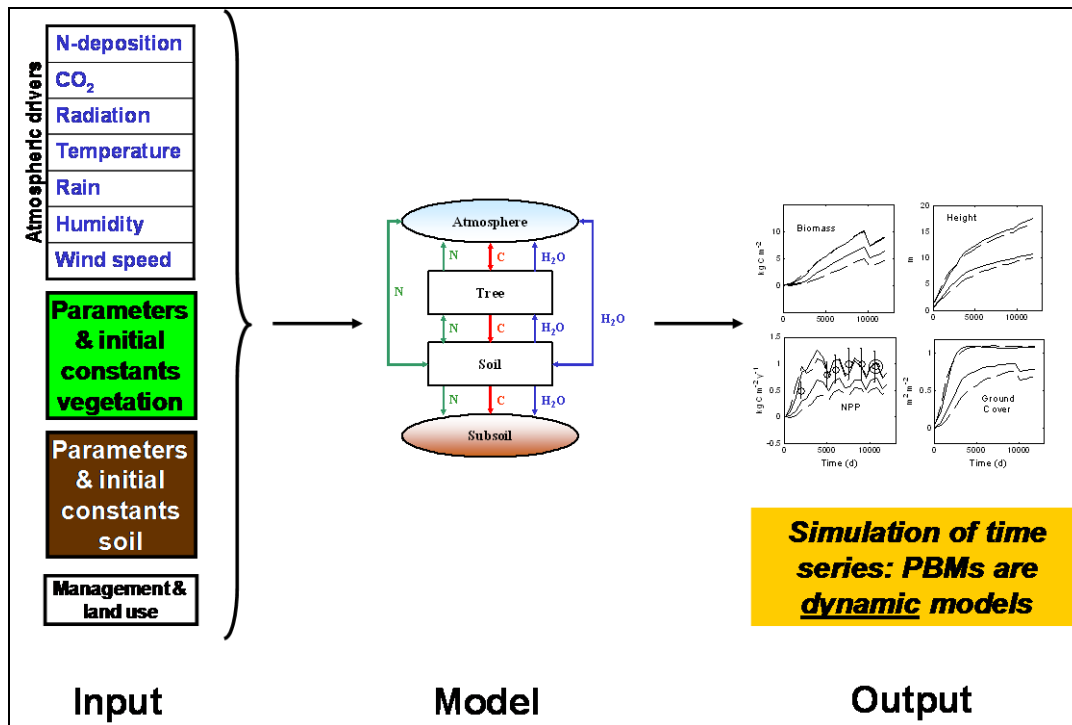


Fig. 1. Common inputs (left) and outputs (right) from process-based ecosystem models.

caused by incomplete knowledge of past and future weather conditions, inaccurate and imprecise soil maps, and poorly known process parameters and initialisation constants of the model's state variables. Levy et al. (2004) quantified the forward propagation of both structural and input uncertainty in forest modelling by feeding multiple PBMs with parameter values sampled from probability distributions that reflected the lack of unanimity in the literature. In their analysis, the structural and input uncertainties were so large that the use of PBMs to predict the impact on forests of atmospheric nitrogen deposition was severely limited.

### 3 PBMs AND UNCERTAINTIES ABOUT SOIL VARIABLES

Soil variables are essential inputs to ecosystem PBMs but soil knowledge is often limited, which poses several problems to the use of the models.

