ELSEVIER

Ecological Modelling 83 (1995) 55–66

ECOLOGICAL MODELLING

Calibration of process-oriented models

P.H.M. Janssen *, P.S.C. Heuberger

RIVM, P.O. Box 1, 3720 BA Bilthoven, Netherlands

Accepted 25 October 1993

Abstract

Model calibration is a critical phase in the modelling process, and the need for a well-established calibration strategy is obvious. Therefore a systematic approach for model calibration is proposed which is guided by the intended model use, and which is supported by adequate techniques, prior knowledge and expert judgement. The success of calibration will be primarily limited by the nature, amount and quality of the available data, in relation to the complexity of the model; additional limitations are the effectiveness of the applied techniques and the availability of time, man- and computer power, adequate expertise and financial resources. These limitations will often preclude a unique calibrated model. As a consequence, calibration studies should provide information on the non-uniqueness and/or uncertainty which will be left in the model (parameters) after calibration, and this uncertainty should be adequately accounted for in subsequent model applications.

Keywords: Calibration; Uncertainty analysis; Validation

1. Introduction

Mathematical models have a wide-spread use in environmental applications. The developed models vary in sophistication and complexity, ranging from simple data-oriented models to complex process-oriented models, depending on the intended aim, the system under study, the available data, expertise, time, man- and computer power and financial resources.

The applied models typically only render an *approximate* description of the system under study and moreover contain various quantities, e.g. parameters, initial and boundary conditions [1], which are *incompletely* known; also the selection of an adequate model structure (i.e. the specific form of the constituting equations) can be a point of discussion. More information on these quantities, which are often not directly measurable, is required to obtain accurate inferences from the model, and for judging its performance adequately (e.g. in the context of a model comparison study). Hence *model calibration* will be required to determine these values accurately from the available measurements, taking into account

* Corresponding author.

[1] When discussing calibration of these quantities in the sequel, no explicit distinction will be made between them; they will all be referred to by the term 'parameters'.

SSDI 0304-3800(95)00084-4

the intended model use and the available prior knowledge.

Model calibration thus becomes a critical phase in the modelling process. Despite its importance, the required activities for calibration are often given little consideration, and in many situations

