Ecological Modelling 280 (2014) 117–128

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Merging validation and evaluation of ecological models to 'evaludation': A review of terminology and a practical approach

Jacqueline Augusiak[a,*], Paul J. Van den Brink[a,b], Volker Grimm[c,d,e]

[a] Wageningen University, Aquatic Ecology and Water Quality Management Group, Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[b] Alterra, Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[c] UFZ, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany
[d] Institute for Biochemistry and Biology, University of Potsdam, Maulbeerallee 2, 14469 Potsdam, Germany
[e] German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany

ARTICLE INFO

Article history:
Available online 29 November 2013

Keywords:
Model validation
Terminology
Decision support
Documentation
Ecological models
Risk assessment

ABSTRACT

Confusion about model validation is one of the main challenges in using ecological models for decision support, such as the regulation of pesticides. Decision makers need to know whether a model is a sufficiently good representation of its real counterpart and what criteria can be used to answer this question. Unclear terminology is one of the main obstacles to a good understanding of what model validation is, how it works, and what it can deliver. Therefore, we performed a literature review and derived a standard set of terms. 'Validation' was identified as a catch-all term, which is thus useless for any practical purpose. We introduce the term 'evaludation', a fusion of 'evaluation' and 'validation', to describe the entire process of assessing a model's quality and reliability. Considering the iterative nature of model development, the modelling cycle, we identified six essential elements of evaludation: (i) 'data evaluation' for scrutinising the quality of numerical and qualitative data used for model development and testing; (ii) 'conceptual model evaluation' for examining the simplifying assumptions underlying a model's design; (iii) 'implementation verification' for testing the model's implementation in equations and as a computer programme; (iv) 'model output verification' for comparing model output to data and patterns that guided model design and were possibly used for calibration; (v) 'model analysis' for exploring the model's sensitivity to changes in parameters and process formulations to make sure that the mechanistic basis of main behaviours of the model has been well understood; and (vi) 'model output corroboration' for comparing model output to new data and patterns that were not used for model development and parameterisation. Currently, most decision makers require 'validating' a model by testing its predictions with new experiments or data. Despite being desirable, this is neither sufficient nor necessary for a model to be useful for decision support. We believe that the proposed set of terms and its relation to the modelling cycle can help to make quality assessments and reality checks of ecological models more comprehensive and transparent.

> "I assert that whenever a dispute has raged for any length of time, especially in philosophy, there was, at the bottom of it, never a problem about mere words, but always a genuine problem about things."
>
> I. Kant (1786)

1. Introduction

Ecological models are increasingly used and needed for supporting environmental decision-making (Schmolke et al., 2010a). Often they are the only way to take into account the relevant spatial and temporal scales and the multitude of processes characteristic to ecological systems. Corresponding experiments can be impossible, and insights from descriptive studies do not necessarily provide enough mechanistic understanding to predict responses of ecological systems to new conditions.

Since models are simplified representations of real systems, a key challenge is, however, to show that the models are realistic enough to meet their intended purpose (Rykiel, 1996). Before we can transfer inferences from model results to the real world,

* Corresponding author at: Department of Aquatic Ecology and Water Quality Management, Wageningen UR (University and Research Centre), P.O. Box 47, 6700 AA Wageningen, The Netherlands. Tel.: +31 317 48 59 58; fax: +31 317 41 90 00.
E-mail address: jacquelinwe.augusiak@wur.nl (J. Augusiak).

0304-3800/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2013.11.009

we have to demonstrate that the model reproduces observations for the right reasons, not just because it has been tweaked via calibration to do the right thing. If models are in fact used without being carefully checked for their validity, they might lead to erroneous decisions. Pilkey and Pilkey-Jarvis (2007) call inappropriate models "useless arithmetics" and find that "these types of applied models are frequently detached from reality – built on oversimplified and unrealistic assumptions about natural processes". Thus, scepticism with regard to using ecological models to support environmental decisions is a healthy attitude. It is up to the modellers to provide evidence and indicators that their model is realistic enough.

An example field of decision making, where scepticism regarding ecological models so far has prevented the use of models, is ecological risk assessment of chemicals, in particular pesticides (Forbes et al., 2009, 2010; Thorbek et al., 2010). Ecological risk assessments are required for pesticides to minimise potentially negative impacts on non-target flora and fauna and, thus, on ecosystems in general. Regulatory decisions on whether or not a certain pesticide can be used are, at least in the lower tiers of the risk assessment, based on highly standardised schemes. They focus on effects on individuals of a set of standard species, observed under standardised conditions in the laboratory.

Mechanistic effect models have long been identified as potentially useful tools to extrapolate the limited findings from standard tests to more realistic conditions such as fluctuating exposure profiles, higher levels of biological organisation, and larger temporal and spatial scales, thus making risk assessments ecologically more relevant (Forbes et al., 2009, 2010; Galic et al., 2010; Pastorok, 2002; Thorbek et al., 2010). Mechanistic effect models comprise ecological and organism-level effect models. They are referred to as 'mechanistic' to clearly separate them from descriptive, or statistical, models, and as 'effect models' to separate them from physico-chemical models describing the fate and exposure of chemicals in the environment.

Despite the high potential of mechanistic effect models to improve the ecological realism of pesticide risk assessment, so far they have not often been used or accepted in regulatory risk assessments. A major obstacle is the doubt as to whether a given model represents the real world sufficiently well, which is reinforced by a lack of clear criteria for assessing a model's realism. Additionally, a comprehensive model assessment is often hampered by the ambiguous application of terminology within and between involved stakeholder groups. Academics, industry, as well as regulators each possess a different set of vocabulary, knowledge, and interests (Hunka et al., 2013; Jakeman et al., 2006), which interferes with both a more productive advancement and communication of methods, and with actually using models to support decision making.

Terminology regarding model assessment has in general proven to be a particular source of confusion (Oreskes et al., 1994a; Rykiel, 1996). To describe general tasks of quality assurance throughout a model's development and application, academics often use the term 'validation' more or less intuitively, due to a lack of a clear and unambiguous definition. Yet, academics are at odds with each other as to what 'validation' should mean in a modelling context, to which degree model validation would be generally feasible, and which methods or criteria should be applied to assess the compliance of a given model with its real counterpart.

This issue has been debated in the context of ecological modelling for the past 50 years and still no commonly accepted language and methodology could be agreed upon (see references in Rykiel, 1996). This makes it very hard to clearly assess and communicate the credibility of models, which in turn makes it difficult, if not impossible, for decision makers, who are usually not trained in assessing whether a model is good enough, to let models influence their decisions. Other domains, e.g. hydrology, economics, meteorology, or environmental engineering, where mechanistic models are being used as well to support decision making, are facing similar problems (Ferson et al., 2008; Gass, 1983; Hodges and Dewar, 1992; Oriade and Dillon, 1997; Refsgaard et al., 2005).

In this article, we review and evaluate the literature concerning the terminology and methodology regarding model validation. We focus predominantly on literature related to ecological models but draw relevant lessons from other scientific fields with relations to regulatory frameworks to provide a pragmatic solution to the above-mentioned challenges. According to the most dominant trends that we could identify, we will propose a common vocabulary for the evaluation of applied ecological models. This can for example assist the risk assessment process by introducing a structured system of language. In particular, we will suggest the new, artificial term, 'evaludation', which is a merger of 'evaluation' and 'validation'.

Evaludation consists of several elements, or steps, that correspond to the different stages of iterative model development forming the 'modelling cycle' (Grimm and Railsback, 2005). They thus serve as the main structuring elements for the suggested terminological system. The modelling cycle consists of the following elements (see also Section 3): formulation of the questions to be addressed; assembly of hypotheses that constitute our conceptual model of the system in question; choice of model structure, i.e. choice and representation of entities, state variables, and processes; implementation of the model via equations and/or a computer programme; model analysis; and communication of model output.

Based on this approach, we will demonstrate that validation is not a binary criterion that is determined once a model's development has been finished. Rather, overall model credibility arises gradually throughout the entire modelling cycle.

2. Terminology and concepts

Mechanistic modelling simplifies real-world processes to understand driving mechanisms well enough so that forecasts of a system's response to certain conditions become feasible. This simplification implies the risk that not all relevant factors were captured or that relevant data are missing. Investigating these deficiencies in detail is not always feasible due to monetary, time, or other constraints. For this and other reasons, models inherently possess a level of uncertainty.