First, the way soils are represented in ecosystem PBMs does not always correspond to the variables that are actually measured. The most common problem is that of soil carbon pools. Total soil carbon can be fairly easily measured, but PBMs tend to subdivide soil organic carbon in pools of different turnover rate. A typical model subdivision consists of three pools: (1) a very slowly decomposing "recalcitrant" soil carbon pool, (2) a pool of soil organic matter that decomposes at intermediate rate, (3) a pool of litter that decomposes fast. Such pools are near-impossible to distinguish in measurement, so even when total carbon is well known, there is still uncertainty about the relative pool sizes. The impact on model performance of an incorrect assumption about carbon distribution over pools may be very large (Fig. 2; Yeluripati et al. 2009), and many modellers have resorted to "spinning-up", i.e. pre-running the model to equilibrium, to stabilise the simulation of carbon dynamics. However, real ecosystems are generally not in equilibrium and a much better method than spinning-up may be to use Bayesian calibration to quantify the joint probability distribution for the three carbon pools (Yeluripati et al. 2009).

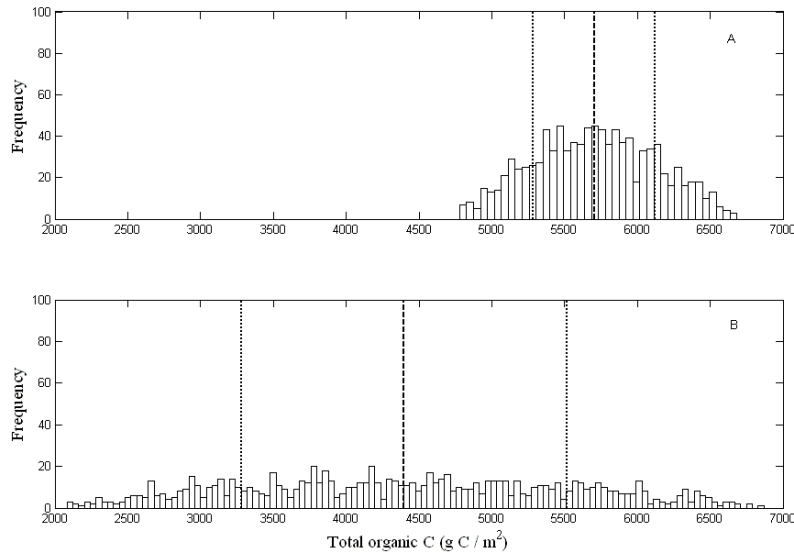


Fig. 2. Forward propagation in time of uncertainty about relative magnitudes of grassland soil carbon pools. The model DAYCENT (Del Grosso et al. 2001) was run from a fixed total of  $6500 \text{ g C m}^{-2}$  at time = 0, but relative magnitudes of the three constituting soil C pools were varied according to prior uncertainty. A) Uncertainty about total soil C after 10 years of simulation. B) Uncertainty after 100 years. [After Yeluripati et al. 2009]

Secondly, in recent years ecosystem modellers have realised that soil information needs to include data on the nitrogen content of the soils, even in cases where the primary model application is to predict carbon dynamics. Modelling studies have shown that the response of ecosystems to environmental change is strongly dependent on the degree of nitrogen saturation of the soil (e.g. Van Oijen & Jandl 2004, Van Oijen et al. 2008). Most current ecosystem PBMs simulate the linkages between the carbon and nitrogen cycles in the soil-vegetation system. However, information on initial soil nitrogen pools is often lacking, or very poor in spatial resolution. Van Oijen & Thomson (in press) used nitrogen data from the global IGBP-DIS dataset to initialise soil nitrogen content across the U.K. for their process-based simulations of nationwide carbon sequestration. Bayesian calibration, using data from some well-researched forest sites, was used to quantify the parameter uncertainty. Output uncertainty was large and not proportional to the values of carbon sequestration itself (Fig. 3). Although this study demonstrated the effectiveness of the Bayesian approach for parameterisation of fairly complex forest PBMs, no effort was made to quantify the uncertainty associated with the soil nitrogen input data – as there was only one source of such data, the IGBP-DIS global dataset - although the relative insensitivity of the model to spatial variation in atmospheric N-deposition suggests that the soil data were overestimates. Uncertainty about the quality of such data strongly affects the confidence we can place in the application of ecosystem PBMs.