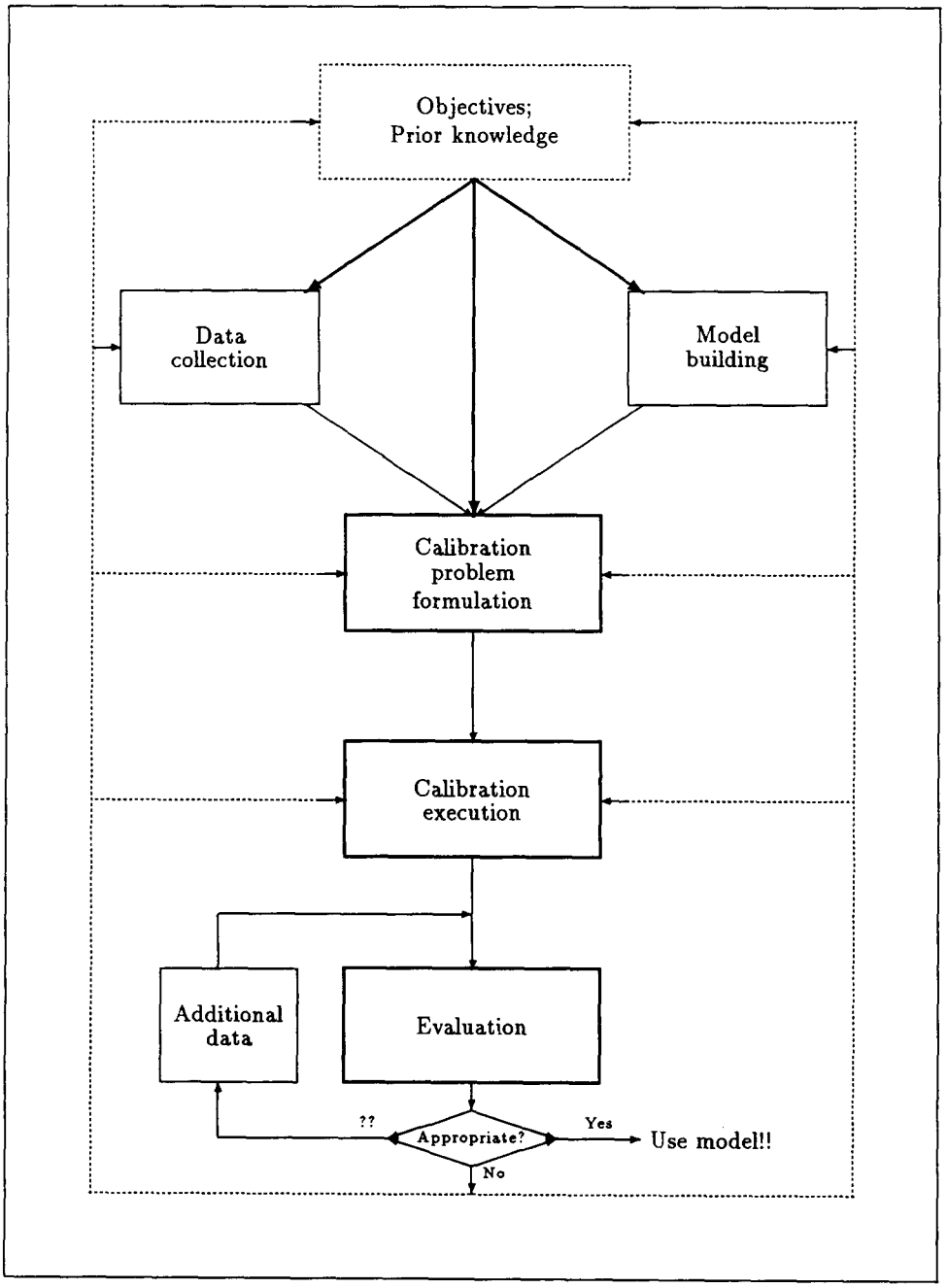

Fig. 1. An outline of the central issues in the calibration process.

the model is calibrated in a trial-and-error fashion by tuning some parameters manually until the model matches the data in some, often vague and subjective, sense ('eye-ball fitting'). Although this approach can be rather successful in applications where the number of parameters is small (< 3), it suffers from the lack of exactness, reproducibility and objectivity. Moreover, it becomes rapidly intractable for larger problems, and renders inconclusive results: clear information on the uncertainty in the calibrated model is lacking, and it is difficult to trace whether a mismatch between data and model is e.g. due to model deficiencies or to incomplete tuning of the parameters.

In the current pursuit for adequately certified and documented models, a well-structured and systematic calibration approach is needed instead, supported by useful guidelines and (automated) techniques. This will, when successful, improve the reproducibility, objectivity and quality of the results, and will substantially contribute to the quality of the calibrated models, in judicious correspondence with experimental or field work. When unsuccessful, it will still render useful diagnostic information on ways to improve the model.

Aim of this paper is to propose such a systematic calibration protocol in more detail, discussing the various choices which are involved, and indicating techniques for performing the calibration. Potential problems and limitations are highlighted, and their consequences are discussed. Focus will be on process-oriented models. The paper provided background information for the calibration and model evaluation activities at the 'Workshop on Comparison of Forest–Soil–Atmosphere Models' (Van Grinsven et al., 1995).

2. An outline of the calibration process

Model calibration is the determination of the model parameters and/or structure on basis of measurements, and prior knowledge. Since the notion of a 'true' model is typically a fiction for real-life applications, *'usefulness' rather than 'truth'* should be the guiding principle in developing and calibrating models. Therefore the ultimate use of the model should be explicitly acknowledged in the calibration process. Viewed against this background, various activities can be distinguished in this process: Experiments have to be designed and/or measurement data have to be collected and analysed. These measurements are compared with the corresponding model predictions. In its most elementary form this comparison is performed mainly *qualitatively* by visually inspecting the agreement between observed data and model predictions; more sophisticated approaches express the agreement between data and model *quantitatively* in terms of *misfit measures*, which typically are functions of the error between measurements and model predictions. The definition of these misfits should reflect the intended use of the model and should concern the model quantities which are deemed important. Subsequently the structure and/or parameters of the model are adjusted such that the 'data-to-model agreement' is satisfactory, thus reducing the initial uncertainty in the model (parameters). Information on this reduced uncertainty should be obtained to enable adequate inference from the model in subsequent applications (e.g. prediction). Having thus determined information on the calibrated model and its uncertainty, its quality should finally be assessed.

The involved issues in this process are indicated in Fig. 1, and will be discussed in more detail in the subsequent subsections.

2.1. The intended purpose for model calibration

Since the models will, by nature, only render an approximate description of reality it is essential to specify the *purpose(s) of the model* (e.g. prediction of short-term or long-term dynamics; simulation of specific dynamical features, e.g. episodes etc.). These objectives should serve as the main guiding principle in building, as well as in calibrating the model. They will indicate which aspects of the model (inputs, outputs, states) are relevant, and what temporal and/or spatial scales, and which levels of detail and accuracy are desired.

In addition to specifying the model purpose(s), it will be important to address questions like

'Why is calibration needed?', and *'What should be calibrated and in which form should it be reported?'* Modelling and calibration thus become *dedicated, tailored activities* rather than quests for the 'ultimate truth' ('pragmatism' instead of 'unattainable utopianism').

2.2. The available prior information

Prior information on the model quantities (e.g. parameter ranges, constraints) should be explicitly incorporated into the process of calibration. The nature and characteristics of this information should be clearly indicated, as well as its reliability and quality.