To reduce the likelihood of a flawed decision due to an uncertain, simplified representation, decision makers usually demand that a model should be validated. Typically, they ask for a comparison of model output with new empirical data to determine whether possible discrepancies render the model too unrealistic for use. Many scientists argue (correctly in our opinion) on the contrary that this approach to validation is too limited for at least three reasons. First, agreement between modelled and empirical data does not necessarily imply that a model is 'correct', but could also result from a combination of 'wrong' input parameters and process representations (Oreskes and Belitz, 2001). Second, this kind of direct validation often is impossible to achieve because such data do not exist, which is rather the rule than the exception in ecological and environmental systems. In fact, this is the reason why models are needed for these systems in the first place. Third, the genuine meaning of the word "validation" does not fully match with the uses of the term in ecological modelling and is accompanied by philosophical discourses about its legitimate usage.

It seems obvious that validation should not be mistaken with 'truth', although the term certainly implies a strong sense of legitimation (Oreskes et al., 1994b; Rykiel, 1996). Decision makers would appreciate having some form of quantifiable certification

that increases confidence in a model's appropriateness for application; or, as a risk assessor of pesticides once asked: 'Isn't there a kind of R-square to assess a model's validity?' This desire is understandable but reflects a lack of knowledge and understanding of how modelling is usually done and should be used, i.e. the modelling cycle. If validation would be defined to depend on only one or a few expressions of error, major flaws in the model structure could still mislead a decision.

However, decision makers cannot be blamed for lack of understanding of the above points. The roots of the controversy around validation reach much deeper and keep confusing modellers as well. One of the main reasons for disagreements concerning semantics and methodological approaches lies in the philosophical views on how science is performed and, in turn, what validation means in science in general.

Logical empiricism, or positivism, dominated scientific conduct between the middle of the 19th to middle of the 20th century. This school of thinking favoured inductive inferences building from singular observations and/or experiments to universal statements such as hypotheses or theories (Barlas and Carpenter, 1990; Refsgaard and Henriksen, 2004). The proposed hypotheses or theories are eventually to be tested in experiments that are designed to confirm or refute the general statement at hand. From a model validation perspective, such an approach would render the process of validation formal and algorithmic. Under such premises models would be assumed to be objective and absolute representations of the modelled system, such that they could only be either true or false. This perspective seems to be taken by many non-modellers.

Critics of this approach (Kuhn, 1962; Popper, 1959) argue that theories can be only falsified and never verified. Typically, they follow a more deductive approach towards science, where inferences are drawn from universal statements, such as theories or hypotheses, to more specified statements. Conclusions are derived logically from several statements, and predictions of empirical patterns must be formulated as deductive consequences from theories or hypotheses. If those conclusions and predictions can be shown to be true, the overarching hypothesis is deemed corroborated or confirmed (Popper, 1959). The larger the wealth of confirming observations the more credible the respective hypothesis is deemed to be. However, no matter the number of confirmations, there is always a chance that an observation can be explained by more than one theory. Furthermore, a single falsifying incident is sufficient to reject the correctness of the scrutinised hypothesis; for example, seeing a single black swan falsifies the theory that all swans are white, which hitherto might have been 'verified' by observing a million white swans (Taleb, 2010).

From this rationalist, deductive perspective, validation becomes a less formal process since a valid model is assumed to be one of several probable representations of a real-world process. Barlas and Carpenter (1990) as well as Oreskes and Belitz (2001) and Oreskes et al. (1994a) argue that one such representation may be preferable over other alternatives, but that no model could claim absolute objectivity as each is also subject to the modeller's subjectivity, view and understanding of the world, and proneness to mistakes. Thus, models are neither true nor false but lie on a continuum of usefulness for which credibility can be built up only gradually (Barlas and Carpenter, 1990; Rykiel, 1996). The question is transferred from whether or not a model holds true to how likely it is to be sufficiently true in the light of accumulated, existing evidence and the model's purpose.

Ecological modellers have discussed model validation since the 1960s. The development of ideas and methodological concepts for validating ecological models underwent several turns since then. Levins stated in 1966 that validation of a model ought to be the generation of testable hypotheses rather than finding that a model is 'true' but he left out any quantifiable measures of assessment. On the other hand, Goodall (1972) suggested that the degree of agreement between a model and its real counterpart would be an appropriate measure, which corresponds to today's most common understanding of validation. He furthermore suggested that model input data and the field data used for comparison should be statistically independent. This line was followed in 1977 by Overton. He viewed modelling as an iterative process of refinements and calibration until the output met specified performance criteria, that is, the model was capable of mimicking a predefined data set. He acknowledged that validation in the sense of absolute truth was not possible, as this approach does not necessarily allow identifying the most appropriate model from a set of candidate models.

Early ideas of evaluating a model according to its purpose were discussed by Holling (1966), May (1973) and Caswell (1976). Caswell distinguished between models used in an engineer-like fashion as predictive tools and models used as tools for scrutinising and testing scientific theory. He furthermore introduced the term 'corroborate' for the latter class of models and 'validate' for the first. He explained this choice by comparing the testing of scientific models with hypothesis testing in which a statement might be scientifically corroborated or refuted, whilst validation, as defined by Goodall (1972), would resemble a form of engineering performance testing. Caswell furthermore claimed that the two different uses would not have to be mutually exclusive. A model could well be predictively valid and be scientifically refuted at the same time. A famous example of such a model is the Ptolemaic model of the solar system, which makes precise predictions of the planets' visible trajectories, but is based on an incorrect view of the structure of the solar system. Understandably, such combinations should preferably be avoided if models are used to predict responses to changes in the environment.

Holling (1978) and Shugart (1984) both shared the view that models resemble complex hypotheses and that validation therefore is impossible to achieve and that only their falsification is possible. Holling went so far as to consider the request for validated models to be inappropriate. He argued that invalidation could be regarded as a tool to establish the limits of a model's credibility to establish a sufficient degree of belief in the model to justify its application. Shugart built on ideas of Goodall (1972) and Overton (1977) and defined model validation as the application of procedures to test a model's agreement with a set of data that is independent from that used for calibrating and parameterising the respective model. Complementary, he defined verification as a test of whether a model can be made correspond with a given data set.

Rykiel (1996) sought a technical and more pragmatic understanding of the term validation. He pointed out that ecological models usually aim to combine theory and practice and that this duality leads to conflicts when model validation is sought to combine hypothesis testing and engineering practice, a conflict which remains unresolved until today. His pragmatism is in line with Beven's (2002) suggestion to extend the philosophical context in which environmental models are viewed. Beven suggests that one should explicitly account for underlying uncertainties and promotes Von Bertalanffy's idea of 'equifinality', e.g. that more than one model can be reliably applied for a given situation. He considers it an option to compare different possible models (different structural models or parameter combinations) and their closeness to predefined performance criteria to gain a more complete understanding of the influence of alternative considerations. The range of plausible models can thus be limited over time as knowledge about the system grows. In contrast, Oreskes et al. (1994a) argue that equifinality would rather pose a source of doubt than help increasing trust in models. Nevertheless, both, Beven and Oreskes et al., share the view that absolute validation of environmental models is impossible to achieve, as environmental systems are open, which complicates strict deductive thinking.

Table 1
Synonyms and definitions used in model testing and validation literature.

Definition	Term	Source
Entire process of forming the decision whether and when a model is suitable to meet its intended purpose by building confidence in model applications and increasing the understanding of model strengths and limitations.	Corroboration	Popper (1959), US-EPA (2009)
	Evaluation	Bart (1995), Borenstein (1998), Committee on Models in the Regulatory Decision Process (2007), Hodges and Dewar (1992), Jakeman et al. (2006), Loizou et al. (2008), Schmolke et al. (2010b)
	Testing	Goodall (1972)
	Validation	Bacsi and Zemankovics (1995), Barlas (1996), Borenstein (1998), Gass (1983), Hodges (1991), Kirchner et al. (1996), Landry et al. (1983), Sargent (2005)
	Verification	Arthur et al. (1999)
Tests to ensure that the 'right model' is being built.	Validation	Aumann (2007), Ormerod and Rosewell (2009)
	Verification	Borenstein (1998)
Assuring that the computer programme and implementation of the conceptual model are correct.	Verification	Aumann (2007), Barlas and Carpenter (1990), Gass (1983), Hodges (1991), Loizou et al. (2008), Oriade and Dillon (1997), Ormerod and Rosewell (2009), Refsgaard and Henriksen (2004), Rykiel (1996), Sargent (2005), Schmolke et al. (2010a), US-EPA (2009), Van Waveren et al. (1999)
Assessment of the implications of errors made in design and implementation for the model output and whether the output behaviour exhibits the required accuracy with regard to the model's intended purpose. The assessment is mainly built on comparing model output to data that were preferably not used for model development.	Validation	Arthur et al. (1999), Beck et al. (1997), Ferson (1996), Gass (1983), Oriade and Dillon (1997), Ormerod and Rosewell (2009), Refsgaard and Henriksen (2004), Rykiel (1996), Van Waveren et al. (1999), Wang and Luttik (2012)
	Verification	Jakeman et al. (2006)
	Substantiation	Borenstein (1998)

Botkin (1993) and Oreskes et al. (1994a,b) focused particularly on the semantics of validation and verification. Their concerns were that the usage of these terms would not agree with their original definitions, which, according to the authors' understandings, would follow the deductive school of thinking. Oreskes et al. (1994b) argued that the slight differences in meaning of various alternative terms for validation (namely corroboration, confirmation, verification) matter and that current usage of these terms would not follow a common school of thinking.