Thirdly, there is the issue of scale, which affects both soil and atmospheric inputs. As described above, ecosystem PBMs are mostly point-support models. Therefore, in many applications of PBMs to problems that affect large geographical areas, the models are only applied to a finite number of specific sites that are considered to be representative of the area. In such methods, there is no true upscaling of the models, and model uncertainties can easily be quantified using Bayesian calibration which can be carried out for each site separately or for all sites simultaneously (Reinds et al. 2008, Lehuger et al. 2009). However, often there is a need for environmental assessment that covers not just a finite number of sites, but complete regions. When PBMs are applied across large regions, the approach tends to be to subdivide space in large matrices of grid cells, and run the models once for each grid cell

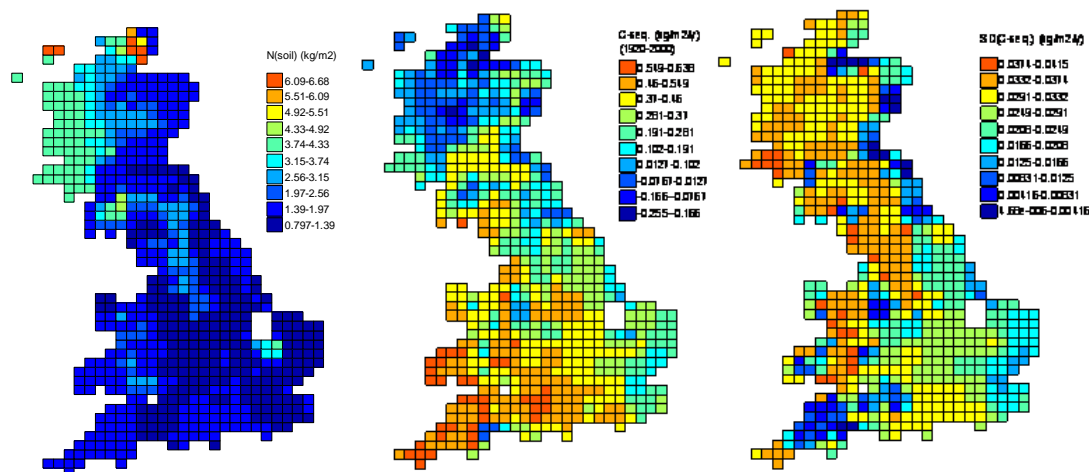


Figure 3: Process-based modeling of C-sequestration in the UK, 1920-2000. Cells are 20x20 km. Left: soil nitrogen content (source: IGBP-DIS). Mid: average annual C-sequestration simulated using forest model BASFOR. Right: uncertainty (S.D.) of outputs shown in mid panel. [Based on: Van Oijen & Thomson, in press]

using as input the average conditions for that cell. In global modelling, the grid cells can be up to 100 by 200 km in size, and in continental modelling the grid cells are often 25 by 25 or 50 by 50 km. Obviously, this approach raises uncertainty about the impact of space discretization and the use of cell-average inputs on model predictions. The following section (based on Van Oijen et al. 2009) discusses methods that have been used or proposed for upscaling of PBMs from the point support to the grid cell level.

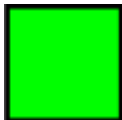
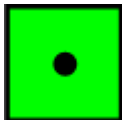
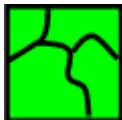
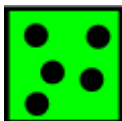
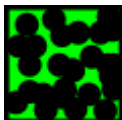
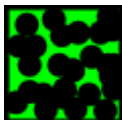
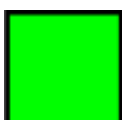
#### 4 AN OVERVIEW OF METHODS FOR UPSCALING PROCESS-BASED MODELS

Grid-cells used in spatial application of ecosystem PBMs often are larger than 20 by 20 km. The problem facing a ‘regional’ (grid-cell level) user of point-support PBMs is that the models tend to be too slow to be run exhaustively across regions of such size, even if environmental information happens to be available for every point in the region. This problem can only be solved by working with approximations, of either region or model. This implies using some form of sampling and/or model simplification. We can distinguish seven methods, each described briefly in Table 1.