2.3. The data

Data from the system under study can originate from various sources, e.g. previous or current field studies, laboratory experiments. Decisions on items like *'Which data are collected?'*, *'When, where, how often and how to measure?'*, should ideally take into account the calibration objectives and the relevance which is attributed to the various quantities (inputs, outputs) for the application at hand. More formal and theoretical expositions on these *experimental design* issues is given by Walter and Pronzato (1990).

The characteristics of the data (quantity, quality, representativity, information contents, temporal and spatial scales etc.) can vary considerably from application to application and essentially defines the boundaries of what can be achieved by model calibration. It is recommended to *pretreat* the data before calibration in order to detect and remove inadequacies (e.g. outliers) and noise, and to discover important features and relationships (e.g. trends). A clear documentation of these activities is required.

2.4. The model (set)

Specification of the features of the employed model(s) will expose some limitations already in an early phase of the calibration. Issues to be addressed are (Tiktak and Van Grinsven, 1995): (a) aim and scope of the model; (b) model principles and major processes involved; (c) model components (i.e. what are the features of parameters, initial and boundary conditions, state variables, model inputs and outputs?); (d) governing equations; (e) technical information (numerical schemes, input data requirements, programming language, hardware requirements, run time etc.); (f) status of the model; (g) documentation.

Moreover, it is strongly recommended to perform a preliminary *model analysis* (sensitivity/uncertainty analysis), in order to (i) check the model operation; (ii) obtain increased insight in the model; (iii) indicate whether calibration is meaningful; (iv) identify important model parameters and outputs (Janssen et al., 1994).

2.5. The formulation of the calibration problem

To enable a systematic treatment of model calibration, the calibration problem should be posed in a formal and explicit sense. This task, and the success of solving the stated problem, will strongly depend on the intended aim and characteristics of the model, the nature, availability and quality of the data and the prior knowledge. These aspects are reflected in the four substeps of the problem formulation:

(i) Specification of employed model structure(s) and parametrization(s)

For process-oriented models the employed model structure(s) and parametrization(s) are often given beforehand. Then this substep amounts to specifying the parameters which are considered candidates for calibration, and to denoting the prior knowledge.

(ii) Specification of misfit measure(s)

Calibration is performed by comparing model results with measurements, and adjusting the structure and/or parameters of the model such that the model results and the observations match adequately. This match between model outputs and data is usually expressed quantitatively in terms of misfit measures. The choice of these misfit(s) should reflect the intended use of the model and should concern model quantitities

which are deemed relevant for the study at hand. Moreover it should also depend on the characteristics of the data and on the prior information on parameters and error structure. Decisions have to be made on which outputs to include in the misfit(s), and how to do this (e.g. transforming, weighting). Certainly trade-offs and compromises will be involved in these choices.

As an example, suppose that the discrepancy between model and data for the kth output ($k = 1, \cdots, q$) is expressed as:

$$C_k(\theta) = \sum_{i=1}^{N} |e_k(i;\theta)|^2 \quad (1)$$

where

$$e_k(i;\theta) := y_k(i) - \hat{y}_{M,k}(i;\theta) \quad (i = 1, \cdots, N) \quad (2)$$

is the error/residual between the ith measurement $y_k(i)$ and the associated model prediction $\hat{y}_{M,k}(i;\theta)$ of the kth output. θ denotes the p-dimensional parameter vector belonging to the subset $\Omega \subset \mathbb{R}^p$ which reflects the parameter constraints on the basis of prior information. When considered appropriate, the various misfits $C_k(\theta)$ for the individual outputs $y_k(\cdot)$ can be combined in one *overall* criterion, e.g. in terms of a weighted sum, or a maximum over all weighted individual misfits

$$C(\theta) := \sum_{k=1}^{q} w(k) \cdot C_k(\theta) \text{ or}$$

$$C(\theta) := \max_{k=1,\cdots,q} |w(k) \cdot C_k(\theta)| \quad (3)$$

where the weighting factors $w(k) > 0$ express the relative importance attributed to the individual outputs/misfits.

The specified misfit(s) can now be employed in two different ways in the subsequent calibration [2]:

(a) In a *direct* way, as a means to directly express the '*degree of approximation*' of the model with respect to the system.

(b) In an *indirect* way, as a basis to express the '*degree of acceptability*' or '*likelihood*' of the model in relation to the available measurements and the prior information on the model parameters. The underlying idea is that there typically will exist many parameter combinations which yield a suitable agreement with the measurements, some of them being more 'likely' than others (Beven and Binley, 1992; Klepper and Hendrix, 1994). In these approaches this is usually quantified by evaluating the '*posterior*' *parameter likelihood* distribution over the parameter-space. This likelihood is often expressed in terms of the misfit measures $C(\theta)$ (Janssen et al., 1995).