The term validation has not been used consistently in the literature. Different authors used different definitions depending on their view of the matter; others had similar meanings in mind but used different synonyms. The same holds for other terms commonly used in relation to evaluating the different stages in the modelling cycle. While Popper (1959) used the term corroboration to describe the process of evaluating a model as a whole, Goodall (1972) named the same process testing. Nowadays, verification usually describes the process of checking a computer code for mistakes (e.g. Rykiel, 1996; Sargent, 2005; Van Waveren et al., 1999). On the other hand, Arthur et al. (1999) described the process of model evaluation with this term, while Borenstein (1998) used the same word for testing whether the correct model has been built, not if it had been built correctly. In contrast, Jakeman et al. (2006) understood verification as a step in which the accurate fit of model results is tested, a step that Borenstein (1998) called substantiation, and a majority of publications validation (e.g. Beck et al., 1997; Gass, 1983; Rykiel, 1996; Van Waveren et al., 1999). These are just a few examples where different authors introduced differing connotations of particular terms. Table 1 gives an overview of the confusing usage of terms and synonyms that can be found in the literature. The term 'validation' has been given virtually any possible meaning in this context (Table 1). A reason for this might be that this term seemingly prejudices expectations of the outcome towards the positive (i.e. the model is valid or the quality is assured), which is one of the major criticisms surrounding the term. Yet, or maybe because of this positive reassurance, the term persistently remains and returns regularly in discussions.

To conclude, there is little agreement on terms and underlying notions in the literature, with the one exception that it has repeatedly been pointed out that the evaluation of a model should depend on its purpose (e.g. Hoover and Perry, 1989; Mankin et al., 1977; Mayer and Butler, 1993; Rykiel, 1996, 1984).

3. Proposed terminology based on the modelling cycle

Many of the discussions listed above focus on general aspects of how validation should be defined, what it should comprise, or how it should be done. Most of them, however, do not consider structured approaches. Schmolke et al. (2010a) demonstrated that a structured documentation of the subsequent modelling steps already would support a more comprehensive assessment of a model. They proposed a generic structure for documenting modelling which is built on the structure of the modelling cycle. We propose a similarly structured approach towards model evaluation.

The central elements in model development are shown in Fig. 1. Typically, basic or applied questions about an environmental system lead to a conceptualisation of the underlying processes. Once a conceptual model has been derived that seems to account for the most relevant processes to answer the question at hand, the conceptual model is translated into a computerised model. Proceeding from the conceptual to the computerised model works in two steps. First, the conceptual model has to be made quantitative and operational so that it can be run on computers (note that we here also refer to mathematically formulated models, which are numerically solved on computers, as computerised models). This step comprises the definition of entities and state variables for characterising the state of the model system; mathematical or algorithmic submodels that represent the processes included in the model; and a schedule of the model's processes. We call this the 'written formulation' of a model. Second, the written formulation has to be translated into a programme that can be run on computers, referred to as the 'implementation of the model'.

At all stages of the cycle, lack of knowledge and good quality data, and human imperfection, unavoidably induce uncertainty. The level of uncertainty can be reduced by applying a standardised evaluation scheme similar to quality assessment protocols (Refsgaard et al., 2005). For such a scheme to be practicable, the different elements in the modelling cycle should be examined separately. To distinguish between these, and to reduce currently prevailing misunderstandings between involved stakeholder groups, we follow the pragmatic recommendations from

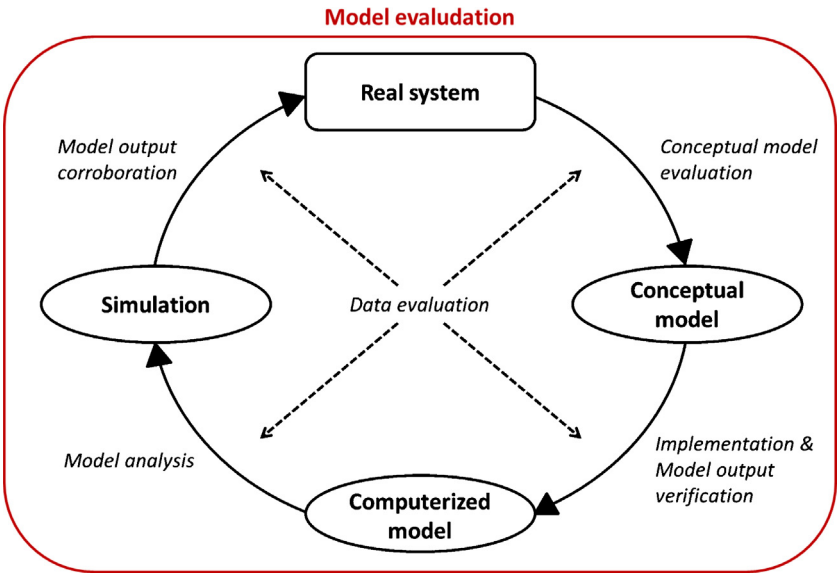

Fig. 1. A simplified representation of the modelling cycle, consisting of the four main steps of model development and their corresponding elements of evaludation. The terms in italics comprise the terminology that we suggest to organise and communicate model evaludation. The four elements of model development were, in the context of model quality assurance, suggested by Refsgaard and Henriksen (2004) and Schlesinger (1979). Their definition of quality assurance corresponds to what we here refer to as 'evaludation' and what so far in ecological modelling usually has been implicitly referred to as evaluation and/or validation.

different scientific fields to split model evaluation into subparts (Fig. 1; Barlas, 1996; Refsgaard and Henriksen, 2004; Rykiel, 1996; Sargent, 2005). The general logical order calls first for a test of the appropriateness of the chosen model structure before testing accuracy of model output.

To combine the imprecise but important term 'validation', and building on its implied meaning for assessing a model's quality, with the more neutral and complementing connotations of 'evaluation', we introduce the new, artificial term 'evaludation'. We define 'evaludation' as 'the entire process of establishing model quality and credibility throughout all stages of model development, analysis, and application'. We suggest this term for several reasons. Firstly, we aim to avoid using 'validation' itself while keeping it still visible. Secondly, we aim to link the understandable request for validity assurance with the more neutral, multi-step process of evaluating the quality of not only the model output but also all other relevant aspects of modelling, which can affect model credibility and validity. Furthermore, a new term implies that it is more likely to be specified when it is used, which avoids misunderstandings and emphasises the multi-criteria character of model assessment.

Evaludation consists of six elements, which are necessary to communicate which uncertainties have to be faced at the different stages of the modelling cycle, which evaluation tools are applied for these elements, and which measures should or could be taken to reduce uncertainties. In the following, we will define the different elements that constitute evaludation and discuss why they are important. Short discussions of possible approaches to tackle the different evaludation steps will be discussed in Section 4.

3.1. Data evaluation

'Data evaluation' is defined as the critical assessment of the quality of numerical and qualitative data used to parameterise the model, both directly and inversely via calibration, and of the observed patterns that were used to design overall model structure. Thus, with data we here not only refer to numerical data, as in some data sheets or spread sheet tables, but also qualitative data, i.e. expert knowledge for which no hard numbers exist. Computer models can take into account such knowledge in the form of probabilistic if-then rules: if a certain state is given, various things may happen with certain probabilities. The term 'data' also refers to patterns (Grimm and Railsback, 2012) or, in economists' terminology, 'stylised facts', which are general trends and signals in data, observations, and empirical knowledge.