The example of Fig. 3 showed the most common approach to PBM upscaling: combining methods (1) and (3). Method (1) strictly requires reparameterisation of the PBM, using regional I-O data, but this is rarely done. An exception is the work of Patenaude et al. (2008), who calibrated the 3PG model using remotely sensed data from a small forested region. Recent years have seen an increasing abundance of data on structural characteristics of vegetation, measured by remote sensing, and on gas-fluxes, measured by eddy-covariance towers. This makes direct regional parameterisation increasingly possible. If the parameterisation is carried out using probabilistic techniques like Bayesian calibration (Patenaude et al. 2008), uncertainty quantification will be included.

The sampling-based methods (2)-(4) require uncertainty quantification to account for assumptions of representativeness of strata or control points. Recent developments in geostatistics may provide methodology. Bayesian kriging, for example, affords a means to quantify uncertainties comprehensively, including the uncertainty of the spatial interpolation parameters (Banerjee et al. 2004).

Table 1: Methods for regional application of point-support process-based models. Each algorithm description ends with a line that shows how to calculate regional totals of model outputs. “y” is model output (stocks, fluxes, etc.) per unit area, “area” is the area of the region, “Total” is the integral of y across the regional area.

Category	Upscaling method	Algorithm
(1)-(3): Methods that use the original model		(1) Reinterpret the point-support model as being a regional one 1. Re-calibrate model for regional use 2. Get average inputs for region 3. Run model for average conditions 4. Total = $y * \text{area}$
		(2) Select representative point 1. Select a representative point 2. Get inputs for point 3. Run model for point 4. Total = $y * \text{area}$
		(3) Stratify into homogeneous subregions 1. Stratify the region 2. Get average inputs per stratum 3. Run model for all strata 4. Total = $\sum(y_i * \text{area}_i)$
(4): Method in which the original model is extended		(4) Run for selected points & interpolate 1. Select control points 2. Get inputs for control points 3. Run model for control points 4. Derive geostatistical interpolation model f 5. Total = $\int y(s)$ where $y(s) = f(y_{1..n}, s_{1..n})$
(5)-(7): Methods that rely on making a new model		(5) Create deterministic metamodel & apply exhaustively across the region 1. Create fast metamodel using training set of multiple point-scale I-O 2. Get inputs across whole region 3. Run model for all points 4. Total = $\text{mean}(y) * \text{area}$
		(6) Create stochastic emulator & apply exhaustively across the region 1. Create fast emulator using training set of multiple point-scale I-O 2-4. As Method (5)
		(7) Summarize model behaviour & embed in regional model with wider scope than point-support model 1. Summarise model behaviour in the form of a regional summary model 2. Embed summary in regional model 3. Run the regional model using regional-scale inputs 4. Total = $y * \text{area}$

Methods (5)-(7) use new models that approximate the original PBM. Probabilistic frameworks for dealing with such ‘models of models’ are still subject of intense research (Goldstein & Rougier 2009, Kennedy & O’Hagan 2001), and there is not yet a generally accepted method. The debate is mainly about how to account for the discrepancy between models and reality.

## 5 DISCUSSION

We have given examples that show both the versatility of PBMs, and their limitations because of the high demands they put on the quality of input data, in particular on soils. Another limitation is persisting model structural uncertainty. Furthermore, the fact that PBMs tend to be too slow, computationally, to be run exhaustively across heterogeneous regions – necessitating the upscaling techniques discussed above – makes it hard to quantify upscaling uncertainty. There is a need for a practical probabilistic framework that produces PDFs for the regional model output, conditional on input distribution (environment and parameters), upscaling assumptions and upscaling method. Clearly, more work needs to be done on the quantification of uncertainty associated with upscaling. As we indicated at several places in our overview, Bayesian approaches have been successfully applied to quantify the uncertainties associated with model inputs and structure (Van Oijen et al. 2005), and we expect that they will also be increasingly applied to spatial modelling of PBMs.