(iii) Definition of the calibration problem

The problem of selecting model(s) which render an appropriate match (misfit) to the measured data, can now be expressed in various forms, e.g.:

(a) As *optimization problem*: If it is appropriate to employ *one overall misfit* function $C(\theta)$ to quantify the mismatch between model and data, and if it is meaningful to look for the model (parameters) which minimizes this misfit, the calibration problem amounts to:

$$\min_{\theta \in \Omega}[C(\theta)] \quad (4)$$

Issues like the (non)uniqueness of the optimal model(parameters), the existence of various local minima of $C(\theta)$, the availability of suitable techniques to solve this problem etc. are of central importance in this context.

(b) As *set-identification problem*: In many situations it is inappropriate to look for one specific (optimal) model, e.g. due to the prominence of errors, uncertainties or variabilities in model and data. Then it is often more adequate to look for a *set* of model parameters rendering *acceptable* misfit(s), e.g. misfit(s) which do not exceed a prescribed level:

$$\{\theta \in \Omega | C_k(\theta) \leq \varepsilon_{\text{tol}}^{(k)};\ k = 1, \cdots, q\} \quad (5)$$

$\varepsilon_{\text{tol}}^{(k)}$ denotes the tolerance level for the kth misfit $C_k(\theta)$ ($k = 1, \cdots, q$), which e.g. refers to the kth output $y_k(\cdot)$ (see Eq. 1). The success of this approach will be largely influenced by the avail-

[2] This subdivision, albeit general, is personally biased and does not pretend to cover the rich diversity of calibration approaches completely.

ability of efficient methods for determining the above set, and for using it in further model computations.

(c) As *multi-criteria problem*: The problem of Eq. 4 can be generalized to the simultaneous minimization of *several* misfit criteria:

$$\min_{\theta \in \Omega}\left[C_1(\theta); \cdots ; C_q(\theta)\right] \quad (6)$$

Complications arise, since decrease of one criterion often leads to increase of another and vice versa. A *trade-off* between the separate misfits $C_k(\theta)$ is involved, and various ways exist to tackle this multi-criteria problem in mathematical terms (e.g. finding a *compromise* fit, determining the *Pareto optimal* set etc.). Issues like the non-uniqueness of minima and the efficiency of computational techniques play a major role in this setting (Nemhauser et al., 1989).

(d) As *explicit evaluation of the posterior likelihood*: Instead of trying to approximate the data by searching for model(s) rendering suitably small or minimal misfit(s), it can also be decided to assess the 'posterior likelihood' for each parameter $\theta \in \Omega$. To make this task manageable, attention should primarily be focused on areas in the parameter space where the posterior likelihood is large or shows considerable variations. The development of efficient techniques for this task and for incorporating the posterior uncertainty in subsequent model computations is a crucial issue.

Certainly alternative formulations can be considered, e.g. by combining some of the above mentioned aspects. E.g. Ferrier et al. (1995) employ an approach which is somehow in between the set-identification and the optimal misfit formulation, since they determine, by applying an adapted optimization procedure, a small set of parameter values which render an acceptable misfit to the data, accounting explicitly for the uncertainty or variability in the measurement data and the fixed parameters.

The ultimate choice amongst the many alternative formulations of the calibration problem should definitely depend on what is considered meaningful and feasible in the given situation. Issues like the uncertainty or variability which is present in the data and the level of accuracy which is required or attainable for the intended model application must be taken into account when making this choice, as well as the availability of adequate numerical techniques to solve the problem.

(iv) Accuracy assessment of the selected model(s)

The accuracy in the calibrated model(s), due to uncertainty in the measurements, should be determined to make meaningful inferences and predictions with the model.

2.6. The solution of the calibration problem

The subsequent step concerns the choice, development, and use of method(s) to solve the formulated problem. Before actually calibrating the model, it is recommended to reduce the number of calibrated parameters first, e.g. by performing sensitivity-, uncertainty- or identifiability analysis, in order to prevent unnecessary computational problems. Parameters whose variations have (nearly) no influence on the model outcomes can e.g. be fixed (see Walter (1987), Janssen and Heuberger (1992) for detailed information).

The solution of the calibration problem stated in subsection 2.5 is typically accomplished in a (semi)automatic fashion by applying appropriate numerical techniques. The choice of these techniques definitely depends on the nature of the formulated problem, e.g. optimization techniques (Nemhauser et al., 1989), set-identification techniques (Walter and Piet-Lahanier, 1990; Klepper and Hendrix, 1994), multi-criteria techniques (Masud and Zheng, 1989; Narula and Weistroffer, 1989), 'posterior likelihood' function evaluation techniques (Beven and Binley, 1992; Klepper and Hendrix, 1994). In choosing amongst the various alternative techniques for these problem classes, aspects like reliability, effectiveness, efficiency and computational demand of the considered methods play a role. Besides, limitations are set by the run time of the model, the number of calibrated parameters and the specific features of the misfit function (e.g. smoothness, non-linearity, multiple optima).