To illustrate these different types of data, consider a census time series of a population of small mammals in a certain area over 30 years. The numerical data are the set of abundances, which are uncertain in themselves because abundance could not be determined directly. Expert knowledge could exist about individual behaviour, for example territoriality, and how it changes, in broad terms, in response to changes in population density. Another pattern could occur in a series of years with bad weather, during which variability of animal abundance differs from that in series of years with good weather. Such kinds of data are important for model development, although they are statistically uncertain.

The appropriateness, accuracy, and availability of data used throughout model development are a major source of uncertainty and often a reason for failed attempts to validate a model (Sargent, 2005). Data are needed for the design of a conceptual model, to deduce relevant theories and derive mathematical and logical relationships that represent the modelled system sufficiently well to fulfil the model's stated purpose. Furthermore, data are needed to fully parameterise and calibrate a model. Finally, data, which were preferably not used for model development and calibration, are needed to test the model's underlying operational assumptions. Without such independent data, confidence in a model can be hard to establish (Rykiel, 1996; Sargent, 2005). However, it should be kept in mind that for ecological systems, independent numerical data often do not exist or even cannot exist. Instead, using additional qualitative data or general patterns that were not considered or known during the model development can and should be used ('pattern-oriented modelling': e.g. Grimm and Railsback, 2012; Grimm et al., 2005).

The quality of available numerical data can be corrupted by measurement errors (e.g. by quality of instruments and frequency of their calibration, data logging, etc.), flawed experimental design (e.g. choice of sampling site, small sampling sizes, etc.), and natural heterogeneity and stochastic variability inherent to environmental systems (Gass, 1983; Wang and Luttik, 2012). Likewise, expert knowledge and the detection of patterns are notoriously prone to

bias and therefore must be treated with particular caution. Another aspect to watch carefully concerns the extrapolation of data from one situation to another. This can include the usage of laboratory data to estimate effects in the field, as well as the usage of data from another climate zone or related, and not the actually studied species.

Data evaluation is needed to ascertain a high level of quality, a point laid out in several quality assurance and control protocols (Refsgaard et al., 2005; US-EPA, 2009; Van Waveren et al., 1999). This is part of the reason why one cannot simply assume that data yield the best testing conditions for a model's structure or output, as data themselves do not always represent the real system sufficiently well. Additionally, experimental data are only gathered during a particular period or in a particular area and therefore represent only one of the many states of the ecosystem (Fagerstrom, 1987; Topping et al., 2012). A model cannot be expected to provide more accuracy and clarity than what has been used to develop it in the first place.

3.2. Conceptual model evaluation

'Conceptual model evaluation' is defined as the critical assessment of the simplifying assumptions underlying a model's design. A conceptual model is our verbal or graphical model of the system of interest with regard to a certain question. As with any element in the modelling cycle, the conceptual model is very simple at first and subsequently develops gradually. Evaluating the conceptual model means to explicitly list, discuss, and justify its most important simplifying assumptions. Typically, assumptions include the choice of spatial and temporal scales; the choice of entities and processes to be represented in the model; considerations concerning stochasticity and heterogeneity; considerations of local versus global interactions; representation of environmental drivers; etc. Furthermore, conceptual model evaluation includes the assessment of whether the structure, underlying theories, concepts, assumptions, and causal relationships are reasonable to form a logically consistent model.

Conceptual model validity is mostly affected by a modeller's subjectivity, incomplete understanding and knowledge of underlying problem entities, and the quality of available data. Different modellers may make different decisions about the kind and form of processes to include in a model. Boesten (2000) found in a comparative study that despite equal starting conditions, i.e. having the same data sets and study objectives defined, different modellers obtained significantly different model results. He identified the expert judgement involved in establishing the process relationships as one of the major causes for this variation. Additionally, incomplete knowledge about the factors that control behavioural aspects of the modelled system, either due to a modeller's lack of awareness of relevant studies or a lack thereof, as well as limitations arising from simplifying assumptions, need to be considered and justified. As another guiding principle, Occam's razor should be applied to ensure that the chosen model complexity does not introduce avoidable uncertainty (Beck et al., 1997; Clark, 2004; Jakeman et al., 2006).

3.3. Implementation verification

We define 'implementation verification' as the critical assessment of (1) whether the computer code for implementing the model has been thoroughly tested for programming errors and (2) whether the implemented model performs as indicated by the model description. This element of evaludation is hence primarily concerned with checking the computer code for errors, bugs, and oversights. However, even an error-free programme code might not actually implement the model as intended or described, which can be due to ambiguities in the model description or due to misinterpretations of ready-to-use procedures in the employed software platform (for an example of the latter, see Grimm and Railsback (2012, Chapter 5)).

Although implementation verification mainly focuses on technical aspects of a modelling exercise, it is also essential for assessing whether a model is of sufficient realism and quality. Tests of independent model predictions (see below) might look promising but without a thorough evaluation of the implementation procedure, the risk could still be high that the model leads to wrong conclusions because the model might not work as we think it does.

3.4. Model output verification

'Model output verification' is defined as the critical assessment of (1) how well model output matches observations and (2) to what degree calibration and effects of environmental drivers were involved in obtaining good fits of model output and data.

Model development always includes a judgement of model output according to observed data and patterns and some criteria of similarity. After all, the purpose of models is to be 'representations' of real systems, even if this representation has to be much simpler than reality. However, just considering 'predicted vs. observed' figures is not sufficient either. Model users need to know how much calibration was involved to make the model fulfil verification quality criteria. The more parameters had to be fine-tuned via calibration, the higher the risk that successful verification was enforced by unrealistic parameter combinations, i.e. by a combination of factors which does not occur like this in the real system. Likewise, a good match of model output and observations might have been imposed by representing strong environmental drivers, for example weather, chemical disturbances, or predation risk so that model output rather reflects the dynamics of the drivers and not necessarily a realistic representation of the system's internal organisation.

Furthermore, when comparing model output to data it is critical to avoid comparing apples with oranges: environmental conditions, initial states of the model world, and data sampling protocols (for example the timing of sampling a population) implemented in the model should match those underlying the data as close as possible (Zurell et al., 2012). For complex models, this can be a major task (Topping et al., 2012).

In general, the task of this element of evaludation is to demonstrate that the individuals and populations represented in the model respond to habitat features and environmental conditions in a way that is sufficiently similar to their real counterparts. What can be considered 'sufficient' cannot be defined from the outset and also depends on other elements of evaludation, on the overall understanding and experience with managing the system in question, and on whether the model is supposed to deliver absolute or relative predictions. In the latter case, a qualitative agreement of model output and data might already be considered sufficient.

3.5. Model analysis

We define 'model analysis' here as the assessment of (1) how sensitive model output is to changes in model parameters (sensitivity analysis) and (2) how well the emergence of model output has been understood. Testing model sensitivity is essential since a good match of model output and data might also be the result of fine-tuning several parameters. The match might vanish as soon as one or more or the calibrated parameters are changed. Model evaluations, which do not include sensitivity analyses, are thus too limited.

Sensitivity analyses identify subsets of parameters that have strong effects on the model outputs. Since parameters represent

the relative contribution of certain processes and feedbacks, we thereby learn which processes are most important for further considerations, which is an important first step to understanding which factors are most important in explaining model behaviour.

Understanding model behaviour is needed to avoid using a model as a black box. If we understand why and how a model produces certain outputs, we can, if the model is well evaludated, cautiously transfer this understanding to the real world, which would often be more important for supporting decisions than any kind of specific numeric model output. Evaludation thus implies that the modeller has tried several possibilities, and documented them, to understand and explain how model output emerged.

3.6. Model output corroboration

'Model output corroboration' is here defined as the comparison of model predictions with independent data and patterns that were not used, and preferably not even known, while the model was developed, parameterised, and verified. The emphasis on new, independent data is important because with data known and used during model development, modelling will often end up with a model reproducing these data. This implies the unavoidable risk that the model has been 'tweaked' to do the right thing for the wrong reasons.

Still, for models of complex systems, making a model match known observations can be difficult and it is a myth to believe that a model could reproduce or forecast any data with just enough model parameters. Thus, model output verification can already indicate whether the internal organisation of the real system has been captured sufficiently well, in particular when verification comprised not only one data set or pattern, but multiple ones.