## 6 REFERENCES

- Banerjee, S., B.P. Carlin & Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman Hall, Boca Raton: 472 pp.
- Del Grosso S.J., Parton, W.J., Mosier, A.R., Hartman, M.D., Brenner, J., Ojima, D.S. & Schimel, D.S. (2001). Simulated interaction of carbon dynamics and nitrogen trace gas fluxes using the DAYCENT model. In: Schaffer, M., et al. (Eds.), *Modeling Carbon and Nitrogen Dynamics for Soil Management*, p. 303-332, CRC Press, Boca Raton, Florida, USA.
- Goldstein, M. & Rougier, J. (2009). Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference* 139: 1221-1239.
- Kennedy, M. C. & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 63: 425-450.
- Lehuger, S., Gabrielle, B., Van Oijen, M., Makowski, D., Germon, J.-C., Morvan, T. & Hénault, C. (2009). Bayesian calibration of the nitrous oxide emission module of an agro-ecosystem model. *Agriculture, Ecosystems and Environment* 133: 208-222.
- Levy, P.E., Wendler, R., Van Oijen, M., Cannell, M.G.R. & Millard, P. (2004). The effects of nitrogen enrichment on the carbon sink in coniferous forests: uncertainty and sensitivity analyses of three ecosystem models. *Water, Air and Soil Pollution: Focus*, 4: 67-74.
- Ogle, K. & Barber, J.J. (2008). Bayesian data-model integration in plant physiological and ecosystem ecology. *Progress in Botany* 69: 281-311.
- Patenaude, G., R. Milne, M. Van Oijen, C.S. Rowland & Hill, R.A. (2008). Integrating remote sensing datasets into ecological modelling: a Bayesian approach. *International Journal of Remote Sensing* 29: 1295-1315.
- Reinds, G.J., Van Oijen, M., Heuvelink, G.B.M. & Kros, H. (2008). Bayesian calibration of the VSD soil acidification model using European forest monitoring data. *Geoderma* 146: 475-488.
- Van Oijen, M., Ågren, G.I., Chertov, O.G., Kellomäki, S., Komarov, A., Mobbs, D.C. & Murray, M.B. (2008). Evaluation of past and future changes in European forest growth by means of four process-based models. In: *Causes and Consequences of Forest Growth Trends in Europe*. Eds. H.P. Kahle et al. EFI Research Reports, Brill publ., Chapter 4.4: 183-199.
- Van Oijen, M. & Jandl, R. (2004). Nitrogen fluxes in two Norway spruce stands in Austria: An analysis by means of process-based modelling. *Austrian Journal of Forest Science* 121: 167-172.
- Van Oijen, M., Rougier, J. & Smith, R. (2005). Bayesian calibration of process-based forest models: bridging the gap between models and data. *Tree Physiology* 25: 915-927.
- Van Oijen, M. & Thomson, A. (in press). Towards Bayesian uncertainty quantification for forestry models used in the U.K. GHG Inventory for LULUCF. *Climatic Change*.
- Van Oijen, M., Thomson, A. & Ewert, F. (2009). Spatial upscaling of process-based vegetation models: An overview of common methods and a case-study for the U.K.. *StatGIS2009*, 16-18 June 2009, Milos, Greece: 6 pp.
- Yeluripati, J.B., Van Oijen, M., Wattenbach, M., Neftel, A., Ammann, A., Parton, W.J. & Smith, P. (2009). Bayesian calibration as a tool for initialising the carbon pools of dynamic soil models. *Soil Biology and Biochemistry* 41: 2579-2583.