Moreover, an efficient method is required to

assess the accuracy of the obtained model. For the minimal misfit problem of Eq. 4, various methods have been proposed in a statistical context (confidence regions). These methods range from approximate approaches on basis of the Hessian of $C(\theta)$ (Seber and Wild, 1989), to computationally intensive statistical re-sampling methods (e.g. bootstrapping; cf. Efron and Tibshirani, 1986), which mimic the uncertainty in the data set by re-sampling the noise and re-calibrating the model. These methods are typically too computer-intensive to be used in the context of the set-identification, multi-criteria or 'posterior likelihood evaluation' problem. Moreover, their use is doubtful if the underlying statistical assumptions do not hold. Therefore a complete accuracy assessment along these lines is often omitted in these situations; it is assumed instead that the parameter set or the posterior likelihood which results from calibration, serves already as an adequate characterization of the remaining uncertainty.

It is obvious that the above-mentioned procedure(s) for model calibration and accuracy assessment must be tested decently on 'synthetic' data from the model, before they can be used with confidence on the real data.

2.7. Evaluation of the calibrated model

The adequacy of the calibrated model should finally be assessed by confronting the model with data, expert opinion, model tests etc. Three major aspects can be discerned:

1. Assessment of the ability of the model to reproduce the system behaviour:

This assessment preferably takes place on a data set which is different from the one used for calibration (*cross-validation*). Various performance measures can be employed for comparing the measurements with the model predictions (see Appendix), and many statistical techniques can be applied e.g. residual analysis, hypotheses tests, goodness-of-fit tests (Reckow et al., 1990). The results of these comparison studies should be interpreted appropriately, bearing in mind the desired accuracy which is required in reproducing the behaviour of the system at hand. Moreover these techniques should be supplemented by less data-oriented approaches, e.g. model analyses (internal testing), expert judgment, peer reviews etc. If feasible, it should also be checked whether the model incorporates the relevant processes in an adequate way, and whether the calibrated parameter values are realistic.

2. Assessment of the suitability of the model for the intended use:

Performance criteria and tests should be employed which reflect the intended use of the model (e.g. short-, medium-, or long-range prediction; prediction of episodes). The relevance of post-audit studies is obvious.

3. Assessment of the 'robustness' of the model for the data set:

If the obtained model varies little when calibrated on different data sets, under varied conditions and perhaps with different misfit functions, then one should feel rather confident that it represents important aspects of the system. Klepper and Slob (1994) propose an approach to study this robustness issue, and to obtain useful diagnostic information for improvement of the model and for delineation of its application domain.

Model quality assessment is not a 'once-and-for-all' activity leading to an absolute and definite judgement on the model's adequacy. Rather it is an ongoing *process*, which is always performed in a certain evolving (but limited) context against which the statements should be expressed and interpreted. In many situations, a thorough testing or (in)validation will be impossible due to the pragmatic reasons (lack of data, time or manpower).

2.8. Closing the loop

If the model can not be adequately assessed, then, if feasible, additional data must be obtained to enable a better judgement on its suitability. On the other hand, if the model appears to be unsatisfactory, it has to be improved. Suggestions for improvement may be obtained from the experiences gained in the previous steps and concern,

e.g. (i) additional experimentation or data collection; (ii) more adequate calibration (e.g. employing other misfit measures, alternative search techniques, new data sets); (iii) model adaptation (e.g. by applying reparametrization, parameter aggregation, model reduction, additional modelling of relevant processes).

If, finally, a satisfactory model has been obtained, it can be used for the intended purposes, e.g. scenario analysis, prediction, uncertainty analysis, decision making, design, optimization, control. Ideally, the uncertainty of the calibrated model should be taken into account for these applications to obtain reliable inferences.

3. Discussion

One of the major problems in calibrating environmental models is the tension or imbalance between the complexity of the employed model(s) and the availability of the data. This tension is often inevitable and typically finds its roots in the different nature, objectives, dynamics and costs of modelling and experimental work, the poor communication and attunement between these activities, the different temporal and spatial scales to which they address etc. A way out of this dilemma is to attune the modelling activities and the measurement activities more closely in future research, e.g. by applying simpler, tailor-made models, collecting more data on subsystem/process level, carrying out more informative experiments etc. This certainly will require a good agreement, co-operation and discipline of the people who are involved in these activities.

But, as long as this dilemma is prevalent in our applications, it sets the limits for an adequate calibration of models. Lack of informative data and/or the use of over-complex models ('surplus' content), often seriously limits the success of calibration and precludes a substantial uncertainty reduction of the parameters, let alone that an unequivocal determination of the parameters is feasible. Any decent calibration study should therefore provide adequate information on the uncertainty which is left in the parameters after calibration; moreover, this uncertainty should be appropriately accounted for in further model applications (Klepper et al., 1991; Beven and Binley, 1992; Summers et al., 1993).