However, even when using several patterns simultaneously to verify if a model is working correctly (Grimm and Railsback, 2012), there is still a risk that the respective model might have been manipulated too much to produce the right behaviour for the wrong reasons. Multiple patterns are sometimes not independent from each other and thus do not necessarily reduce this risk (Latombe et al., 2011). In contrast, with new data and patterns, this risk is being eliminated. Not knowing these data or information and patterns makes artificial imposing of rules and tweaking impossible.

One form of new data consists of results from new and specifically designed experiments and field studies. This corresponds to one of the most common interpretations of validation: a model is considered 'valid' if it made predictions that were confirmed by subsequent experiments. However, as mentioned earlier, such experiments or field studies are usually unfeasible for most ecological systems. Thus, for model output corroboration we usually have to resort to comparing model predictions with data and patterns that already exist, but have not been known or used by the modeller. The guide to finding such data and patterns can be the model itself. Does it predict any striking features, regularities, or patterns, which are robust and seem to emerge from the interaction of the key processes in the model? If so, can we find corresponding data or patterns in the literature, existing data bases, or can we confirm them via expert judgement ("This is exactly what I have observed.")?

One example for this approach is a model of natural beech forests in Central Europe (Neuert et al., 2001; Rademacher et al., 2004). In this model, canopy trees were represented as individuals with certain ages and crown sizes. However, information concerning age and size were never used during model development and calibration. Rather, local stand structure was assigned to three so-called 'developmental stages', which take into account the leaf cover in four different height classes. Temporal and spatial patterns regarding the developmental stages, which were know while the model was developed, were used for model output verification, and the model was published and used for its original purpose (Neuert et al., 2001; Rademacher et al., 2004), which was the estimation of the sizes of forest reserves that would be needed to enable natural spatio-temporal forest dynamics.

In a follow-up project, the age structure of model canopy trees was analysed as well as the spatial distribution of very large and old trees (Rademacher et al., 2001). The two patterns found in this analysis were that neighbouring canopy trees on average differed in age by 60 years and that 80% of all trees older than 300 years had a tree of similar size within a distance of less than 40 m. This pattern was confirmed after re-analysing the existing literature (Rademacher et al., 2001).

In the current literature, model output corroboration based on patterns identified in the model is the exception rather than the rule. This might partly be due to limited resources, but the main reason seems to be that the term 'prediction' is often used in a very broad sense, which blurs the distinction between verification and corroboration of model output. The term 'prediction' should be used only for new, independent – secondary – predictions that forecast something new, either results to be obtained in the future or patterns to be detected in the existing data and knowledge. Data used for verification should then be referred to as 'model output', as no prediction is involved because the data and patterns 'predicted' were already used in the model development.

A clearer distinction between model output verification and corroboration could actually lead to more systematic attempts of model output corroboration. And it should be noted that a model can still be considered realistic and fit enough to meet its purpose, even if corroboration was not possible due to lack of resources or data.

4. Evaludation: planning and approaches

Considering the vast complexity of environmental issues that can be addressed with modelling approaches and the diverse set of modelling concepts, it is not possible to establish a detailed, fool-proof protocol for evaludating a model or declaring whether or not it can be deemed fit for application. Nevertheless, the systematisation of checking the different building blocks of a model throughout its lifecycle and evolution ensures reduced uncertainties, and, maybe more importantly, an easier communication of the capabilities and limitations of a model so that decision-makers feel more confident about using it. Some general concepts and considerations can help to add more structure to the task of model evaludation. Especially ideas derived from general quality assessment and control frameworks, as well as experiences from common practice, can help to establish a more consistent procedure. Some discussions on Good Modelling Practice and existing regulatory protocols in a number of fields have already succeeded in establishing a first, rough guidance (Rykiel, 1996; US-EPA, 2009; Van Waveren et al., 1999). Refsgaard et al. (2005) also provide a review of existing quality assessment guidelines and Matott et al. (2009) give an extensive overview of approaches to analysing model uncertainties.

In the following, we mainly list recommendations given in these contributions. It should be noted, that we focus on the overall scope and rationale of the methods discussed, not on technical details, which are described in the corresponding literature.

An important first question in evaludation is: Who is to carry out the evaludation? A common answer would be 'the model user', but we claim that this perspective would be inefficient. As a matter of fact, most modellers perform all steps of model evaludation anyway because they are integral parts of model development, analysis, and testing, and modellers are usually no less interested in evaludating their models than decision makers trying to use the models or their

output. Thus, the correct answer is: both model developers and users.

Often, model developers might be biased and tend to overstate the structural realism of their models. It is thus advisable to either include potential model users in the model development process to establish model acceptance, as has been tried in the CREAM project (Grimm et al., 2009), or to follow what is called an "independent verification and validation" approach. The latter is derived from computer science and refers to an evaludation carried out by an external party, which was not previously involved in model development.

It is furthermore crucial to consider the timing of evaludation points. Evaludation measures can be taken either while the model is being developed, or after a model has been completely coded and parameterised. Common practice and experience favour evaludation to take place throughout the model development to reduce costs imposed by errors or misjudgements made early on. This corresponds also to our framework for model evaludation (Fig. 1), which emphasises the iterative nature of model development: design and parameterisation of a model and its submodels are revised when the model did not pass certain performance criteria.

During the early stages of model development, sufficient time should also be invested in defining performance criteria and benchmarks. Benchmarks are metrics that allow an evaluation model output compared to empirical observations. Thus, they support defining meaningful points of reference for model output verification and corroboration (Jakeman et al., 2006; Kirchner et al., 1996). In many cases, goodness-of-fit parameters or confidence intervals are used as quantitative performance criteria to assess the statistical agreement of observed versus modelled data in form of a hypothesis test. A thorough understanding of the applied metrics is needed for this step to avoid potential misinterpretation due to a misunderstanding of a metrics' weaknesses (Bennett et al., 2013). A coefficient of performance, e.g., can be strongly influenced by low sample sizes or outliers, which in return could be a relevant feature of the investigated system. Another tool that is frequently used for qualitative benchmarking is the visual inspection of graphs that trace, for example, the behaviour of model entities.

Currently there seems to be trend in the modelling literature to require increasingly sophisticated statistical tests, in particular Bayesian methods. This trend is laudable, but should not lead to an underestimation of face validation, which is defined by Klügl (2008) as: "all methods that rely on natural human intelligence" (p. 39). Examples listed by Klügl include: "structured walk-throughs, expert assessments of descriptions, animations or results". Klügl accordingly concludes: "face validity shows that processes and outcomes are reasonable and plausible within the frame of theoretic basis and implicit knowledge of system experts or stakeholder. Face validation may be applied from the early phases of the simulation study under the umbrella of conceptual validations. It is often also called plausibility checking". This way of comparing model output to data is thus an integral tool for the evaludation steps "Conceptual model evaluation" and "Model output verification".

Finding the right benchmarks, or metrics, often is part of the problem to be solved in ecology and environmental systems. It can furthermore be necessary to adjust or extend the set of performance criteria. New knowledge or understanding gained during the modelling process can enforce changes not only to the conceptual or computerised model, but also to the way it is analysed. For defining suitable and representative benchmarks it is important to take natural stochasticity into account by using confidence intervals and by focussing on a set of benchmarks. The latter is the core idea of pattern-oriented modelling, i.e. to use multiple patterns for model output verification and corroboration, not just only one (Grimm et al., 2005; Grimm and Railsback, 2012).

4.1. Data evaluation

At this step, a list of all parameters used in a model should be compiled with a description from which sources the parameter values were taken. Additionally, the parameter's units and where exactly (page number, Table number) in a publication they were found need to be provided. If multiple data sources exist for the same parameter it should be mentioned how much the corresponding values differed and whether the differences are caused by different environments, sampling protocols, or other reasons. If no hard data should exist for a given parameter, it should be noted on what grounds the parameter 'guesstimation' was based, e.g. expert knowledge, data from similar species, theoretical considerations, etc.

Essentially, when assessing the quality of the data and patterns used, not only do the measurement protocols need to be evaluated but conclusions drawn from the data should be challenged as well. In some instances, wrong interpretations of data caused delays of model development (Holling, 1978).

The main question in data evaluation is whether the available data are sufficient to support the choice of the model to be applied, and to ensure that the data are sufficiently characteristic of the system to be modelled to provide meaningful insights and comparisons to observations. It is therefore helpful to address these questions as early as possible in the modelling cycle and not postpone them until the end.

4.2. Conceptual model evaluation

There are hardly any specific testing strategies available to confirm conceptual model validity. Frequently, structural inconsistencies are only disclosed later, during model analysis. For example, for spatial processes like movement, visual model output of the implemented model can be decisive in spotting inconsistencies.