These requirements inevitably have important consequences for the way in which the calibration problem should be formulated and solved, and for the techniques which should be applied. It will be desirable to develop a (semi-)automatic calibration environment with appropriate tools for (multi-criteria) optimization, set-determination, 'posterior likelihood' evaluation, sensitivity and uncertainty analysis, accuracy assessment etc., which also provides adequate means for visualisation. The computational demands of the developed techniques will, for the time being, often preclude an adequate calibration of large, computationally demanding models with many parameters. However, it is expected that the rapid developments on the hard- and software front (e.g. parallel processing) will steadily widen the range of problems which can be appropriately treated.

But these issues are definitely only part of the complete story. The heart of the matter is formed by the choices which have to be made during the various stages of the calibration process, and which have been discussed in the previous sections. We believe that these choices should be, at the very best, a mixture of sound judgement, insight, prior knowledge and the availability of suitable techniques. This decision process should be guided by the intended model use. Its final success is determined by various factors, e.g. the effectiveness of the applied techniques, the amount and quality of the measurement data, the available time, man- and computer power, expertise, financial resources. The involved choices cannot be fully automated; subjective judgement will definitely remain an inevitable and indispensable part of the whole decision process.

Appendix 1

Performance measures for 'model-data' comparison

When comparing model predictions with observed data or with results of a bench-mark model,

qualitative as well as *quantitative* techniques should be employed. While qualitative techniques are typically based on visual inspection of the results (scatterplots, pairs of time series, histograms, cumulative distribution functions etc.), quantitative techniques try to express the agreement between model and data numerically in terms of the outcomes of certain performance

Table 1
Performance measures for comparing model predictions and observations

Criterion	Symbol	Formulation
Average error	AE	$\frac{\Sigma_{i=1}^{N}(P_i - O_i)}{n} = \bar{P} - \bar{O}$
Normalized average error	NAE	$\frac{(\bar{P} - \bar{O})}{\bar{O}}$
Fractional mean bias	FB	$\frac{(\bar{P} - \bar{O})}{\frac{1}{2}(\bar{P} + \bar{O})}$
Relative mean bias	rB	$\frac{(\bar{P} - \bar{O})}{S_O}$
Fractional variance	FV	$\frac{(S_P^2 - S_O^2)}{\frac{1}{2}(S_P^2 + S_O^2)}$
Variance ratio	VR	$\frac{S_P^2}{S_O^2}$
Kolmogorov–Smirnov	KS	$\max_x(\lvert F_P(x) - F_O(x)\rvert)$
Root mean square error	RMSE	$\sqrt{\frac{\Sigma_{i=1}^{N}(P_i - O_i)^2}{N}}$
Normalized RMSE	NRMSE	$\frac{\text{RMSE}}{\bar{O}}$
Index of agreement	IoA	$1 - \frac{\Sigma_{i=1}^{N}(P_i - O_i)^2}{\Sigma_{i=1}^{N}(\lvert P_i'\rvert + \lvert O_i'\rvert)^2}$
Alternative index of agreement	AIoA	$1 - \frac{\Sigma_{i=1}^{N}\lvert P_i - O_i\rvert}{\Sigma_{i=1}^{N}(\lvert P_i'\rvert + \lvert O_i'\rvert)}$
Mean absolute error	MAE	$\frac{\Sigma_{i=1}^{N}\lvert P_i - O_i\rvert}{N}$
Normalized mean absolute error	NMAE	$\frac{\text{MAE}}{\bar{O}}$
Maximal absolute error	MaxAE	$\max_{i=1,\cdots,N}(\lvert P_i - O_i\rvert)$
Median absolute error	MedAE	$\text{median}(\lvert P_i - O_i\rvert)$
Upper quartile absol. error	UppAE	75-th percentile$(\lvert P_i - O_i\rvert)$
Ratio of scatter	RS	$\frac{\left[\Sigma_{i=1}^{N}(O_i - \bar{O})^2\right]}{\left[\Sigma_{i=1}^{N}(P_i - \bar{O})^2\right]}$
Modelling efficiency	ME	$\frac{\left[\Sigma_{i=1}^{N}(O_i - \bar{O})^2 - \Sigma_{i=1}^{N}(P_i - O_i)^2\right]}{\left[\Sigma_{i=1}^{N}(O_i - \bar{O})^2\right]}$
Regression quantities	α, β, R^2	$O_i = \alpha + \beta \cdot P_i + \varepsilon_i$

P_i and O_i denote the predicted value and observed value i; $\bar{P}$, $\bar{O}$ and S_P^2, S_O^2 are their means and variances; F_P, F_O denote the cumulative empirical distribution functions; $P_i' = P_i - \bar{O}$ and $O_i' = O_i - \bar{O}$.

measures. It is often difficult, or time-consuming, to judge the significance of these outcomes objectively. Moreover the various measures typically only highlight specific aspects of the system and the model. Therefore a judicious combination of several techniques should be employed for a thorough model assessment. The results should be used and interpreted with due care, taking into account the features of the data (quantity, quality, uncertainty, variability), and the level of accuracy which is required for the intended model application. It is obvious that insight, intuition and sound judgement play an important role in this process.

In Table 1 various common performance measures are presented. These measures quantitatively express the deviation between model predictions P_i and observations O_i (the index i refers e.g. to time instances). Their application requires that a choice is made on *which quantities* to compare and on the *temporal* and *spatial* scale on which they should be considered. It is strongly recommended to evaluate the performance measures on a data set which is different or independent from the data set used for calibrating the model (*cross-validation*). The measures in Table 1 can be subdivided into five groups:

1. The first group compares predictions and observations on an *average* level (e.g. over the whole time span): they express the bias in the average values of model predictions and observations. Different ways of normalizing the average error (AE) result in the dimensionless measures NAE, FB and rB. The displayed measures only render a rough and incomplete impression of the 'model-data' discrepancy, since averaging smooths out the dynamic features. The signs of the measures denote, albeit crude, over- or under-prediction. Outliers can have considerable effects.

2. The next three measures refer to a comparison between predictions P_i and observations O_i on a *population level*, i.e. in terms of their variances or (empirical) distributions. Application of these measures is only appropriate if the involved quantities can be considered as *random* samples from an underlying distribution reflecting e.g. their variability (e.g. spatial heterogeneity). VR and FV represent two ways of comparing the variances. The Kolmogorov–Smirnov measure (Press et al., 1986), is a non-parametric measure expressing the difference between the empirical cumulative distribution functions of model predictions and observations.

3. The subsequent group of measures compares the predictions P_i and observations O_i on an *individual level*, and tries to express the 'spread' in $P_i - O_i$:

(a) The RMSE measures $P_i - O_i$ in a *quadratic* sense; therefore it is rather *sensitive to outliers*. If the model accurately describes the noise-free data, the RMSE should be approximately equal to the standard deviation of the measurement noise. However, if model-errors are significant, it is more difficult to objectively assess the agreement between model and data on basis of the RMSE.

As an alternative, Willmott (1981) proposes a related measure (Index of Agreement, IoA) to express this agreement more directly. This dimensionless index has limits 0 (indicating no agreement) and 1 (indicating perfect agreement). *IoA* can thus be viewed as a standardized (by the variability in the predictions and observations about the observed mean) measure of the mean squared error. It should, however, be used and interpreted with due care. The IoA will e.g. invariably yield 0 if all P_i' and O_i' have opposite signs, irrespective of the sizes of the deviations. The IoA then falsely indicates that there is absolutely no agreement between model predictions and observations.

An alternative way of standardizing the RMSE is by dividing it by the mean $\overline{O}$ of the observations. This results in the NRMSE and renders a kind of coefficient of variation of the discrepancies $P_i - O_i$ around the mean $\overline{O}$.

(b) Accounting for the deviations $P_i - O_i$ in *absolute value* sense, renders the mean absolute error (MAE). This measure is less sensitive to outliers than the RMSE, and unlike the average error (AE), does not allow for compensations of positive and negative discrepancies. Dividing the MAE by the observed mean $\overline{O}$, the dimensionless measure NMAE is obtained, which is preferred over the MAE.

Similar to the index of agreement (IoA), a

different standardization can be obtained by the Alternative Index of Agreement (AIoA), which has also limits 0 and 1. This index should also be used and interpreted with care.

(c) The *maximum absolute error* (MaxAE) is most sensitive to outliers, and in fact is a *worst-case* measure. Measures which are far less sensitive to outliers are the *median* of the absolute errors (MedAE), and the upper quartile of the absolute errors (UppAE). Certainly other levels are possible (e.g. 90%); they all focus on somewhat different aspects of the 'data-to-model' agreement.

Notice that in situations where the compared quantities *vary substantially* over the considered time-span (e.g. seasonal variations, episodes with low and very high levels) the above mentioned performance measures tend to overemphasize periods where the higher values occur. To obtain a more evenly analysis of the discrepancy over the complete span of values, it is recommended to apply the measures also to the relative discrepancies $|P_i - O_i|/|O_i|$.

4. The scatter ratio and modelling efficiency measures (Loague and Green, 1991) in the fourth group relate the model predictions P_i and observations O_i to a *'nominal'* or *'bench-mark'* situation, namely to the *mean of the observations* $\overline{O}$. The scatter ratio (RS) expresses the fraction of the overall scatter in the observations which is explained by the model. Its optimal value is 1. This measure only renders information on the agreement of the overall scatter. It tells nothing about the agreement between the individual observations O_i and the model predictions P_i.