Especially for models with numerous entities or processes, the conceptual model becomes more difficult to evaluate. Under such circumstances, the option of evaluating several alternative conceptual models should be considered (Beven, 2006; Refsgaard et al., 2006; Troldborg et al., 2007). Later phases in model development may reveal major flaws in one or more of the alternatives, meaning that the underlying conceptual model has to be rejected. Alternative models may focus on alternative conceptual models with one or a few key processes, or behaviours, differing while keeping other parts of the model unchanged.

Consultation of expert knowledge in the form of a peer review process with or without a scoring system can be another helpful measure at this stage (Landry et al., 1983; Van der Sluijs et al., 2005) but requires the involvement of experts, potentially from various fields, which needs time and organisational preparation.

4.3. Implementation verification

Most of the testing at the stage of code verification involves techniques such as structured walkthroughs, correctness proofs, or an examination of programme structure properties (Sargent, 2005). Ferson (1996) suggests also using specifically designed software that detects common errors in computer code, such as dimensional and unit consistency, correlation matrices, constraints imposed by the biological domain (e.g., negative species abundance is not possible), or realisations of mathematical equations. Argent (2004) and Loizou et al. (2008) exemplify the automatic generation of model codes or equations. Scheller and Mladenoff (2006) and Scheller et al. (2010) demonstrate how current techniques from computer science can be used to manage and verify complex simulation models in ecology.

Table 2
Summary of evaludation terminology.

Term	Definition
Evaludation	The entire process of assessing model quality and establishing model credibility throughout all stages of model development, analysis, and application.
Data evaluation	The assessment of the quality of numerical and qualitative data used to parameterise the model, both directly and inversely via calibration, and of the observed patterns that were used to design overall model structure, whereby not only the measurement protocols need to be evaluated but conclusions drawn from the data should be challenged as well.
Conceptual model evaluation	The assessment of the simplifying assumptions underlying a model's design and forming its building blocks, including an assessment of whether the structure, essential theories, concepts, assumptions, and causal relationships are reasonable to form a logically consistent model.
Implementation verification	The assessment of (1) whether the computerised implementation the model is correct and free of programming errors and (2) whether the implemented model performs as indicated by the model description. The aim is to ensure that the modelling formalism is accurate.
Model output verification	The assessment of (1) how well model output matches observations and (2) to what degree calibration and effects of environmental drivers were involved in obtaining good fits of model output and data. The aim is to ensure that the individuals and populations represented in the model respond to habitat features and environmental conditions in a sufficiently similar way as their real counterparts.
Model analysis	The assessment of (1) how sensitive model output is to changes in model parameters (sensitivity analysis), and (2) how well the emergence of model output has been understood. The aim is to understand the model and be able why which output is being produced to avoid drawing the wrong conclusions from model output.
Model output corroboration	The comparison of model predictions with independent data and patterns that were not used, and preferably not even known, while the model was developed, parameterised, and verified. This step strengthens a model's credibility by proving that the model is capable of predicting/reproducing pattern and data that could not have influenced the model development.

4.4. Model output verification

Several authors viewed it as crucial that performance criteria should be established early in the model development phase against which the model output can then be measured (Crout et al., 2008; Jakeman et al., 2006; Refsgaard et al., 2005; US-EPA, 2009; Van Waveren et al., 1999). In any case the various criteria used for claiming that model is realistic enough should be communicated and justified. The choice of these criteria will be influenced by the overall quality of available data and the design of the conceptual model. Therefore, thorough performance of the previous steps of evaludation can reduce the effort required for model output verification.

4.5. Model analysis

Model analysis can be performed by a multitude of quantitative and qualitative methods. However, which particular approaches are acceptable may depend on the domain in which decisions are intended to be supported by the model, and on the model's purpose.

A method that is regularly performed to evaluate a model is a sensitivity analysis where the model's response to changes in model inputs is explored, i.e., computer programme is executed under different conditions to investigate how a model's response can be apportioned to changes in model inputs, i.e., parameter values and initial conditions (Saltelli et al., 2000). Sensitivity analysis is recommended as the principal evaluation tool for characterising the most and least important sources of uncertainty in environmental models. Local sensitivity analysis, where one parameter is varied a little at a time, is easy to perform but does not capture interaction between parameters and their processes and is restricted to linear effects. Global sensitivity analysis, where parameters are varied over their meaningful range and all possible parameter combinations are sampled, is usually only feasible for a small number of parameters due to run time limitations. In sensitivity experiments, one parameter is varied over its entire range, which can be combined with a second parameter in a contour plot. Sometimes, statistical models like ANOVA, GLMs, boosted regression trees or structured equation modelling can help to summarise the results of global sensitivity analyses.

In general, to understand a model, controlled simulation experiments are needed. The design of simulation experiments should follow the same principles as those of real experiments: keep all factors constant except one or two; explore simplified scenarios in which, for example, the environment is homogeneous and constant, or where some processes are de-activated; try different output metrics (also referred to as summary statistics, observations, or 'currencies', Railsback and Grimm, 2012); etc. The overall approach is to try and understand simplified versions first and then gradually increase the number of possibly confounding factors.

Sensitivity analyses and the testing of alternative model formulations are the most consistently applied methods for model analyses.

4.6. Model output corroboration

Similar approaches can be used as for model output verification. The only principal difference is that we here compare model output to new, independent data and patterns. Such data can sometimes be obtained from new experiments and field studies but more often will be taken from existing literature and expert knowledge. In the latter case, a first step is to identify patterns in the model, which can be considered independent, or secondary, predictions, as the model was not designed to reproduce these patterns.

5. Documentation

Documenting major elements of iterative model development and evaludation is crucial in communicating assumptions, justifications, and findings. Aber (1997) identified cryptic model descriptions as another source for mistrust in a model and Van Waveren et al. (1999) highlight in their Good Modelling Practice Handbook that all steps and actions taken ought to be described in a way understandable for the decision maker.

Schmolke et al. (2010a) proposed the TRACE (transparent and coherent ecological modelling) documentation scheme. The purpose of a TRACE document, which would usually be provided as a supplement or appendix, is to provide additional evidence that the model has been carefully designed and thoroughly tested and analysed. The basic idea of TRACE was to introduce a common terminology and document structure, so that modellers and model users know exactly what to document and where to put and look for the different elements of evidence that a model is fit for its intended purpose. TRACE thus not only provides a common terminology and structure, but is also a checklist for model developers

and users to make sure they addressed all important elements of model evaludation (see Grimm et al., this volume).

While a model is developed, the different steps and activities performed throughout the different stages of the modelling cycle should be documented in a modelling notebook, which corresponds to notebooks or journals kept in laboratories. If TRACE terminology is used for the entries in the notebook, it will be easy and efficient to extract the relevant information from the modelling notebook and assemble a TRACE document when the model is delivered.

TRACE has been tested in about 10 modelling projects (Grimm et al., this volume). It turned out that TRACE, as originally described by Schmolke et al. (2010a) was not ready for being used. Grimm et al. (this volume) present an update of TRACE and its rationale. The overall idea remains the same, but more specific guidelines for producing and reading TRACE documents were formulated. Most importantly, TRACE terminology and document structure was completely changed and now follows the terminology introduced here, including the six elements of evaludation. The focus of TRACE thus shifts from documentation, which is not necessarily linked to a specific purpose, to evaludation, which we here defined as 'the entire process of establishing model quality and credibility throughout all stages of model development, analysis, and application'. Consequently, a main purpose of TRACE documents is to report all elements of a model's evaludation. In addition, TRACE documents include a detailed problem formulation, a full model description, and a description and justification of the environmental scenarios explored with the model (for details, see Grimm et al., this volume).

6. Concluding remarks

Confusing terminology is one of the main obstacles to get a good understanding what model validation is, how it works, and what it can deliver. Attempts to clarify terminology were criticised by Hodges (2008): "There is a repeated call for ecological terminology to be standardised and for terms to be defined more concretely. These calls for the standardisation of definitions are based on faulty premises about the way language conveys meaning." (p. 35). We agree that terminological discussions can turn into hair-splitting exercises, but we hold, following the quote of Immanuel Kant that we chose as a motto for this article, that terminology discussions are also about genuine problems, not just words (see also Jax (2008)).