This latter aspect is of concern in the modelling efficiency measure (ME). The ME in fact quantifies the relative improvement $(\Psi_{nom} - \Psi_{model})/\Psi_{nom}$ of the employed model over the 'nominal' or 'bench-mark' situation $\overline{O}$. The feature Ψ for which the improvement is studied is the variation in the model residuals $O_i - P_i$ or $O_i - \overline{O}$. Any positive value of ME can be interpreted as improvement over the 'nominal' situation $\overline{O}$; the closer to $+1$ the better.

5. Finally *linear regression* can be used to compare the model-predictions and the observations. The deviation of the intercept term α, and the slope β, in the regression relationship $O_i = \alpha + \beta \cdot P_i + \varepsilon_i$, from 0 and 1 respectively, and the deviation of the R^2 of regression from 1 are useful indicators for the model-data agreement (Thomann, 1982; Flavelle, 1992). The regression results should, however, be carefully used and interpreted, especially if the underlying conditions are not completely fulfilled or if there is much variability in observations or model predictions (Reckow et al., 1990).

References

Beven, K.J. and Binley, A.M., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Processes, 6: 279–298.

Efron, B. and Tibshirani, R.J., 1986. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat. Sci., 1: 54–77.

Ferrier, R.C., Wright, R.F., Cosby, B.J. and Jenkins, A., 1995. Application of the MAGIC model to the Norway spruce stand at Solling, Germany. Ecol. Model., 83: 77–84.

Flavelle, P., 1992. A quantitative measure of model validation and its potential use for regulatory purposes. Adv. Water Resour., 15: 5–13.

Janssen, P.H.M. and Heuberger, P.S.C., 1992. The role of sensitivity analysis and identifiability analysis in model calibration. RIVM report nr. 723001007, RIVM, Bilthoven, the Netherlands (in Dutch).

Janssen, P.H.M., Heuberger, P.S.C. and Sanders, R., 1994. UNCSAM: a tool for automating sensitivity and uncertainty analysis. Environ. Software, 9: 1–11.

Janssen, P.H.M., Heuberger, P.S.C. and Klepper, O., 1995. Calibration and validation of process-oriented models: general considerations and guidelines. RIVM report, RIVM, Bilthoven, the Netherlands, in preparation.

Klepper, O. and Hendrix, E.M.T., 1994. A method for robust calibration of ecological models under different types of uncertainty. Ecol. Model., 74: 161–182.

Klepper, O. and Slob, W., 1994. Diagnosis of model applicability by identification of incompatible data sets illustrated on a pharmacokinetic model for dioxins in mammals. In: J. Grasman and G. van Straten (Editors), Predictability and Nonlinear Modeling in Natural Sciences and Economics. Kluwer, Dordrecht, pp. 527–540.

Klepper, O., Scholten, H. and van de Kamer, J.P.G., 1991. Prediction uncertainty in an ecological model of the Oosterschelde estuary. J. Forecast., 10: 191–209.

Loague, K. and Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. J. Contaminant Hydrol., 7: 51–73.

Masud, A.S.M. and Zheng, X., 1989. An algorithm for multiple-objective non-linear programming. J. Oper. Res. Soc., 40: 895–906.

Narula, S.C. and Weistroffer, H.R., 1989. A flexible method for nonlinear multicriteria decisionmaking problems. IEEE Trans. Syst. Man Cybern., 19: 883–887.

Nemhauser, G.L., Rinnooy Kan, A.H.G. and Todd, M.J. (Editors), 1989. Handbooks in Operations Research and Management Science. Vol. 1: Optimization. Elsevier, Amsterdam.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., 1986. Numerical Recipes: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Reckow, K.H., Clements, J.T. and Dodd, R.C., 1990. Statistical evaluation of mechanistic water-quality models. J. Environ. Eng., 116: 250–268.

Seber, G.A.F. and Wild, C.J., 1989. Nonlinear Regression. John Wiley, New York.

Summers, J.K., Wilson, H.T. and Kou, J., 1993. A method for quantifying the prediction uncertainties associated with water quality models. Ecol. Model., 65: 161–176.

Thomann, R.V., 1982. Verification of water quality models. J. Environ. Eng. Div., 108: 923–940.

Tiktak, A. and Van Grinsven, J.J.M., 1995. Review of sixteen forest–soil–atmosphere models. Ecol. Model., 83: 35–53.

Van Grinsven, J.J.M., Driscoll, C.T. and Tiktak, A. (Editors), 1995. Workshop on Comparison of Forest–Soil–Atmosphere Models. Preface. Ecol. Model., 83: 1–6.

Walter, E., 1987. Identifiability of Parametric Models. Pergamon Press, Oxford.

Walter, E. and Piet-Lahanier, H., 1990. Estimation of parameter bounds from bounded-error data: a survey. Math. Comput. Simul., 32: 449–468.

Walter, E. and Pronzato, L., 1990. Qualitative and quantitative experiment design for phenomenological models – a survey. Automatica, 26: 195–213.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr., 2: 184–194.