We therefore devised a standard set of terms related to validation that we derived from a literature review and from a consideration of the different elements of iterative model development (summary in Table 2). We believe that this set of terms and its relation to the modelling cycle can help to make model assessment more comprehensive and transparent. Our distinction of different evaludation steps offers a generic checklist, which makes it easier for modellers and model users to organise model evaluation and its communication.

We want to point out, however, that there can be reasons to perform one or more evaludation steps to only some limited degree. Reasons for this may include a lack of data, limited time and other resources, or ambiguities in the problem formulation. Our advice for such circumstances is to document and discuss possible limitations, their reasons, and how they could be overcome in the future. This enables model users to understand that limitations do not reflect oversights but limitations that the modeller could not overcome at a given time. It should also be noted that even despite partly performed evaludation steps, models can add important information to a decision-making process. Similarly, a fully performed evaludation does not guarantee that a model is good enough for application in a decision-making context. Thus, evaludation does not provide a yes/no criterion for whether a model can support decision-making. Rather, the different evaludation steps add to the 'weight of evidence' (Weed, 2005) that a model is fit for its purpose. For specific fields of application, it might be well possible to provide more respective guidance and requirements for the different evaludation steps, but such guidance can only be based on a joint activity of all stakeholders involved. We believe that the evaludation scheme that we presented here, and its documentation in TRACE documents, will facilitate such activities.

TRACE documents are a tool to put evaludation into practice and get it established, both for modellers and model users. Modellers will always benefit from keeping a modelling notebook, preferably on a daily basis. If they use TRACE terminology, which is based on the terminology introduced here, modellers can at a later stage easily assemble TRACE documents, which provide, in a structured and standardised way, various kinds of evidence that their model was well designed and thoroughly tested and analysed. Modellers will thus directly profit from using and keeping a modelling notebook and from providing the kind of information that decision makers need to see to assess whether or not they can use model output as a basis for their decisions.

To conclude, we believe the suggested terminology and framework can ultimately contribute to establish an advanced culture of model development and evaluation, so that in the future better models are developed and actually used to support more environmental decisions in a productive and robust way.

Acknowledgements

We thank Pernille Thorbek and two anonymous reviewers for their valuable comments on this article. We acknowledge financial support by the European Union under the 7th Framework Programme (project acronym CREAM, contract number PITN-GA-2009-238148).

References

Aber, J.D., 1997. Why don't we believe the models? Bulletin of the Ecological Society of America 78, 232–233.
Argent, R.M., 2004. An overview of model integration for environmental applications – components, frameworks and semantics. Environmental Modelling and Software 19, 219–234.
Arthur, J.D., Gröner, M.K., Hayhurst, K.J., Holloway, C.M., 1999. Evaluating the effectiveness of independent verification and validation. Computer 32, 79–83.
Aumann, C.A., 2007. A methodology for developing simulation models of complex systems. Ecological Modelling 202, 385–396.
Bacsi, Z., Zemankovics, F., 1995. Validation: an objective or a tool? Results on a winter wheat simulation model application. Ecological Modelling 81, 251–263.
Barlas, Y., 1996. Formal aspects of model validity and validation in system dynamics. System Dynamics Review 12, 183–210.
Barlas, Y., Carpenter, S., 1990. Philosophical roots of model validation: two paradigms. System Dynamics Review 6, 148–166.
Bart, J., 1995. Acceptance criteria for using individual-based models to make management decisions. Ecological Applications 5, 411–420.
Beck, M.B., Ravetz, J.R., Mulkey, L.A., Barnwell, T., 1997. On the problem of model validation for predictive exposure assessments. Stochastic Hydrology and Hydraulics 11, 229–254.
Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., 2013. Characterising performance of environmental models. Environmental Modelling and Software 40, 1–20.
Beven, K., 2002. Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system. Hydrological Processes 16, 189–206.
Beven, K., 2006. A manifesto for the equifinality thesis. Journal of Hydrology 320, 18–36.
Boesten, J.J.T.I., 2000. Modeller subjectivity in estimating pesticide parameters for leaching models using the same laboratory data set. Agricultural Water Management 44, 389–409.
Borenstein, D., 1998. Towards a practical method to validate decision support systems. Decision Support Systems 23, 227–239.
Botkin, D.B., 1993. Forest Dynamics: An Ecological Model. Oxford University Press, New York, NY, USA.
Caswell, H., 1976. The validation problem. In: Patten, B. (Ed.), Systems Analysis and Simulation in Ecology. Academic Press, New York, NY, pp. 313–325.

Clark, J.S., 2004. Why environmental scientists are becoming Bayesians. Ecology Letters 8, 2–14.
Committee on Models in the Regulatory Decision Process, 2007. Models in Environmental Regulatory Decision Making. National Academies Press, Washington, DC.
Crout, N., Kokkonen, T., Jakeman, A.J., Norton, J.P., Anderson, R., Assaf, H., Gaber, N., Gibbons, J., Holzworth, D., Mysiak, J., Reichl, J., Seppelt, R., Wagener, T., Whitfield, P., 2008. Good modelling practice. Environmental Modelling and Software 3, 15–32.
Fagerstrom, T., 1987. On theory, data and mathematics in ecology. Oikos 50, 258–261.
Ferson, S., 1996. Automated quality assurance checks on model structure in ecological risk assessments. Human and Ecological Risk Assessment 2, 558–569.
Ferson, S., Oberkampf, W.L., Ginzburg, L., 2008. Model validation and predictive capability for the thermal challenge problem. Computer Methods in Applied Mechanics and Engineering 197, 2408–2430.
Forbes, V.E., Hommen, U., Thorbek, P., Heimbach, F., van den Brink, P.J., Wogram, J., Thulke, H.-H., Grimm, V., 2009. Ecological models in support of regulatory risk assessments of pesticides: developing a strategy for the future. Integrated Environmental Assessment and Management 5, 167–172.
Forbes, V.E., Calow, P., Grimm, V., Hayashi, T., Jager, T., Palmqvist, A., Pastorok, R., Salvito, D., Sibly, R., Spromberg, J., Stark, J., Stillman, R.A., 2010. Integrating population modeling into ecological risk assessment. Integrated Environmental Assessment and Management 6, 191–193.
Galic, N., Hommen, U., Baveco, J.M., Van den Brink, P.J., 2010. Potential application of population models in the European ecological risk assessment of chemicals II. Review of models and their potential to address environmental protection aims. Integrated Environmental Assessment and Management 6, 338–360.
Gass, S.I., 1983. Decision-aiding models: validation, assessment, and related issues for policy analysis. Operations Research 31, 603–631.
Goodall, D.W., 1972. Building and testing ecosystem models. Mathematical Models in Ecology, 173–194.
Grimm, V., Railsback, S.F., 2005. Individual-Based Modeling and Ecology. Princeton University Press, Princeton, pp. 480.
Grimm, V., Railsback, S.F., 2012. Pattern-oriented modelling: a "multi-scope" for predictive systems ecology. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 298–310.
Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H.-H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-oriented modeling of agent-based complex systems: lessons from ecology. Science 310, 987–991.
Grimm, V., Ashauer, R., Forbes, V.E., Hommen, U., Preuss, T.G., Schmidt, A., Van den Brink, P.J., Wogram, J., Thorbek, P., 2009. CREAM: a European project on mechanistic effect models for ecological risk assessment of chemicals. Environmental Science and Pollution Research 16, 614–617.
Hodges, J.S., 1991. Six (or so) things you can do with a bad model. Operations Research 39, 355–365.
Hodges, K.E., 2008. Defining the problem: terminology and progress in ecology. Frontiers in Ecology and the Environment 6, 35–42.
Hodges, J.S., Dewar, J.A., 1992. Is it you or your model talking? – A framework for model validation. Rand, Santa Monica, CA.
Holling, C.S., 1966. The functional response of invertebrate predators to prey density. Memoirs of the Entomological Society of Canada 98, 5–86.
Holling, C.S., 1978. Adaptive Environmental Assessment and Management. John Wiley & Sons, New York, NY.
Hoover, S., Perry, R.F., 1989. Simulation: A Problem-Solving Approach. Addison-Wesley, Reading, MA.
Hunka, A.D., Meli, M., Thit, A., Palmqvist, A., Thorbek, P., Forbes, V.E., 2013. Stakeholders' perspective on ecological modeling in environmental risk assessment of pesticides: challenges and opportunities. Risk Analysis 33, 68–79.
Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling and Software 21, 602–614.
Jax, K., 2008. Concepts, not terms. Frontiers in Ecology and the Environment 6, 178–179.
Kirchner, J.W., Hooper, R.P., Kendall, C., Neal, C., Leavesley, G., 1996. Testing and validating environmental models. Science of the Total Environment 183, 33–47.
Klügl, F., 2008. A validation methodology for agent-based simulations. In: Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC'08), pp. 39–43.
Kuhn, T.S., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Landry, M., Malouin, J.-L., Oral, M., 1983. Model validation in operations research. European Journal of Operational Research 14, 207–220.
Latombe, G., Parrott, L., Fortin, D., 2011. Levels of emergence in individual based models: coping with scarcity of data and pattern redundancy. Ecological Modelling 222, 1557–1568.
Levins, R., 1966. The strategy of model building in population ecology. American Scientist 54, 421–431.
Loizou, G., Spendiff, M., Barton, H.A., Bessems, J., Bois, F.Y., D'Yvoire, M.B., Buist, H., Clewell, H.J., Meek, B., Gundert-Remy, U., Goerlitz, G., Schmitt, W., 2008. Development of good modelling practice for physiologically based pharmacokinetic models for use in risk assessment: the first steps. Regulatory Toxicology and Pharmacology: RTP 50, 400–411.
Mankin, J.B., O'Neill, R.V., Shugart, H.H., Rust, B.W., 1977. The importance of validation in ecosystem analysis. In: Innis, G. (Ed.), New Directions in the Analysis of Ecological Systems, Part 1. The Society for Computer Simulation, La Jolla, CA, pp. 63–71.
Matott, L.S., Babendreier, J.E., Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resources Research 45, W06421.
May, R.M., 1973. Stability and Complexity in Model Ecosystems, Vol. 6. Princeton University Press, Princeton, pp. 235.
Mayer, D.G., Butler, D.G., 1993. Statistical validation. Ecological Modelling 68, 21–32.
Neuert, C., Rademacher, C., Grundmann, V., Wissel, C., Grimm, V., 2001. Struktur und Dynamik von Buchenurwäldern: Ergebnisse des regelbasierten Modells BEFORE. Naturschutz und Landschaftsplanung 33, 173–183.
Oreskes, N., Belitz, K., 2001. Philosophical issues in model assessment. Model validation. Perspectives in Hydrological Science 23, 23.
Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994a. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263, 641–646.
Oreskes, N., Belitz, K., Shrader-Frechette, K., Sterman, J.D., Rykiel, E.J.J., 1994b. The meaning of models. Science 264, 331.
Oriade, C.A., Dillon, C.R., 1997. Developments in biophysical and bioeconomic simulation of agricultural systems: a review. Agricultural Economics 17, 45–58.
Ormerod, P., Rosewell, B., 2009. Validation and verification of agent-based models in the social sciences. In: Squazzoni, F. (Ed.), Epistemological Aspects of Computer Simulation in the Social Sciences. Springer, Berlin, Heidelberg, pp. 130–140.
Overton, S., 1977. A strategy of model construction. In: Hall, C., Day, J. (Eds.), Ecosystem Modeling in Theory and Practice: An Introduction with Case Histories. John Wiley & Sons, New York, NY, pp. 49–73.
Pastorok, R.A., 2002. Ecological Modeling in Risk Assessment: Chemical Effects on Populations, Ecosystems, and Landscapes. Lewis Publishers, Boca Raton, FL, USA.
Pilkey, O.H., Pilkey-Jarvis, L., 2007. Useless Arithmetic: Why Environmental Scientists Can't Predict the Future. Columbia University Press, New York.
Popper, K., 1959. The Logic of Scientific Discovery. Hutchingson & Co., London, pp. 513.
Rademacher, C., Neuert, C., Grundmann, V., Wissel, C., Grimm, V., 2001. Was charakterisiert Buchenurwälder? Untersuchungen der Altersstruktur des Kronendachs und der räumlichen Verteilung der Baumriesen in einem Modellwald mit Hilfe des Simulationsmodells BEFORE. Forstwissenschaftliches Centralblatt 120, 288–302.
Rademacher, C., Neuert, C., Grundmann, V., Wissel, C., Grimm, V., 2004. Reconstructing spatiotemporal dynamics of Central European natural beech forests: the rule-based forest model BEFORE. Forest Ecology and Management 194, 349–368.
Refsgaard, J.C., Henriksen, H.J., 2004. Modelling guidelines – terminology and guiding principles. Advances in Water Resources 27, 71–82.
Refsgaard, J.C., Henriksen, H.J., Harrar, W.G., Scholten, H., Kassahun, A., 2005. Quality assurance in model based water management – review of existing practice and outline of new approaches. Environmental Modelling and Software 20, 1201–1215.
Refsgaard, J.C., Van der Sluijs, J.P., Brown, J., Van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Advances in Water Resources 29, 1586–1597.
Rykiel, E.J.J., 1984. Modelling agroecosystems: lessons from ecology. In: Lowrance, R., Stinner, B.R., House, G.J. (Eds.), Agricultural Ecosystems: Unifying Concepts. John Wiley & Sons, New York, NY, pp. 157–178.
Rykiel, E.J.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229–244.
Saltelli, A., Tarantola, S., Campolongo, F., 2000. Sensitivity analysis as an ingredient of modeling. Statistical Science 15, 377–395.
Sargent, R.G., 2005. Verification and validation of simulation models. In: Proceedings of the 2005 Winter Simulation Conference, Syracuse NY, USA, Orlando, Florida.
Scheller, R.M., Mladenoff, D.J., 2006. An ecological classification of forest landscape simulation models: tools and strategies for understanding broad-scale forested ecosystems. Landscape Ecology 22, 491–505.
Scheller, R.M., Sturtevant, B.R., Gustafson, E.J., Ward, B.C., Mladenoff, D.J., 2010. Increasing the reliability of ecological models using modern software engineering techniques. Frontiers in Ecology and the Environment 8, 253–260.
Schlesinger, S., 1979. Terminology for model credibility. Simulation 32, 103–104.
Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010a. Ecological models supporting environmental decision making: a strategy for the future. Trends in Ecology and Evolution 25, 479–486.
Schmolke, A., Thorbek, P., Chapman, P., Grimm, V., 2010b. Ecological models and pesticide risk assessment: current modeling practice. Environmental Toxicology and Chemistry 29, 1006–1012.
Shugart, H.H., 1984. A Theory of Forest Dynamics: The Ecological Implications of Forest Succession Models. Springer-Verlag, New York, NY.
Taleb, N.N., 2010. The Black Swan: The Impact of the Highly Improbable. Random House, New York, pp. 480.
Thorbek, P., Forbes, V.E., Heimbach, F., Hommen, U., Thulke, H.-H., Van den Brink, P.J., Wogram, J., Grimm, V. (Eds.), 2010. Ecological Models for Regulatory Risk Assessments of Pesticides: Developing a Strategy for the Future. Society of Environmental Toxicology and Chemistry (SETAC) and CRC Press, Pensacola and Boca Raton, FL, USA.
Topping, C.J., Dalkvist, T., Grimm, V., 2012. Post-hoc pattern-oriented testing and tuning of an existing large model: lessons from the field vole. PLoS One 7, e45872.
Troldborg, L., Refsgaard, J.C., Jensen, K.H., Engesgaard, P., 2007. The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. Hydrogeology Journal 15, 843–860.

US-EPA, 2009. Guidance on the Development, Evaluation, and Application of Environmental Models. Environmental Protection EPA/100/K-99.

Van der Sluijs, J.P., Craye, M., Funtowicz, S., Kloprogge, P., Ravetz, J., Risbey, J., 2005. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. Risk analysis an Official Publication of the Society for Risk Analysis 25, 481–492.

Van Waveren, R.H., Groot, S., Scholten, H., Van Geer, F.C., Wösten, J.H.M., Koeze, R.D., Noort, J.J., 1999. Good modelling practice handbook. In: Fourth International Conference Hydro Informatics 2000, p. 165.

Wang, M., Luttik, R., 2012. Population level risk assessment: practical considerations for evaluation of population models from a risk assessor's perspective. Environmental Sciences Europe, 24.

Weed, D.L., 2005. Weight of evidence: a review of concepts and methods. Risk Analysis 25, 1545–1557.

Zurell, D., Grimm, V., Rossmanith, E., Zbinden, N., Zimmermann, N.E., Schröder, B., 2012. Uncertainty in predictions of range dynamics: black grouse climbing the Swiss Alps. Ecography 35, 590–603.